On measuring and theorising mathematical identity

Eivind Kaspersen

On measuring and theorising mathematical identity

Doctoral Dissertation at the University of Agder

University of Agder
Faculty of Engineering and Science
2018

Doctoral Dissertations at the University of Agder 188

ISSN: 1504-9272
ISBN: 978-82-7117-887-1

Trykk: Wittusen & Jensen
Oslo

Preface

In this study, I have tried to understand how mathematical identity can be measured and theorised. During the work, I have benefitted from moral and intellectual assistance. To this, I am grateful.

First, I would like to express my sincere gratitude to my supervisors, Birgit Pepin, Pauline Vos, and Svein Arne Sikko, for the support during this research. I am particularly grateful for your encouragement when the study shifted, for me, quite dramatically, from empirical to theoretical and methodological.

I also wish to thank my colleagues at the Norwegian University of Science and Technology. From you, I have learned a lot about research in general.

My sincere thanks also go to the University of Agder and the Norwegian University of Science and Technology for economic and other material support.

Except for my supervisors, the published papers and the chapters in this thesis have been reviewed by a number of scholars in a variety of fields, including psychometrics, philosophy, mathematics, linguistics, and, surely, mathematics education. Due to these reviews, my writing has improved in content and rhetoric. I have also learned how difficult it is to communicate to a broad audience.

Last but not the least, I would like to thank my family in Kristiansand and Namsos, and my family in Trondheim: Marianne, Silje, and Oliver. This thesis is dedicated to you.

Eivind Kaspersen
Trondheim, January 2018

Abstract

One strand of the research on mathematics education is the study of identity (teacher identity, ethnic identity, etc.). In this strand, I have focused on mathematical identity. Over the past decades, the number of studies on mathematical identity has increased, and these studies have illustrated how the construct can be applied to understand both personal and social aspects of how humans relate to mathematics.

While it is recognised that the construct is important, studies on mathematical identity suffer from some challenges. One problem is that it has proven difficult to measure mathematical identity, mainly, due to methodological and theoretical issues. This particular problem is the topic of this thesis.

A better understanding of the measurement of mathematical identity would be beneficial for several reasons. For instance, measurement has been an important tool in the history of science. Hence, an instrument for measuring mathematical identity that is compatible with principles of measurement could assist researchers, for example, in understanding identity development. Moreover, a theoretical perspective on mathematical identity that is consistent with measurement would contribute to a better understanding of the construct as such.

To address these issues, I have taken a mixed-methods approach. Rasch Measurement Theory was applied to develop and validate an instrument for measuring mathematical identity. Rasch Measurement Theory is a psychometrical theory that claims to be consistent with principles of interval measurement. From a sample of 133 students in teacher education (TE) and 185 science, technology, engineering, and mathematics (STEM) students, 20 items have been concluded to be productive for measurement. Also, qualitative data provide illustrative examples of the characteristics.

Informed by the obtained measures, I have sought a theoretical perspective on mathematical identity. A premise of the theorisation is that it is possible to measure mathematical identity. That is, within specific contexts, it is possible for some persons to relate more strongly with mathematics than others.

In addition to this premise, the theorisation builds on two assumptions. First, mathematical identity is assumed to be relational since measures, in general, are relational. Second, the locus of identity—whether mathematical identity is mostly situated or context-free—is an empirical question that can be studied from the process of measurement.

From these assumptions, mathematical identity is defined to be a relative position between persons and the social structure of being mathematical within the activity in which they participate and to which

they contribute. The social structure is a set of characteristics and its internal structure, and the social structure is defined to be person-independent.

From these definitions, some theoretical insights follow. First, arguments are made that the distinction between structural and personal change is an arbitrary point of perspective. In short, the process of measurement cannot distinguish between structural and personal identity development. Second, the comparison of mathematical identities does not require structural equivalence. What is required is that, if mathematical identity is to be compared between two contexts, there must exist a subset of characteristics that are approximately equal, in content and structure, between the contexts. The relative size of the subset is of little importance. Third, mathematical identities do not exist in isolation. Consequently, it is possible for one person to relate more strongly and, at the same time, more weakly to mathematics than another person, without contradiction.

A measuring perspective on mathematical identity is a unifying framework, that is, a framework that captures social (structural) and personal aspects of identity simultaneously. As an illustration of the social aspect, structural differences between the STEM and TE contexts have been studied. The results point to some particular structural differences, although the overall conclusion is that mathematical identity is ‘practically context-free’ between these contexts. ‘Practically context-free’ means that, while the structural differences are interesting in themselves, they hardly affect personal measures when they are not accounted for.

As an example of the personal aspect, I have studied the association between self-reported mathematical identities and average grades in university mathematics courses. The result indicates that the variables correlate poorly. Nevertheless, there seems to be an association, as the average grade of persons with ‘strong’ mathematical identities was reported to be about one grade higher than amongst students with ‘low’ mathematical identities.

The most significant contribution of this study is that the theorisation of mathematical identity has led to the conclusion that, when it is measured, mathematical identity is a relative position rather than something people have. Moreover, mathematical identity implies the existence of a social structure. This structure can be measured, and it is not assumed to be static. It is true that at least one point must be static, but this point is arbitrary. From this, I conclude that there is no ontological or epistemological difference between the development of structural and personal identities.

Contents

0 List of abbreviations

ANOVA	Analysis of Variance
ASSIST	The Approaches and Study Skills Inventory for Students
CHAT	Cultural-historical activity theory
C→M	Category implies measure
DIF	Differential Item Functioning
ICC	Item Characteristic Curve
INFIT MNSQ	Inlier-sensitive mean-square statistic
IRT	Item Response Theory
JMLE	Joint Maximum Likelihood Estimation
M→C	Measure implies category
OUTFIT MNSQ	Outlier-sensitive mean-square statistic
PCA	Principal Component Analysis
PCM	Partial Credit Model
RMT	Rasch Measurement Theory
RSM	Rating Scale Model
STEM	Science, technology, engineering, and mathematics
TE	Teacher education
ZSTD	Standardised fit statistics

1 Introduction

1.1 Rationale and background

In this introduction, I present the rationale and background for my research. I also describe my research process, which has undergone fundamental changes.

The starting point for this PhD was to conduct research on students' transitions from being students at the university to becoming professionals in the world of work. Doing research into this should assist in better understanding, and promoting participation and engagement in mathematics education in higher education, for example, in fields of teacher education (TE) and university disciplines such as science, technology, engineering, and mathematics (STEM). In particular, I initially aimed at doing research on the negotiating of identities in students' transitions to the world of work. At a later point in this chapter, I explain more about this, but at this stage, it suffices to say that in my research I focus on student identities in higher education. Moreover, I concentrate on two populations: (1) TE students, and (2) STEM students.

A person can 'have' many different identities: a national identity, a religious identity, a gender identity, and so forth (e.g., Gee, 2000). In my research, I have focused on *mathematical identities*. With "having a mathematical identity" I roughly mean: having a research mathematician's identity, for instance, being seized by yet unsolved mathematical problems and trying to invent own methods to solve such problems (Burton, 1998; Nardi, 2007; Wood, Petocz & Reid, 2012). Initially, I intended to use measurement as a tool for studying how students' mathematical identities changed in the transition to the world of work. Thus, early on, I searched instruments for measurement in the research literature about identities.

However, I experienced two problems. First, I did not find instruments that could measure mathematical identity on a single dimension. No existing instrument would allow some persons to identify more strongly with mathematics than others, it appeared. This problem led to the development of an instrument that I discuss in the first paper of this thesis. During the process of instrument validation, I experienced a second, more pressing, problem, namely, with theorising the outcomes of the measurements in line with existing theories on identity (e.g., Gee, 2000; Holland, Lachicotte, Skinner, & Cain, 2001; Sfard & Prusak, 2005; Wenger, 1998). I explain these problems in more detail later, but in short, they caused a shift in emphasis for my study. From a focus on studying student transition empirically, and using measurement as a tool, my research changed to developing better understandings of theoretical and methodological underpinnings of the measurement of mathematical identity.

I did not start this research with a clean sheet. Rather, the first move in my study was highly influenced by cultural-historical factors. To be specific, in 2012 I applied for a position as a PhD student in a project that aimed at better understanding Norwegian TE and STEM students' transitions from compulsory school to higher education, and further, from higher education to the world of work. This Norwegian transition project was closely connected to the TransMaths-project in the UK (see www.transmaths.org), which had produced significant insights into students' trajectories in and through mathematics programmes (e.g., the special issue in 'Research in Mathematics Education', introduced by Wake (2011)).

Outcomes of the TransMath project in the UK include an understanding of how different classroom experiences relate to mathematical identities, how leading identities shape students' motives for mathematical activities (Black et al. 2010), and the association between relevant variables—disposition, perceived support, perceived transitional gap, and pedagogy—in the transition to university (Pampaka, Pepin, & Sikko, 2016). The TransMaths project followed a mixed-methods design, and several instruments were constructed and validated for measuring relevant variables (e.g., Pampaka et al., 2013). Theoretically, the project mainly took a sociocultural stance, and my study has been influenced by both the methodological approaches to measurement and the sociocultural theories of identity. It is the tension I experienced between sociocultural theories and theories on measurement that later caused the shift in my research.

1.2 An introduction to identity

1.2.1 Identity

On the personal aspect of learning, the construct of identity has gained increased attention over the last decades. Sfard and Prusak (2005, p. 15) claimed "that the notion of identity is a perfect candidate for the role of 'the missing link' in the researchers' story of the complex dialectic between learning and its sociocultural context". However, there are considerable differences on how to conceptualise identity. These differences were explained by Darragh (2016) who made an approximate distinction between those who perceive identity as an action (often situated in a context) and those who see it as an acquisition. Similarly, Cote and Levine (2014) distinguished between theories that perceive identity as context-free and those that regard it as mostly situated.

Hannula et al. (2016) and Darragh (2016) documented that many recent studies on identity in the field of mathematics education rely on sociocultural theories, that is, theories on the situated end of the spectrum. One recent example is the study by Williams (2011) who applied Holland et al.'s (1998) framework of figured worlds to understand the narratives of two teachers who were teaching students before their transitions to higher education.

1.2.2 Measuring mathematical identities

As described above, my research was situated within an emerging research area of transition and identity studies, which, again, was situated within the broader field of sociocultural studies. Hence, the point of departure of my research was influenced by the documented challenges regarding TE and STEM students' transitions to the world of work. In particular, studies on negotiations of personal identities in transition, documented by the TransMaths project (e.g., Black et al., 2010; Hernandez-Martinez et al., 2011; Williams, 2011), inspired the research. Accordingly, the original research question of the study was: "How are identities negotiated in STEM and TE students' transitions to the world of work?"

My approach to this question was to explore the possibility of using measurements, and accordingly, an instrument for measuring mathematical identity was sought. Specifically, the development of the instrument relied on three primary sources: (1) related instruments, in particular the Approaches and Study Skills Inventory for Students (ASSIST) (Entwistle, 1997), (2) existing literature on the understanding of mathematics (e.g., Hiebert, 1986; Skemp, 1987), and (3) members of mathematical communities (e.g., PhD students in STEM-related subjects). A more detailed description of these sources will be discussed later.

As a result of these influences, an instrument for measuring the extent to which students are working conceptually with mathematics (i.e., how deeply they are working with mathematics) was validated. The technical validation is discussed in detail in the first paper of this thesis.

1.2.3 Theorising measured identities

When I was working on the second paper, I was exploring *mathematical identities* beyond technical validity. Then, I experienced a problem, namely, how measures (i.e., outcomes of a technically valid measurement through a questionnaire) could be conceptualised as identities. Specifically, most sociocultural theories conceptualise identity as 'complex', which obviously complicates the measurement of identity since measures are required to be uni-dimensional (e.g., Thurstone, 1954). Another aspect is that identities are, in some frameworks, conceptualised as motions, for example, in communities of practice where identity is seen as “a constant becoming” (Wenger, 1998). In contrast, classical psychometrical techniques produce static, in-the-moment measures. Moreover, identities in sociocultural theories are, for the most, seen as situated. However, Thurstone (1954) required measures to be invariant.

On the surface, measures of mathematical identities seemed to fit within an acquisition interpretation of identity. However, my data rejected an acquisition perspective on identity as, for instance, the analysis of mathematical identities indicated that identity is not entirely context-free. That is, the characteristics of being mathematical, represented by the items, structured differently in the STEM and TE contexts. This result supported a sociocultural perspective. In the thesis, I refer to a set of characteristics of being mathematical and its internal structure as ‘a social structure of being mathematical’. Later, I will explain more about the social structure and its properties.

In conclusion, I experienced the relationship between data and theories of identity as problematic. Accordingly, the emphasis of the research changed from empirical to theoretical and methodological. A fair summary is, therefore, to say that the study has moved dialectically between empirical data collection, analysis, methodology, theory, and writing.

Other contributors to the shifting emphasis are elements in the research community such as papers and books, persons with whom I have collaborated, other people that I have met at conferences and seminars, and reactors/reviewers. Of particular influence was one reviewer of the second paper who offered profound suggestions and questions about the philosophical foundation of measured mathematical identities.

When I was engaging in theoretical issues regarding mathematical identity, I realised that method is an integral part of a theory. As I explain in more detail later, there exist several, incommensurable methods for measurement in the social sciences. These methods have been developed within different paradigms, that is, in communities that perceive philosophies of measurement differently. Thus, in the thesis, I explain the relationship between principles of measurement and principles of mathematical identity.

One consequence of the shifting emphasis is a shift in the scope of the study. Initially, the study was situated firmly within mathematics education. In the present form, although my results are anchored in illustrations from mathematics education, the scope has shifted, and I consider the thesis to contribute also to general education knowledge, including research methodology and, in particular, educational measurement, and the more general field of identity research.

1.3 Research questions

After having gone through a dialectical, non-linear, research process, and after having faced difficulties in conceptualising the results from the beginning of my study, I was guided by the research questions stated below. The overarching research question is:

How can mathematical identity be measured and theorised?

This research question is subdivided into the following sub-questions:

1. How do theories on measurement inform a measuring perspective on mathematical identity?
2. What is a paradigmatic method for measuring mathematical identity?
3. What are theoretical principles of mathematical identity?
4. Which empirical questions can be asked within a framework for measuring mathematical identity?
    a. What are characteristics of mathematical identity in Norwegian TE and STEM contexts?
    b. How can mathematical identity measures provide information on how much the social structure of being mathematical differs across the STEM context and the TE context?
    c. What is the association between STEM students' self-reported mathematical identities and average grades in mathematics courses?

1.4 Short overview of the thesis

The four sub-questions answer the main research question. The sub-questions are answered throughout the thesis with the support of three papers, Papers I, II, and III (Kaspersen, 2015; Kaspersen, Pepin, & Sikko, 2016; Kaspersen, Pepin, & Sikko, in print). The papers are comprehensively summarised later, and they are included in the Appendices.

In chapter 2, I review and summarise theories on identity. I have chosen to focus on a limited selection that I believe represent studies on mathematical identities. The selection has been influenced by Darragh's (2016) review and includes identities in cultural worlds (Holland et al., 2001), identities in communities of practice (Wenger, 1998), discursive theories on identity (Gee, 2000), narrated identities (Sfard & Prusak, 2005), and identities in cultural-historical activity theory (CHAT) (e.g., Stetsenko & Arievitch, 2004).

In this thesis, I argue that a theoretical perspective on measured mathematical identities must be compatible with theories on measurement in general. Thus, in chapter 3, I present general principles of measurement as formulated by Thurstone (e.g., 1954). In short, I conclude that a measuring perspective on mathematical identity must consider: dimensionality, additivity, invariance, and relativity. Moreover, I connect these principles with existing theories on identity, that is, those presented in chapter 2. I conclude that there is a need for a framework on mathematical identity that addresses principles of measurement explicitly. In this chapter, I answer research question 1.

The need for a measurable mathematical identity, however, is not a claim that existing theories are ‘false’. I discuss the notion of ‘truth’ more closely in the first part of chapter 4 where I situate the study within a (neo)-pragmatist philosophy, mainly influenced by Putnam (e.g., 1981). In short, I postulate that theory-pluralism is no more of a problem than the co-existence of multiple geometries.

Later in chapter 4, I discuss philosophical differences between two commonly applied theories of measurement—Rasch Measurement Theory (RMT) and Item Response Theory (IRT). Specifically, I show how proponents of RMT see principles of measurement as requirements, as opposed to those of IRT who regard them as assumptions. Moreover, I explain how I interpret principles of mathematical identity as requirements. On this ground, I answer research question 2 when I conclude that the application of RMT is a paradigmatic method for the measurement of mathematical identity.

In chapter 5, I describe the design of the study and methods of data collection and analysis. Moreover, I discuss issues of validity and ethics.

In chapter 6, I present the papers in this thesis. First, I summarise Paper I, which reports on the validation of an instrument for measuring the extent to which students work conceptually with mathematics. Evidence for invariance was sought between TE and STEM students. In the paper, I answer research question 4a when I discuss 20 characteristics[1] of mathematical identity.

In Paper II, I address research question 3 when I discuss theoretical principles of mathematical identity. I then provide some empirical examples and illustrate a key result of the theorisation, namely, the person-independent property of the social structure[2] of identity. I also provide some qualitative examples of mathematical identities, and I discuss research question 4b, that is, structural differences between TE and STEM students’ mathematical identities.

[1] There is a strong connection between ‘item’ and ‘characteristic’. In this thesis, I define a characteristic to be an ideal feature or quality. An item, in this thesis, is the translation from a characteristic to a statement in a questionnaire to which persons respond. Consequently, mathematical identities are inferred indirectly from persons’ answers to the questionnaire.

[2] The word ‘structure’ appears repeatedly throughout the thesis. I use this word generically, and it must, therefore, be understood in the context in which it is applied. For example, a ‘structure’ in community of practice explains how a community builds meaning, and this is a different structure than the ‘structure’ of mathematical identity, which is based on ordering questionnaire items.

In Paper III, I discuss the association between self-reported mathematical identities and average grades in university mathematics courses. From measures obtained by the instrument validated in Paper I, 361 STEM students were categorised as having a ‘low’, ‘medium’, or ‘high’ mathematical identity, and the paper illustrates how the mean average grade of students with high mathematical identities was significant and about one grade higher than students with lower mathematical identities. In the paper, I answer research question 4c.

In the last chapter, following a brief summary, I discuss the theoretical insights of the research. I claim: (1) that the distinction between structural and personal change is an arbitrary point of perspective, (2) that the comparison of mathematical identities does not require structural equivalence, and (3) that mathematical identities do not exist in isolation. The third claim implies the possibility of one person having both a stronger and a weaker mathematical identity than another person, without contradiction.

Subsequently, I position mathematical identity between two extremes, one that postulates that identity can never be measured and the other that claims that identity can always be measured. Thereafter, I discuss limitations and challenges of the study, before I conclude the thesis.

Writing is, inevitably, linear—one sentence after another—and therefore, it has been a challenging task to describe, in words, a process that has not been linear. Data collection and analysis, reading and writing, and research questions and findings: All have evolved dialectically in a non-linear way. In particular, chapter 4 and 6 connect so tightly that I wish the reader could read both chapters simultaneously. Specifically, in a few cases in chapter 4, I have found it necessary to bring forward some of the results that I later explain in chapter 6. Although I have made my best efforts in making the argument as linear as possible, my suggestion for the reader is to read the papers, in their order of appearance, before the full thesis.

2 Literature

In this chapter, I summarise some of the most influential theories of identity that have been applied in mathematics education research. There exist many theories on identity, and it would be impossible to summarise them all. Thus, inspired by Darragh (2016), I have relied on a limited number of theories, all of which have been used extensively in mathematics education studies. First, I discuss two theoretical frameworks which Darragh (2016) highlighted as representatives of participative perspectives on identity: identities in cultural worlds (Holland et al., 2001) and identities in communities of practice (Wenger, 1998). Subsequently, I summarise Gee's (2000) discursive perspective and Sfard and Prusak's (2005) narrated identities. Finally, I discuss identities in CHAT.

This chapter aims to identify and review common perspectives on identity, not to compare them—that I do in the next chapter. Therefore, I present the following perspectives separately, although many of them are related.

2.1 Identities in cultural worlds

In their seminal book, *Identity and Agency in Cultural Worlds*, Holland et al. (2001) discussed a theory on identity that was influenced by Marxist theories (e.g., Bakhtin, 1981; Vygotsky, 1978). Holland et al. (2001) defined identity in the following way:

> People tell others who they are, but even more important, they tell themselves and then try to act as though they are who they say they are. These self-understandings, especially those with strong emotional resonance for the teller, are what we refer to as identities. (p. 3)

Holland et al. (2001) acknowledged both the cultural and the personal aspects of identity, and the authors presented four contexts in which identities are produced: figured worlds, the negotiation of positions, the space of authoring selves, and the play worlds.

The first of these contexts, the figured world, was defined as an 'as if' realm. The authors used the example of virtual reality as one way of understanding figured worlds. In a game—for example, a computer game—the players act in a virtual 'as if' world, which is constrained by artefacts, rules and possibilities. These cultural artefacts exist both within the software and in the social world of multiple players. Just like the 'as if' world of computer games have their own distinct rules and cultural artefacts, so do other figured worlds, for example, the figured world of mathematics.

One feature of figured worlds is that they are abstractions. They are simplified “realms of interpretation in which particular characters and actors are recognised, significance is assigned to certain acts, and particular outcomes are valued over others” (Holland et al., 2001, p. 52).

Moreover, although figured worlds are distinct from activities, they share the property of being socio-historic, and hence, constantly reproduced. The simplified abstractions of interpretation are constantly negotiated, and the reproduction of figured worlds might be affected by the reproduction of activities. For example, the emergence of computers has changed mathematical activities, but it has also changed the figured world of mathematics, such as the discourse over what it means to be mathematical or what counts as a legitimate proof.

Alternative notions of the abstract structure are ‘figurative worlds’ and ‘narratized/dramatized worlds’. Holland et al. (2001) referred to the Webster’s Third International Dictionary meaning of figurative, that is, “transferred in sense from literal or plain to abstract or hypothetical; representing or represented by a figure” (p. 52). This notion illustrates how figured worlds are shaped by real activities and, subsequently, abstracted into a hypothetical and abstract realm. The narratized/dramatized notions link to the fact that the world is a narrative, often with some standard plots. One example was provided in Skinner’s (1990) studies in Naudada, a Hindu community in central Nepal. For the women in Naudada, a narrated image of ‘good women’ existed.

> Girls are good, hard-working, and obedient daughters.
> Eventually they marry, leaving their natal homes (*maita*) for the homes of their husbands (*ghar*).
> At their *ghark*, good daughters-in-law are obedient, respectful, and diligent in their household and agricultural duties, laboring from dawn to dark for their in-laws.
> As wifes, women devote themselves to their husbands, seeing to their needs and obeying their demands.
> A good woman bears sons to carry the patriline.
> As she gives birth to and raises sons, she attains more status in the household.
> After the marriage of her own sons, she directs the activities of the daughter-in-law.
> A good woman dies before her husband does. (Holland et al., 2001, p. 54)

The second aspect of identification is the idea of positional identities, which has to do with the ways in which people position themselves relative to socially identified others, their sense of social place, and entitlement.

> Relational identities have to do with behavior as indexical claims to social relationships with others. They have to do with how one identifies one’s position

> relative to others, mediated through the ways one feel comfortable or constrained, for example, to speak to another, to command another, to enter into the space of another, to touch the possessions of another, to dress for another, or…to enter the kitchen of another. (Holland et al., 2001, p. 127)

Regarding the situatedness of relational identities, Holland et al. (2001) took no firm position. Sometimes, relational identities are bound to specific contexts, and at other times, they cut across several activities. Holland and colleagues rejected reducing the significance of identity to only the cultural or only the personal element.

Positional identities develop over time, and just like figured worlds are both structuring and being structured, so are people's positions.

> The development of social position into a positional identity—into dispositions to voice opinions or to silence oneself, to enter into activities or to refrain and self-censor, depending on the social situation—comes over the long term, in the course of social interaction. (Holland et al., 2001, pp. 137-138)

The third context in which identities develop is the space of authoring selves, a construct that draws upon Bakhtin's dialogism (e.g., 1981). In effect, people are authoring their identities. However, the means of authoring are culturally bound. That is, the social context provides individuals with both constraints and opportunities for their narratives.

The fourth context in which identities are produced is the play world. Holland et al. (2001) argued that play consists of culturally shaped artefacts and positions. When people act in play, the cultural world becomes internalised, from the interpersonal to the intrapersonal. This perspective relates to Vygotsky's (1978) idea that "every function in the child's cultural development appears twice: first on the social level, and later, on the individual level; first *between* people (*interpsychological*), and then *inside* the child (*intrapsychological*)" (p. 57).

2.2 Identities in communities of practice

In the previous paragraph, I summarised the participative perspective on identity in cultural worlds from Holland et al. (2001). Another participative perspective to identity is Wenger's (1998) theory of identities in communities of practice. The theory builds upon four explicit propositions.

1. We are social beings. Far from being trivially true, this fact is a central aspect of learning.
2. Knowledge is a matter of competence with respect to valued enterprises—such as singing in tune, discovering scientific facts, fixing machines, writing poetry, being convivial, growing up as a boy or a girl, and so forth.

3. Knowing is a matter of participating in the pursuit of such enterprises, that is, of active engagement in the world.
4. Meaning—our ability to experience the world and our engagement with it as meaningful—is ultimately what learning is to produce. (Wenger, 1998, p. 4)

From these propositions, Wenger (1998) conceptualised identity as being relative to 'communities of practice'. A community of practice is the relation between practice and community on three dimensions, as illustrated in Figure 1.

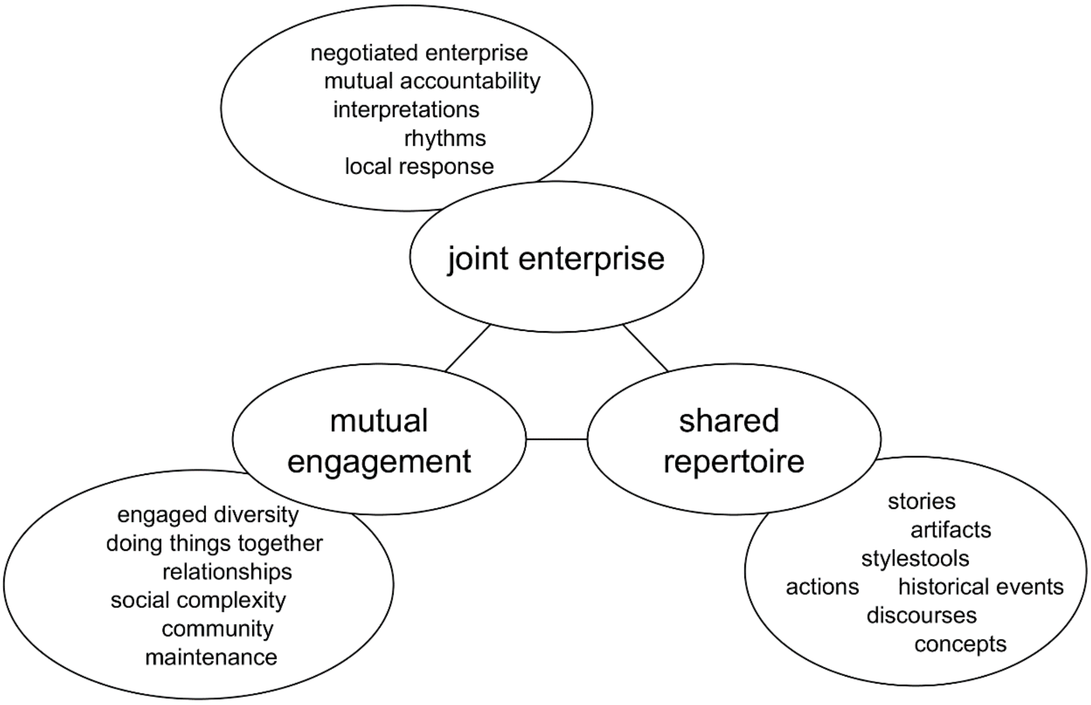


Figure 1. Dimensions of practice as the property of a community

The first dimension is *mutual engagement* amongst participants in a concrete practice. Such practices are not abstractions; rather, they exist in the concrete communities in which participants negotiate meaning. The second dimension is *joint enterprise*. The joint enterprise comprises instrumental aspects, such as the making of physical products, but also personal aspects—being productive, having fun, making a career, and so forth. Joint enterprises are not static. They are constantly negotiated in the community of practice. The third dimension is the *shared repertoire,* which includes "routines, words, tools, ways of doing things, stories, gestures, symbols, genres, actions, or concepts that the community has produced or adopted in the course of its existence, and which have become part of its practice" (p. 83).

Wenger considered identity as a concept that connected the individual with the social. Specifically, identity in communities of practice involves a focus on people; however, the focus comes from a social perspective.

Wenger discussed four issues of identity. First, there is a clear connection between practice and identity, as practice involves the negotiation of being a person in that community. Thus, the negotiation of meaning is also a negotiation of identity. Accordingly, Wenger proposed five characterisations of identity.

> - Identity as *negotiated experience*. We define who we are by the ways we experience our selves through participation as well as by the ways we and others reify our selves.
> - Identity as *community membership*. We define who we are by the familiar and the unfamiliar.
> - Identity as *learning trajectory*. We define who we are by where we have been and where we are going.
> - Identity as *nexus of multimembership*. We define who we are by the ways we reconcile our various forms of membership into one identity.
> - Identity as *a relation between the local and the global*. We define who we are by negotiating local ways of belonging to broader constellations and of manifesting broader styles and discourses. (Wenger, 1998, p. 149)

As can be seen from these characterisations, identities in communities of practice are not static traits, like personal characteristics, but rather a constant process of becoming.

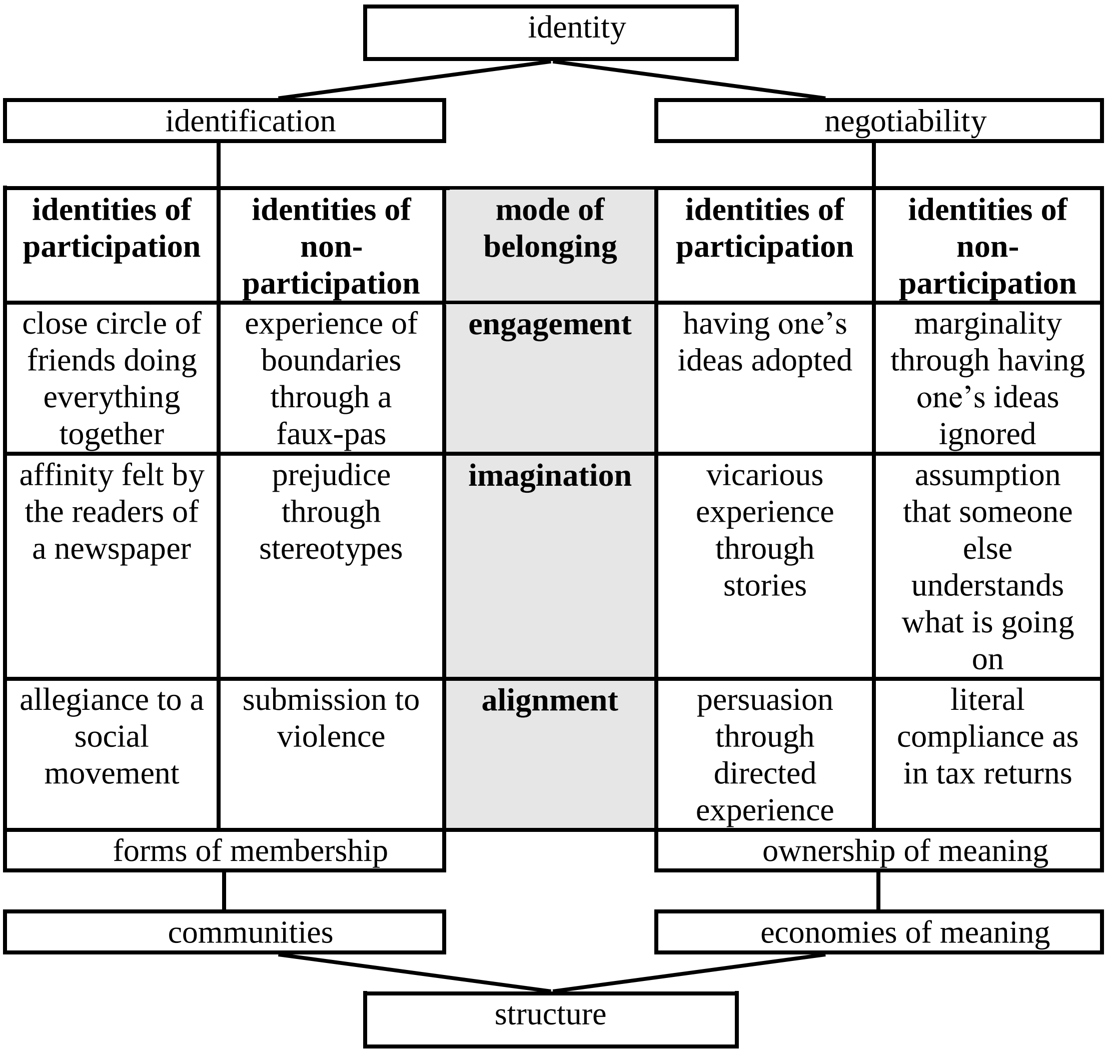
The second issue is that the negotiation of identity includes both participation and non-participation. Wenger distinguished between two forms of non-participation: peripherality and marginality. One way of distinguishing the two is to consider their trajectories. Peripherality often has an inbound path; newcomers who have not yet learned the discourse of the community. By contrast, marginality has an outbound path, in which some participants are marginalised by others (Wenger, 1998, p. 166).

The mix of participation and non-participation shapes fundamental aspects of our lives, such as:

> 1. how we locate ourselves in a social landscape
> 2. what we care about and what we neglect
> 3. what we attempt to know and understand and what we choose to ignore
> 4. with whom we seek connections and whom we avoid
> 5. how we engage and direct our energies
> 6. how we attempt to steer our trajectories. (Wenger, 1998, pp. 167-168)

Third, to understand identity, Wenger argued that identity can be considered as three modes of belonging. *Engagement* is the ongoing negotiation of meaning, the formation of trajectories, and the unfolding of histories of practice (p. 174). *Imagination* is “looking at an apple seed and seeing a tree. It is playing scales on a piano, and envisioning a concert hall. It is entering a temple and knowing that the ritual you are performing is performed and has been performed by millions throughout the world” (p. 176). *Alignment* means to connect to participants in the community, to connect and coordinate energy on the broader enterprise.

Finally, a crucial aspect of identity is that identity formation is a process of both identification and negotiability. The ecology of identity was illustrated by Wenger (1998, p. 190) as in Figure 2.



identities of participation	**identities of non-participation**	**mode of belonging**	**identities of participation**	**identities of non-participation**
close circle of friends doing everything together	experience of boundaries through a faux-pas	**engagement**	having one’s ideas adopted	marginality through having one’s ideas ignored
affinity felt by the readers of a newspaper	prejudice through stereotypes	**imagination**	vicarious experience through stories	assumption that someone else understands what is going on
allegiance to a social movement	submission to violence	**alignment**	persuasion through directed experience	literal compliance as in tax returns

Figure 2. Social ecology of identity

2.3 A discursive theory on identity

Another perspective that has gained much attention is the discursive perspective to identity. Gee (2000, p. 99) is a representative of this position, and he defined identity as “being recognized as a certain ‘kind of person’, in a given context”. He proposed four different, yet related, forms of identity; namely, (N)ature-, (I)nstitution-, (D)iscourse-, and (A)ffinity-identity. Gee (2000, p. 100) summarised these forms of identity as in Table 1.

Table 1. Four ways to view identity

Process		Power	Source of power
1.Nature identity: a state	developed from	forces	in nature
2.Institution-identity: a position	authorized by	authorities	within institutions
3.Discourse-identity: an individual trait	recognized in	the discourse/ dialogue	of/with ‘rational’ individuals
4.Affinity-identity: experiences	shared in	the practice	of ‘affinity groups’

N-identity is “a state developed from forces in nature”. One example is to be recognised as a mathematical genius. In literature, or in the movies, mathematical geniuses are often portrayed as if they were born with some inhuman ‘feel for numbers’. Being a mathematical genius, in this sense, is an identity that develops from biology. A person does not choose to be a mathematical genius; he or she is born as such.

However, being a mathematical genius is an N-identity only as far as it is recognised as such. For instance, some discourses, or some communities, might reject the assertion that the genius characteristics have developed only from biology. Some might even claim that biology plays a marginal role, and therefore, that mathematical genius is not a nature-given identity.

I-identity is a “position authorised by authorities within institutions”. One example, discussed by Gee (2000), is to be recognised as a professor. Being a professor is a position that has been authorised by authorities within the academic community, predominantly other professors. No one is born as a professor, and one cannot appoint oneself to be one.

Gee (2000) argued that the I-identity can be seen as a continuum regarding the extent to which people consent to their identities. Some people feel the identity as a calling, for instance, someone who has long strived to become a professor, and finally has been accepted. Others might feel that the identity has been imposed on them, and Gee used the examples of prisoners or people who have been diagnosed with ADHD.

D-identity is an “individual trait recognised in the discourse/dialogue of/with ‘rational’ individuals”. One example is to be recognised as being mathematical. As a D-identity, being mathematical is a trait. However, it is not a trait that a person is born with, nor is it something that has been authorised. Rather, the trait is something that is recognised by other members—rational individuals—in a community.

Arguably, several discourses in different communities exist. As such, someone who is recognised as being mathematical in one context might not be recognised as such in another. Therefore, being mathematical as a D-identity is a relative identity. Accordingly, two people who are both recognised as being mathematical within different contexts do not necessarily share the same set of characteristics.

A-identity is “the set of experiences shared in the practice of ‘affinity groups’”. For example, being a member of a mathematical community is one A-identity. In this community, members share a set of experiences, such as being intrigued by specific problems, visiting the same internet-forums, sharing stories, a way of talking, and so forth.

It might seem that A-identity is distinct from N-, I-, and D-identities in the way that people can, to a larger extent, choose who they want to be recognised as. However, Gee (2000) stressed that A-identities could be affected by others—for instance, companies which try to bond their customers into certain A-identities.

These different forms of identity do not exist in isolation. It is not the case that some identities are N-identities, some are A-identities, and so forth. Instead, the forms of identity blend, and are constantly negotiated. Again, Gee (2000) used the example of ADHD. Being recognised as ADHD can be seen as an N-identity, as something people are born with. Alternatively, ADHD can be regarded as an I-identity, something imposed on a person by an authority, based on some test or observation. Moreover, ADHD is a D-identity when people, in their discourse, recognise a person as such. Finally, a person can identify him or herself with some affinity group of ADHD.

Gee (2000) discussed some historical aspects of identity. That is, in pre-modern societies, there was an emphasis on I-identities. These identities were often connected with N-identities. The authorities—often the church—decided who should have which position. These positions were often God-given. In ‘modern’ societies, people relied less on identities authorised by others or by nature, and more on D-identities. However, the search for a D-identity was more available for the elite than for people without time and resources who were still constrained to their I-identities. In ‘postmodern’ societies, the A-identity has become more important, due to changes (e.g., socioeconomic changes) in society.

2.4 Narrated identities

The perspectives on identity mentioned so far have been criticised for lacking consensus and clarity. Specifically, Sfard and Prusak (2005) complained that the definitions of identity in the literature were vague. As a solution, they proposed identity to be a series of narratives, instead of narratives being representations of identities.

> We suggest that identities may be defined as collections of stories about persons or, more specifically, as those narratives about individuals that are *reifying*, *endorsable*, and *significant*. (Sfard & Prusak, 2005, p. 16)

> The reifying quality comes with the use of verbs such as *be*, *have*, or *can* rather than *do*, and with the adverbs *always*, *never*, *usually*, and so forth, that stress repetitiveness of actions. A story about a person counts as *endorsable* if the identity-builder, when asked, would say that it faithfully reflects the state of affairs in the world. A narrative is regarded as *significant* if any change in it is likely to affect the storyteller's feelings about the identified person. The most significant stories are often those that imply one's membership in, or exclusions from, various communities. (Sfard & Prusak, 2005, pp. 16-17)

Sfard and Prusak (2005) argued further that, if identity is a story, then every identity can be represented as the triplet ${}_{B}A_{C}$. In this representation, A is the identified person, B the author, and C the recipient. Accordingly, Sfard and Prusak (2005) summarised the individual's different identities in the following way.

> ${}_{A}A_{C}$ = an identifying story told by the identified person herself. This story we call A's *first-person* identity (1st P).
> ${}_{B}A_{A}$ = an identifying story told to the identified person. This story we call A's *second-person* identity (2nd P).
> ${}_{B}A_{C}$ = a story about A told by a third party to a third party. This story we call A's *third-person* identity (3rd P). (Sfard & Prusak, 2005, p. 17)

In this framework, ${}_{A}A_{A}$—the reifying, endorsable, and significant story that people tell themselves about themselves—is defined as 'identity', when no further specifications are made.

Sfard and Prusak (2005) proposed some consequences of perceiving identities as narratives. First, when identities are seen as stories, the researcher overcomes the problem of representations. That is, narratives are not approximate representations of 'something else', narratives *are* identities. In this way, Sfard and Prusak (2005) explicitly distinguished themselves from Wenger (1998) who saw identity as an experience.

Another consequence of narrated identities is that people might tell different stories—the ${}_{B}A_{A}$ and the ${}_{A}A_{C}$ identities, for instance, might look very different. On this issue, Sfard and Prusak (2005) showed little concern, since they claimed that the process of identification is the aim of the research, not the identities themselves. With this objective in mind, each story is valuable, they asserted.

Sfard and Prusak (2005) distinguished between ‘actual’ and ‘designated’ identities. Actual identities are narratives that present the current state of things. Designated identities are stories about what is expected, for instance in the future. From this, learning is the process of closing the gap between the actual and designated identities, Sfard and Prusak (2005) claimed.

In sum, like the other perspectives I have discussed so far, Sfard and Prusak (2005) rejected the idea of identity as a ‘thing in the world’ (p. 21). The greatest distinction between Sfard and Prusak (2005) and the former theoretical frameworks, then, is how identities, in the narrative perspective, is *equated* with narratives. Stories do not represent identities, they are identities.

2.5 Identities in cultural-historical activity theory

All generations of CHAT are related to Vygotsky (e.g., 1978) who built a psychology that explained individuals and the social context by a unifying framework. In doing so, stimulus-response processes were replaced with complex mediated acts, whereby signs and tools served as mediating links.

According to this psychology, the origin of signs and tools is the external world. That is, Vygotsky (1978) explained how external tools become psychological tools from the process of ‘internalisation’. My understanding of this process is that every conscious psychological tool has an external version. For instance, if someone applies some conscious psychological representation to solve the task $\frac{1}{2}+\frac{1}{3}$, then this representation can be explained or illustrated externally (e.g., in words or as a drawing). This is because the representation, whatever it is, has its origin in the external world. However, the reverse is not true: To say that all psychological tools have an external version is not the same as saying that all external tools have a psychological version. For instance, I can visualise that I press the buttons ‘17’ and ‘$\sqrt{}$’ on a calculator, but when I do, I see no output, for example, the fifth decimal to $\sqrt{17}$. This is because I have not internalised the functioning of a calculator. To me, and probably most others, a calculator is an external tool exclusively.

Vygotsky did not emphasise identity or related constructs such as ‘the self’ or ‘personality’ (these constructs were picked up in later generations of CHAT). Nonetheless, a plausible inference of his psychology is that identity—insofar as it is a conscious, mediating, construct—originates in the external world.

In continuing the work of Vygotsky, Leont'ev (1978) had a more collective view of activity and used the collective hunt as a metaphor. Specifically, Leont'ev (1978) defined activity as

> the non-additive, molar unit of life for the material, corporeal subject. In a narrower sense (i.e., on the psychological level) it is the unit of life that is mediated by mental reflection. The real function of this unit is to orient the subject in the world of objects. (p. 3)

Leont'ev (1978) did mention 'personality' relative to the activity, and when he did, he continued the thoughts of Vygotsky, namely, that the direction of personality development is in the order from the external to the psychological, that is, from the activity to the subject. Consequently, in order to understand personality, the chief task is to understand the activity, Leont'ev (1978) asserted.

Later, Engeström (1987) developed a framework for understanding the activity (Figure 3), clearly being influenced by Vygotsky and Leont'ev. This framework provides a tool for structuring important aspects of the activity, and relationships between them, in addition to the relationships between multiple activities.

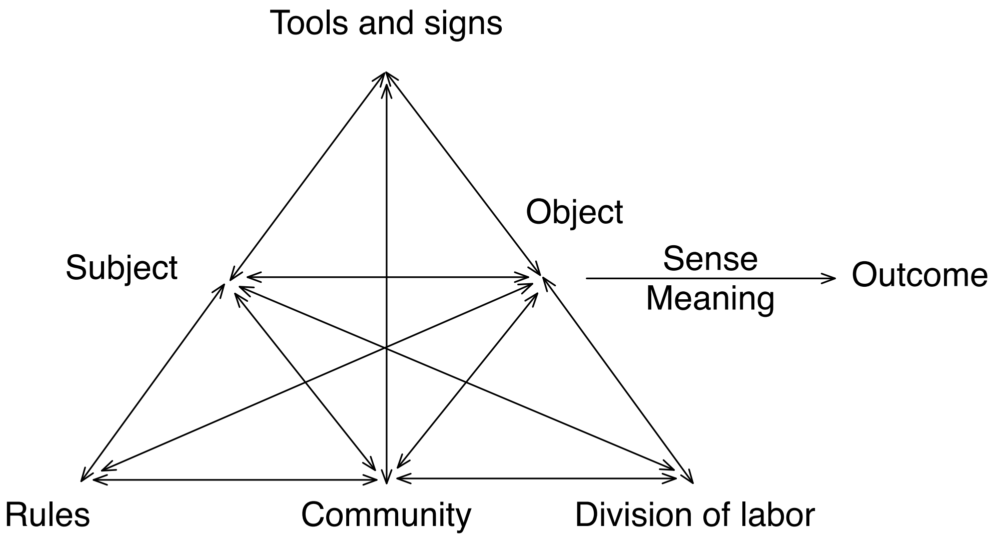


Figure 3. The structure of an activity system as presented in Engeström (1987, p.78)

Stetsenko and colleagues (e.g., Stetsenko, 2013; Stetsenko & Arievitch, 2004; Vianna & Stetsenko, 2011) endorsed the efforts of Vygotsky, Leont'ev, Engeström, and others, to overcome the Cartesian view of identity. However, they—Stetsenko and colleagues—claimed that the resilience to the Cartesian view had marginalised human agency. Notably, Stetsenko and Arievitch (2004) criticised how identity development previously had been portrayed more or less unidirectional, that is, from the activity to the subject.

In response to this critique, Stetsenko and Arievitch (2004) maintained that identity development is dialectical: Humans change activities and activities change humans. Accordingly, identities do not develop in relatively passive participation, but rather, in the relatively agentic process of activity transformation. In short, Stetsenko and Arievitch (2004) valued contribution more than participation.

Common to all generations of CHAT is the materialist view, with explicit roots to Karl Marx. Consequently, structures of activities are always centred around real objects—there exists no activity without an object (Leont'ev used the example of a prey in the collective hunt). Since identity can be understood only in relation to the structure of the activity, for instance, as modelled by Engeström (1987), identity is, inevitably, situated in a cultural-historical context.

Another common theme in CHAT is that uniqueness is a property of subjectivity (e.g., identity, personality, or the self). This property was explicitly expressed by Stetsenko and Arievitch (2004, pp. 476-477) when they claimed that their interpretation of the self in CHAT allows 'us', amongst others, "to define the self as a subject of unique constellation of activities in the real world reflected in a person's 'leading activity'".

2.6 Concluding remarks

In this chapter, I have summarised key theoretical perspectives on identity. I have focused the summary on theories that are frequently applied in mathematics education research, and for the most, these theories perceive identity as more or less socially constructed. This chapter has been descriptive. In the next chapter, I will compare the different perspectives, albeit not directly, but by relating them to theories of measurement.

3 Theoretical framework

In this chapter, I discuss fundamental principles in theories of measurement as they were presented by Thurstone (1928, 1954, 1959). These principles guided the development of RMT and related approaches, such as IRT, but also assisted me in developing the theoretical perspective on mathematical identity that I, later on, describe in this thesis. Subsequently, in this chapter, I discuss existing theories on identity—those presented in the former chapter—in light of Thurstone's requirements. With this chapter, I address research question 1: How do theories on measurement inform a measuring perspective on mathematical identity?

3.1 Principles of measurement

Thurstone (e.g., 1959) claimed that principles of measurement should be the same regardless of the nature of the constructs being measured, for instance, whether they are physical or psychological. Consequently, Thurstone advocated requirements of measurements in the social sciences that, he claimed, were identical to those in the natural sciences.

Thurstone never managed to operationalise his requirements. This accomplishment was made some years later, first by Rasch (1960, 1961) followed by other researchers, such as Wright and Stone (1979) and Andrich (1978).

In this chapter, however, I do not dwell on methodical issues—these I discuss later. Instead, I return to Thurstone's general accounts on measurement. Specifically, I consider three proposed requirements: uni-dimensionality, additivity, and invariance—the same conditions that Andrich (1989) applied when he considered "the distinction between assumptions and requirements in measurement in the social sciences". Moreover, I discuss the property of measures as relational.

3.1.1 Uni-dimensionality

When X and Y are measured successfully, the interpreter concludes that X is either more than Y, less than Y, or about equal to Y. Moreover, the observer can conclude that there is some measured difference between X and Y. Accordingly, a measure is not about 'pure difference', rather, it is a certain magnitude of difference. A measure is about sameness and distinction; it is about more and less, higher and lower, stronger and weaker.

Such distinctions, Thurstone claimed, are sensible only when the measured constructs—weight, depression, IQ, and so forth—are uni-dimensional.

> When we discuss opinions, about prohibition for example, we quickly find that these opinions are multidimensional, that they cannot all be represented in a linear continuum. The various opinions cannot be completely described merely as 'more' or 'less'. They scatter in many dimensions, but the very idea of measurement implies a linear continuum of some sort, such as length, price, volume, weight, age. (Thurstone, 1954, p. 534)

No measure, however, is observed directly—not even height. Instead, measures are inferred indirectly from multiple observations, each supposed to belong to the same, latent, dimension. And in the end, each observation is reduced to a dichotomous, 'yes' or 'no' judgement.

For instance, when we measure the height of a person, we compare the person with a ruler that consists of a finite number of marks, each supposed to belong to the same, abstract, dimension 'length'. And although we might think that we make a direct comparison between the person and the ruler, the comparison is, actually, a series of indirect, dichotomous, judgements: For each mark, we make a 'taller than/shorter than' opinion, appreciating that, occasionally, we make some mistakes.

> In almost every situation involving measurement there is postulated an abstract continuum such as volume or temperature, and the allocation of the thing measured to that continuum is accomplished usually by indirect means through one or more indices. Truth is inferred only from the relative consistency of the several indices, since it is never directly known. (Thurstone, 1928, p. 533)

> It is clear that the linear continuum which is implied in a 'more and less' judgement may be conceptual, that it does not necessarily have the physical existence of a yardstick. (Thurstone, 1928, p. 535)

No construct is truly uni-dimensional since the concept of linearity is merely an abstract idea. Consequently, one is forced to decide as to whether the construct is 'sufficiently' linear, and this decision depends on the problem at hand. For some purposes, a 'straight line' drawn with a pencil is sufficiently linear to be called a straight line, although we know for sure that what we have drawn is not a perfectly straight line when we observe it carefully. For other purposes, the same 'line' is too thick, or too curved to be called a straight line.

The same goes for all theoretical constructs, Thurstone argued. If the construct is concluded to be sufficiently linear for the problem at hand, it can be measured. Otherwise, the construct must be separated into multiple dimensions and measured one at a time.

> It will be conceded at the outset that an attitude is a complex affair which cannot be wholly described by any single numerical index. For the problem of measurement this statement is analogous to the observation that an ordinary table is a complex affair which cannot be wholly described by any single numerical

> index. So is a man such as a complexity which cannot be wholly represented by a single index. Nevertheless we do not hesitate to say that we measure the table. The context usually implies what it is about the table that we propose to measure. We say without hesitation that we measure a man when we take some anthropometric measurement of him. (Thurstone, 1928, pp. 530-531)

Thus, the only situations when a construct is theoretically unmeasurable, not considering methodical or ethical difficulties, would be cases when the talk about finite dimensions is nonsense, that is, when the construct is nonlinear and at the same time too complex or too fluid to be split into a finite number of dimensions.

3.1.2 Additivity

If A, B, and C are measured to be 1.0, 2.0, and 4.0, respectively, then the distance between B and C is, exactly, twice the distance between A and B, ignoring measurement errors. This is the additivity requirement proposed by Thurstone (e.g., 1959).

What has been proven particularly problematic, however, is that counts are not additive. Consider, for example, the following mathematics test.

1. 3 + 4 =
2. 9 + 7 =
3. 12 – 32 =
4. Prove Fermat's last theorem.

Imagine, that person A got 1 point on this test (perhaps he solved the first task correctly), person B got 2 points (she solved the first and the second task correctly), and person C got 4 points.

The difference in raw score between B and C is, indeed, twice the distance between A and B. Nevertheless, from what we can infer from the limited data, most would agree that the real distance between B and C is probably more than twice the distance between A and B. This can be seen from the fact that just about anyone would get from 1 to 2 in a relatively short amount of time. As we know, hardly anyone would ever get from 2 to 4, no matter how hard they tried. This example illustrates how raw scores are not measures.

3.1.3 Invariance

The third requirement of measurement is that of invariance. In essence, this requirement means that an instrument should not be affected by who or what it measures.

> A measuring instrument must not be seriously affected in its measuring function by the object of measurement. To the extent that its measuring function is so affected, the validity of the instrument is impaired or limited. If a yardstick

> measured differently because of the fact that it was a rug, a picture, or a piece of paper that was being measured, then to that extent the trustworthiness of that yardstick as a measuring device would be impaired. Within the range of objects for which the measuring instrument is intended, its function must be independent of the object of measurement. (Thurstone, 1928, p. 547)

In practice, most productive instruments function in a limited range only. A 'regular' ruler, for example, is fit for measuring the length of certain planks but unfit for measuring the size of an electron or the distance to the nearest star. Likewise, instruments that measure cosmic distances might be unsuitable for measuring the length of a plank.

It is worth noting, however, that these differences could be due to practical shortcomings only. That is, a regular ruler is *impractical* for measuring the distance to the nearest star. But the fact that an instrument is impractical is not a violation of invariance.

3.1.4 Measures as relational

The requirements of measurement, discussed so far, build on the assumption that measures in the social sciences are, methodologically, equivalent to physical measures. Consequently, since physical measures are relational, so are measures in the social sciences (later, I make a similar claim, when I argue that mathematical identity must be relational).

The relational assumption means that the value of a person means nothing unless we specify the qualitative nature of this measure, its origin and unit length. Consequently, the height of a person is a value relative to the structure of the marks on the ruler, just as a psychometrical measure is a measure relative to the marks (e.g., attitude statements) on the instrument.

> The only way in which we can identify the different attitudes (points on the base line) is to use a set of opinions as landmarks, as it were, for the different parts or steps of the scale. The final scale will then consist of a series of statements of opinion, each of which is allocated to a particular point on the base line. (Thurstone, 1928, p. 540)

Hence, there must also exist a zero-point and a unit length, and these values are arbitrary.

> Before we can put numbers into these parameters, we must define an arbitrary origin which may be taken as the mean value that one of the stimuli projects on the continuum. As a unit of measurement we may choose arbitrarily the standard deviation of the dispersion which that stimuli projects on the subjective continuum. When that has been done, similar numerical values can be assigned to all of the other specimens that have entred into the comparative judgements. (Thurstone, 1954, p. 49)

Since items take the form of marks on a ruler, there might be some discrepancies between empirical and ‘true’ measures. This approximation was also acknowledged by Thurstone.

> It must be recognized that there is a discrepancy, some error of measurement as it were, between the opinion or overt action that we use as an index and the attitude that we infer from such an index. But this discrepancy between the index and the ‘truth’ is universal. When you want to know the temperature of your room, you look at the thermometer and use its reading as an index of temperature just as though there were no error in the index and just as though there were a single temperature reading which is the ‘correct’ one for the room. (Thurstone, 1928, p. 532)

Furthermore, when these approximate measures are compared with a theory, there will always be some discrepancies between theory and measure. Then, the question is whether such divergences are due to theory-data misfit or due to measurement errors. If there were a theory-data misfit, then researchers would question the theories. If there were measurement errors, one would accept the errors as within the ‘acceptable range’ and, possibly, make efforts to improve the instrument.

One problem, however, is to distinguish measurement errors from theory inaccuracies, and, unfortunately, the history of science provides no external criterion for deciding which is which.

> Scientific practice exhibits no consistently applied or consistently applicable external criterion. ‘Reasonable agreement’ varies from one part of science to another, and within any part of science it varies with time. What to Ptolemy and his immediate successors was reasonable agreement between astronomical theory and observation was to Copernicus incisive evidence that the Ptolemaic system must be wrong. Between the times of Cavendish (1731-1810) and Ramsay (1852-1916), a similar change in accepted chemical criteria for ‘reasonable agreement’ led to the study of the noble gases. (Kuhn, 1977, p. 185)

So far, I have summarised Thurstone’s (1928, 1954, 1959) requirements of measurement. In short, Thurstone argued that these requirements are general. Measures of length, weight, attitude, or level of anxiety are, methodologically, the same thing. Therefore, these requirements will be essential when I, in later chapters, discuss the measurement of mathematical identity. For now, I link Thurstone’s requirements to those theories of identity that were discussed in the former chapter.

3.2 The compatibility between theories of identity and principles of measurement

Since I aim at a deeper understanding of how mathematical identities can be measured, it would be helpful to understand how existing frameworks relate to theories of measurement. Therefore, in this section, I compare theories of identity with Thurstone's requirements of measurement. For the most, I have chosen to focus on theories described in the previous chapter since these are the most commonly applied frameworks in mathematics education research.

The easiest way to assess compatibility between theories is proposition-by-proposition. When their underlying propositions agree, theories are compatible, even if they focus on different consequences. If on the other hand, propositions contradict each other, the theories are considered incompatible, even if some of the consequences happen to coincide (Euclidean and elliptic geometries, for example, are incompatible although there exist results that are true in both). On a general note, since consequences do not imply propositions, there is nothing wrong with the coexistence of multiple, incompatible, theories that all have the consequences that identity can be measured.

However, as Sfard and Prusak (2005) pointed out, theories on identity suffer from vaguely stated propositions and definitions. Thus, in some cases, it has been difficult to compare theories on identity with principles of measurement directly. In some cases, I have, therefore, compared requirements of measurement with general descriptions of identity, and not their underlying assumptions.

3.2.1 Identity and uni-dimensionality

Commonly applied theories in mathematics education tend to treat identity as something multidimensional. Nonetheless, most of these theories portray identity as being about sameness and distinction, as opposed to the unique aspects of people. One specific example was expressed by Gee (2000).

> When any human being acts and interacts in a given context, others recognize that person as acting and interacting as a certain 'kind of person' or even several different 'kinds' at once…. A person might be recognized as being a certain kind of radical feminist, homeless person, overly macho male, 'yuppie,' street gang member, community activist, academic, kindergarten teacher, 'at risk' student, and so on and so forth, through countless possibilities. (Gee, 2000, p. 99)

Being recognised as 'a certain kind of person', I argue, is compatible with the uni-dimensional requirement of measurement, insofar as it is possible to be more or less similar to the abstract 'kind' (e.g., if one could say that person A is 'a radical feminist' but not as radical as person B). That is not to say that such identities *are* uni-dimensional. Rather, it means that some persons are more alike than others, for example, that two individuals are the same 'kind of person', and therefore, that identity must consist of a finite number of dimensions.

This point of sameness and distinction was also expressed by Wenger (1998), although more implicitly.

> An identity, in this sense manifests as a tendency to come up with certain interpretations, to engage in certain actions, to make certain choices, to value certain experiences—all by virtue of participating in certain enterprises. (Wenger, 1998, p. 153)

However, Wenger (1998) also provided some pointers that identities might not be about sameness and distinction, specifically when he argued that identities are non-categorical.

> Identity is not merely a category, a personality trait, a role, or a label; it is more fundamentally an experience that involves both participation and reification. Hence, it is more diverse and more complex than categories, traits, roles, or labels would suggest. (Wenger, 1998, p. 163)

When Wenger's statements are compared in isolation, they appear to contradict each other. That is, it is hard to see how identity can be a tendency but not a category[3]. Thus, I agree with Sfard and Prusak (2005) that Wenger could have been more explicit on his account of identity. Nonetheless, I find Wenger's last description—that identity is non-categorical—to be more representative when I interpret his theory as a whole. If this interpretation is accurate, then Wenger's perspective would be incompatible with the uni-dimensional requirement of measurement.

Most theories perceive identity as multidimensional, although they describe this multiplicity differently. One specific example is how Axelsson (2009) recognised mathematical identity to comprise four dimensions: self-perceived mathematical knowledge, ability, motivation and anxiety (p. 387).

More than being multidimensional, most theories acknowledge that people have multiple identities. Black et al. (2010), for example, argued that people "have a collection of identities upon which to draw at any one moment" (p. 58). Gee (2000) also claimed multiplicity when he argued that each identity has four strands: nature, institution, discourse, and affinity.

In addition, Sfard and Prusak (2005) explained 'multiplicity' on the identified person, the author, and the recipient.

[3] The word 'category' appears repeatedly throughout the thesis. I emphasise that the word is generic, and it must, therefore, be understood in the context in which it is applied. For example, a 'category' Wenger talked about is different from a response 'category' or Kant's description of 'categories' as pure, *a priori*, knowledge.

> As a narrative, every identifying story may be represented by the triple ${}_{B}A_{C}$, where A is the identified person, B is the author, and C the recipient. Within this rendering it becomes clear that multiple identities exist for any person. (Sfard & Prusak, 2005, p. 17)

In sum, I do not consider ‘multiplicity’ to contradict measurement, insofar as multiplicity can be conceptualised as a set of dimensions, each of which can be measured. For natural reasons—namely, that theories on identity were not conceptualised for measurement—few theories discuss the nature of the multiplicity in terms of measurement. Therefore, it is difficult to say whether the selected set of theories contradict the uni-dimensional requirement. Nonetheless, it seems that Wenger’s (1998) perspective is the one that would be most difficult to force into a set of single dimensions.

3.2.2 Identity and additivity

Identities are frequently conceptualised as positions, for example, as expressed by Holland et al. (2001).

> It is important, in understanding positioning, to pay attention to the fact that positional identities develop heuristically over time. (Holland et al., 2001, p. 137)

However, some contrasting views exist, for example, in the writings of Wenger (1998), who conceptualised identity as a trajectory as opposed to a position.

> A community of practice is a field of possible trajectories and thus the proposal of an identity. It is a history and the promise of that history. It is a field of possible pasts and of possible futures, which are all there for participants, not only to witness, hear about, and contemplate, but to engage with. (Wenger, 1998, p. 156)

Few authors address additivity explicitly—again, probably because they were not constructed for the purpose of measurement. Nevertheless, I argue that theories that conceptualise identities as positions, in one form or another, might be compatible with the requirement of additivity. In contrast, I consider Wenger’s (1998) conceptualisation as incompatible with the additive requirement.

> In the same way that meaning exists in its negotiation, identity exists—not as an object in and of itself—but in the constant work of negotiating the self. (Wenger, 1998, p. 151)

> As such, it [identity] is not an object, but a constant becoming. (Wenger, 1998, pp. 153-154)

Although I believe that theories that see identities as positions might be compatible with the additivity requirement of measurement, I do not consider the fact that positions constantly change to be a theoretical problem (although I, in the last chapter, argue that it might have methodical implications). However, a theory must allow identity to be a static picture of this movement. Change would then be a change of positions and require at least two observations. On this point, Sfard and Prusak (2005) were quite explicit when they claimed that "identifying is an attempt to overcome the fluidity of change by collapsing a video clip into a snapshot" (p. 16).

3.2.3 Identity and invariance

The requirement of invariance is highly connected with the structural/agency debate, that is, the debate on the situatedness of identity. In effect, measures require some overlap in the structures—they require that the structure of what is being measured is not entirely situated. Thus, any theory on identity that claims no such overlap to exist would be theoretically incompatible with measurement. To be compatible with the requirement of invariance, a theory must, as a minimum, acknowledge that situatedness is an empirical question.

This point was most explicitly addressed by Holland et al. (2001) who claimed that the situatedness of identity varies.

> Relational identities and the cultural artefacts through which they are claimed may be specific to a figured world. They may have to do with one's honor in the Algerian peasant village Bourdieu tells us about, or one's attractiveness in the sphere of gender relations on college campuses in the United States, or one's machismo in the Nicaraguan village. Other positional identities and markers may, however, be less specific and cut across such worlds. (Holland et al., 2001, pp. 129-130)

The existence of identities that cut across figured worlds, I believe, is one significant distinction between Holland et al. (2001) and identities in CHAT that are defined to be entirely situated within the concrete activity. On this issue, Wenger (1998) agreed with Holland et al. (2001), it seems, when he emphasised 'communities' as the structural frame of reference, however, without the requirement of material objects at the centre.

I continue the discussion on invariance in the last chapter. For now, I suffice to conclude that any theory that acknowledges that there might exist an intersect between social structures across activities, no matter how small, might be compatible with the invariance principle.

3.2.4 Identity and relativity

Most theories on identity in the field of mathematics education conceptualise identity as relational, for instance, as expressed by Holland et al. (2001), Gee (2000), and Wenger (1998).

> We have attempted to articulate the relation of person and society in a way that makes light of neither social life nor the world of the psyche. At the same time, we reject a dichotomy between the sociological and the psychological. 'Person' and 'society' are alike as sites, or moments, of the production and reproduction of social practices. But there is a substantiality to both sides. We object to an anti-essentialism that rotely rejects any sense of durability or predisposition in social life. (Holland et al., 2001, p. 270)

> At one period of history, or in one society, certain combinations result in recognition of a certain sort, while at a different period of history, or in a different society, the same combinations would be unrecognizable or recognized differently…. The combinations (words, deeds, ways of interacting, values, beliefs, etc.) that got one recognized as a saint in the medieval church would, today, in many places, get one institutionalized as a mental patient. (Gee, 2000, p. 110)

> The concept of identity serves as a pivot between the social and the individual, so that each can be talked about in terms of the other. (Wenger, 1998, p. 145)

However, one consequence of a relational perspective on identity is hardly discussed in the literature, namely, that for something to change, something else must be held constant. This is a property that Durkheim recognised.

> For objectivity depends upon the existence of a constant and identical point of reference to which the representation can be referred, and which makes it possible to eliminate everything that is variable and subjective. (Durkheim, 1972, pp. 65-66)

It should be noted here that measurements require one point to be *held* constant, not that it *is* constant. I believe that this difference implies some consequences for how identity can be interpreted, and I will come back to this issue in the last chapter.

Several authors claim that there is an analytical difference between studying social structure and personal positions within the structure, although these facets must be understood relative to each other. Some specific examples were provided by Holland et al. (2001).

> Another facet of lived worlds, that of power, status, relative privilege, and their negotiation, and another facet of lived identities, that of one's self as entitled or as disqualified and inappropriate, must also receive theoretical attention. In order

> to highlight these facets, we make an analytical distinction between aspects of identities that have to do with figured worlds—storylines, narrativity, generic characters, and desire—and aspects that have to do with one's position relative to socially identified others, one's sense of social place, and entitlement. (Holland et al., 2001, p. 125)

> The first context of identity is the *figured world*…. The second context, then, is *positionality*. It is less a separate 'second context' than a (separable) counterpart of figuration. (Holland et al., 2001, p. 271)

In contrast, Wenger (1998) warned against focusing on either the person or the social structure.

> Issues of identity are an integral aspect of a social theory of learning and are thus inseparable from issues of practice, community, and meaning. (Wenger, 1998, p. 145)

> It is therefore a mistaken dichotomy to wonder whether the unit of analysis of identity should be the community or the person. The focus must be on the process of their mutual constitution…. In a duality it is the interplay that matters most, not the ability to classify. (Wenger, 1998, p. 146)

In sum, I consider Thurstone's requirements as more compatible with Holland et al. (2001) than with Wenger (1998) on this issue.

3.3 Concluding remarks

In this chapter, I have addressed research question 1. Specifically, I advocate that a measuring theory on mathematical identity, that is, a theory that allows some persons to identify more strongly with mathematics than others, must consider four aspects: dimensionality, additivity, invariance, and relativity.

It is evident that few theories of identity were constructed for the purpose of measurement. Therefore, the fact that such theories avoid addressing Thurstone's requirements of measurement is by no means a complaint. Nonetheless, I have experienced that it is difficult to decide whether existing theories of identity are compatible with measurement.

However, even if it is difficult to compare theories proposition-by-proposition, I conclude that some ideas on identity are incompatible with theories of measurement. The way Wenger's (1998) conception of identity as a 'constant becoming' challenges measures as static positions is one such example. Other theories seem more compatible. The theory of Holland et al. (2001), in particular, appears more consistent with Thurstone's requirements than Wenger's (1998) theory, although Holland et al. (2001) did not address measurement explicitly.

In summary, I claim that a perspective on mathematical identity that addresses requirements of measurement explicitly is desirable. This will be the topic of the following chapters.

4 Methodology

The argument in this chapter is influenced by a (neo)-pragmatic philosophy on truth and theory, which I describe in the first section. In this first section, I discuss theory-pluralism, and I explain the existence of objective truth relative to a defined theoretical frame.

In this chapter, I address research question 2: What is a paradigmatic method for measuring mathematical identity? If we accept the existence of objective truth, then I claim that some methods of measuring mathematical identity are truly better than others, insofar as the theoretical frame is defined. This issue, I discuss in the subsequent sections. First, I describe two commonly applied theories of measurement—RMT and IRT—both of which relate to principles of interval measurement. I then explain the incommensurability thesis, before I build on Andrich (2004) when I claim that RMT and IRT are situated in incommensurable paradigms. The main philosophical difference is the interpretation of principles of measurement: IRT sees them as assumptions, RMT as requirements.

In the final section, I explain how I perceive the principles of measurement as requirements, and, therefore, I conclude that the application of RMT is a paradigmatic method for measuring mathematical identity.

4.1 Pragmatism, truth, and theory

Quite early in my study, I was intrigued by Charles S. Peirce, William James, and John Dewey, frequently referred to as the classical pragmatists. Later, I was influenced by Thomas Kuhn and Hilary Putnam. Amongst these philosophers, James was the first to use the term ‘pragmatism’, and he proposed the following maxim:

> To attain perfect clearness in our thoughts of an object, then, we need only consider what conceivable effects of a practical kind the object may involve—what sensations we are to expect from it, and what reactions we must prepare. Our conception of these effects, whether immediate or remote, is then for us the whole of our conception of the object, so far as that conception has positive significance at all. (James, 1977, pp. 377-378)

This maxim was the extension of one presented by Peirce some years earlier.

> Consider what effects, that might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of these effects is the whole of our conception of the object. (Peirce, Hartshorne, & Weiss, 1935, p. 1)

From these theses, it is evident that a pragmatic view connects validity with practical consequences of knowledge and theory. To illustrate with two extreme and hypothetical examples: A theory can be perfectly logical, and, yet, the theory will be useless, and hence, according to the maxims, no true theory at all, if accepting/rejecting the theory does not lead to some practical consequences. Conversely, a theory might be proven incomplete, and nevertheless, it can in some sense be true if it is found to be useful (Newton's theory of gravity is an example of this kind).

It is worth noting that Putnam claimed that he was no pragmatist, although he was, admittedly, inspired by this philosophy. Likewise, Peirce refused James' extension, and therefore, claimed that he, Peirce, was no pragmatist either (he preferred the term pragmatisism to make the difference). The aim of this chapter, however, is not to provide a historical description of pragmatism as such. Therefore, I will not pursue the Peirce/James controversy, nor other controversies in the pragmatic discussion any further. Instead, I will interpret the word 'pragmatism' in the broader sense when I describe some ideas on truth and theory that have had a practical impact on my study. In particular, I have been influenced by Putnam's pragmatic realism.

Before the classical pragmatists, Kant proposed a philosophy that introduced a human element to knowledge. His *Critique of Pure Reason* is immensely difficult to read—even harder to understand—and, thus, I will not pretend that I follow his arguments in length. It is evident, however, that his ideas of categories have influenced some of the pragmatic thinkers.

In essence, Kant believed that humans have *a priori* knowledge, that is, categories of interpretation that precede experience. These categories shape the way we understand the world. This view—that knowledge implies a great deal of human interpretation—is particularly visible in Putnam's writings. That is, Putnam rejected the extreme realist perspective in which knowledge is seen as a 'copy' of an external reality and, consequently, that the role of science is to document natural facts as they present themselves. If we turn the attention to mathematical identity, I agree with Putnam when I claim that science cannot represent identity *as it is*; the signs we use to represent mathematical identity do not correspond to some unrelated object.

> In an internalist view also, signs do not intrinsically correspond to objects, independently of how those signs are employed and by whom. But a sign that is actually employed in a particular way by a particular community of users can correspond to particular objects *within the conceptual scheme of those users.* 'Objects' do not exist independently of conceptual schemes. *We* cut up the world

> into objects when we introduce one or another scheme of description. Since the objects *and* the signs are alike *internal* to the scheme of description, it is possible to say what matches what. (Putnam, 1981, p. 52)

An extreme relativist, then, might claim: If the world does not reveal itself *as it is*, that is, if knowledge and theories are merely human constructions that depend on our categories of interpretations, innate or culturally appropriated, then anyone could propose just about any theory. Moreover, since we do not have access to the external reality, any such theory would be equally valid.

However, in agreement with Putnam (1981), I do not believe that a rejection of extreme realism implies the endorsement of radical relativism. It is certainly true, most pragmatists would agree, that there exists no objective rule for determining when a theory is right and when it is wrong. Identity, for example, *is* not, in its absolute existence, a narrative, as Sfard and Prusak (2005) proposed. Nor *is* it a constant becoming as described in the theory of communities of practice (Wenger, 1998), or, what I later propose in this thesis, a relative position.

Nevertheless, even if any argument that tries to decide which perspective that mirrors reality most accurately is doomed to fail, it is not the case that every theory should be valued equally. This is because, according to most pragmatists, knowledge is not *entirely* a human construction. There also is a contribution from nature. It is evident, for example, that Euclidean geometry is a human construct. However, the fact that houses do not collapse is an indication that there also must be something fundamentally right about this geometry, and we can make this claim without asserting that the universe *is* Euclidean, just as we can make a similar argument to advocate that there is something fundamentally right about other geometries, that is, in situations when they work.

To discuss an example closer to our field: When people remember mathematical constructs—algorithms for example—more readily when they understand the algorithms than when they have learned them by rote, it indicates that there is something fundamentally right about principles of reform mathematics, although these principles were constructed by humans.

This is why I, in this thesis, claim that mathematical identity, seen as a relative position, is not entirely a human construction. When we observe that there exist characteristics that align on a single dimension, and that some of these characteristics are person-independent, it means that there is also something fundamentally right about perceiving mathematical identity in this way.

The rejection of the extreme realist and the extreme relationalist positions implies that knowledge is a collection of theories that evolves relative to human experience, as opposed to a growing body of facts. This evolutionary perspective was mainly advocated by Kuhn and Dewey, for example, when Dewey discussed the method of science.

> [The method of science] (1) regards all statements as provisional or hypothetical till submitted to experimental test; (2) endeavours to frame its statements in terms which will themselves indicate the procedures required to test them; and (3) never forgets that even its assured propositions are but the summaries of prior inquiries and testings, and therefore subject to any revision demanded by further inquiries. (cited in Bacon, 2012, p. 54)

As I discussed in chapter 2, the field of identity research consists of a variety of incompatible theories. From my perspective, such multiplicity of theories is no more of a problem than the co-existence of multiple geometries, each usable in certain situations. What is most important is that we have enough geometries to answer our needs, that is, to build bridges, travel to the moon, understand gravity, and so forth. Likewise, I believe that it is more important to have enough theories on mathematical identity to answer relevant questions regarding how students, or people in general, relate to mathematics than it is to unite around one theory of everything. This is why I claim that a theoretical perspective on measured mathematical identities is not taking the place of other theories, but rather, extending the kind of questions that can be asked. If we choose to see identity as relative positions, then both personal identities and social structures can be measured, I claim.

Although practical consequence—what a theory 'can do'—is highly valued amongst pragmatists, internal consistency is also an ideal. In effect, this means that incomplete theories are acceptable as far as they are found useful in one way or another. However, such incompleteness will usually leave a feeling of unease—the theories are accepted, but only in anticipation of better theories. Thus, I partly agree with Sfard and Prusak's (2005) criticism that theories of identity, in general, are too vague, for example, when they lack precise definitions. However, since the value of a theory is a function of both internal consistency and practical consequence, I am not willing to discard every theory that applies vague definitions. Wenger's (1998) framework, for example, was particularly criticised by Sfard and Prusak (2005), but nevertheless, has been proven to be extremely useful in many settings.

The evolutionary nature of theory development means that knowledge evolves dialectically, for instance, between theories, experience, technology, and categories of interpretation. To illustrate, it is no coincidence that Rasch measurement increased in popularity parallel to the rapid development of personal computers since many of the analyses (e.g., maximum likelihood estimations) rely on fast computer calculations. Moreover, without technical possibilities of measuring mathematical identities, I doubt that I would be looking for a theoretical perspective that allowed it to be. But that is not all. From my prior education, I have gained sociocultural categories of interpretation. That is, I have been, and I still am, inclined to interpret experiences in a sociocultural frame. I believe this is one of the reasons why I see mathematical identity as relational as opposed to a static personal trait.

Since multiple theories are accepted, truth cannot exist in isolation. Instead, truth exists relative to some theory, and when it does, I agree with Putnam (1981) that it takes an objective nature. This means that there exists some truth to the sum of angles in a triangle. However, this truth is related to some particular geometry. When the geometry, say, the Euclidean geometry, is selected, then truth no longer depends on what people might think, that is, the truth is no longer conventional (although the choice of geometry might be). If for instance, the majority believed that the sum of angles in a Euclidean triangle was equal to three right angles, then the majority would simply be wrong, convention or not.

In the extension of this argument, I claim that some methods are truly better than others. Likewise, some research questions truly do make sense, however, only relative to a theoretical frame. Questions about the strength of mathematical identity, for instance, *are* not meaningful or meaningless in themselves. Since there exist multiple theories on identity, such questions can be meaningful relative to some theories and meaningless relative to others.

Some of these ideas were also discussed more recently by Radford (2008, p. 322) (although I have not seen that he relates himself explicitly to pragmatism) who argued that a theory is a flexible triplet $\tau = (P, M, Q)$ consisting of principles, methodologies, and paradigmatic research questions. The principles of a theory (P) was considered by Radford (2008) as a relational system, as opposed to a set. Moreover, the associated methodologies (M) support the underlying philosophies of (P). Finally, a theory is defined by the range of questions, (Q), which can be asked. Moreover, the triplet is interrelated in a specific way. For instance, the methodology must be compatible with the basic principles, and suitable for answering the research question. Likewise, the research question must be formulated in a way that is consistent with the principles, and in a way that can be answered from the associated methodology.

4.2 Philosophy of measurement

The pragmatic stance I have discussed so far affects how I regard the nature of the questions that I ask and their associated answers. That is, to the question of a paradigmatic method of measuring mathematical identity, there exists no absolute answer detached from theoretical frames. However, when theoretical boundaries are made explicit, then an answer does exist, and when it does, the answer is objective. This means that the truth value of the answer depends on rational reasoning, not convention.

The key to understanding a paradigmatic method for the measurement of mathematical identity, then, is to understand the basic principles of two theories: (1) a theory of measurement, and (2) a theory of mathematical identity. A paradigmatic method for measuring mathematical identity must be compatible with both.

In this section, I consider the first relationship, that is, the relationship between theory and methods of measurement. Specifically, I discuss two commonly applied methods for measurement in the social sciences, the application of RMT and IRT, both claiming to be consistent with Thurstone's principles of measurement. After a description of RMT and IRT, I explain how these theories differ in philosophy, that is, how they interpret Thurstone's principles differently. This difference has significant consequences, much deeper than technical accuracy, when I in later chapters discuss the second relationship, that is, between the method of measurement and principles of mathematical identity.

4.2.1 Rasch measurement theory

RMT is a psychometrical theory that claims to be consistent with requirements of interval measurement. That is, Rasch (1960, 1961) formulated a model that was additive as opposed to ordinal. This additive property means that a person, or an item, that is measured to be two units is not 'somewhere in between', rather, 'exactly in the middle' of one and three units, if we exclude measurement errors. Moreover, the model requires the data to be uni-dimensional. As such, standardised residuals between data and model are considered to be random noise when data from an instrument fit the Rasch model. Also, Rasch demonstrated how persons and items could be conditionally separated and, consequently, how person measures and item measures could be estimated independently.

In effect, Rasch measurement is analogous to physical measurement. Thus, anything that, in general, is true for physical measures must, in theory, be true for psychometrical measures. A few examples illustrate this point. If the physical height of a person A is measured with a ruler, using even numbers only (including one arbitrary odd number, say 1), and the height of another person B is measured with a ruler using odd numbers only (including one arbitrary even number, say zero), then one can compare the heights of person A and B directly, even if there exist only two common points of reference, that is, zero and one. Equally important: Changing instruments—measuring person A on the odd ruler and person B on the even ruler—does not affect our conclusion, again, ignoring measurement errors.

If we continue this analogy, some more points become apparent. The inclusion or exclusion of marks on the ruler does not affect the height of the person, only the precision of the measure. Moreover, if we measure the height of individuals, the inclusion or exclusion of marks in the range 1-2 meters would affect the accuracy more than what marks in the range 2-3 meters would.

In the Rasch paradigm, every argument in this analogy is true. Accordingly, both items and people appear on the same variable—items being analogous to marks on the ruler. Moreover, just like marks on the ruler have a thickness—sometimes quite thin, like the rulers in students' pencil cases, other times thicker, like the ones teachers use on the blackboard—so does each item have a thickness that represents the standard error of its location.

Moreover, there is a distance between each pair of adjacent items, and this distance is not required to be uniform. At this point, however, a digression is appropriate. On a ruler, the one in your pencil case, for example, the distances between the marks are as close to uniform as the manufacturers managed to make them. While this might be something of an ideal in psychometrics, it is practically difficult to add, say, one unit to an existing item. Consequently, the items in psychometrical instruments often have a non-uniform distribution. The practical consequence of this distribution can be understood as similar to the hypothesised impact of erasing some marks on a (physical) ruler.

If we were to decide whether a person, measured with a ruler, was shorter or taller than, for example, 1.60m, then the likelihood of our conclusion being 'taller' would be a function of the distance between 1.60 and the true height of the person. If the true height of the person were exactly 1.60, then the likelihood of our conclusion being 'taller' would be the same as being 'shorter', namely, 50%. The taller the person, relative to 1.60, the more likely we would conclude the person to be taller than 1.60. The shorter the person, relative to 1.60, the less likely we would be to conclude her to be taller than 1.60.

The dichotomous Rasch model expresses this relationship: If δ is the measure of an item, then, the likelihood of the person, of measure β, exceeding—answering correctly, agreeing, or the like—to the item is:

$$\mathrm{P(X = 1)} = \frac{\exp(\beta - \delta)}{1 + \exp(\beta - \delta)} \quad (1)$$

When $\beta = \delta$, the likelihood equals 50%. As β increases, relative to δ, the probability gets closer to one. When β decreases, the likelihood gets closer to zero.

Andrich (1978) extended the Rasch model to rating scale data and formulated the probability of a person, with measure β, responding in category x, on an item with measure δ, and m inter-category thresholds—τ_k being the kth threshold location. This is known as the Rating Scale Model (RSM). By convention $\sum_{k=0}^{m} \tau_k = 0$.

$$\mathrm{P(X} = x) = \frac{\exp(x(\beta - \delta) - \sum_{k=0}^{x} \tau_k)}{\sum_{n=0}^{m} \exp(n(\beta - \delta) - \sum_{k=0}^{n} \tau_k)} \quad (2)$$

In addition to the RSM, there exist multiple variations of the Rasch model, including the partial credit model (PCM) (Masters, 1982) and multi-faceted models (Linacre, 1989). However, since I, in my study, have relied on the RSM, none of these alternatives will be discussed in depth.

There are multiple algorithms for estimating the parameters in the model. Such algorithms include the Normal Approximation algorithm, Conditional Maximum Likelihood estimation, and Joint Maximum Likelihood Estimation (JMLE).

Since the choice of algorithm, in most cases, has little practical consequence (Linacre, 1999), I made the selection based on convenience—the data in the study was calibrated from JMLE which happened to be implemented in the Winsteps software, that is, the software I used in the analysis. The algorithm can be studied in detail in Wright and Masters (1982). Thus, for now, I sketch an intuitive explanation only.

The JMLE process is successful when, for each item and person, the observed score and the expected score based on the current parameter estimates are the same. This is obtained by improving the parameter estimates by means of Newton-Raphson iteration. Initial estimates for the parameters are established by, for example, setting every parameter to zero, or using the Normal Approximation algorithm with the original data matrix to calculate approximate, initial, values. Moreover, a constraint must be introduced, for example, shifting the values so that the mean item difficulty is zero. Subsequently, an expected matrix, a variance matrix, and a residual matrix are constructed from the initial Rasch estimates. For each item and each person, new values are estimated by correcting the initial values by the sum of residuals divided by the sum of the model variances of the expected observations. The true logic of this procedure lies in the Newton-Raphson algorithm. However, intuitively, a person's (or an item's) initial value is corrected so that, if he, in general, score higher than what is expected by the initial value, the value is shifted to be slightly less, and vice versa. When every persons' and every items' measures are corrected, the original constraint is again made (e.g., setting the item mean to zero again), and the procedure continues until the corrections are 'small enough', where the criterion for 'small enough' is set by the researcher, usually too small to be visible on any graphical output.

4.2.2 Item Response Theory

IRT developed parallel to RMT. Both theories responded to problems with 'raw score measurement' (e.g., classical test theory). Three models are typically used within the IRT-paradigm: 3PL, 2PL, and 1PL, whereby 1PL and 2PL are special cases of the 3PL model.

In dichotomous cases, the 3PL model expresses the likelihood of person j, with measure Θ_j, getting a score of 1 (e.g., answering correctly to an ability test, or agreeing with some statement) on item i, with measure b_i.

$$\mathrm{P}(\mathrm{X}=1)=c_i+(1-c_i)\frac{\exp\left[1.7a_i\left(\Theta_j-b_i\right)\right]}{1+\exp\left[1.7a_i\left(\Theta_j-b_i\right)\right]} \quad (3)$$

The constant, 1.7, is a scaling parameter that eases the interpretation of a for those who are familiar with the normal ogive metric. However, there are no other reasons for including this constant, and thus, it can be omitted easily (DeMars, 2010, p. 14).

The b parameter is the item difficulty, analogous to the δ parameter in the Rasch model. Figure 4 illustrates the case of two items with different b but similar a and c.

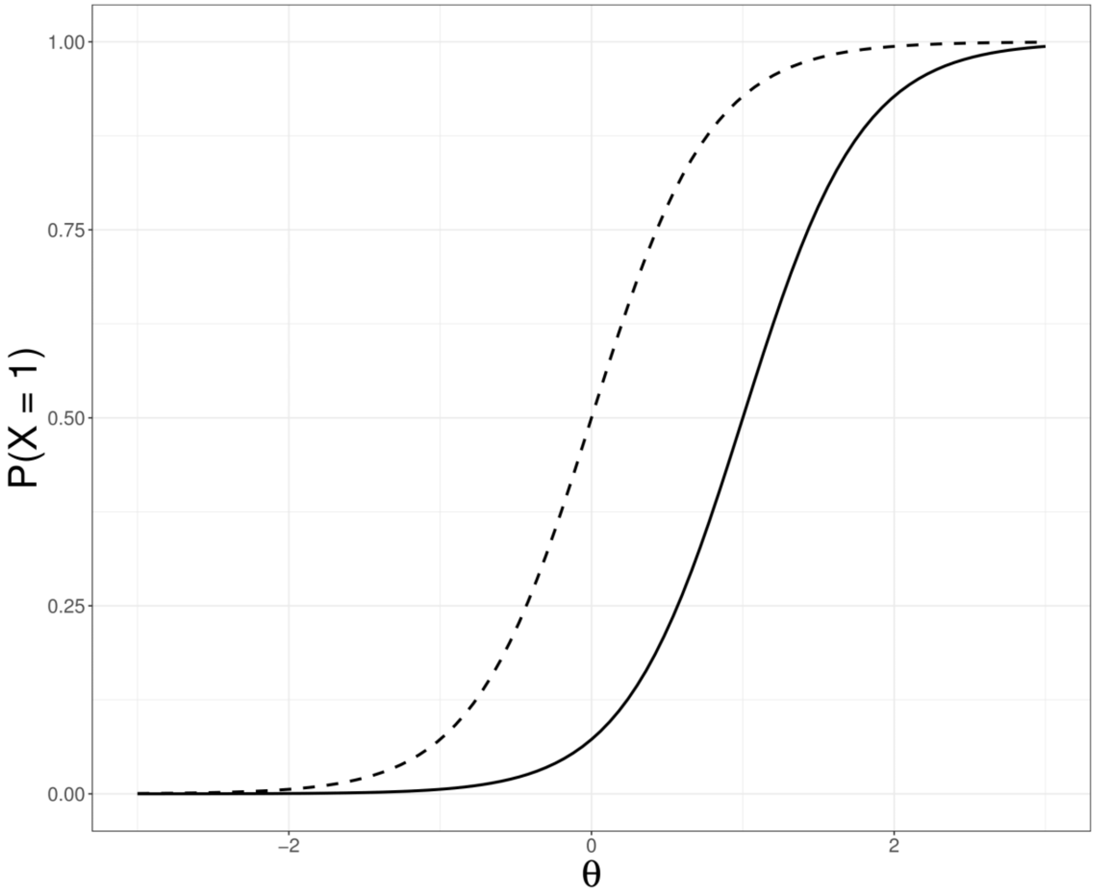


Figure 4. Two items with different difficulties, *b*

The *a* parameter is the item discrimination, that is, the slope of the item characteristic curve (ICC). When *a*=0, there is no discrimination, and the item is, therefore, equally difficult for all ‘ability levels’. When *a* is sufficiently large, the discrimination is perfect, in the sense that (almost) every person with a measure below *b* will ‘fail’ and (almost) every person with a measure higher than *b* will ‘succeed’ on this item. Figure 5 illustrates these hypothetical examples.

In practice, however, the value of *a* typically ranges from 0 to 3 (DeMars, 2010, p. 21). Figure 6 illustrates a realistic variety of discriminations to three items with similar *b* and *c*. Notice how the order of the difficulties changes when person measures go from weak to strong.

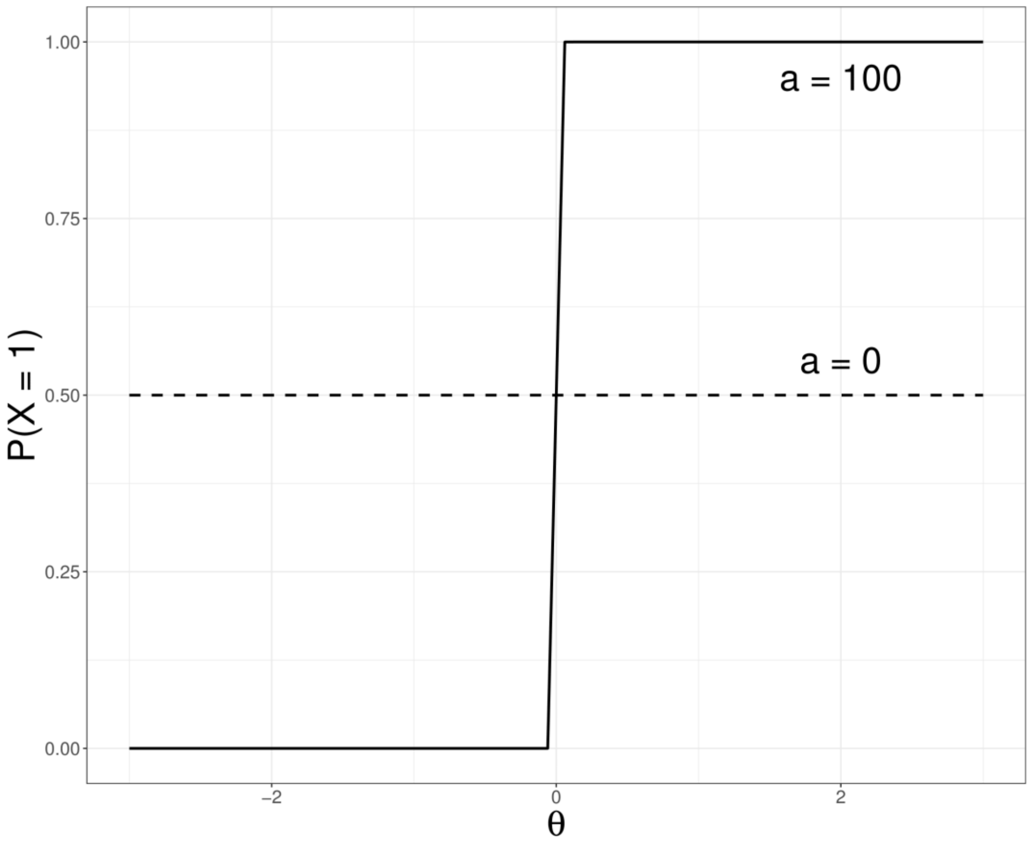


Figure 5. No discrimination vs ‘perfect’ discrimination

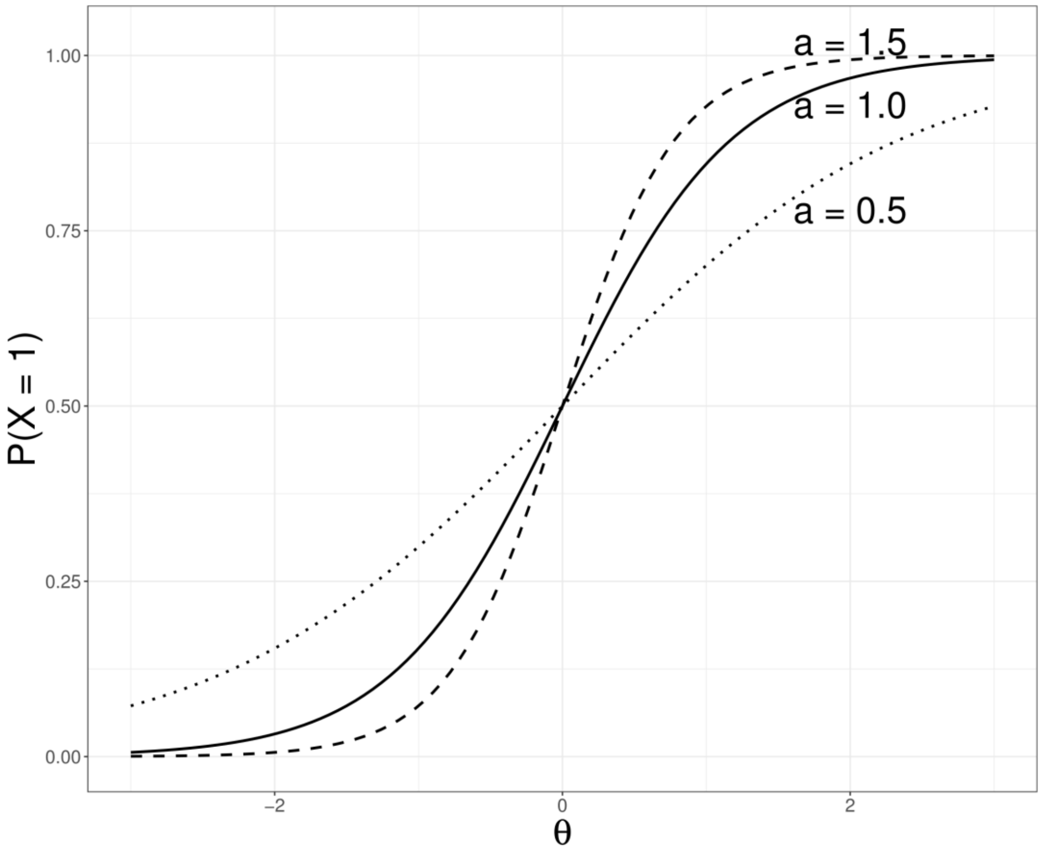


Figure 6. Items with different discrimination, *a*

The *c* parameter is the lower asymptote parameter, and it is frequently referred to as the ‘guessing parameter’. This is because it accounts for guessing on a test, or any phenomena that cause similar shapes of the ICCs. To illustrate, consider two items on a mathematics test. One item is an open question with little chance of guessing the correct answer if one does not know it, and the other item is a multiple choice item. If students were graded on this test, then it is likely that some of them would guess on some of the questions for which they did not know the answer (and, for the sake of the argument, we assume that everyone who did not know the answer guessed on every item for which they did not know the answer). Then, both items in the example could be equally difficult and discriminate equally—they would have the same *a* and *b* parameters. Nonetheless, persons with low abilities (these are the ones most likely to guess) would be more likely to answer correctly to the multiple choice item than the open question. Figure 7 illustrates the case of guessing. Notice how every person is more likely to answer correctly to the items which allow for guessing, although the difficult parameter, *b*, is similar. Moreover, the difference in probability of answering correctly is greatest amongst persons with low measures, since they are more likely to guess than persons with high measures.

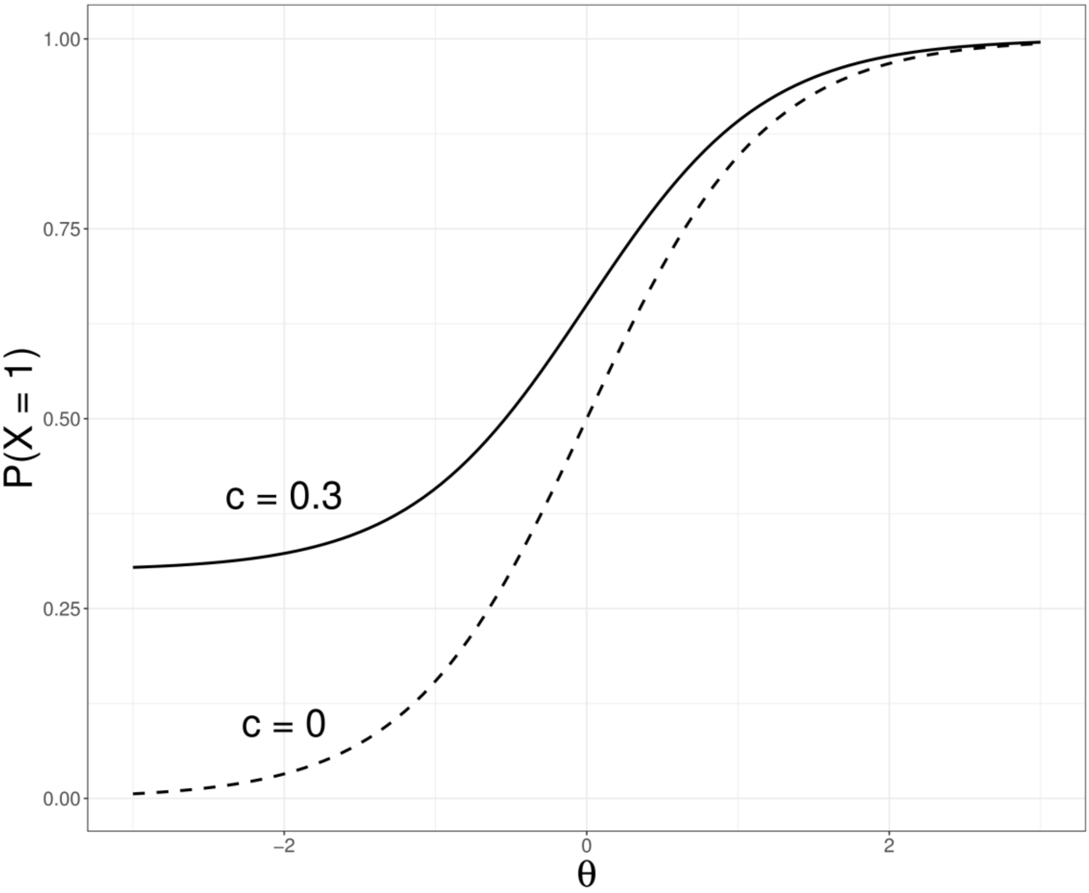


Figure 7. Item with no guessing vs item with some guessing

It has been argued that IRT modelling is sometimes more productive when no more parameters than necessary are included (DeMars, 2010, p. 29). Thus, when there are few indicators of guessing, the c parameter can be constrained to 0, leading to the 2PL model:

$$\mathrm{P(X = 1)} = \frac{\exp\left[1.7a_i\left(\Theta_j - b_i\right)\right]}{1 + \exp\left[1.7a_i\left(\Theta_j - b_i\right)\right]} \quad (4)$$

Finally, when the discrimination, a, is constrained, the 1PL becomes:

$$\mathrm{P(X = 1)} = \frac{\exp\left[1.7\left(\Theta_j - b_i\right)\right]}{1 + \exp\left[1.7\left(\Theta_j - b_i\right)\right]} \quad (5)$$

The intuitions behind IRT and RMT fit statistics are quite similar. Specifically, in each paradigm, the discrepancies between data and model are analysed. A common technique in the IRT paradigm is to group persons with approximately similar Θ and assess residuals between actual and expected responses.

In the RMT paradigm, the sum of squared standardised residuals is, most frequently, assessed over individuals' β (for interpretive reasons, this sum is divided by the number of persons). Moreover, an alternative to this fit statistic is the information-weighted statistic that puts less emphasis on residuals when $|\beta - \delta|$ is relatively large and more emphasis on residuals when $|\beta - \delta|$ is close to 0 (e.g., Wright & Stone, 1979).

Mathematically, the 1PL model is equivalent to the Rasch model, except for the constant 1.7 which is arbitrary, and thus, does not have any practical consequence other than interpretation. Nevertheless, the Rasch model and the 1PL model differ in philosophy. Andrich (1989, 2004) applied Kuhn's notion of paradigms (Kuhn, 1970) and the role of measurement in science (Kuhn, 1977) to discuss these philosophical differences. In the following sections, I draw on this argument to illustrate how RMT and IRT are incommensurable theories.

4.2.3 Paradigms and the incommensurability thesis

In effect, a paradigm is a set of generalisations, beliefs, and values of a community of specialists. Typically, persons within a paradigm share philosophical principles such as ontologies, epistemologies, ethics, and methodologies (Creswell & Clark, 2011, p. 39). Kuhn (1970, p. 110) pointed out that "since no two paradigms leave all the same problems unsolved, paradigm debates always ask which problems are more significant to have solved".

Sometimes, from the starting point of a paradigm, a revolution generates a new paradigm that is incommensurable with the old one (Kuhn, 1970, p. 103). Three aspects of the incommensurability between paradigms were discussed by Kuhn (1970).

First, there are often disagreements on which problems to solve. When paradigm A works with the problems a_1, a_2, a_3, and paradigm B considers problems b_1, b_2, b_3, it is hard to agree on which paradigm works best, since there are no shared problems. If both paradigms recognise the differences in these problems, the paradigms can live 'peacefully' side-by-side, each respecting (though not agreeing with) the other paradigm as a legitimate one. Sometimes, however, the different paradigms can have the false impression that they are working on the same problems. For instance, paradigm A can believe that problem b_1 is the same as a_1, and vice versa. This is likely to happen if one of the paradigms—say paradigm B—is extricated from another paradigm—say paradigm A (Kuhn, 1970, p. 148).

Indeed, when one paradigm is derived from another, the second aspect of incommensurability is likely to appear, that is, the problem with equal terminology used in different ways. Specifically, when a new paradigm emerges, the terminology is often transferred from the old paradigm. Since the new paradigm has a different worldview than the old one, however, the meanings of this terminology can change (Kuhn, 1970, p. 148).

The third aspect of the incommensurability thesis is that different paradigms practice their trades in different worlds. Consequently, not only do they perceive the world differently, but they also perceive relationships differently (Kuhn, 1970, p. 150).

A lot has been said about how research communities deal with controversies. Collins (1975), for instance, argued that not only evidence but also rhetoric is necessary to reach consensus. Moreover, Engelhardt and Caplan (1987) asserted that closure could take place via loss of interest, force, consensus, sound argument, and negotiation (Hess, 1997, p. 99). Kuhn (1970) described 'resolutions of revolutions' to be difficult because different paradigms dictate which problems are important and which criteria to use to decide which methods work best to solve them:

> If there were but one set of scientific problems, one world within which to work on them, and one set of standards for their solution, paradigm competition might be settled more or less routinely by some process like counting the number of problems solved by each. But, in fact, these conditions are never met completely.

> The proponents of competing paradigms are always at least slightly at cross-purposes. (Kuhn, 1970, pp. 147-148)

However, people do convert to new paradigms for different reasons, for instance, the promise to solve problems that led to the ‘crisis’, better precision, simplicity, prospective consequences, or, inevitably, faith (Kuhn, 1970).

4.2.4 Kuhn’s incommensurability thesis—the case of IRT and RMT

Although IRT and RMT seem very similar, even equivalent in the case of 1PL, there are fundamental, philosophical differences. One difference is the ontology of Thurstone’s (1959) principles of measurement: uni-dimensionality, additivity, and invariance. In IRT these are referred to as assumptions (e.g., DeMars, 2010), as opposed to RMT where they are known as requirements (e.g., Rasch, 1960). The distinction between assumptions and requirements is fundamental. Requirements *must* exist while assumptions *do* exist and should be accounted for if they are violated. This distinction was explicitly discussed by Andrich (1989).

Consequently, the ontologies of data and model are also different. In the case of IRT, the 1PL is one of many possible models. As such, it is something amendable. In the case of the RMT, the Rasch model is not a model that explains the data but a definition of measurement, and therefore cannot be changed.

Again, these different interpretations have epistemological consequences when there are discrepancies between the data and the model. In effect, IRT conforms to traditional modelling and works with the model (holding the data sacred), whereas RMT works with the data (holding the model sacred).

In summary, it can be argued that IRT and RMT are anchored in two different paradigms. To understand why proponents of IRT and RMT never managed to agree, I will discuss both traditions in terms of the three aspects of the incommensurability thesis (Kuhn, 1970). A more extensive account can be found in Andrich (2004).

First, proponents of RMT and proponents of IRT seem to be working on different problems, typical for contrasting paradigms:

> The proponents of competing paradigms will often disagree about the list of problems that any candidate for paradigm must resolve. Their standards or their definitions of science are not the same. (Kuhn, 1970, p. 148)

Many of these problems, however, seem to be ‘translations’ of problems that yield both paradigms. The problem with ‘guessing’ and similar phenomena illustrate the point. This issue is evident, regardless of whether the researcher chooses the IRT or the RMT. If we consider how the problem is formulated in the literature, however, we find problems that appear to be ‘translations’ of the problem with guessing. In the case of IRT, the problem with guessing is an expression of the kind: How can we account for guessing? In the RMT, the problem is a formulation of the kind: How can we eliminate guessing? Both paradigms agree that guessing is a problem, but neither of them accepts how the problem is formulated by the conflicting paradigm.

To account for guessing means that guessing is allowed to appear in the data. This is a violation of the requirement of invariance and, according to RMT, cannot happen if the numbers are to be called measures. Conversely, to eliminate guessing means ‘fixing the data to fit the model’, which, according to IRT, is not an option as this is merely cheating with the data.

As we have seen, in IRT, the *c* parameter is included when guessing is observed. In contrast, Waller (1976) proposed a solution in which the data can be reduced in such a way that all responses to items ‘too hard’ for the respondents are removed. On the assumption that persons only guess on hard items, Waller (1976) asserted that most of the guessing would disappear from the data through this procedure. Another account of ‘guessing’ in the RMT was discussed by Andrich, Marais, and Humphry (2012).

A second issue with conflicting paradigms is the interpretations of vocabulary.

> Since new paradigms are born from old ones, they ordinarily incorporate much of the vocabulary and apparatus, both conceptual and manipulative, that the traditional paradigm had previously employed. But they seldom employ these borrowed elements in quite the traditional way. (Kuhn, 1970, p. 149)

Much of the vocabulary in the ‘new’ paradigm (RMT) also appears in the ‘traditional’ paradigm (IRT). The best example is the Rasch model, which is formulated equivalently to the 1PL model but employed differently. What in IRT is ‘one possible model’, is in RMT ‘a definition of measurement’.

Another example is ‘fit statistics’. In IRT ‘fit statistic’ is an indicator of how well the model fits the data. In RMT ‘fit statistic’ is an indicator of how well the data fits the model. The vocabulary is so similar that it is hard to see the difference, but the consequences of bad fit statistics are quite different: In IRT, if the model does not fit the data, one has to fix the model; in RMT, if the data does not fit the model, one has to fix the data.

The problem used to be that proponents of IRT frequently referred to the Rasch model as a particular case of the IRT models. From this stance, not recognising the philosophical differences, it may seem strange why proponents of RMT were so hesitant to include more parameters in the model. Divgi (1986) reflected this view when he considered the Rasch model as "a special case of a more general model" (p. 284). Furthermore, Traub (1983) reacted on the a priori considerations that proponents of the RMT were making. However, when Traub (1983) reacted to the Rasch model, he referred to unidimensionality, invariance, and additivity as assumptions, but these are assumptions only in the IRT paradigm. In the RMT paradigm, they are requirements; proponents of RMT do not hope that they occur, they require them to do so. This serves as an example of how slightly different words, but with fundamentally different meanings, can make competing paradigms 'talk through each other'. Today, these philosophical differences are much more transparent.

A third issue of the incommensurability thesis is how paradigms operate in different worlds.

> In a sense that I am unable to explicate further, the proponents of competing paradigms practice their trades in different worlds…. In some areas they see different things, and they see them in different relations one to the other. (Kuhn, 1970, p. 150)

I assert that this aspect is closely related to the different ontologies. When IRT sees assumptions, RMT sees requirements. When IRT sees a model as something amendable, RMT sees it be a definition of measurement. When IRT sees data as sacred, RMT sees it as something that can be modified.

Moreover, the relationship between model and data is entirely different in the competing paradigms. It is as if they stand on different positions, looking at each other. Proponents of IRT 'stand on the data' and look for an appropriate model, whereas proponents of RMT 'stand on the model' and look for relevant data.

Over the past years, it seems as if the controversy has settled to an 'agreement to disagree'. This is evident in more recent publications where authors from both traditions have contributed (e.g., Nering & Ostini, 2010). Why none of the paradigms has managed to persuade the whole field is difficult to answer. One reason might be that the case is not settled. Kuhn (1970) suggested that a paradigm shift can last for a generation, so it is possible that only one of the paradigms will survive in the future (until it is replaced by yet another one).

In this section, I have explained philosophical differences between RMT and IRT. In the next section, I build on these differences when I discuss what I consider to be the paradigmatic method for measuring mathematical identity.

4.3 A paradigmatic method for the measurement of mathematical identity

If ‘truth’ exists relative to a theoretical frame, then, to understand a paradigmatic method of measuring mathematical identity, we must specify the interpretation of the principles of measurement. According to the description in the former section, the decisive matter is whether the principles are assumptions or requirements. In this thesis, I interpret the principles as requirements for two reasons. First, in Thurstone’s original writings, the principles were referred to as requirements and not assumptions. This is a common argument in the RMT paradigm. More importantly, the analysis of the empirical data rejected the assumption that mathematical identity is entirely context-free. This, and similar results, I believe, should be visible from the process of measurement, and not hidden in model parameters. On this ground, I answer research question 2 when I conclude that the application of RMT is a paradigmatic method for the measurement of mathematical identity.

However, although I am much in favour of the logical arguments posed by proponents of the RMT, I appreciate that, at the time this thesis is written, there is no consensus on the interpretation of Thurstone’s principles. Accordingly, I maintain that the interpretation is arbitrary. Hence, when I say that the application of RMT is a paradigmatic method for measuring mathematical identity, this assertion is not absolute, rather, a premise for the subsequent theorisation of principles of mathematical identity. Specifically, the interpretation of Thurstone’s principles does not only lead to a specific method of measurement, but also forces specific theoretical interpretations of mathematical identity.

To illustrate with a concrete example: In Paper II, I argue that there exist social structures of being mathematical and that these structures are person-independent (I will explain more about this in chapter 6). The ‘choice’ of method, that is, the interpretation of Thurstone’s principles affects how the person-independence is understood—whether it is an assumption or requirement.

If the principle is an assumption, then empirical contradictions will force actions to account for the discrepancies or, alternatively, lead to the conclusion that mathematical identity cannot be measured after all, insofar as the assumption is proven false. If on the other hand, the social structure is required to be person-independent, then empirical ‘contradictions’ will cause a reduction of the structure.

When I assert that the application of RMT is the paradigmatic method for measuring mathematical identity, it follows that the social structure of being mathematical is required, not assumed, to be person-independent. Consequently, in Paper II, I *define* the social structure of being mathematical as the person-independent subset of every characteristic of being mathematical. Thus, any ‘contradiction’ is not a contradiction at all but an indication that some elements are not socially structured.

A consequence of the latter interpretation is that the social structure is not assumed to be equal across contexts, for example, between countries or institutions. What is required is that, if one is to compare measures across context A and B, there must exist a subset of structure A that is similar—in content and structure—to a subset of structure B, whereas the relative size of these subsets is of little importance. In Paper II, I advocate that, when no such subset exists, then mathematical identity is entirely situated between contexts A and B. When structure A is equivalent to structure B, mathematical identity is entirely context-free.

4.4 Concluding remarks

In this chapter, I have discussed research question 2. I have asserted that the measurement of mathematical identity must be compatible with both principles of measurement and principles of mathematical identity. However, there are two theories of measurement that are frequently applied in the social sciences, RMT and IRT, both of which claim to be consistent with principles of measurement. The main philosophical difference between these two approaches is how IRT sees the principles as assumptions, whereas RMT sees them as requirements. Since I interpret Thurstone’s principles to be required, as opposed to assumed, I claim that the application of RMT is a paradigmatic method for the measurement of mathematical identity.

There are theoretical consequences of perceiving principles of measurements as requirements, as opposed to assumptions. This is something that will be discussed in chapter 6 and 7. Before that, I will, in the next chapter, present the methods that I have applied in this study.

5 Methods

In this chapter, I describe the methods I have applied in the empirical parts of the study. First, I describe the research design and data collection. I then present quantitative and qualitative methods and issues of validity. I end the chapter with a discussion on ethical issues.

5.1 Design and data collection

The empirical part of my study had a mixed-methods design, whereby the study is not a mixture of quantitative and qualitative data only, but also one of research site and theory. Methodological textbooks (e.g., Creswell & Clark, 2011) present a variety of mixed-methods designs, all depending on the weight of quantitative and qualitative data, the sequence of the study, and method of merging the data. I find it difficult, however, to position this study within any broad category. In the end, the study has moved dialectically between quantitative and qualitative data and also between theory and data.

5.1.1 Sources for instrument development

The development of an instrument for measuring mathematical identities relied on three primary sources. First, related instruments were examined, and I was mainly influenced by the ASSIST instrument (Entwistle, 1997) that identifies three learning styles: the deep approach, the surface apathetic approach, and the strategic approach. Concrete examples of items that were influenced by the ASSIST instrument are:

- I struggle with putting math problems aside, and
- when I work with a problem, I pause along the way to reflect on what I am doing.

A second source was the literature on the understanding of mathematics. That is, in their classical writings, Skemp (e.g., 1987) and Hiebert (e.g., 1986) discussed ways of knowing and understanding mathematics, for instance, relational (or conceptual) and instrumental (or procedural) understanding/knowledge, and how these ways of understanding are related. Subsequently, much has been said about teaching and learning mathematics for understanding. For example, Kilpatrick, Swafford, and Findell (2001) identified five strands of mathematical proficiency:

- conceptual understanding,
- procedural fluency,
- strategic competence,

- adaptive reasoning,
- productive disposition. (Kilpatrick et al., 2001, p. 5)

Another example is Carpenter and Lehrer (1999) who proposed five mental activities that promote understanding:

- constructing relationships,
- extending and applying mathematical knowledge,
- reflecting about experiences,
- articulating what one knows, and
- making mathematical knowledge one's own. (Carpenter & Lehrer, 1999, pp. 20-21)

Examples of items informed by the literature are:

- I take the initiative to learn more about math than what is required at school/work, and
- math ideas I hear or learn about help me inspire new trains of thoughts.

In addition to external instruments and the literature on mathematical knowledge, information was sought amongst members of mathematical communities. That is, characteristics of working 'deeply with mathematics' were discussed with colleagues of mine at the University. In addition, I corresponded with mathematicians and PhD students in STEM-related subjects (pure mathematics in particular). An example of a response from a lecturer in STEM-related courses is presented below:

> Students who work 'deeply' with mathematics often try to solve the problem in different ways, and they also pursue methods that do not work to find out why they fail. Students who work 'on the surface' are happy when the task is solved. In fact, they avoid multiple solutions since this will only cause unnecessary confusion. To a great extent, these students search 'recipes' that (they believe) can be used in every situation…. Regarding formulas, those who work 'deeply' with mathematics are curious about where the formula comes from. Those who work on the surface do not consider this at all…. Another impression I have is that students who work 'deeply' are more likely to visualise and make sketches, diagrams, etc., as aids for understanding the problem….

Examples of items that were suggested by members of mathematical communities are:

- when I try to use a method that doesn't work, I spend time to find out why it didn't work, and
- if I forget a formula or method, I try to derive it myself.

It should be noted that most items could be related with multiple sources.

5.1.2 Rounds of piloting and participants in the study

Rounds of piloting were conducted to increase validity and reliability. In the first pilot, 50 items were administered to 88 TE students in their second year of education. The students were studying to become teachers in grades 1 to 10 (elementary and lower secondary school). I will not discuss the analysis of this pilot in detail but suffice to say that it followed the same framework as the final analysis, a framework that will be discussed later. As a result of the first analysis, and in particular, as a consequence of the items' locations on the variable, nine new items were formulated intended to fill the greatest gaps on the variable. Subsequently, another pilot was conducted with 45 TE students in their third year of education, before a final set of 30 items were chosen.

Table 2. Participants in the study

Pilot 1			
	88 TE		
Pilot 2			
	45 TE		
DP 1			
	185 STEM		
		72 Calculus 2	–1 N/A grade
		48 Pre-calculus	–1 N/A grade
		65 Calculus 3	
DP 2			
	187 STEM		
		125 (Norm.) final year students	–6 N/A grade
		12 Cryptography	–1 N/A grade
		50 Calculus 3	–2 N/A grade
Total	505		

Note. N/A did not report grades. DP = Data point. (Norm.) final year students means students who started their education in 2010

Following the rounds of piloting, the instrument was administered to a convenient sample of 185 STEM students at a Norwegian university. For the last sub-study (Paper III), an additional sample of 187 STEM students was selected. In Table 2, I explain which courses the students attended. It should be noted that students in the same courses attended a variety of study programmes. I discuss more about this in the last chapter.

Missing data include 11 respondents who did not report their grades. Therefore, the sample in Paper III is reported to be 361. The sample is summarised in Table 2.

5.1.3 Comparing web-based and paper-based questionnaires

Most respondents answered to a traditional paper questionnaire, except the subsample of 125 students in their normalised final year of education who, for practical reasons, responded to a web-based version of the instrument. To study whether the difference between web-based and paper forms affected the structure of the responses significantly, the measures of the final year subsample were analysed twice: first relative to the structure of mathematical identity calibrated on all STEM students, and then relative to the structure calibrated on final year students only. Thus, the normalised final year students (*n* = 125) were compared with themselves relative to two structures: one in which paper-based responses were included and one in which paper-based responses were excluded. The correlation between students' measures relative to these two structures was *r* = 1.00. Thus, from this analysis, there is no evidence that the form of the instrument affected personal responses significantly.

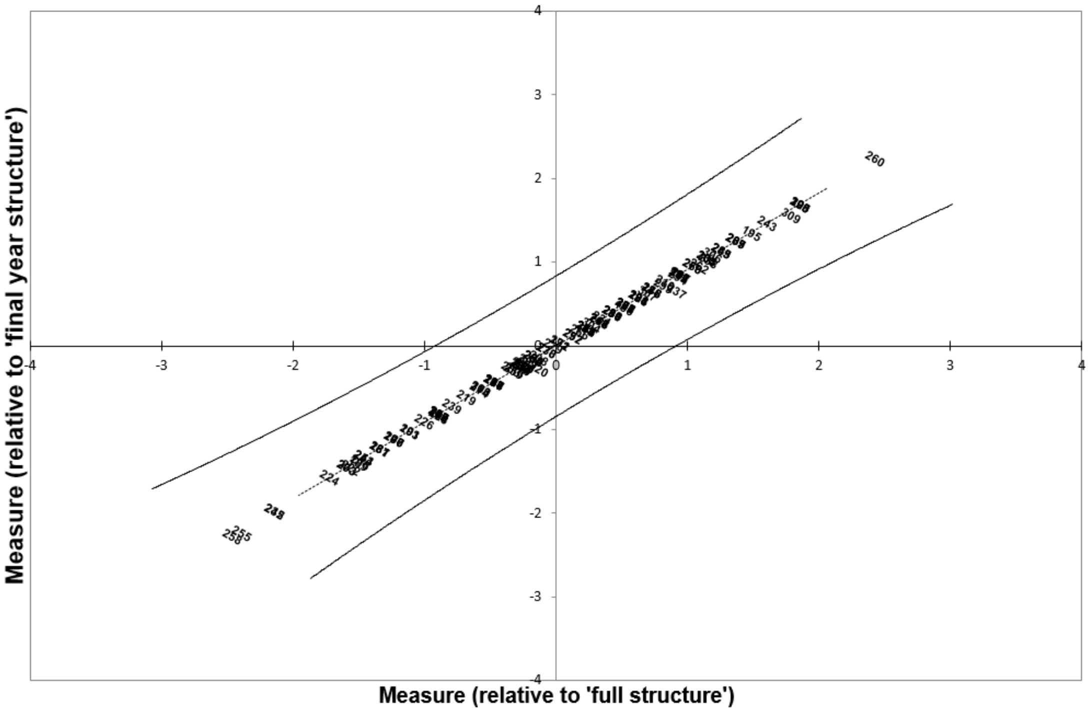


Figure 8. Person measures of the 125 students taking the web-based questionnaire relative to: 'full structure' vs 'final year structure'

The items were administered in random order. Due to missing data being relatively unproblematic in the RMT paradigm (i.e., relative to other statistical techniques), as it is evident in the calibration algorithm (see, e.g. Wright & Stone, 1979), the pilot data were included in the subsequent full analysis.

5.1.4 Methods for the qualitative study of four STEM students and their mathematical identities

Four STEM students were interviewed and asked about how they perceived their time at the University, which project they were currently engaged in, how they worked with mathematical problems, and how they learned new mathematics. Two of the interviewees were selected on the basis of having high, the other two on the basis of having lower measures. All interviews lasted for 60-90 minutes, and the interviews were conducted in the University canteen, as all interviewees preferred this location. The design of the interview, including the forms of the questions, was influenced by guidelines presented by Kvale and Brinkmann (2009). Specifically, the interviews aimed to get some detailed illustrations on characteristics of mathematical identity, rather than a broad narrative. The interview-guide is presented in Appendix G.

The first student, called ‘student A’ in Paper II, studied pure mathematics. He attended the University straight from upper secondary, and he claimed that he was “very motivated”, although he had experienced a slight drop in motivation the last year. In lower secondary, he attended a “special maths class” that consisted of students that were especially interested in mathematics. At the time of the interview, he was working on a project with the aim of improving the validity of mathematical models when there is significant uncertainty in empirical observations.

The second student, ‘student B’, started his higher education as a TE student before he shifted to pure mathematics. His interest in mathematics was influenced by activities in TE, he said. At the time of the interview, he was working on a project on the modelling of efficient maintenance of machines (e.g., in oil platforms).

The third student, ‘student Y’, was studying physics. He claimed that he got very high grades in upper secondary, but that he was shocked when he faced university mathematics. During the interview, he referred to mathematics as “a necessary evil” and “a tool for solving physical problems”. When the student talked about his friends, he characterised them as “more clever than him”.

The fourth student, ‘student Z’, was studying computer science. He told that he had “never been particularly bright in mathematics”, but in upper secondary, he got an interest in computer programming. On his spare-time, he was involved in sports, and this occupied much of his time and energy, he said. He explained how he found it difficult to start studying mathematics when he came home from training. At the time of the interview, he was working on a project with the aim of developing a web application. There were five students in the project, and ‘student *Z*’ was responsible for what he called “the basic programming”. He said that it was much to do, but nothing particularly demanding. Moreover, the student explained how “the others” were able to apply mathematics to construct more efficient scripts, whereas his main strategy was to find solutions on the internet.

In this study, I consider the profiles of the respondents to be of little importance. This is because the aim of the interviews was not to understand individuals as such, rather, to find qualitative illustrations of the characteristics, for example, what ‘struggling with putting mathematics aside’ might ‘look like’ in practice. When I make this claim, I am making the assumption that the meaning of, say, ‘struggling with putting mathematics aside’, does not depend on the profile of individuals, for example, whether they are males or females, or if they have high or low measures. When I, later, discuss differences between TE and STEM students, I illustrate how this assumption is not always true. Thus, future studies might assess which aspects—gender, nationality, age, and so forth—that affect the meaning of characteristics.

5.2 Quantitative analysis and validity

In my study, I have relied on a framework for validity, presented by Wolfe and Smith (2006a, 2006b) who extended Messick’s (1995) validation framework with two aspects of evidence put forth by the Medical Outcomes Trust (MOT). I indicate this as a *technical* validity, because the analysis is done through calculations and checks against criteria for the different types of validity: content validity, substance validity, generalisability, structural validity, external validity, and responsiveness. In this section, I explain the methods I have applied, not their results. The results will be discussed in the next chapter and Paper I (pp. 128-134).

To ensure *content validity*, Infit and Outfit Mnsq and Zstd were used to assess data-to-model fit. Outfit Mnsq is a statistic based on the mean of squared standardised residuals, and Infit Mnsq is an information-weighted sum (Bond & Fox, 2003, p. 238).

When person *n* responds to item *i*, the standardised residual of this response is defined to be:

$$z_{ni} = \frac{\left(x_{ni} - E(X_{ni})\right)}{\left(Var(X_{ni})\right)^{1/2}} \qquad (6)$$

where,
x_{ni} is the observed response,
$E(X_{ni})$ is the expected response, and

$Var(X_{ni})$ is the variance (e.g., Wu & Adams, 2012, p. 341).

Following from this, the Outfit Mnsq statistic is defined to be:

$$\text{Outfit Mnsq} = \frac{\sum_n z_{ni}^2}{N}$$
$$= \frac{1}{N}\sum_n \frac{\left(x_{ni}-E(X_{ni})\right)^2}{Var(X_{ni})}, \quad (7)$$

where N is the number of respondents (p. 341).

The Infit Mnsq statistic is defined to be:

$$\text{Infit Mnsq} = \frac{\sum_n z_{ni}^2 W_{ni}}{\sum_n W_{ni}}, \qquad (8)$$

where W_{ni} is the variance for the response of person *n* on item *i*. Infit Mnsq is weighted so that the statistic is less affected by outliers, as compared to the Outfit Mnsq.

If every response equals the expected response, then Outfit Mnsq and Infit Mnsq will be zero. In practice, however, few responses equal the expected value. Thus, given 1.0 as the modelled expected value, every Outfit Mnsq/Infit Mnsq below 1.0 is considered to be over-fitting, and every value over 1.0 under-fitting.

From the formulas presented by Smith, Schumacker, and Busch (1998, p. 78), critical values for Infit Mnsq and Outfit Mnsq were set to 1.1 and 1.3 respectively. By convention, critical values for |Zstd| were set to 2.0, and the cut-value for item-measure correlation was set to .40.

To find evidence for *substantive validity*, Linacre’s (2002) eight aspects for well-functioning rating scales were evaluated.

1. Each rating scale category should contain more than 10 observations,
2. the shape of each rating scale distribution should be smooth and unimodal,
3. the average respondent measure associated with each category should increase with the values of the categories,
4. the category Outfit Mnsq fit statistics should be less than 2.0,
5. step calibrations should advance,
6. ratings should imply measures, and measures should imply ratings,
7. step difficulties should advance by at least 1.4 logits, and
8. step difficulties should advance by less than 5.0 logits.

Furthermore, several post hoc categorisations were considered to assess the empirical consequences of merging response categories. In addition, person fit statistics, and qualitative judgements about items' placements on the variables were evaluated.

To find evidence for the *generalisability* aspect of validity, analysis of invariance was conducted through differential item functioning (DIF): "the loss of invariance of item estimates across testing occasions" (Bond & Fox, 2003, p. 309). That is, the comparison between two items should be independent of which persons are being used in the calibration (Rasch, 1961, pp. 331-332). Specifically, DIF analysis was conducted to test the hypothesis that item measures remained invariant between institutions (TE and STEM), personal positions, and data points.

Evidence for *structural* validity was sought in the examination of the dimensionality of the items by principal components analysis (PCA) of standardised residuals. Moreover, person measures based on items from different sub-dimensions were correlated to examine the impact on those dimensions, and the degree of local dependency between items was assessed.

As evidence of *external* validity, theory-based predictions were evaluated. Specifically, measures from the mathematical identity instrument were compared with self-reported average grades on mathematics courses with the prediction that high measures are positively related to attainment. This study is discussed in Paper III.

To seek evidence for *responsiveness* validity, the person-item map was examined to assess whether the items were well targeted for the sample.

The Winsteps (Linacre, 2006) software was used in the analysis. An additional analysis was conducted using the RUMM2030 software (Andrich, Sheridan, & Luo, 2010) to ensure that the choice of software did not have any practical effect on the outcome. I will not discuss the RUMM2030 analysis any further, but I conclude that the consequences of choosing Winsteps are at the level of presentation, that is, the kind of statistics, graphs, and tables that are presented in the papers.

5.3 Qualitative analysis, trustworthiness, generalisability, and importance

The analysis of the interview data was conducted using codes from the quantitative data (i.e., the characteristics of mathematical identity) in addition to general theoretical codes, that is, *tools*, *rules*, *community*, *division of labour*, and *objective* (Engeström, 1987). The NVIVO software was used in the qualitative analysis.

Since Wolfe and Smith’s (2006a, 2006b) aspects of validity are mostly related to quantitative studies, a supplementary framework was considered to find evidence for validity of the qualitative data, namely, Schoenfeld’s (2007) aspects of trustworthiness, generalisability, and importance. In this section, I discuss these aspects.

First, Schoenfeld addressed the descriptive power of research, which is “the capacity of theories or models to represent ‘what counts’ in ways that seem faithful to the phenomena being described” (p. 83). He illustrated this issue with an example.

> Consider a typical related-rates problem that involves a ladder sliding down the side of a building. The building is assumed to be (or explicitly stated to be) vertical and the ground horizontal. In the diagram representing the situation, the ladder, the building, and the ground are represented as *lines* that comprise parts of a right triangle. What matters for the purposes of the desired analysis are their lengths, and the way the ladder is moving. That information, properly represented and analyzed, enables one to solve the given problem; that information and nothing else is represented in the diagram and the equations one derives from it. What does not matter (indeed, what would be distracting in this context) includes how many rungs are on the ladder or how much weight it might support. In a different context, of course, such things would matter quite a bit. (Schoenfeld, 2007, p. 83)

In my study, I propose a measuring perspective on mathematical identity. I am, therefore, emphasising issues related to measurement, including the dimensionality of mathematical identity. Consequently, I am ignoring other aspects that I consider of less importance regarding measurement. Such elements include uniqueness of data, that is, data that happen only once and which describe unique qualities of individuals or activities. When I do this, I am not disregarding their importance. However, regarding measurement, these aspects are like rungs on the ladder.

On another note on the descriptive power: In Schoenfeld’s example, the reduction is not an exclusion of redundant information only, but also a simplification of the information that matters. When he said, for example, that the building is assumed to be vertical, this assumption is clearly false since no building is perfectly vertical. Nonetheless, the assumption—that is, the simplification—is warranted insofar as not making the assumption would disrupt the analysis without any practical gain.

In my study, I make similar simplifications. For instance, when I claim that mathematical identity consists of a finite number of invariant dimensions, I am aware that this assumption, strictly speaking, must be wrong. This is because linear dimensions are nothing but abstract ideas—they can never be shown. The only time we can ‘see’ a line is when it is not a line (this holds for a psychometrical dimension as well—it can only be seen when it is multidimensional). Nonetheless, I claim that this simplification is warranted, because, if the multidimensionality has practical significance, I propose the inclusion of more dimensions to capture this complexity.

Another simplification is the idea of social structure as person-independent. I will explain this point later, but in short, it means that a social position does not depend on personal positions. This idea is clearly a reduction since no social positions are fully person-independent.

Schoenfeld proposed one more issue related to trustworthiness, namely, explanatory power—"the degree to which a characterization of some phenomenon explains how and why the phenomenon functions the way it does" (p. 83). This thesis is, admittedly, more descriptive than it has explanatory power. I explain key elements of mathematical identity, but the perspective provides few clues to why mathematical identity develops the way it does.

However, I do believe that descriptive and explanatory powers relate. That is, I consider the conceptualisation of a measurable mathematical identity to be a potential tool for further explorations on how identities develop. In effect, I agree with Kuhn (1977) that the role of measurement in scientific work is to describe phenomena—not explain them. Measures, in general, provide little information on the nature of the measures, that is, explanatory power. To this end, qualitative work is usually necessary.

> Only a minuscule fraction of even the best and most creative measurements undertaken by natural scientists are motivated by a desire to discover new quantitative regularities or to confirm old ones. (Kuhn, 1977, p. 187).
>
> New laws of nature are…very seldom discovered simply by inspecting the results of measurements made without advance knowledge of those laws.... Because nature itself needs to be forced to yield the appropriate results, the route from theory or law to measurement can almost never be travelled backwards. (Kuhn, 1977, p. 197)

One particular case illustrates this point. I will later show how the characterisic 'taking the time to find better methods' was *structured* differently in the TE and STEM contexts. This result was shown by measures. However, the measures provided no clue as to why this characteristic was structured differently. Qualitative interpretations, however, suggested one explanation: TE students were expected to search for multiple methods whereas tasks in the STEM context often required one particular method, either explicitly ("use Newton's method to solve…") or implicitly, as described by one of the participants.

Y: Here is another thing I don't like about mathematics. Usually, it is given what method you should use. And, if you use other methods, well, you rarely have multiple options, it's only one.

Kuhn (1977) explained further how measurements have been particularly significant in cases when they show anomalies to existing theories.

> To the extent that measurement and quantitative technique play an especially significant role in scientific discovery, they do so precisely because, by displaying serious anomaly, they tell scientists when and where to look for a new qualitative phenomenon. To the nature of that phenomenon, they usually provide no clues. (Kuhn, 1977, p. 205)

In sum, I consider this thesis to have descriptive power, and that it also has the potential of explanatory power when it is connected with qualitative interpretations.

Another aspect of trustworthiness, as explained by Schoenfeld (2007), is prediction and falsification. Drawing on Popper (1963), every good theory should be testable: It should be possible to prove a theory wrong.

The theoretical result that I present in this thesis is more a collection of theoretical definitions, concepts, and principles than it is a falsifiable theory. However, I believe that the proposed perspective relates to falsification, and the empirical results illustrate this. For example, when I claim that "taking the time to find better methods is *structurally* easier in the TE context than in the STEM contetxt", this claim is falsifiable by applying the theoretical framework and the associated methods in this thesis. Another researcher could replicate the study to prove this general claim to be false.

Earlier in this chapter, I argued that 'truth' exists only in relation to a theoretical frame. For example, there exist results that are true in quantum theory but false in Einstein's general theory, and vice versa. Likewise, there are results that are true in Euclidean geometry but false in elliptic geometry. Hence, since a claim can be both true and false, I consider 'falsification' of truth claims to be situated. That is, a claim can be falsifiable as long as it is either tested within the same theoretical framework as the original claim or tested against practical consequences.

Schoenfeld (2007) distinguished between four forms of generalisability. *Claimed* generalisability is "the set of circumstances in which the author of that work claims that the findings of research apply". *Implied* generalisability is "the set of circumstances in which the authors of that work appear to suggest that the findings of the research apply". *Potential* generalisability is "the set of circumstances in which the results of the research might reasonably be expected to apply". *Warranted* generalisability is "the set of circumstances in which the authors have provided trustworthy evidence that the findings do apply" (p. 88).

My claimed generalisability in this thesis is that there exist person-independent social structures of being mathematical in at least two contexts—the TE and the STEM context. I also claim that the structure of mathematical identity is relatively stable between these contexts. These claims, I believe, are warranted, since I provide evidence of the findings.

A potential generalisability, one that is not supported with evidence, is that the proposed framework in this thesis is applicable in the broader field of identity research. For any identity to which some persons relate more strongly than others, the results in this thesis might apply. When this is true, personal identities and social identities are measured simultaneously.

The aspect of importance is related to the question "why should one care?" (Schoenfeld, 2007). In effect, this aspect is something that has evolved together with the study itself. As I was studying the most commonly applied theories on identity (e.g., Sfard & Prusak, 2005; Wenger, 1998), I experienced difficulties in measuring such identities. Thus, the importance aspect is related to the experienced incompatibilities between theories on identity and theories on measurement.

One reason why I perceived this to be a problem was less scientific, I admit. Namely, at an early stage, at the time when I was working on Paper I, my 'common sense'—my predisposed categories of interpretation—told me that some persons relate to mathematics more strongly than others. This is because, without any particular theory in mind, I could think of people I knew whom I was quite sure identified strongly with mathematics. Correspondingly, I could think of people whom I assumed identified poorly with mathematics. If this intuition was right, there should be a theoretical perspective that allowed it to be.

On this thought, it should be noted that a consequence of the subsequent theorisation is that it is not possible for one person to *have* a stronger mathematical identity than another. The theoretical perspective rejects the common sense view that initiated it (this will be discussed in more detail in the last chapter). One important aspect of this thesis, therefore, is that it provides a measuring perspective on mathematical identity without implying that identity is something we have.

5.4 Ethical considerations

Ethical considerations were, in this study, grounded in national guidelines for research (Kalleberg et al., 2006). The guidance consists of 46 aspects, which I will not recite in full.

Written and oral consent was gathered at each data point. Appendix D shows the written form. I emphasised that the participants could withdraw their interviews at any time without giving any reason or excuse. If they did, I promised that all data would be deleted and that they would not be mentioned in any form. In addition, I informed the participants that they would be anonymous, but I could not promise that no one would recognise who they were (for example, friends might 'see' that it was them by the way they talked or the way they were described). The data collection has been reported and approved by the Norwegian Centre for Research Data. According to this report, all data that can be linked to individuals will be deleted after the project.

All interviewees volunteered for participation. That is, in the questionnaire, they replied that they were willing to be interviewed. One benefit of this is that it is likely that the respondents were particularly motivated for participation. A drawback is that the options for sample selection were few. Consequently, the profiles of the interviewees were quite similar: All of them were Norwegian, male students in their early or mid-20's.

One critical issue, which I consider to be a general problem, is that the focus of the study has changed. Accordingly, the information that was provided on the consent forms and informed orally did not represent the final thesis adequately. Hence, initially, the respondents never exactly knew what they agreed—neither did I. Consequently, at the end of the study, I sent an e-mail to the interviewees in which I explained the shift of focus.

5.5 Concluding remarks

In this chapter, I have described methods of data collection and analysis. Moreover, I have discussed ethical issues. The aim is that readers should be able to replicate, and possibly falsify general claims I make in this thesis.

6 Summary of the papers

In this chapter, I summarise three papers: (1) Paper I is a journal paper on the validation of an instrument for measuring mathematical identity, (2) Paper II is a journal paper on the theorisation of mathematical identity, and (3) Paper III is a conference paper that discusses the association between STEM students' self-reported mathematical identities and average grades in mathematics courses.

6.1 Using the Rasch model to measure mathematical identity (summary of Paper I)

In Paper I, I address research question 4a, which asked: What are characteristics of mathematical identity in Norwegian TE and STEM contexts? To answer this question, I discuss the technical validation of an instrument for measuring mathematical identity. (See also paragraph 5.2.) At the time the paper was written, I used the term "the degree to which students' are working conceptually with mathematics". Later, I realised that the latent variable was broader than 'working conceptually'. Thus, in this section, I refer to the measures as mathematical identities.

In the initial phase, the instrument development was informed by the literature (e.g., Carpenter & Lehrer, 1999), persons with expected knowledge about the variable (e.g., teacher-educators), and existing instruments (Entwistle, 1997). These sources were discussed in the previous chapter, paragraph 5.1.

Two parameterisations of the polytomous Rasch model, the PCM (Masters, 1982; Masters & Wright, 1997), and RSM (Andrich, 1978), are typically used when items have more than two options. In short, the RSM assumes that distances between thresholds do not depend on the item. That is, the distance between, for example, 'sometimes' and 'often' does not depend on the question asked. The PCM does not make this assumption. I followed guidelines presented by Linacre (2000, p. 768), and the RSM was chosen for three reasons:

1. All items in the instrument were intended to share the same scoring structure.
2. If the PCM had been selected, some of the items would have had less than ten responses to some of the categories, which is considered to be problematic.
3. The correlation between person measures, when the different approaches were tested, was close to 1—leaving little room for argument as to whether to reject the RSM.

From the JMLE, item and person measures were estimated, and the results are shown in Figure 9 and Table 3.

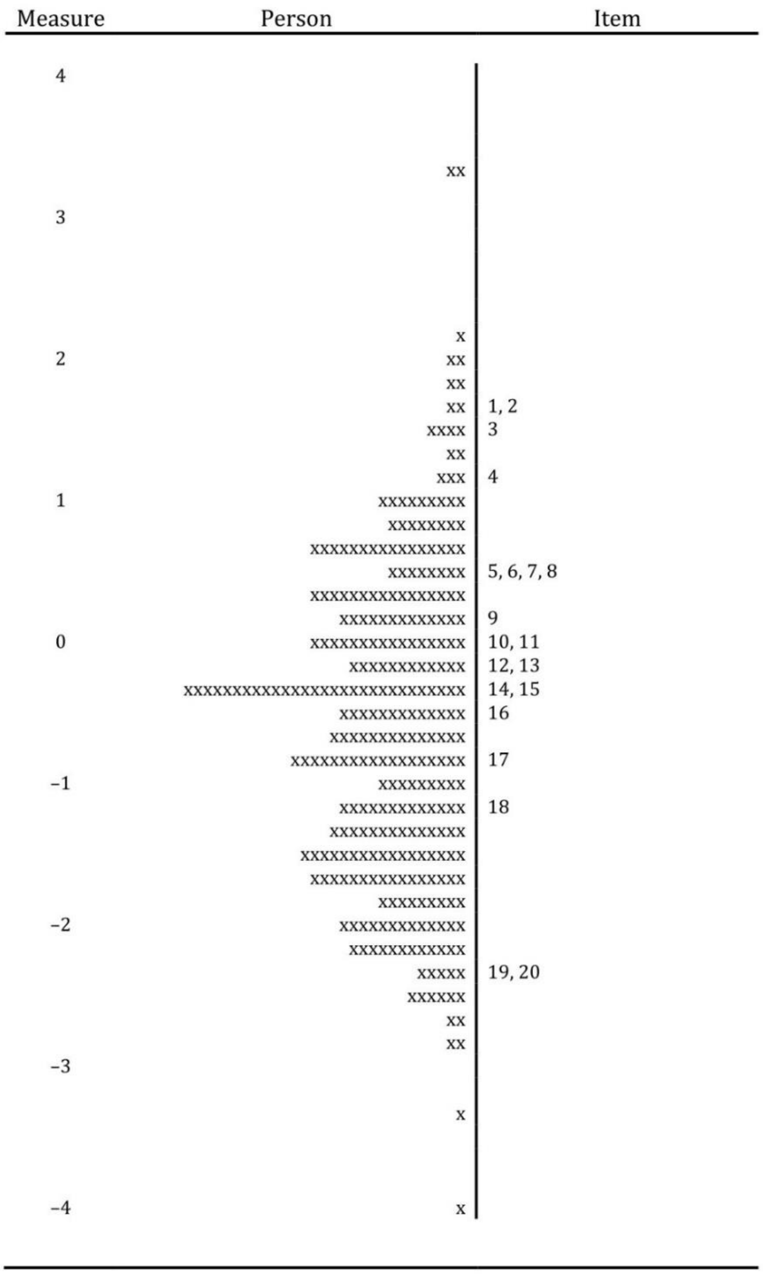


Figure 9. Person-item map

The person reliability (analogous to Cronbach’s alpha) was concluded to be .86. This index is primarily affected by the same principles that increase precision on a physical instrument (e.g., a ruler). Thus, from Figure 9, we can hypothesise how reliability might improve. If items on the right-hand side are analogous to marks on a ruler, then it appears that there would be some measurement errors when measuring persons in the range –2 to –1 logits. The same can be said about persons at both extremes. Thus, including more items in these locations would, most likely, improve reliability. We also see that adding items in the range –1 to 1 logits would hardly affect reliability since persons in this range are already measured precisely.

In addition to the possible improvement of reliability, Figure 9 provides information on the responsiveness validity, that is, the ability to measure persons in the future. It is evident that, if a sample with greater variance is being measured, then more items at both extremes would improve precision.

Table 3. Item statistics

	INFIT		OUTFIT		
Measure	MNSQ	ZSTD	MNSQ	ZSTD	Item
1.71	.93	−.7	.86	−1.3	1. Takes the initiative to learn more.
1.69	.89	−1.3	.86	−1.5	2. Takes time to find better methods.
1.42	.98	−.2	.93	−.7	3. Thinks of times when methods don't work.
1.10	**1.23**	**2.7**	**1.21**	**2.4**	4. Struggles with putting problems aside.
.58	1.09	1.2	1.11	1.4	5. Derives formulas.
.55	1.12	1.4	1.12	1.3	6. Likes to discuss math.
.54	.99	−.1	.99	−.1	7. Makes his/her own problems.
.45	.91	−1.0	.89	−1.3	8. New ideas lead to trains of thoughts.
.19	**1.25**	**2.9**	**1.25**	**2.9**	9. (x) Likes to be told exactly what to do.
.08	**1.23**	**2.7**	**1.20**	**2.4**	10. Finds out why methods wouldn't work.
−.06	.89	−1.4	.88	−1.5	11. Finds out why formulas/algorithms work.
−.21	.94	−.7	.94	−.7	12. Studies proofs until they make sense.
−.24	.71	−4.1	.72	−3.9	13. Considers different possible solutions.
−.27	.84	−2.0	.86	−1.8	14. Moves back and forth between strategies.
−.35	1.08	1.0	1.08	1.0	15. Wants to learn more things.
−.51	.96	−.4	.99	−.1	16. Pauses and reflects.
−.82	**1.22**	**2.3**	**1.22**	**2.3**	17. Visualizes problems.
−1.20	.75	−3.4	.78	−3.0	18. Can explain why solutions are correct.
−2.27	.98	−.2	1.02	.3	19. Connects new and existing knowledge.
−2.38	1.06	.7	1.04	.5	20. Keeps trying.

Note. Item 9 reversely coded

The increase of *n* would affect reliability indirectly. That is, items at both ends have greater measurement errors than items around zero due to the person distribution. The increase of persons with extreme measures would increase the precisions of these items, and hence, increase reliability.

The analysis of fit-statistics (*content* validity) showed that four items (4, 9, 10, and 17) were slightly under-fitting with Infit Mnsq in the range of 1.22–1.25 logits, and the PCA (*structural* validity) showed 1.7 unexplained variance (Eigenvalue units) in a second dimension. Moreover, analysis of the ICCs indicated some problems with the category structure. This was confirmed in a rating scale analysis (*substantive* validity). Specifically, empirical responses tended to deviate from the modelled responses when person measures were high. Figure 10 provides one illustrative example in which individuals with measures around 4 logits stronger than the measure of item 13 gave unexpectedly low responses on this item. One possible solution to this problem is to change the last category from ‘always/almost always’ to ‘always’. A more detailed analysis, including ICCs of all items, is presented in Appendix A.

Two items, ‘taking the time to find better methods’ and ‘connecting new and existing knowledge’, were found to have DIF between TE and STEM students (*generalisability* validity). In the psychometrical tradition, DIF is regarded problematic. However, this problem is based on the idea that the variable that is being measured is context-free. In my thesis, I claim that the measure of mathematical identity is the measure of both persons and social structure. Thus, the DIF between TE and STEM students point at specific structural differences between the TE and the STEM context.

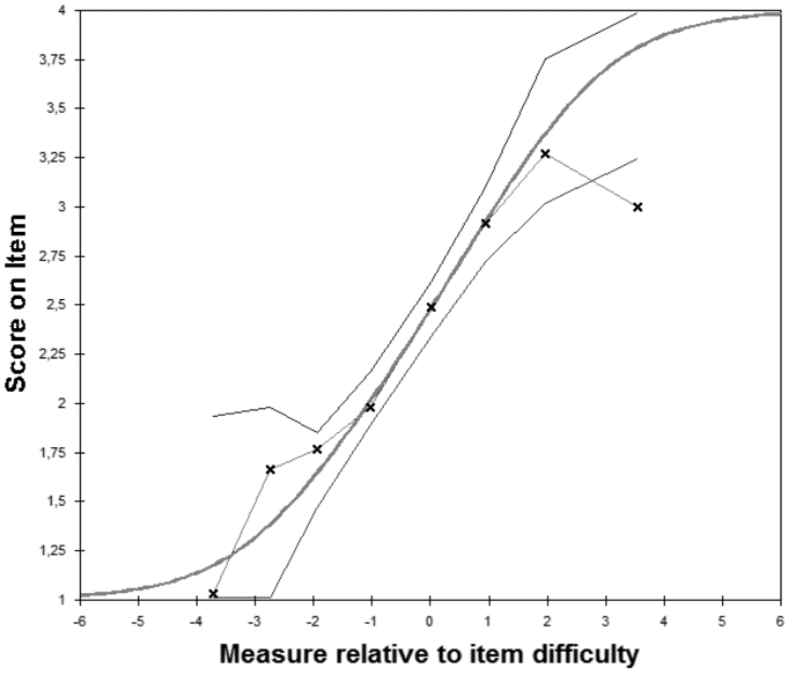


Figure 10. Item Characteristic Curve for Item 13

The final items, in the language they were administered, are provided in Appendix F. The English translation is presented in Appendix E.

It is worth noting that the characteristics discussed in this paper are not concluded to be exclusive, nor static. Later, I discuss these properties further when I explain how structural flux is a general property of mathematical identity.

6.2 Theorising the measuring of mathematical identity (summary of Paper II)

In Paper II, I address research questions 3 and 4b. In the paper, I present three problems with the measurement of mathematical identity. First, there is no consensus on philosophical issues on identity. Thus, it seems impossible to measure identity as such. Identity *is* not measurable, nor *is* it unmeasurable. The second problem is that most theories conceptualise identity as complex. However, as I have discussed in chapter 4, in the RMT paradigm, a measure is required to be uni-dimensional. In the IRT paradigm, measures are assumed to be uni-dimensional. The third problem is that most theories conceptualise identity as more or less situated, whereas measures are required (or assumed) to be invariant.

According to these perceived problems, I question: (1) which theoretical perspective on identity that is consistent with the requirements of measurement; (2) what the characteristics of STEM students' mathematical identities are; and (3) how measures of mathematical identity can provide information on how much the social structure of being mathematical differ across the STEM context and the TE context.

In effect, the theorisation is based on the following question: "If mathematical identity can be measured, then what are the principles of mathematical identity?" That is, a premise of the theorisation is that mathematical identity can be measured: Mathematical identity is, therefore, about sameness and distinction as opposed to the unique qualities of individuals.

From this, I have grounded the theorisation on the assumption that mathematical identity is relational. This assumption is based on the fact that measures, in general, are relational. Hence, if mathematical identity can be measured, it must consist of both individuals and some structured background, I argue. Consequently, I have been searching theoretical concepts that can describe the body of reference that personal identities are measured relative to. An initial attempt was to use the collective and object-oriented 'activity' as described in CHAT and summarised in section 2.5.

One dispute in identity research is the locus of identity, that is, whether it is mostly context-free or context-bound. My response to this argument is that no measures can be assumed to be context-free. Invariance is an empirical question, and thus, there is no general law that dictates the locus of identity—the locus *is* not positioned on some particular point between context-free and context-bound. Consequently, I assume the locus of identity to be an empirical question that can vary between activities, or even within the same activity.

If this assumption is true, then the situatedness of identity is, itself, situated. Therefore, we need empirical data on at least two contexts before we can make any conclusion on the situatedness[4] of mathematical identity.

One problem, then, is the following: If it is true that identities are measured in context, and if we need measures from at least two contexts for studying the level of invariance, then the initial attempt of using 'the activity' as the structured background is problematic, insofar as an activity is something specific and, by definition, situated.

My proposed solution is to use a more general structure—something that, at least hypothetically, can be similar across contexts. Accordingly, I define mathematical identity to be *where persons position themselves relative to the social structure of being mathematical within the activity in which they participate and contribute*. I then define the *social structure of being mathematical* as a person-independent set of characteristics (of being mathematical) and its internal structure. The 'internal structure' means the relative distances between each characteristic and the arbitrary zero point within each dimension, applying the arbitrary unit length. The social structure is, therefore, another empirical question—its content and structure must be proven person-independent. Also, the structure can be multidimensional, although I, in this paper, consider a uni-dimensional case only.

The methodology paragraph of the paper builds on Paper I, and I will, therefore, not discuss it much further. In addition to quantitative data, four STEM students were interviewed and asked about: how they perceived their time at the university; which project they were currently engaged in; how they worked with mathematical problems; and how they learned new mathematics. Transcripts of the data were analysed using the item characteristics in addition to concepts described in CHAT. Unlike Sfard and Prusak (2005), I do not consider identity to *be* narratives. Instead, what people say are indications of a latent variable. From this, relative positions are determined by the researcher using techniques of objective measurement. Consequently, when I say "the positioning of individuals", I do not mean this literally (unless someone analysed themselves).

[4] When I, in this thesis, discuss 'situatedness', I consider structural and not personal situatedness. That is, situatedness is the degree to which the structure of being mathematical is similar across activities. I believe that the relationship between structural and personal situatedness is yet another empirical question. Therefore, I consider situatedness, in general, to be a function of both structure and persons.

In the article, I discuss the person-independent feature of the social structure, and this is illustrated in Figure 11. That is, 40 persons with the strongest mathematical identities were, in the second column, removed from the analysis. In the third column, 40 persons with the weakest identities were removed. The figure illustrates how the social structures of mathematical identity are more affected by the change of activity than the change of individuals. This is a property that is required, and not assumed, by the theoretical frame. Additional examples of this property are discussed in Appendix B.

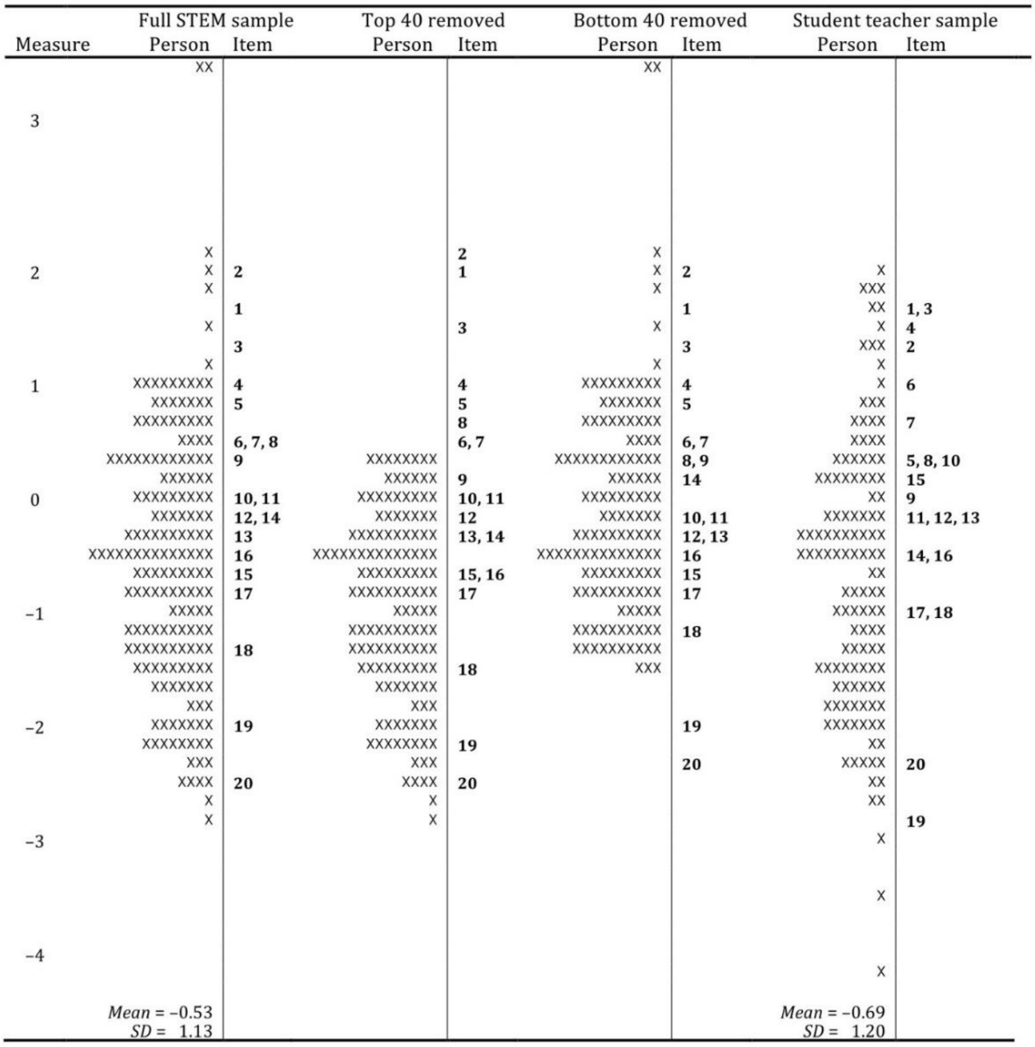


Figure 11. Persons and item measures on the same variable

Further, in the paper, I discuss three levels of the characteristics of being mathematical. Characteristics on the lower end seem to be related to working with mathematical problems, such as visualising problems, being able to explain solutions, connecting new and existing knowledge, and so forth.

> Y: Em, when I first started to look at Padè approximation, I was reading an article that was quite mathematical. And I realised, wow, I cannot remember any of this. So, I looked up in some of the books we had in the basic courses, where we had about sequences. Took a quick look. It wasn't very deep, just a recap on the concepts. When I had a basic understanding, I went back on the article.

Characteristics around zero include making his/her own problems, disliking being told exactly what to do, finding out why methods would not work, studying proofs until they make sense, deriving formulas, and so forth.

> B: One typical thing is integration by parts, which I have derived a million times in my life. You can learn the formula by heart, but I never bothered. So, it's like, we have U and we have V, and…[writes]…there you go. And I often do this with other problems too.

Characteristics on the higher end include struggling with putting problems aside, thinking of times when methods would not work, taking the time to find better methods, and so forth. A general impression is that positive affective elements (e.g., joy) appeared on this end of the variable.

> A: I have had the problem [with a simulation algorithm] for a while, three weeks maybe. I think about the problem everywhere. I think about it when I am at home, on my spare time, and when I relax...Yesterday, I was at a café with my girlfriend...We were talking about something, and then I just started to think about it.

The paper ends with two remarks, one on the issue of missing data, the other on directions for future research. I consider the latter to be of most importance. That is, measured mathematical identities are most frequently concerned with persons (e.g., Axelsson, 2009). However, I have argued in this study that the social structure of being mathematical also can be measured. In fact, an implication of the proposed framework is that it is impossible to measure people without the consideration of a person-independent social structure. Consequently, I believe future research could benefit from measuring and comparing structures of being mathematical across contexts.

6.3 The association between engineering students' self-reported mathematical identities and average grades in mathematics courses (summary of Paper III)

Paper III is an 8-page conference paper, and in the paper I answer research question 4c: What is the association between STEM students' self-reported mathematical identities and average grades in mathematics courses? The study attempts to shed light on how conventional means for assessing students (i.e., exams) reflect mathematical identities.

The rationale for this short study is that specific mathematical competencies are transferred poorly, for example, from higher education to the world of work (e.g., Rystad, 1993). One reason might be that the mathematics is hidden in 'black boxes' more frequently in workplaces than in educational settings (Williams & Wake, 2007). Thus, a common argument is that students need to develop more general characteristics that relate to mathematics (e.g., Hoyles, Wolf, Molyneux-Hodgson, & Kent, 2002), and I hypothesise mathematical identity to be one representation of such characteristics.

The sample in the study included 361 Norwegian STEM students. The classification of these students is listed in Table 4.

Table 4. STEM sample

Course	*n*
Introductory mathematics course	47
Calculus 2	71
Calculus 3	113
Cryptography	11
Variety of courses (final year)	119
Total	361

The participants responded to the instrument discussed in Paper I, and after the validation of the instrument, the respondents were categorised as having either low (measures lower than -1), medium (measures between -1 and 1), or high (measures above +1) mathematical identities (all measures were reported in logit units). Subsequently, a Welch's ANOVA (analysis of variance) was conducted to compare the association between mathematical identities and self-reported average grades in mathematics courses at the university (from grade F=1 to grade A=15).

The Welch's ANOVA showed that the association between mathematical identity and self-reported average grade was significant, $F(2, 110.79)=31.966$, $p=0.000$. Moreover, the mean of the self-reported average grade amongst students with high mathematical identities was about one grade higher than those with low mathematical identities. The Games-Howell test showed that the difference was significant between all groups with low-medium as the least significant ($p=0.001$).

Moreover, Figure 12 shows how there was an unequal variance between the groups. That is, the variances decreased with the increase of mathematical identity, which means that high mathematical identities are associated with high self-reported average grade, whereas there seems to be no limit to how low mathematical identities students can have and still get high grades. Additions to the analysis are presented in Appendix C.

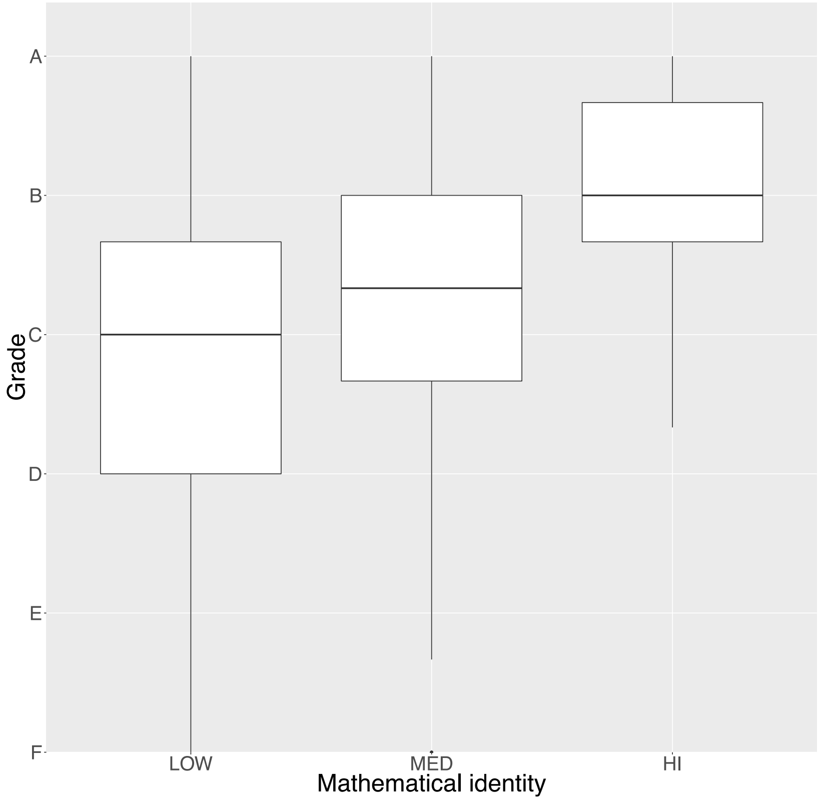


Figure 12. The relationship between self-reported mathematical identity and average grade in university mathematics course

The paper concludes with some obvious limitations, namely that the average grades were self-reported and, most certainly, biased. Thus, I suggest a replicate of the study using actual grades. Moreover, I propose that future studies examine the causal relationship between grades and mathematical identity.

Finally, I suggest that replicates of the study could exchange ‘grade’ with other institutional measures such as PISA or TIMSS results. Specifically, such studies could provide some clue to the question of whether some countries ‘teach to the test’, in which case one might hypothesise that a relatively great proportion of students in these countries would appear at the ‘top left corner’—that is, students with weak mathematical identities, yet, high measures of attainment.

6.4 Concluding remarks

In this chapter, I have summarised three papers, one mostly methodical, one mostly theoretical, and one mostly empirical. I have argued that mathematical identity is a relationship between personal and social positions, and that this relationship can be measured. Moreover, I have claimed that structural positions do not depend on personal positions, and personal positions do not depend on structural positions. I have also discussed the method of measuring mathematical identity. Finally, in an empirical study, I have suggested that mathematical identity is related, yet poorly correlated, with attainment. That is, strong mathematical identity seems to imply high achievement, whilst the reverse is not true.

7 Discussion and conclusion

In this chapter, I first present a summary of this thesis. In this summary, I answer the research questions that I raised in the first chapter. Subsequently, I discuss the results: their implications, how they relate to the wider context, some limitations and challenges; and, finally, I provide some suggestions for further research.

7.1 Summary

In this thesis, I have addressed and answered four research questions, all related to the overarching question: How can mathematical identity be measured and theorised?

First, I have argued that a measuring perspective on mathematical identity must be compatible with theories on measurement in general. Thus, in chapter 3, I claimed that a framework of mathematical identity must consider four aspects explicitly: dimensionality, invariance, additivity, and relativity.

Second, in chapter 4, I built on a pragmatic philosophy when I claimed that, within a defined theoretical framework, some methods are truly better than others. Subsequently, I have followed the argument of Andrich (2004) and discussed philosophical differences between two theories of measurement in the social sciences: RMT and IRT. I have shown how the ontology of principles of measurement—whether they are assumptions or requirements—distinguishes RMT and IRT. The principles of measurement in this study were interpreted as Thurstone's (e.g., 1954) original account, namely, that the principles are requirements and not assumptions. For this reason, I have concluded that the application of RMT is a paradigmatic method of the measurement of mathematical identity.

Third, informed by theories of measurement and empirical data, I have proposed a framework of measured mathematical identities. In addition to the premise that mathematical identity can be measured, I have argued that there are two underlying assumptions of mathematical identity: (1) identity is relational, and (2) the locus of identity is an empirical question—the locus is not fixed at some point between entirely context-free and completely context-bound. Nor is the locus static.

The first assumption implies the need for a body of reference—what I call a social structure. The second assumption implies that, although invariance is required, this principle is not absolute. I am not assuming that there exists one universal structure that is invariant across all contexts. On the contrary, I acknowledge that mathematical identity is in flux and that this applies to both persons and social structure.

From these assumptions, I have defined mathematical identity to be *where persons position themselves relative to the social structure of being mathematical within the activity in which they participate and contribute*. Following this definition, I have argued that the *social structure of being mathematical* is required, and not assumed, to be person-independent. In Paper II, this principle is illustrated with an empirical example.

Finally, the empirical data and analyses in this study have contributed, not only to inform methodology and theory but also with examples of the range of research questions that can be asked within the framework.

In Paper I, I have validated an instrument for measuring mathematical identity. More than technical validity, this paper contributes to a qualitative understanding of the characteristics of mathematical identity. These characteristics were explored further in Paper II, where I provided qualitative and empirical illustrations.

Furthermore, in Paper II, I have discussed how a measuring perspective on mathematical identity measures both persons and social structure. Technically, there is no difference between persons and characteristics in the process of measurement. Consequently, comparisons are comparisons of positions, and these positions can be held by persons or characteristics. Specifically, I have shown how RMT can identify structural differences between the TE and the STEM context.

In Paper III, I have suggested that mathematical identity might be related to attainment (grades) although the two variables correlate poorly.

7.2 Theoretical insights

In this section, I put forward three claims: I argue (1) that the distinction between structural and personal change is an arbitrary point of perspective, (2) that the comparison of mathematical identities does not require structural equivalence, and (3) that mathematical identities do not exist in isolation.

(1) The distinction between structural and personal change is an arbitrary point of perspective.

In 1929 Edwin Hubble discovered that the universe is expanding in every direction. A consequence of this discovery is that the earth is the centre of the universe. However, if an alien made the same observation from another planet, in another galaxy, it would draw a similar conclusion, namely, that this other planet was at the centre of the universe. This is because the universe has no stationary absolute—the centre is not at one particular point in space, it is everywhere. It is all a matter of perspective.

One implication of this study is that mathematical identity exhibits a similar property. That is to say, if identity is in constant flux, as many theories suggest (e.g., Wenger, 1998), then this motion must be seen as relative to something else. There must, therefore, exist one point that is static. However, there is no external criterion for determining whether something is constant without introducing another constant. Accordingly, the static point of reference is arbitrary; it is everywhere, it is all a matter of perception.

If this is true, then the choice of the static point can affect whether a change is interpreted as a personal or as a structural one. If we constrain the point to be at the centre of a group of people at any given time, then what we typically perceive as the development of people will be interpreted as a structural change. At the same time, if we choose the point to be at the centre of the structure, then the same movement will be construed as personal development. Since the static point is arbitrary, we must allow both interpretations without contradiction. A consequence of this study is, therefore, that the only measurable difference between personal change and the change of social structure is the arbitrary point of perception.

(2) The comparison of mathematical identities does not require structural equivalence.

Another implication of the study is that mathematical identity development is the movement of both people and structure relative to the same static point of reference. Consequently, since people are not measured in relation to a structure as a whole, but rather to one point in the structure, there is no need for structures to be equivalent across time and space to compare identities.

To illustrate this point, we can imagine that we were to compare the locations of two persons in a physical structure, for example, a city. Then, for each dimension, we would need one common point of reference and one standard unit length.

Imagine, further, that person A was standing at one end of the city, say the east end, and that we knew his location based on a map that covered the east end and the centre of the city. Imagine also that person B stood on the opposite end, that is, the west end of the city, and we knew her position based on a map that covered the west end and the centre of the city. Consequently, since there was an overlap between map A and map B (the centre of the structure), positions of person A and person B could be compared directly, even if map A and map B were different. Moreover, the relative size of the intersect would be of little importance. What would be more important is that each map was as detailed as possible.

This analogy holds for social structures of being mathematical. If a researcher is to compare personal positions, there must be some overlap in social structures, but in principle, one shared point of reference and one standard unit length are sufficient for each dimension.

At this point, a methodical comment is necessary. That is, due to the nature of social structures and the means that are used to assess them, the relative errors of measurements of social structures are, typically, greater than those of physical structures, although one could find exceptions to this assertion. Consequently, when I say that one common point of reference is sufficient, this claim holds true only when the location of this point contains no measurement error.

In practice, however, measurement errors do occur, and therefore, more common points of reference are appropriate, although the only reason is to reduce the effect of measurement errors. There is nothing fundamentally wrong with comparing personal mathematical identities, even when most of the elements in the social structures are different.

(3) Mathematical identities do not exist in isolation.

So far, I have discussed two implications: First, persons (and characteristics in a social structure) are compared in relation to a point and not a structure as a whole; second, the choice of this shared point of reference is arbitrary.

If we accept these implications, I claim that there is a third consequence. Namely, in the comparison of mathematical identities from one point in time or one activity to another, we can choose different points of references and different unit lengths. In some—perhaps most—cases, the choices will lead to no practical consequence; it will be what Celsius and Fahrenheit are to the measurement of temperature. However, there might be particular cases in which different conclusions can be made, all depending on our choices. Hence, at least hypothetical cases exist in which person A has, at the same time, a stronger and a weaker mathematical identity than person B. According to the proposed framework, this result is not a contradiction.

This last consequence turns the argument back to the original starting point. That is, the theorisation of mathematical identity started with the intuition that it is possible for one person to have a stronger mathematical identity than another. Subsequently, a theoretical perspective was proposed. An implication of this view, however, is that it is possible, at least in some special cases, to make different conclusions, all equally valid, without changing anything but the arbitrary point of perception.

To conclude, I agree with critics who refute that a person can *have* a mathematical identity in isolation. This critique, however, does not affect the discussion of whether mathematical identity can be measured. The fact that “no one has a mathematical identity” merely means that no stationary absolute to mathematical identity exists, and, although a zero point is a requirement of measurement, this point is not required to be absolute. Accordingly, a person *has* a mathematical identity no more than a planet *has* a location in the universe. Nonetheless, the position of both can be estimated.

7.3 A synthesis of two extremes

I have defined mathematical identity to be a relational position between people and social structure. Moreover, I have defined the social structure to include the person-independent subset only. That is, an element in the social structure should not depend on the position of individuals. If it does, it is not an element in the social structure after all.

If we accept this definition, then the status of an element is unknown before the analysis. The element could either be included in the person-independent subset or not. In an empirical study, it is possible that a researcher concludes—let us assume rightly—that no items are person-independent. If this is the case, it is true that mathematical identity, in this particular situation, cannot be measured.

It is worth noting that this conclusion rests on the fact that the measures of the items change significantly when different subsets of the population are used in the analysis. Hence, it is the process of measuring mathematical identity that leads to the conclusion that the construct, in this particular case, cannot be measured in the first place.

Accordingly, I consider the perspective discussed in this thesis to be a synthesis of two extremes: (one) that mathematical identity can never be measured, and (two) that the construct can always be measured. From the perspective of this study, measurability of mathematical identity depends on results from the process of measurement, and the conclusion, whatever it is, is fluid and empirical.

At this point, a distinction can be made between theory and practice. In theory, most practitioners of Rasch measurement would agree that both people and structures are measured. In practice, however, individuals and structures are treated differently. Most frequently, the measures of structures are matters of validity while measures of individuals are issues of results.

In Paper I, I complied with this tradition. However, I no longer believe that structure represents the instrument and persons are whom we measure. On the contrary, I perceive the measurement of mathematical identity as the measurement of the relation between individuals and social structure. Therefore, I find the process of measurement to be a powerful tool even—or perhaps, in particular—in cases when measurements should ‘fail’ in the traditional sense. If for example, mathematical identity is situated between two activities, say, STEM activities in Norway and Japan, then there will be structural differences between these activities, and some of these differences can be shown by measure. Hence, ‘failure’ of comparing measures across activities will be an indication that there is something qualitatively different between the activities, and the measures of the structures indicate what these structural differences are.

7.4 Implications for research

This thesis is mostly theoretical and methodological. However, the questions that I have answered were initiated from empirical problems related to transitions of mathematical identities. I this section, I explain how the result of this thesis is relevant for future research and studies on transitions in particular.

One theoretical insight of the thesis is that the distinction between structural and personal change is an arbitrary point of perspective. This is important because the study of the development of measured mathematical identities, for example in transition, cannot make absolute claims about personal and structural change. From the theoretical perspective in this thesis, there is no philosophical difference between structural and personal change. Technically, this point is evident in the symmetry of the Rasch model, as the model cannot distinguish between persons and items.

When the arbitrary zero point is chosen, the measurement of mathematical identity is the measurement of both individuals and structure. Future studies of transition could benefit from this and map the structure of mathematical identity in multiple contexts. As I conclude in Paper II, a critical analysis of education could examine how the structure of school mathematical identities is structured relative to mathematical identities in other contexts.

Another theoretical insight is that the comparison of mathematical identities does not require structural equivalence. This insight is of practical importance for two reasons. First, mathematical identities between contexts can be compared even when there exists evidence that the contexts are structurally different. Second, structural differences is not a limitation, rather, the outcome of the measurement of mathematical identity. Thus, the claim that context A is structurally different from context B is an empirical claim that can be addressed by the process of measurement.

The third theoretical insight is that mathematical identities do not exist in isolation. It is theoretically possible, without contradiction, for one person to have both a stronger and a weaker mathematical identity than another person. Even if such cases might be rare, I believe that empirical evidence of this point would be valuable because it would illustrate the relational property of mathematical identity.

The measurement of mathematical identity is not only a measure of position, but also one of the level of misfit. Studies on identities in development and transition can benefit from this. For instance, when persons transfer from context A to context B, assuming that the structures of mathematical identities are not equivalent in the two contexts, it is possible to examine how people negotiate their positions relative to the two contexts. One hypothesised situation is that people, regardless of their identities, adapt to new situations rather quickly. In such cases, when persons (transferring from context A to context B) are measured in context B, they will, most probably, show *less* misfit relative to the structure in context B than if their responses were measured relative to the 'old' context A. Another hypothetical situation is that people adapt to the new context rather slowly, or possibly, not at all. In these cases, when persons (transferring from context A to context B) are measured in context B, they will, most probably, show *more* misfit relative to the structure in context B than if their responses were measured relative to the 'old' context A. Accordingly, a hypothesis worth following is that, while measures inform the 'strength' of mathematical identity, the study of misfit provides information to which structure persons are positioned.

7.5 Implications for practice

The intended audience of this study is researchers in mathematics education. Consequently, most implications are implications for research. Nevertheless, some results have potential impact for teachers.

First, there is an increased focus on research-based teaching, that is, on teachers doing research on their practice. One particular example is a selection of Norwegian schools (called 'university schools') which collaborate closely with universities. From this collaboration, teachers get insights into the practices of research. When teachers do research in this context, the main outcome is improved practice, not published papers.

For teachers doing research on their practice, results from this thesis might be useful. For example, teachers could compare their students' grades with their mathematical identities and possibly observe how a 'high performer' is not, necessarily, the same as another 'high performer'. Such observations might add to teachers' reflections about the outcome of teaching and learning and also to the awareness that standard tests in mathematics do not measure 'everything'.

Second, teachers could gain from the qualitative parts of this thesis. Specifically, the thesis suggests that affective characteristics—struggling to put mathematics aside, liking to discuss mathematics, and so forth—are structured on the higher end of mathematical identity. These are characteristics that distinguish persons with 'strong' from persons with 'medium' mathematical identities. Thus, if one aim of education is that students should develop strong mathematical identities, then teachers would gain from an awareness of the qualitative nature of strong mathematical identities.

Third, a property of persons with strong mathematical identities is how they frequently talked about problems as something of a positive value—something you would rather have than not have, something you would be happy to share with your friends, like sharing a game or a puzzle. In contrast, persons with lower mathematical identities frequently talked about problems as something with a negative value—something you would rather not have, like weeds in your garden, and thus, something you would be happy to remove from your friends' shoulders. Hence, one possible implication of this thesis is that students need to learn how to appreciate *having* problems and not only having solved them. Moreover, I hypothesise that, for some students, it would be valuable to learn about solved and unsolved mathematical problems. In many cases, mathematical problems are relatively easy to understand, although a solution is difficult. For example, a special case of the millennium-prize 'P versus NP problem', which asks whether every problem that can be verified quickly (in polynomial time) can also be solved quickly, is the following: If someone proposes a solution to a Sudoku-problem, we can check the answer relatively quickly, but there are no known algorithms that can solve a Sudoku quickly.

7.6 Generalisations, limitations and challenges

A general claim in this thesis is that there exist social structures of being mathematical in at least two contexts—the TE and the STEM context, and that the structure of mathematical identity is relatively stable between these contexts. Future studies could improve generalisability in two ways. First, I suggest that future studies measure mathematical identities in a broader range of contexts, for example, in different workplaces or countries. Second, additional research could apply more fine-grained analyses of the contexts that I have already studied. For instance, in this study, I have measured identities in the TE and the STEM context. There is no limit to how many sub-contexts these, or any other context, consist of. A simple example is how the STEM context consists of pure mathematicians, civil engineers, computer scientists, and so forth. Existing studies have documented particular problems that engineering students encounter when studying pure mathematics (e.g., Harris, Black, Hernandez-Martinez, Pepin, & Williams, 2015), and it is, therefore, worth studying whether there are structural differences in mathematical identities between such sub-groups.

This thesis is a response to particular problems with the measurement of mathematical identity. However, the implications have the potential of being applicable to studies of identities in general. A potential generalisability, therefore, is that the properties of mathematical identity apply to every identity to which some persons relate more strongly than others. To illustrate with one example posed by Gee (2000). If ‘being a feminist’ is an identity, if it is true that some persons are more or less ‘radical feminists’ than others, and if we interpret principles of measurement as requirements, then, it follows that the principles of mathematical identity discussed in this thesis apply to feminist identities. Hence, being a feminist, in this case, is a relational position, and the study of feminist identities is the study of both social structure and personal positions.

Some constraints in this research relate to the empirical data. An obvious example is how ‘grades’ in Paper III were self-reported, and as a consequence, the results of this sub-study are only suggestive. Another example is sample size. From a relational perspective on identity, there are two samples: persons and characteristics in the structure. In future studies, an increase of both will most likely improve reliability.

Another limitation is that the focus of the study shifted as the theoretical problems with the measurement of mathematical identities became increasingly prevalent. While it can be argued that such a change of focus is a general feature of research, it nonetheless had an effect. Specifically, much of the data were gathered to understand mathematical identities in transition and not the nature of mathematical identities as such.

In the rest of this section, I will turn the attention to more general challenges that follow as a consequence of the results of the study. I do propose possible explanations to some of these challenges, but they are not connected with empirical data. They must, therefore, be considered as suggestions for further investigation.

7.6.1 Face validity

The instrument that was validated in Paper I consists of 20 items. These items—their qualitative content and internal structure—are indicators of mathematical identity. In effect, the items show a fragmented picture of ‘reality’. Accordingly, it is clear that 20 items do not ‘fully represent’ mathematical identity. This apparent problem cannot be solved easily since the inclusion of any set of items would lead to the same conclusion. Mathematical identity can never be fully represented.

This conclusion does not imply that the search for more characteristics is worthless, and I assert two main reasons why I encourage future studies to search for more characteristics and locate them in the structure.

First, the precision of a measure is more accurate when the structure is relatively detailed than when it is relatively fragmented. This is particularly true when the object of measurement is positioned close to elements that appear only in the detailed structure. In the RMT, this analogy is mathematically accurate: The inclusion of an item increases person reliability and decreases measurement errors, in particular when the item is located close to individuals and when there are few other items nearby.

Second, items that appear on the detailed map but not on the fragmented map do not increase precision only, they also provide qualitative information about the ‘real’ world. For instance, every new item would improve our qualitative understanding of mathematical identity.

I emphasise that mathematical identity is fluid. Hence, it is entirely probable that the qualitative nature of a structure in one context is different from that in another. A set of items can, therefore, never represent mathematical identity *as it is*, only how it appears in the observed context.

7.6.2 Mathematical identity and common sense

One of the most pressing challenges of this study is that mathematical identity is relational. This feature is a challenge because humans appear to often think in absolute, and not relational, terms (in contrast to his own laws of physics, Isaac Newton maintained that there had to be a universal absolute). By common sense, it is hard to accept that a person can have both a stronger and a weaker mathematical identity than another person. Likewise, it is difficult not to think of development as a separation of individuals and social structure. However, this study implies that there are no means of distinguishing personal and structural change. Therefore, before an arbitrary stationary is defined, mathematical identity development is not personal and structural, but rather person-structural.

7.6.3 Exposure time

When I look back at the data collection in this study, I see that measures of social positions might be affected by a phenomenon similar to the exposure time of a camera. That is to say, instruments of measurement try to determine the position of a subject at a given time, *t*. However, instruments are not exposed to data at a singular point in time, but rather an interval including *t*. The length of this interval can be thought of as the exposure time of the instrument.

To be concrete: In the case of this study, when people responded to the frequency—from never to always—of a set of characteristics, they must have thought of a time interval for these features to yield (if not, they would have replied ‘never’ to virtually any characteristic since, at the time of data collection, they were responding to a questionnaire and not working with mathematics). If the average interval was, say, the last month, then we could say that the exposure time of the instrument was one month.

The fact that the exposure time in this study is unknown, and possibly varying between individuals, is a problem that could have been solved by including a preface such as ‘During the last month…’ to every item in the instrument. This solution, however, would not affect the core of the problem, namely, that there *is* an exposure time, and that it might be quite high, possibly weeks or months.

I hypothesise some consequences of exposure time that future research might consider. In general, I believe that the effects of the phenomenon can be explained similarly to those in physical experiments. If that is so, the problem is a function of exposure time and the pace of mathematical identity development. Put briefly; I suspect that when people (or structure) remain relatively stable and the exposure time is short, people are measured more accurately than if they undergo rapid development and are measured with an instrument with longer exposure time.

Another possible effect is that the change between several data points might look smoother when the exposure time is long in relation to what it would look like if the exposure time were shorter. Hence, it might be that mathematical identity development by measure looks more continuous than it is. Figure 13 illustrates this with a synthetic example.

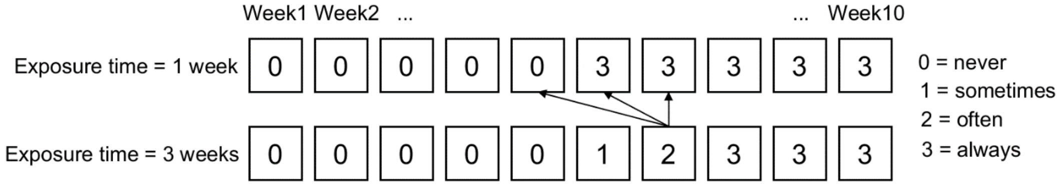


Figure 13. In the first row, a person with a sudden mathematical identity development has responded to an item based on an overall impression of the last week. In the second row, the same person responded based on an overall impression (in this example, the mean) of the last three weeks. The development looks smoother in the latter case

7.6.4 The measurement of mathematical identity in motion

There is another aspect of the measurement of moving objects (or subjects) that is scarcely discussed in the literature; something that is relevant if we assume that mathematical identities are measured in motion. That is, in ‘perfect’ (i.e., synthetic) data which fit the Rasch model, the expected response string of people is, in the dichotomous case, a string of 1’s followed by an interval of scattered responses, followed by a string of 0’s (including occasional unexpected responses). The scattered area of uncertainty is expected, but only to a certain extent.

However, when observations fit the Rasch model perfectly, for example in synthetic data, there is an implicit assumption that the objects of measurement ‘remain still’ at the time of observation. If the objects of measurement do move, this assumption holds true if the motion is practically continuous and relatively slow. If the exposure time of the instrument is close to zero, the assumption holds true when the motion is continuous, never mind the pace.

A problem is that sociocultural theories tend to emphasise how development is scattered and contradictory (e.g., Engeström, 1987). If this is true, then mathematical identity development might not be a continuous change of position. Consequently, if participants are measured as they undergo rapid change, one might expect more contradictory observations than if the individuals were relatively stable at the point of measurement.

If it is true that individuals in motion respond more contradictorily than relatively stable persons, then the area of uncertainty will likely increase (if this is not the case, it would suggest that mathematical identity development is not contradictory by nature after all).

Since the area of uncertainty affects the person fit statistics, one hypothesis that is worth following is that, all else being equal, the fit statistics of individuals in rapid development are greater than amongst those who change more slowly. If this hypothesis can be verified, researchers can get valuable, albeit inaccurate, information about motion from single observations.

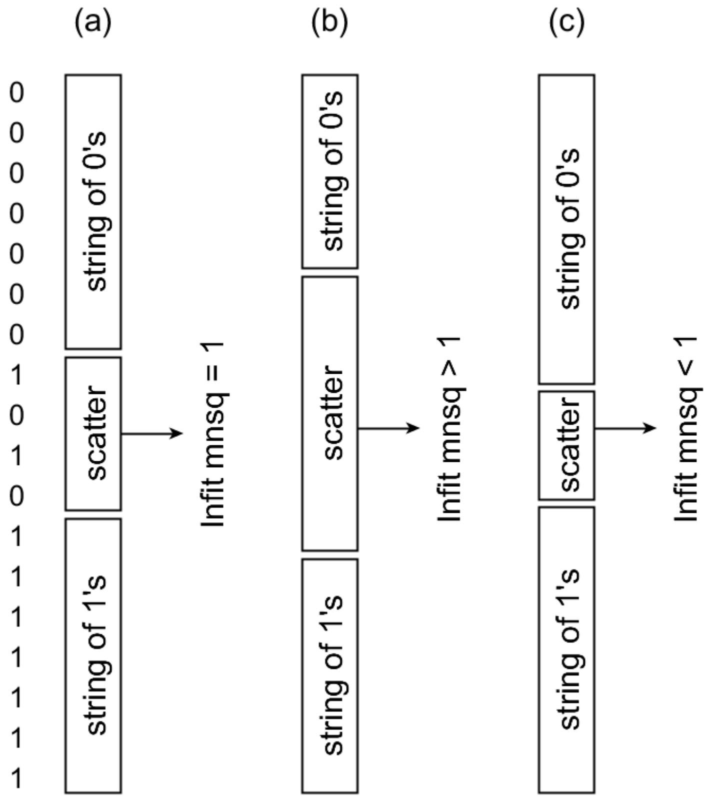


Figure 14. (a) When the data fit the Rasch model, Infit Mnsq = 1, caused by the scattered area of uncertainty. (b) When there are more contradictory responses than expected, the scattered area is likely to be larger, producing larger Infit Mnsq. (c) When there are less contradictory responses than expected, the scattered area is likely to be shorter, producing smaller Infit Mnsq. Measures in (c) are most accurate. There might be occasional 'holes' in the strings at both ends, but these have little effects on the Infit Mnsq statistic

7.7 Concluding remarks

In this study I have explored a measuring perspective on mathematical identity. In short, I have worked from the premise that it is possible for one person to relate more strongly to mathematics than another person.

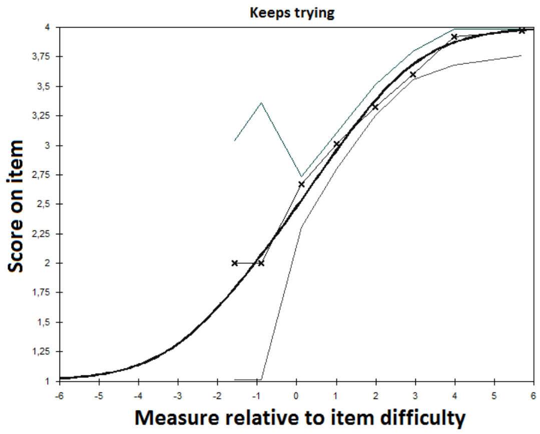
From this I have presented definitions and principles of mathematical identity, a paradigmatic method for the measurement of identity, and some empirical results that illustrate the range of questions that can be asked within the framework.

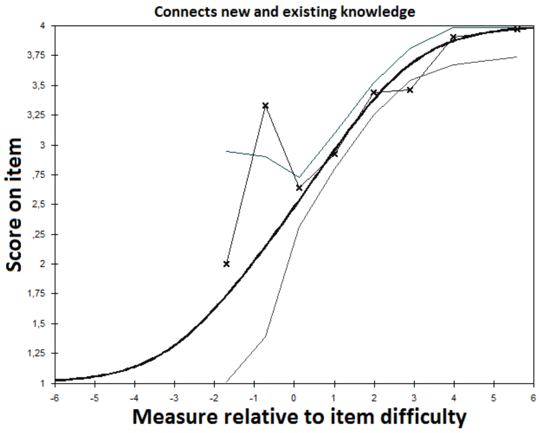
The most significant contribution of this study is that the theorisation of mathematical identity has led to some very specific properties. When it is measurable, mathematical identity is a relative position rather than something people 'have'. Moreover, mathematical identity implies the existence of a social structure. Neither the social structure nor individual positions are assumed to be static. Instead, a defined static point of reference is required, but this point is arbitrary. As a consequence, there is no ontological or epistemological difference between the development of structural and personal identities.

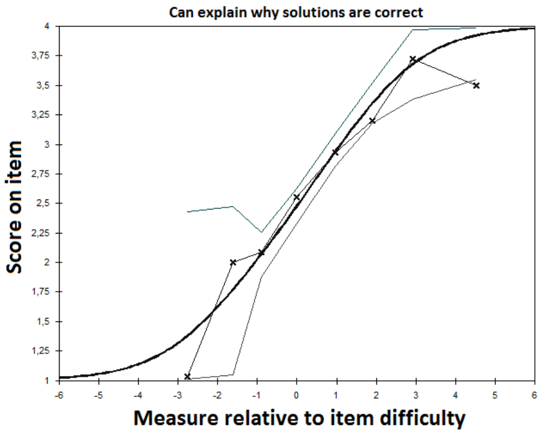
8 Appendices

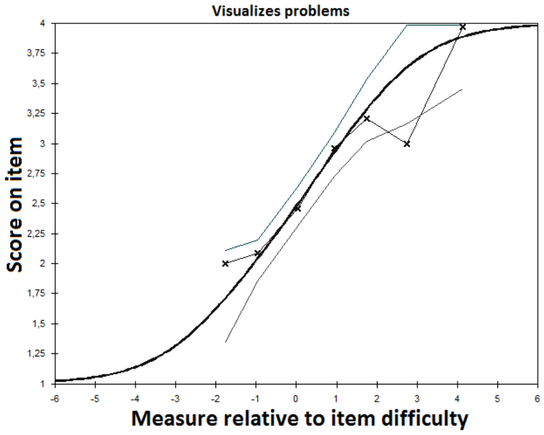
8.1 Appendix A: Additional analyses, Paper I

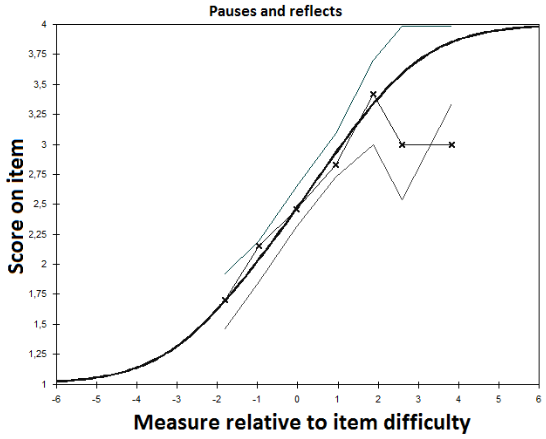
ICCs of the items in the first study are presented in the following Figures. Most deviations from the model appear when person measures relative to item measures are high.

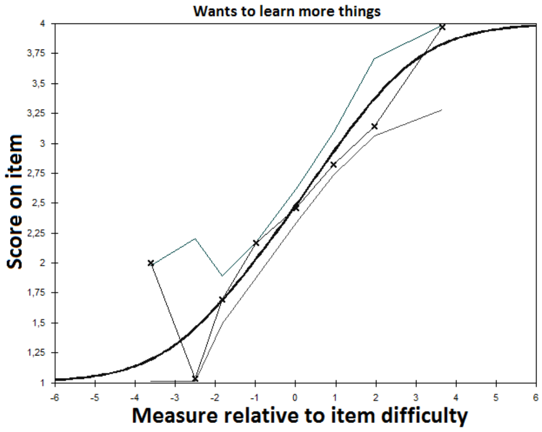

Keeps trying

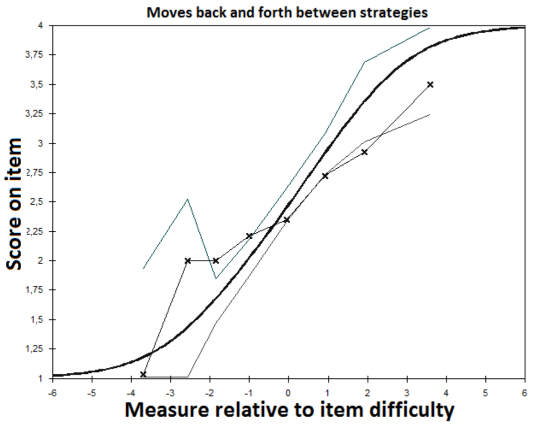

Connects new and existing knowledge

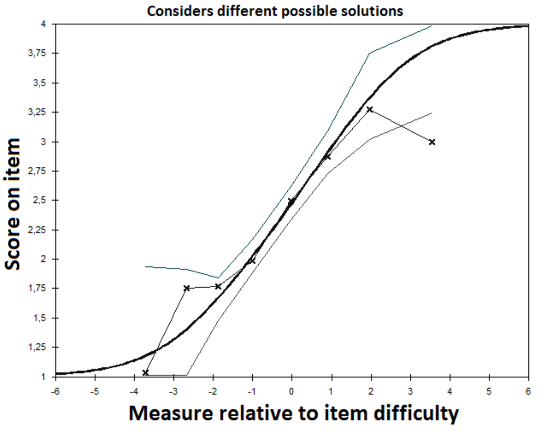

Can explain why solutions are correct

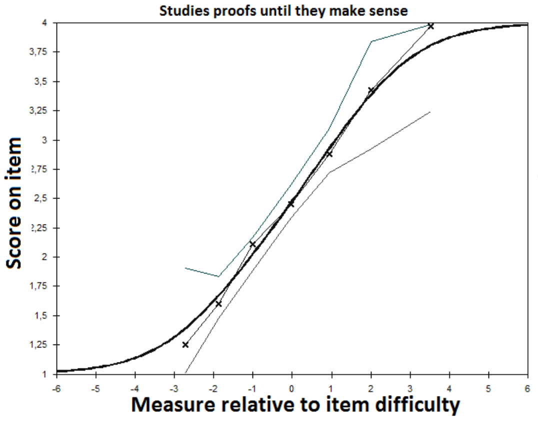

Visualizes problems

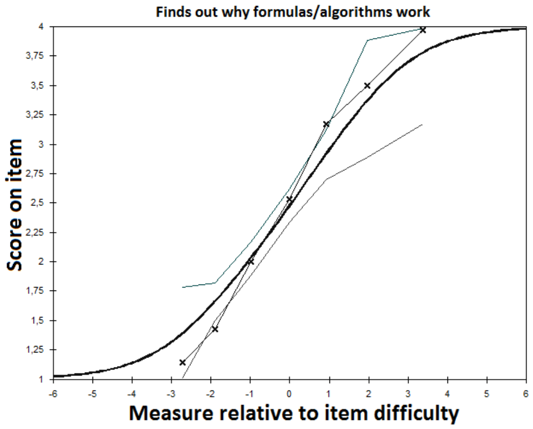

Pauses and reflects

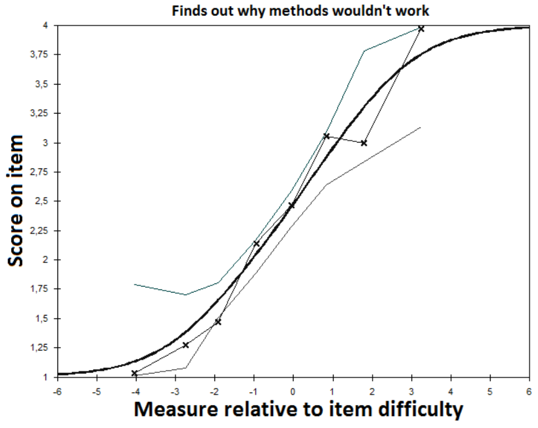

Wants to learn more things

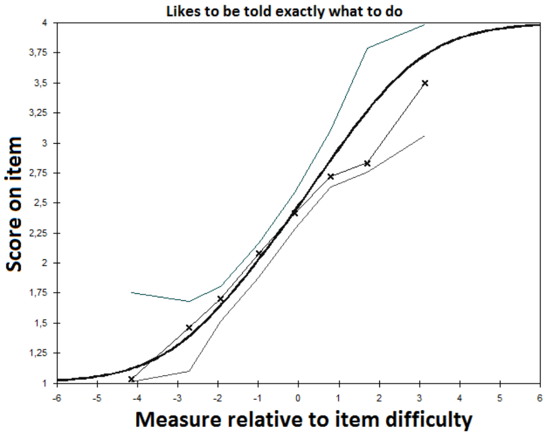

Moves back and forth between strategies
Score on item
4
3,75
3,5
3,25
3
2,75
2,5
2,25
2
1,75
1,5
1,25
1
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
Measure relative to item difficulty

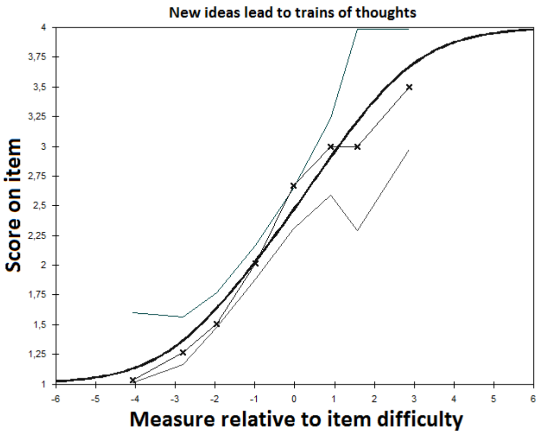

Considers different possible solutions
Score on item
4
3,75
3,5
3,25
3
2,75
2,5
2,25
2
1,75
1,5
1,25
1
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
Measure relative to item difficulty

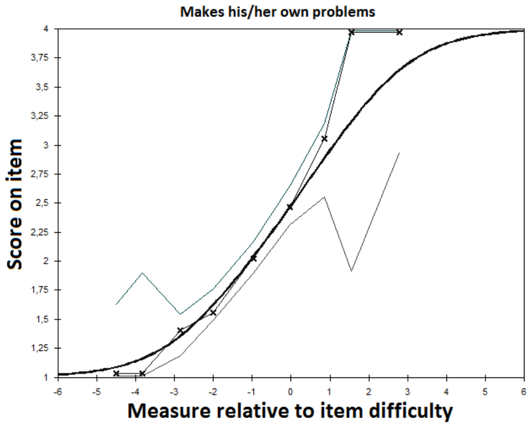

Studies proofs until they make sense
Score on item
4
3,75
3,5
3,25
3
2,75
2,5
2,25
2
1,75
1,5
1,25
1
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
Measure relative to item difficulty


Finds out why formulas/algorithms work
Score on item
4
3,75
3,5
3,25
3
2,75
2,5
2,25
2
1,75
1,5
1,25
1
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
Measure relative to item difficulty


Finds out why methods wouldn't work
Score on item
4
3,75
3,5
3,25
3
2,75
2,5
2,25
2
1,75
1,5
1,25
1
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
Measure relative to item difficulty


Likes to be told exactly what to do
Score on item
4
3,75
3,5
3,25
3
2,75
2,5
2,25
2
1,75
1,5
1,25
1
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
Measure relative to item difficulty


New ideas lead to trains of thoughts
Score on item
4
3,75
3,5
3,25
3
2,75
2,5
2,25
2
1,75
1,5
1,25
1
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
Measure relative to item difficulty


Makes his/her own problems
Score on item
4
3,75
3,5
3,25
3
2,75
2,5
2,25
2
1,75
1,5
1,25
1
-6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6
Measure relative to item difficulty

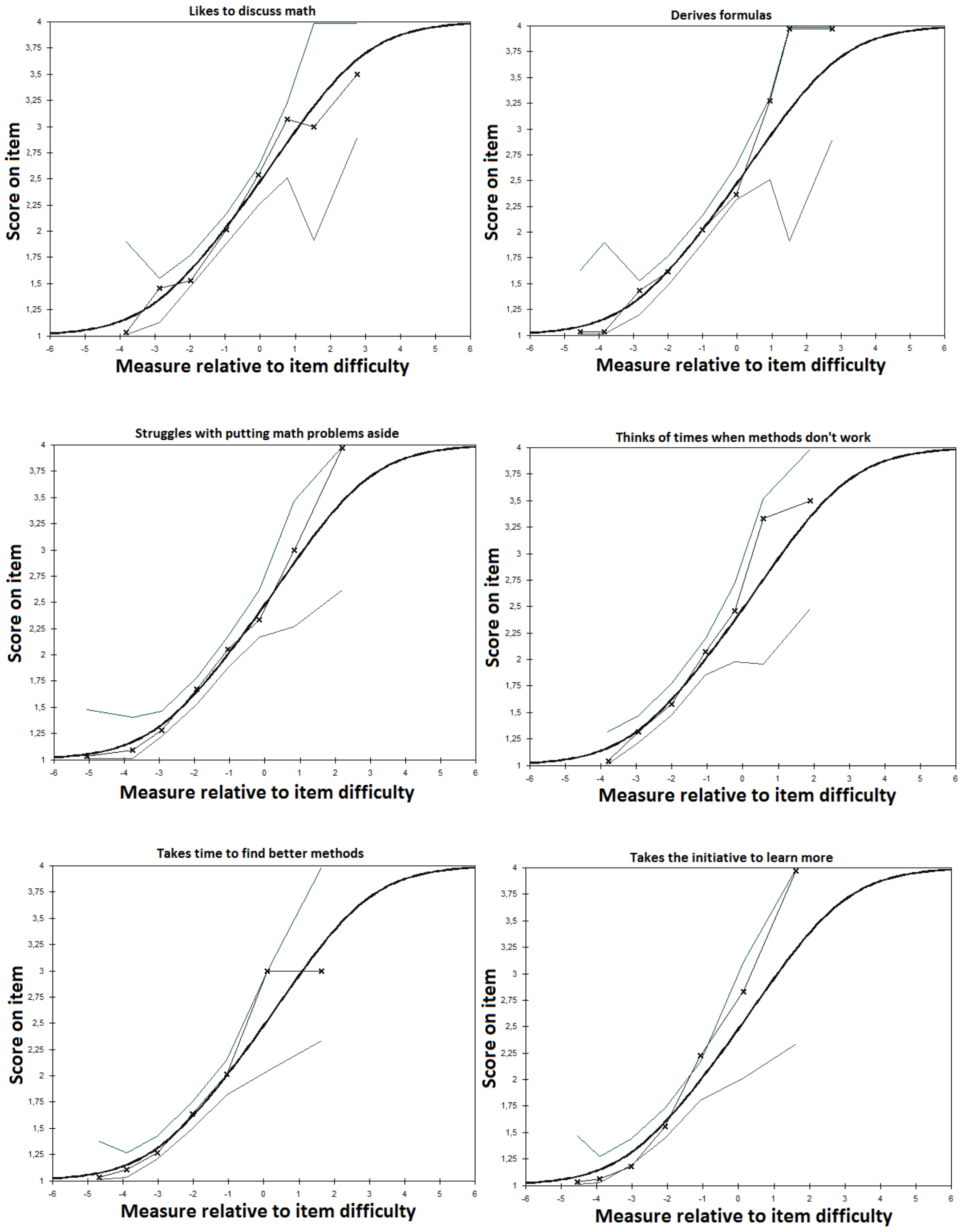


Figure 15. Item Characteristic Curves, Paper I

The category statistics are summarised in Table 5.

Table 5. Category statistics

Category	Observation	Infit mnsq	Outfit mnsq	Threshold	M→C	C→M
1	1147	0.92	0.93	None	0.76	0.42
2	2175	0.96	0.92	-2.13	0.56	0.74
3	1457	0.98	1.00	0.10	0.52	0.56
4	515	1.19	1.20	2.03	0.63	0.28

In the paper, I questioned the low C→M value of 0.28 in the fourth category. The recommended cut-value is 40% (Linacre, 2002). After the publication, however, I have questioned this static cut-value (Kaspersen, Pepin, & Sikko, 2016) for the following reason:

M→C(*x*) is the percentage of measures expected to produce observations in category *x* that are observed in this category. When the data fits the Rasch model perfectly, we expect random deviations from the expected observations. For the most, unexpected observations occur in adjacent categories. Then, the likelihood of making an unexpected observation in category *c* is a function of the distance between *c* and the expected observation. In general, distances from the extreme categories to other categories are greater than distances from categories in the centre to other categories. Thus, in most cases, the expected M→C should be greater in extreme categories than in categories around the centre.

C→M(*x*) is the percentage of observations in category *x* that belong to measures that were expected. Roughly, we expect that some persons drift to unexpected—for the most adjacent—categories. Considering two categories, for example, 3 and 4, we expect that a proportion of persons expected to respond in category 3 actually respond in the fourth category, and vice versa. Therefore, the C→M depends on the distribution of observations. When there are more observations in category 3 than in category 4, there will be more drifters from 3 to 4 than the other way around if the proportion of drifters is equivalent between the categories. Consequently, the C→M is expected to be relatively low in categories with relatively few observations.

These inferences imply that category statistics should be compared with their expected values, not the static cut value of 40%. To produce the expected values, I have conducted 1000 simulations in R (R Core Team, 2017). That is, for each person and each item, an observation was simulated based on the RSM. From the simulated matrix, an analysis was conducted, and category statistics were stored. This procedure was replicated 1000 times leading to the results in Table 6 and 7.

The simulated values confirm that expected values vary: The M→C values are, indeed, greater in the extreme categories, and the C→M values are affected by the observation distribution.

If we compare the empirical results in Table 5 with the expected values, we see that the empirical C→M value of 0.28 in the fourth category is not as bad as it seemed at first, although the value is still less than expected. Moreover, the empirical M→C value of 0.63 in the same category is not as good as it seemed (compared with 0.40) with estimated *p*-value less than 0.05. The conclusion is that the last category is the most problematic.

Table 6. Simulated M→C

Category	Mean	Min	Max	5%	95%
1	0.70	0.66	0.74	0.68	0.72
2	0.56	0.54	0.59	0.55	0.58
3	0.53	0.49	0.57	0.51	0.55
4	0.71	0.64	0.78	0.67	0.75

Table 7. Simulated C→M

Category	Mean	Min	Max	5%	95%
1	0.44	0.39	0.49	0.41	0.47
2	0.72	0.68	0.75	0.70	0.73
3	0.56	0.51	0.60	0.54	0.58
4	0.35	0.28	0.42	0.32	0.39

Critical values for Infit Mnsq and Outfit Mnsq were estimated based on the formula presented in (Smith et al., 1998, p. 78):

$$MS(WT) = 1 + \frac{2}{\sqrt{x}} \quad (9), \text{and}$$

$$MS(UT) = 1 + \frac{6}{\sqrt{x}} \quad (10),$$

where *x* = sample size. Consequently, critical values were set to Infit Mnsq = 1.1 and Outfit Mnsq = 1.3. However, Smith et al. (1998) illustrated how the Type I error rate is affected by contextual factors (e.g., sample size). Thus, I have conducted 1000 simulations to test the critical values that were published in the paper. For each analysis, the minimum and maximum Infit Mnsq and Outfit Mnsq values were recorded. The results, summarised in Table 8, indicate that at least one Infit Mnsq value above 1.26 and at least one Outfit Mnsq value above 1.24 are expected in 5% of the times when data fit the Rasch model perfectly. In conclusion, the cut value of 1.1 that was reported in Paper I is good for 'flagging', but not sufficient evidence of true misfit.

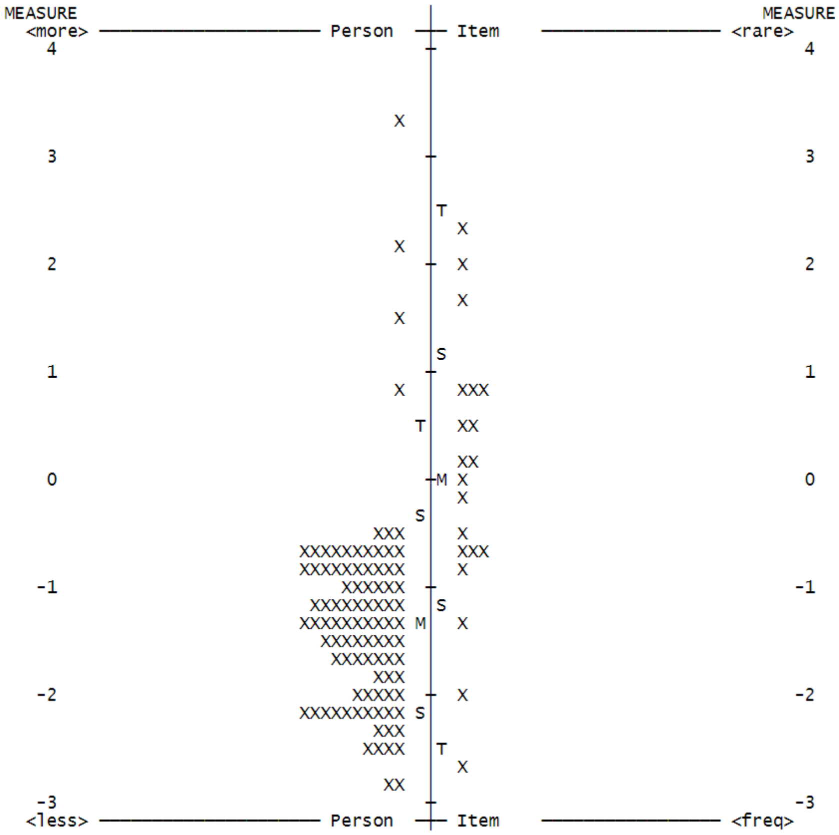
Table 8. Simulated fit statistics

		Mean	Min	Max	5%	95%
Infit	Min	0.85	0.72	0.95	0.79	0.90
Mnsq	Max	1.17	1.05	1.50	1.09	1.26
Outfit	Min	0.85	0.70	0.95	0.79	0.90
Mnsq	Max	1.16	1.06	1.40	1.10	1.24

8.2 Appendix B: Additional analyses, Paper II

In Paper II, I illustrated the person-independent property of mathematical identity from a reduction of the sample. That is, I removed 40 persons with strongest measures, and subsequently, 40 persons with lowest measures from the analysis, and showed how the structure of identity remained relatively stable.

After the publication, I have conducted more extreme reductions. First, I have reduced the sample to include the ‘bottom half’ only, except four persons, uniformly spread in the ‘upper half’ that were kept in the analysis to prevent a serious drop in reliability. The reduced sample is illustrated in Figure 16.



```
MEASURE                                                  MEASURE
 <more> ------------------ Person -+- Item    -------------- <rare>
   4                               +                              4
                                   |
                                   |
                                X  |
                                   |
   3                               +                              3
                                   |
                                   |T
                                   |  X
                                X  |
   2                               +  X                           2
                                   |
                                   |  X
                                X  |
                                   |
                                   |S
   1                               +                              1
                                X  |  XXX
                                  T|
                                   |  XX
                                   |  XX
   0                               +M X                           0
                                   |  X
                                  S|
                              XXX  |  X
                       XXXXXXXXXX  |  XXX
                       XXXXXXXXXX  |  X
  -1                         XXXXXX +                            -1
                        XXXXXXXXX  |S
                       XXXXXXXXXX M|  X
                         XXXXXXXX  |
                          XXXXXXX  |
                              XXX  |
  -2                        XXXXX  +  X                          -2
                       XXXXXXXXXX S|
                              XXX  |
                             XXXX  |T
                                   |  X
                               XX  |
  -3                               +                             -3
 <less> ------------------ Person -+- Item    -------------- <freq>
```

Figure 16. Reduced sample: ‘Upper half’ removed. M = mean, S = one standard deviation, T = two standard deviations

Figure 17 shows that, except item 3, there was no significant difference in the structure of mathematical identity between the full STEM sample and the reduced sample. A similar conclusion was made when the ‘lower half’ was removed.

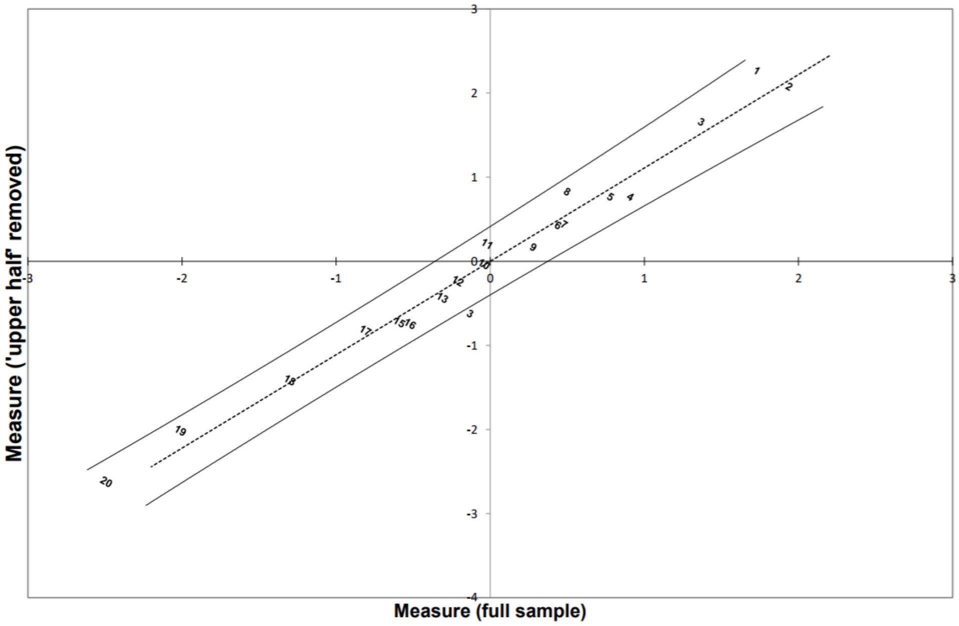
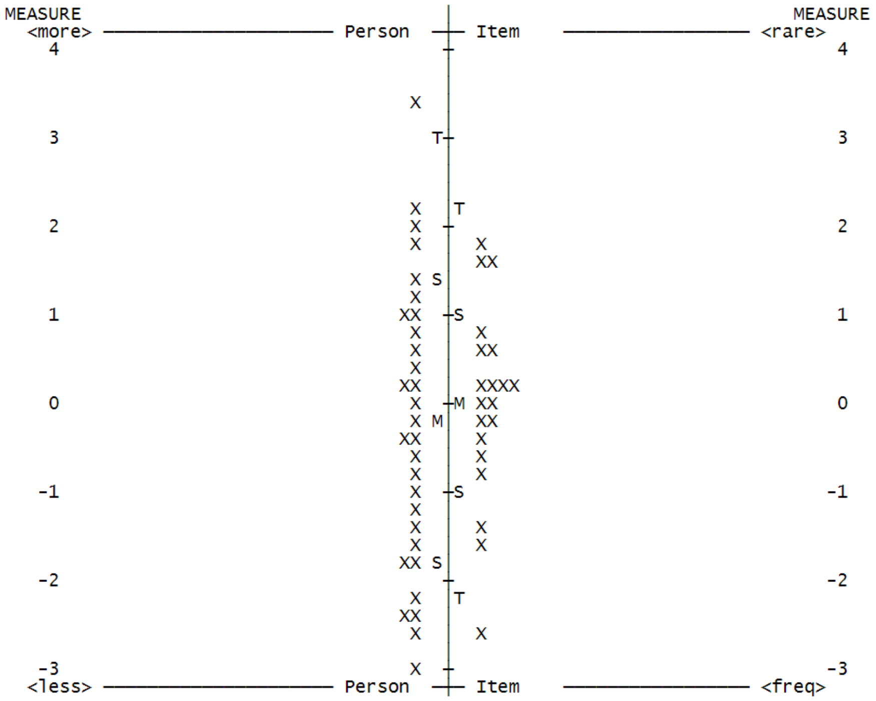


Figure 17. The structure of mathematical identity calibrated on the full STEM sample and a reduced sample with ‘upper half’ removed

To explore person-independence further, I have made a reduction based on size. That is, if my inferences in this thesis are correct, then the structure of mathematical identity should not be affected by sample size, if we ignore random measurement errors. To assess this hypothesis, I have reduced the sample to include a small sample of 30 persons only. The sample was uniformly spread along the variable, as illustrated in Figure 18.



```
MEASURE                                 |                               MEASURE
  <more> ---------------------- Person -+- Item    ------------------ <rare>
    4                                   +                                   4
                                        |
                                        |
                                     X  |
                                        |
    3                                  T+                                   3
                                        |
                                        |
                                        |
                                     X  |T
    2                                X  +                                   2
                                     X  |  X
                                        |  XX
                                     X S|
                                     X  |
    1                               XX  +S                                  1
                                     X  |  X
                                     X  |  XX
                                     X  |
                                    XX  |  XXXX
    0                                X  +M XX                               0
                                     X M|  XX
                                    XX  |  X
                                     X  |  X
                                     X  |  X
   -1                                X  +S                                 -1
                                     X  |
                                     X  |  X
                                     X  |  X
                                    XX S|
   -2                                   +                                  -2
                                     X  |T
                                    XX  |
                                     X  |  X
                                        |
   -3                                X  +                                  -3
  <less> ---------------------- Person -+- Item    ------------------ <freq>
```

Figure 18. Person-item map, reduced sample

Figure 19 shows that there was no significant difference in the structure of mathematical identity between the full STEM sample and the reduced sample.

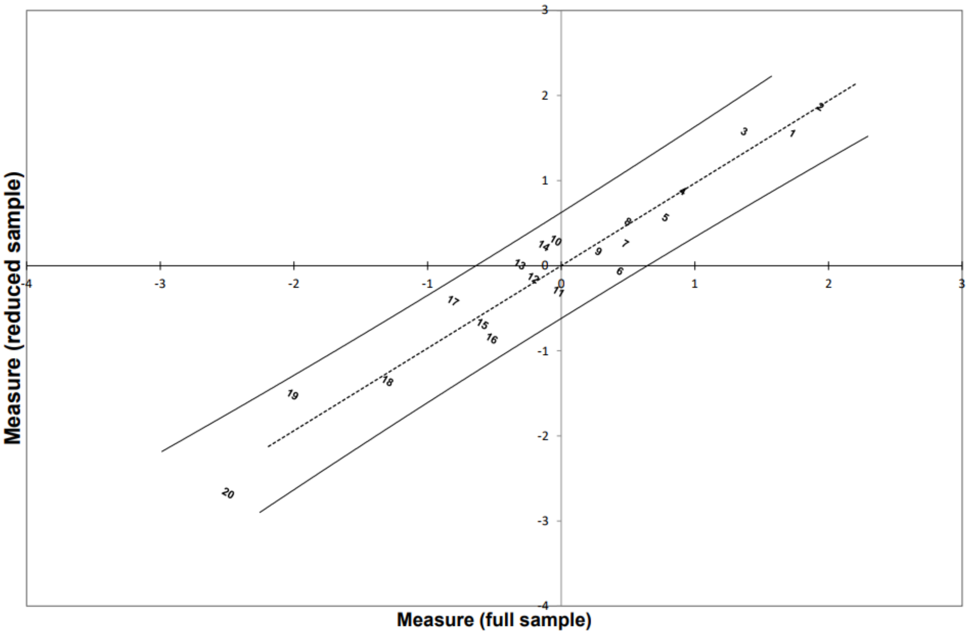


Figure 19. Differential test functioning. Full sample vs reduced sample

To assess the importance of the difference, the reduced sample was measured twice: first relative to the ‘global structure’ (from the full sample), and then relative to the ‘local structure’ (from the reduced sample). Figure 20 illustrates a convenient consequence of the person independent property of the structure of mathematical identity: individual measures hardly depend on sample size.

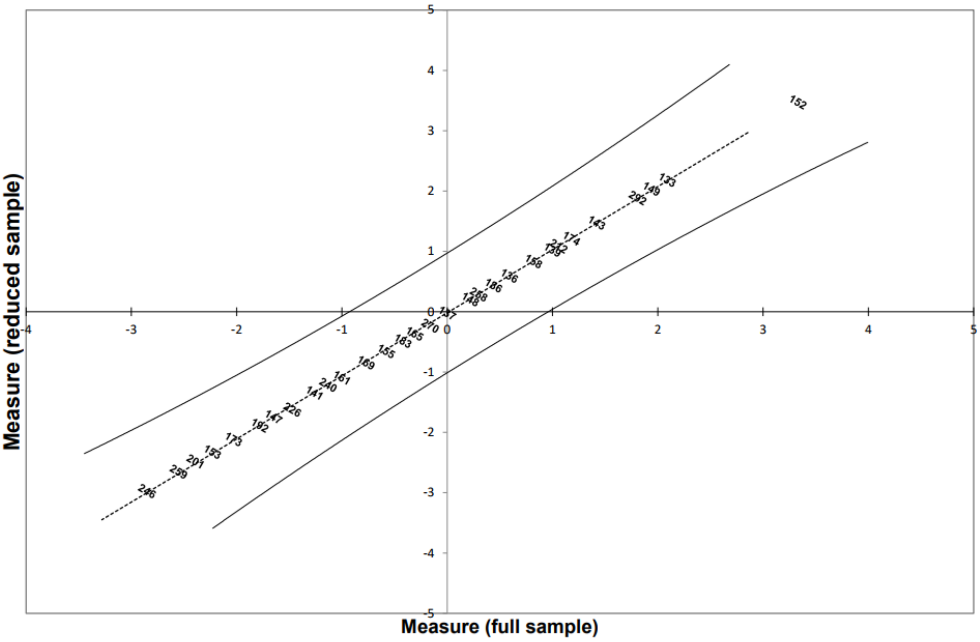


Figure 20. Person measures relative to: full sample structure vs reduced sample structure

8.3 Appendix C: Additional analyses, Paper III

The analysis in Paper III is based on an extended sample of STEM students, 361 in total. In effect, the item measures presented in Paper III (p. 5) are not equivalent to those in Paper I. To see if this divergence is due to chance or a flux in the structure of identity, I have conducted a differential test analysis in which item measures based on STEM responses in Paper I are plotted against their measures in Paper III.

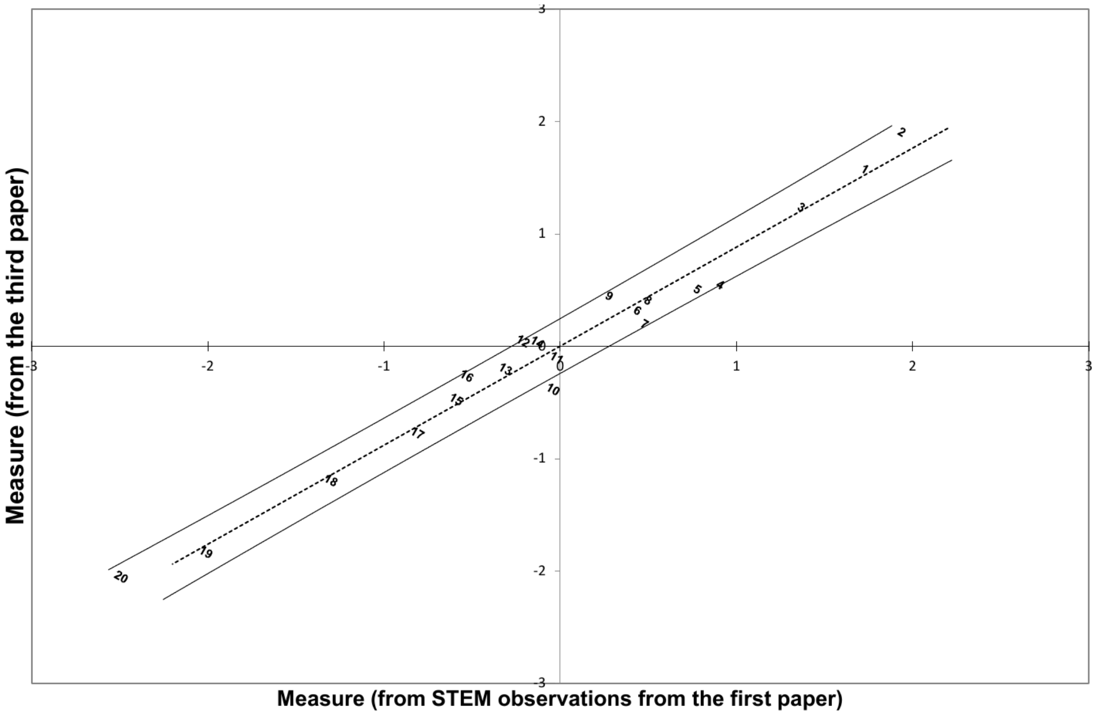


Figure 21. Differential test functioning

In sum, differences in the structures are within the expected range, except for the case of item 10: “Finding out why methods do not work”. Moreover, the impact of the differences is negligible. That is, the differences in positions hardly affect the estimation of personal positions, as can be seen in Figure 22.

The category statistics are presented in Table 9. The general impression is the same as for Paper I: C→M in the fourth category is much lower than 40%, but this is expected due to the distribution of observations.

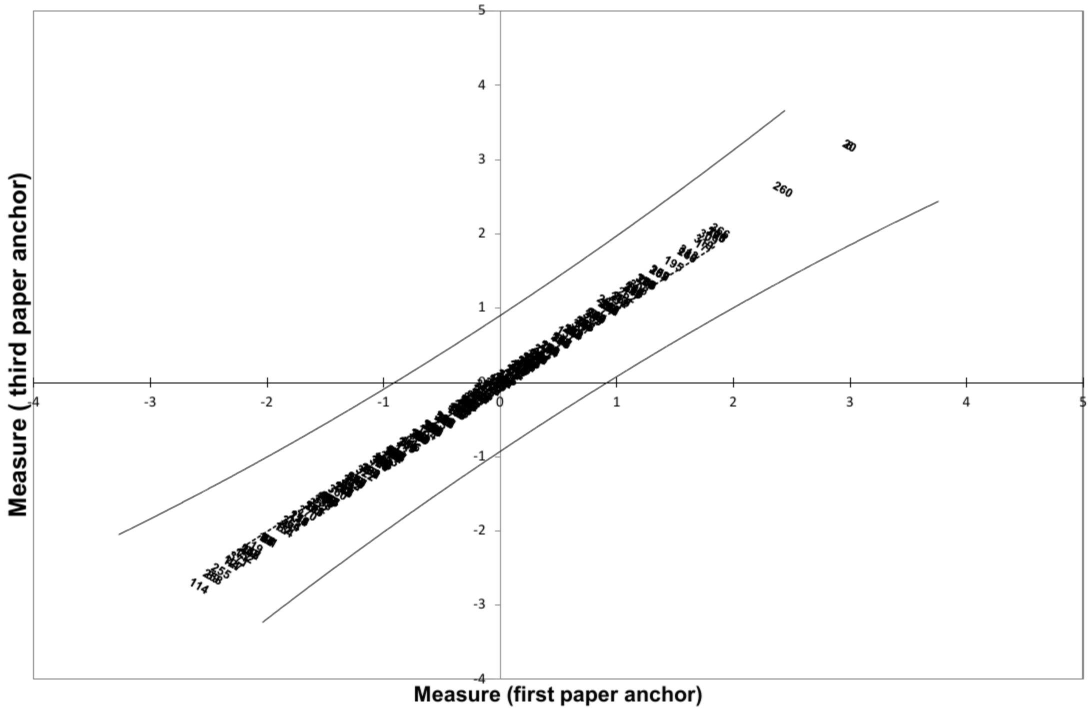


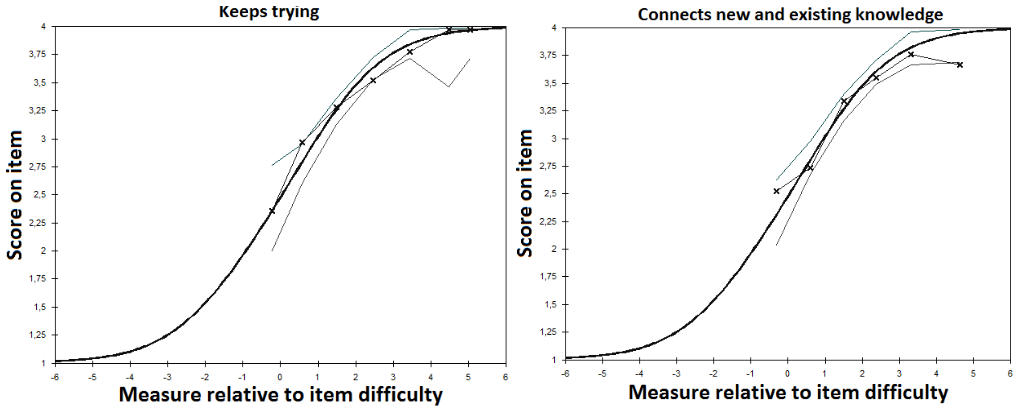
Figure 22. Person measures relative to: Paper I structure vs Paper III structure

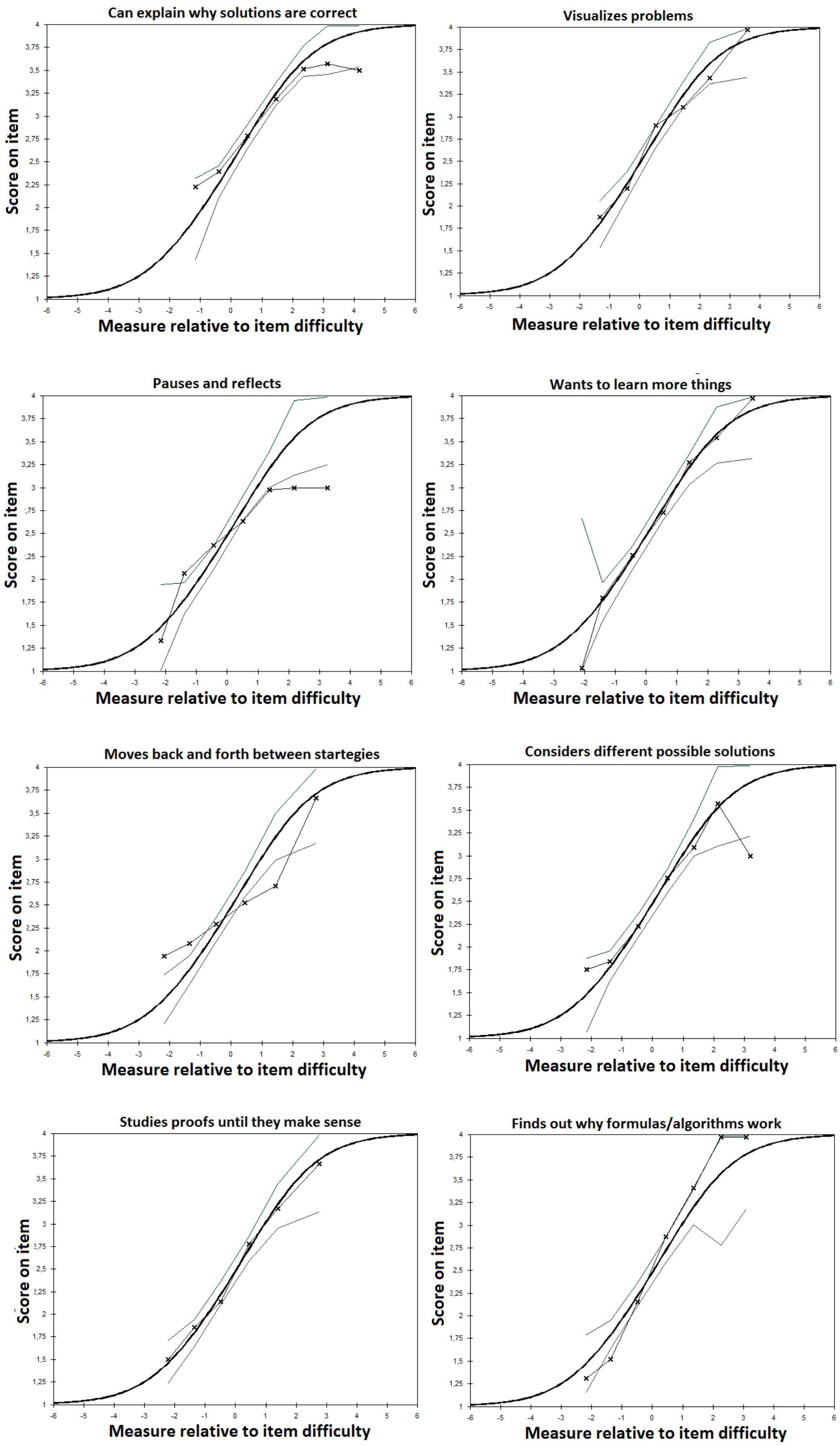
Table 9. Category statistics, fourth paper

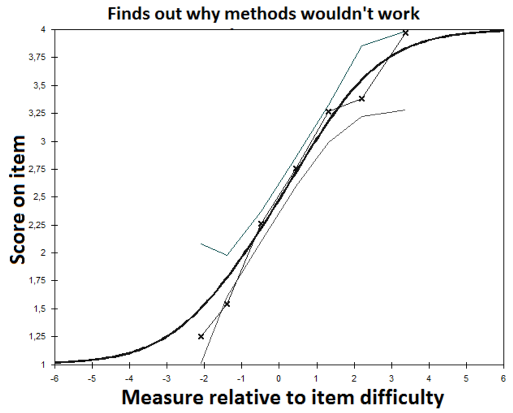
Category	Obs.	Infit mnsq	Outfit mnsq	Thr.hold	M→C|exp.	C→M|exp.
1	1397	0.93	0.95	None	0.72|0.68	0.35|0.32
2	2715	0.96	0.93	-1.81	0.52|0.56	0.69|0.71
3	2152	0.96	0.98	0.10	0.48|0.53	0.58|0.61
4	962	1.13	1.13	1.71	0.61|0.69	0.25|0.28

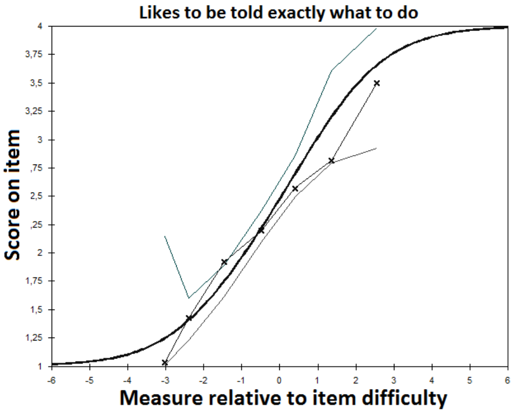
Note. The expected values are estimated based on 1000 simulations

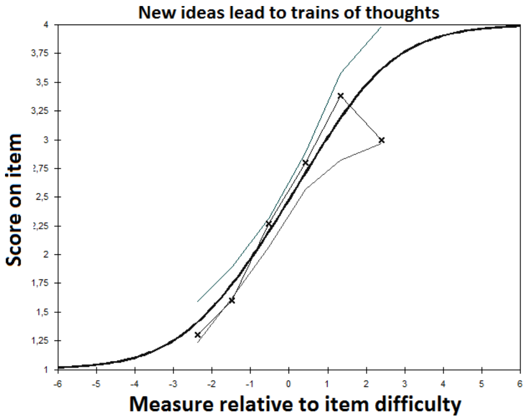
The ICCs from the third study are presented in Figure 23. The main impression is that category 4 is the most problematic, as it was indicated in Paper I.

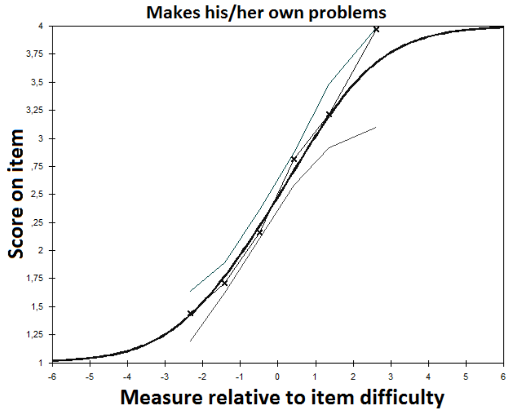


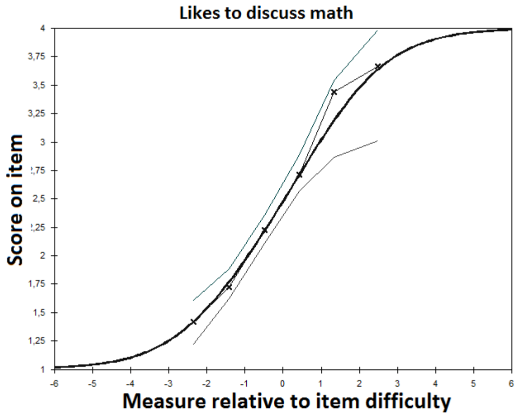

Can explain why solutions are correct
Visualizes problems
Pauses and reflects
Wants to learn more things
Moves back and forth between startegies
Considers different possible solutions
Studies proofs until they make sense
Finds out why formulas/algorithms work
Score on item
Measure relative to item difficulty
4
3,75
3,5
3,25
3
2,75
2,5
2,25
2
1,75
1,5
1,25
1
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6

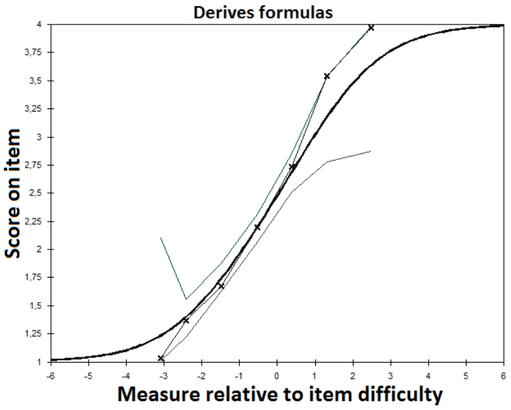

Finds out why methods wouldn't work
Score on item
Measure relative to item difficulty
1
1,25
1,5
1,75
2
2,25
2,5
2,75
3
3,25
3,5
3,75
4
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6

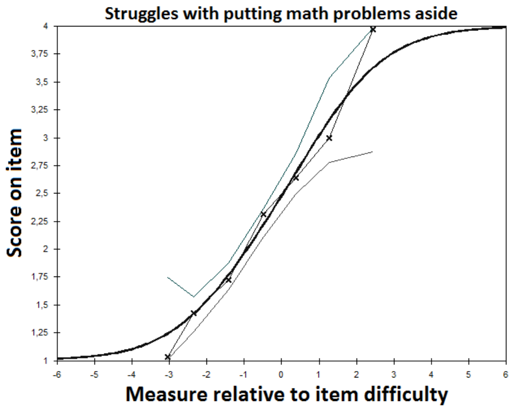

Likes to be told exactly what to do
Score on item
Measure relative to item difficulty
1
1,25
1,5
1,75
2
,25
2,5
,75
3
3,25
3,5
3,75
4
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6

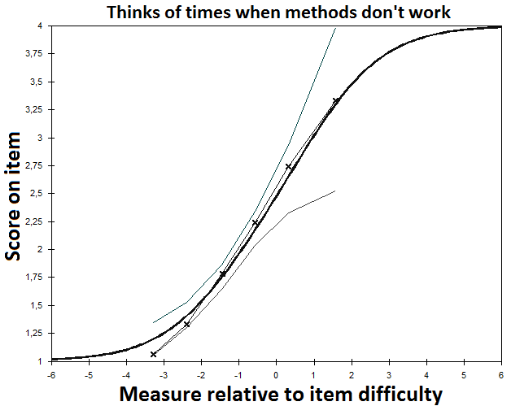

New ideas lead to trains of thoughts
Score on item
Measure relative to item difficulty
1
1,25
1,5
1,75
2
2,25
2,5
2,75
3
3,25
3,5
3,75
4
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6


Makes his/her own problems
Score on item
Measure relative to item difficulty
1
1,25
1,5
1,75
2
,25
2,5
,75
3
3,25
3,5
3,75
4
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6


Likes to discuss math
Score on item
Measure relative to item difficulty
1
1,25
1,5
1,75
2
2,25
2,5
2,75
3
3,25
3,5
3,75
4
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6


Derives formulas
Score on item
Measure relative to item difficulty
1
1,25
1,5
1,75
2
2,25
2,5
2,75
3
3,25
3,5
3,75
4
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6


Struggles with putting math problems aside
Score on item
Measure relative to item difficulty
1
1,25
1,5
1,75
2
2,25
2,5
2,75
3
3,25
3,5
3,75
4
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6


Thinks of times when methods don't work
Score on item
Measure relative to item difficulty
1
1,25
1,5
1,75
2
2,25
2,5
2,75
3
3,25
3,5
3,75
4
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6

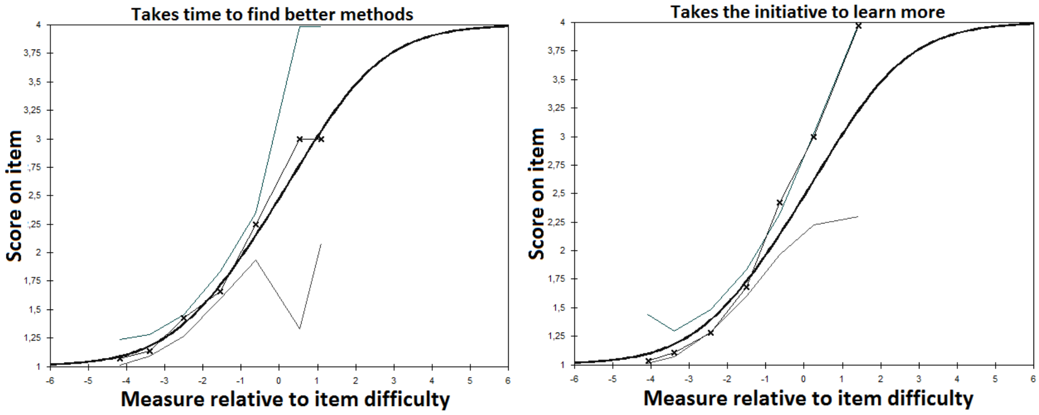
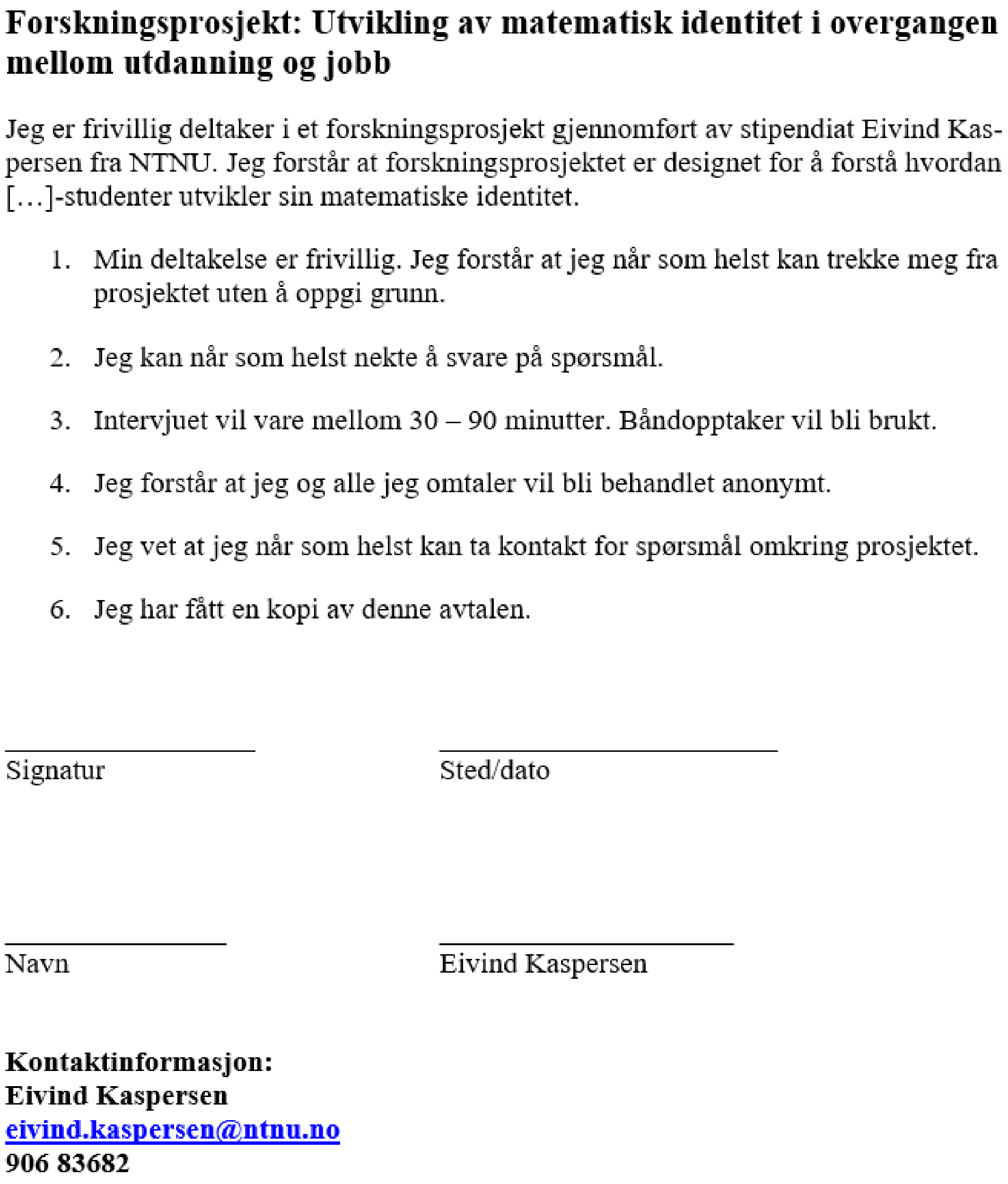


Figure 23. Item Characteristic Curves, Paper III

8.4 Appendix D: Letter of consent



Forskningsprosjekt: Utvikling av matematisk identitet i overgangen mellom utdanning og jobb

Jeg er frivillig deltaker i et forskningsprosjekt gjennomført av stipendiat Eivind Kaspersen fra NTNU. Jeg forstår at forskningsprosjektet er designet for å forstå hvordan […]-studenter utvikler sin matematiske identitet.

1. Min deltakelse er frivillig. Jeg forstår at jeg når som helst kan trekke meg fra prosjektet uten å oppgi grunn.
2. Jeg kan når som helst nekte å svare på spørsmål.
3. Intervjuet vil vare mellom 30 – 90 minutter. Båndopptaker vil bli brukt.
4. Jeg forstår at jeg og alle jeg omtaler vil bli behandlet anonymt.
5. Jeg vet at jeg når som helst kan ta kontakt for spørsmål omkring prosjektet.
6. Jeg har fått en kopi av denne avtalen.

Signatur

Sted/dato

Navn

Eivind Kaspersen

Kontaktinformasjon:
Eivind Kaspersen
eivind.kaspersen@ntnu.no
906 83682

8.5 Appendix E: Instrument, English translation

Never/almost never(1), Sometimes(2), Often(3), Always/almost always(3), Don’t know(9)					
1. I take the initiative to learn more about math than what is required at school/work.	1	2	3	4	9
2. When I learn a new method, I take time to find out if I can find a better method.	1	2	3	4	9
3. When I learn a new method, I try to think of situations when it wouldn’t work.	1	2	3	4	9
4. I struggle with putting math problems aside.	1	2	3	4	9
5. If I forget a formula or method, I try to derive it myself.	1	2	3	4	9
6. I get engaged when someone starts a mathematical discussion.	1	2	3	4	9
7. When I learn something new, I make my own problems.	1	2	3	4	9
8. Math ideas that I hear or learn about help me inspire new trains of thoughts.	1	2	3	4	9
9. When I learn a new method, I like to be told exactly what to do.	1	2	3	4	9
10. When I try to use a method that doesn’t work, I spend time to find out why it didn’t work.	1	2	3	4	9
11. When I learn a new formula/algorithm, I try to understand why it works.	1	2	3	4	9
12. When I face a proof, I study it until it becomes meaningful.	1	2	3	4	9
13. When I face a math problem, I consider different possible ways I can solve it.	1	2	3	4	9
14. When I work with a math problem, I move back and forth between various strategies.	1	2	3	4	9
15. When I learn something new, it makes me want to learn more things.	1	2	3	4	9
16. When I work with a problem, I pause along the way to reflect on what I am doing.	1	2	3	4	9
17. If I get stuck on a problem, I try to visualize it.	1	2	3	4	9
18. I can explain why my solutions are correct.	1	2	3	4	9
19. I try to connect new things I learn to what I already know.	1	2	3	4	9
20. If I immediately do not understand what to do, I keep trying.	1	2	3	4	9

8.6 Appendix F: Instrument, Norwegian translation

Aldri/nesten aldri(1), Noen ganger(2), Ofte(3), Alltid/nesten alltid(3), Vet ikke(9)					
1. Jeg tar initiativ til å lære mer om et matematisk emne enn skole/jobb legger opp til.	1	2	3	4	9
2. Når jeg lærer en ny metode, bruker jeg tid på å se om jeg kan finne en bedre metode.	1	2	3	4	9
3. Når jeg lærer en ny metode, prøver jeg å finne situasjoner hvor denne ikke virker.	1	2	3	4	9
4. Jeg har problemer med å legge fra meg matematiske oppgaver.	1	2	3	4	9
5. Dersom jeg har glemt en formel/metode, prøver jeg å utlede den selv.	1	2	3	4	9
6. Jeg blir engasjert når noen starter en matematisk diskusjon.	1	2	3	4	9
7. Når jeg lærer noe nytt, stiller jeg meg selv egne spørsmål som jeg jobber med.	1	2	3	4	9
8. Matematiske ideer jeg leser eller hører om setter meg på sporet av egne tankerekker.	1	2	3	4	9
9. Når jeg lærer en ny matematisk metode, liker jeg å bli fortalt nøyaktig hva jeg skal gjøre.	1	2	3	4	9
10. Hvis jeg prøver på en metode som ikke fører frem, bruker jeg tid på å finne ut hvorfor denne ikke virker.	1	2	3	4	9
11. Når jeg lærer en ny metode/algoritme, prøver jeg å finne ut hvorfor den virker.	1	2	3	4	9
12. Når jeg kommer over et matematisk bevis/forklaring, studerer jeg det til det gir mening.	1	2	3	4	9
13. Når jeg møter et matematisk problem, tenker jeg over om det finnes flere måter å løse oppgaven på.	1	2	3	4	9
14. Når jeg jobber med et matematisk problem hopper jeg mellom ulike strategier.	1	2	3	4	9
15. Når jeg lærer noe nytt, fører det til at det er flere ting jeg ønsker å finne ut.	1	2	3	4	9
16. Når jeg jobber med en oppgave, stopper jeg opp underveis og reflekterer over hva jeg gjør.	1	2	3	4	9
17. Hvis jeg står fast, prøver jeg å visualisere problemet.	1	2	3	4	9
18. Jeg kan forklare hvorfor løsningen min er rett.	1	2	3	4	9
19. Jeg prøver å koble det jeg lærer opp mot det jeg vet fra før.	1	2	3	4	9
20. Jeg fortsetter å prøve meg frem selv om jeg ikke får det til med en gang.	1	2	3	4	9

8.7 Appendix G: Interview guide

1. Introduction
    a. Explain the purpose of the study.
    b. Explain anonymity and how the interviewee can withdraw the interview at any time.
    c. Explain that the estimated time is 60-90 minutes. The interviewee can abort at any time.
    d. Ask if the interviewee has any questions about the study or the interview.
    e. Written and oral consent

2. Introductory questions [approximately 15-20 minutes each]
    a. Can you start by telling me how you have perceived your time here at the University?
    b. Are you currently engaged in any projects? Can you tell me about it?
    c. Can you tell me how you go about when you work with mathematical problems?
    d. Can you tell me how you go about when you learn new mathematics?

3. Probing questions [when interviewee makes general statements]
    a. Do you have a concrete example?
    b. Could you say something more about that?
    c. Do you have further examples?

4. Specifying questions
    a. [when interviewee explains something loosely]
       Going back to this [e.g., situation], could you explain it once again, now as detailed as possible? Any detail could be important.
    b. [pick up on notes]
       You talked (earlier) about […]. Can you say something more about that?

5. Interpreting questions [when something is unclear]
    a. You then mean that…?
    b. Is it correct that…?
    c. I did not quite understand […]. Could you explain it again?

9 References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561-573.

Andrich, D. (1989). Distinctions between assumptions and requirements in measurement in the social sciences. In J. A. Keats, R. Taft, R. A. Heath, & S. H. Lovibond (Eds.), *Proceedings of the XXIVth international congress of psychology, 4, 7-16: Mathematical and theoretical systems*. North Holland: Elsevier Science.

Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? *Medical Care, 42*(1), 1-7.

Andrich, D., Marais, I., & Humphry, S. (2012). Using a theorem by Andersen and the dichotomous Rasch model to assess the presence of random guessing in multiple choice items. *Journal of Educational and Behavioral Statistics, 37*(3), 417-442.

Andrich, D., Sheridan, B., & Luo, G. (2010). Rasch models for measurement: RUMM2030. Perth, Western Australia: RUMM Laboratory Pty Ltd.

Axelsson, G. B. M. (2009). Mathematical identity in women: The concept, its components and relationship to educative ability, achievement and family support. *International Journal of Lifelong Education, 28*(3), 383-406.

Bacon, M. (2012). *Pragmatism: An introduction*. Cambridge, UK: Polity.

Bakhtin, M. (1981). *The dialogic imagination: Four essays by M. M. Bakhtin*. Austin, TX: University of Texas Press.

Black, L., Williams, J., Hernandez-Martinez, P., Davis, P., Pampaka, M., & Wake, G. (2010). Developing a 'leading identity': The relationship between students' mathematical identities and their career and higher education aspirations. *Educational Studies in Mathematics, 73*(1), 55-72.

Bond, T., & Fox, C. M. (2003). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Oxford, UK: Routledge.

Burton, L. (1998). The practices of mathematicians: What do they tell us about coming to know mathematics? *Educational Studies in Mathematics, 37*(2), 121-143.

Carpenter, T. P., & Lehrer, R. (1999). Teaching and learning mathematics with understanding. In E. Fennema & T. A. Romberg (Eds.), *Mathematics classrooms that promote understanding* (pp. 19-32). Mahwah, NJ: Lawrence Erlbaum Associates.

Collins, R. (1975). *Conflict sociology*. New York, NY: Academic Press.

Cote, J. E., & Levine, C. G. (2014). *Identity, formation, agency, and culture: A social psychological synthesis*. New York, NY: Psychology Press.

Creswell, J. W., & Clark, V. L. P. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage.

Darragh, L. (2016). Identity research in mathematics education. *Educational Studies in Mathematics, 93*(1), 19-33.

DeMars, C. (2010). *Item response theory*. New York, NY: Oxford University Press.

Divgi, D. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement, 23*(4), 283-298.

Durkheim, E. (1972). *Emile Durkheim: Selected writings* (A. Giddens Ed.). New York, NY: Cambridge University Press.

Engelhardt, H. T., & Caplan, A. L. (1987). *Scientific controversies*. New York, NY: Cambridge University Press.

Engeström, Y. (1987). Learning by expanding. *Helsinki: Orienta-Konsultit Oy*.

Entwistle, N. (1997). The approaches and study skills inventory for students (ASSIST). *Edinburgh: Centre for Research on Learning and Instruction, University of Edinburgh*.

Gee, J. P. (2000). Identity as an analytic lens for research in education. *Review of Research in Education, 25*, 99-125.

Hannula, M. S., Di Martino, P., Pantziara, M., Zhang, Q., Morselli, F., Heyd-Metzuyanim, E., . . . Jansen, A. (2016). *Attitudes, beliefs, motivation, and identity in mathematics education*. Switzerland: Springer International Publishing.

Harris, D., Black, L., Hernandez-Martinez, P., Pepin, B., & Williams, J. (2015). Mathematics and its value for engineering students: What are the implications for teaching? *International Journal of Mathematical Education in Science and Technology, 46*(3), 321-336.

Hernandez-Martinez, P., Williams, J., Black, L., Davis, P., Pampaka, M., & Wake, G. (2011). Students' views on their transition from school to college mathematics: Rethinking ‘transition’as an issue of identity. *Research in Mathematics Education, 13*(2), 119-130.

Hess, D. J. (1997). *Science studies: An advanced introduction*. New York, NY: New York University Press.

Hiebert, J. (1986). *Conceptual and procedural knowledge: The case of mathematics*. Hillsdale, NJ: Erlbaum.

Holland, D., Lachicotte, W., Skinner, D., & Cain, C. (1998). *Identity and agency in cultural worlds*. Cambridge, MA: Harvard University Press.

Holland, D., Lachicotte, W., Skinner, D., & Cain, C. (2001). *Identity and agency in cultural worlds*. Cambridge, MA; London, UK: Harvard University Press.

Hoyles, C., Wolf, A., Molyneux-Hodgson, S., & Kent, P. (2002). Mathematical skills in the workplace. *The Science, Technology and Mathematics Council.*

James, W. (1977). *The writings of William James: A comprehensive edition* (J. J. McDermott Ed.). London, UK: The University of Chicago Press.

Kalleberg, R., Balto, A., Cappelen, A., Nagel, A., Nymoen, H., Rønning, H., & Nagell, H. (2006). Forskningsetiske retningslinjer for samfunnsvitenskap, humaniora, juss og teologi. *Oslo: De nasjonale forskningsetiske komiteer*, 5-35.

Kaspersen, E. (2015). Using the Rasch Model to measure the extent to which students work conceptually with mathematics. *Journal of Applied Measurement, 16*(4), 336-352.

Kaspersen, E., Pepin, B., & Sikko, S. A. (2016). Measuring student teachers’ practices and beliefs about teaching mathematics using the Rasch model. *International Journal of Research & Method in Education, 40*(4), 421-442.

Kaspersen, E., Pepin, B., & Sikko, S. A. (2017). Measuring STEM students’ mathematical identities. *Educational Studies in Mathematics, 95*(2), 163-179.

Kaspersen, E., Pepin, B., & Sikko, S. A. (in print). The association between engineering students’ self-reported mathematical identities and average grades in mathematics courses. *Proceedings of the Tenth Congress of the European Society for Research in Mathematics Education (CERME10).*

Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academies Press.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago, IL: The University of Chicago Press.

Kuhn, T. S. (1977). *The essential tension: Selected studies in scientific tradition and change*. Chicago, IL: The University of Chicago Press.

Kvale, S., & Brinkmann, S. (2009). *Interviews: Learning the craft of qualitative interviewing*. Thousand Oaks, CA: Sage Publications.

Leont'ev, A. N. (1978). *Activity, consciousness, and personality*. Englewood Cliffs, NJ: Prentice-Hall, Inc,.

Linacre, J. M. (1989). *Multi-faceted Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (1999). Understanding Rasch measurement: Estimation methods for Rasch measures. *Journal of Outcome Measurement, 3*(4), 382-405.

Linacre, J. M. (2000). Comparing ‘Partial Credit Models’ (PCM) and ‘Rating Scale Models’ (RSM). *Rasch Measurement Transactions, 14*(3), 768.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

Linacre, J. M. (2006). WINSTEPS: Rasch measurement computer program. Chicago, IL: Winsteps.com.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. v. d. Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101-121). New York, NY: Springer.

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.

Nardi, E. (2007). *Amongst mathematicians: Teaching and learning mathematics at university level*. New York, NY: Springer.

Nering, M., & Ostini, R. (2010). *Handbook of polytomous item response theory models*. New York, NY: Routledge.

Pampaka, M., Pepin, B., & Sikko, S. A. (2016). Supporting or alienating students during their transition to Higher Education: Mathematically relevant trajectories in the contexts of England and Norway. *International Journal of Educational Research, 79*, 240-257.

Pampaka, M., Williams, J., Hutcheson, G., Black, L., Davis, P., Hernandez-Martinez, P., & Wake, G. (2013). Measuring alternative learning outcomes: Dispositions to study in higher education. *Journal of Applied Measurement, 14*(2), 197-218.

Peirce, C. S., Hartshorne, C., & Weiss, P. (1935). *Collected papers of Charles Sanders Peirce* (Vol. 2). Cambridge, MA: Harvard University Press.

Popper, K. (1963). *Conjectures and refutations: The growth of scientific knowledge*. London: Routledge, Keagan & Paul.

Putnam, H. (1981). *Reason, truth and history* (Vol. 3). New York, NY: Cambridge University Press.

Radford, L. (2008). Connecting theories in mathematics education: Challenges and possibilities. *ZDM, 40*(2), 317-327.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago, IL: University of Chicago Press.).

Rasch, G. (1961). *On general laws and the meaning of measurement in psychology.* Paper presented at the Proceedings of the fourth Berkeley symposium on mathematical statistics and probability.

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.

Rystad, J. (1993). Alt glemt på grunn av en ubrukelig eksamensform? En empirisk undersøkelse av Matematikk 2 eksamen ved NTH. *UNIPED*(2-3), 29-15.

Schoenfeld, A. H. (2007). Method. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 69-107). Charlotte, NC: Information Age Publishing.

Sfard, A., & Prusak, A. (2005). Telling identities: In search of an analytic tool for investigating learning as a culturally shaped activity. *Educational Researcher, 34*(4), 14-22.

Skemp, R. R. (1987). *The psychology of learning mathematics*. London: Psychology Press.

Skinner, D. (1990). Nepalese children's understanding of themselves and their social world. *(Unpublished doctoral dissertation). The University of North Carolina at Chapel Hill*.

Smith, R. M., Schumacker, R. E., & Busch, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement, 2*(1), 66-78.

Stetsenko, A. (2013). The challenge of individuality in cultural-historical activity theory: 'Collectividual' dialectics from a transformative activist stance. *Outlines: Critical Practice Studies, 14*(2), 07-28.

Stetsenko, A., & Arievitch, I. M. (2004). The self in cultural-historical activity theory: Reclaiming the unity of social and individual dimensions of human development. *Theory & Psychology, 14*(4), 475-503.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*(4), 529-554.

Thurstone, L. L. (1954). The measurement of values. *Psychological Review, 61*(1), 47-58.

Thurstone, L. L. (1959). *The measurement of values*. Chicago, IL: The University of Chicago Press.

Traub, R. (1983). A priori considerations in choosing an item response model. *Applications of Item Response Theory, 57*, 70.

Vianna, E., & Stetsenko, A. (2011). Connecting learning and identity development through a transformative activist stance: Application in adolescent development in a child welfare program. *Human Development, 54*(5), 313-338.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wake, G. (2011). Introduction to the Special Issue: Deepening engagement in mathematics in pre-university education. *Research in Mathematics Education, 13*(2), 109-118.

Waller, M. I. (1976). Estimating parameters in the Rasch model: Removing the effects of random guessing. *Educational Testing Service*(1), 1-17.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge, UK: Cambridge University Press.

Williams, J. (2011). Teachers telling tales: The narrative mediation of professional identity. *Research in Mathematics Education, 13*(2), 131-142.

Williams, J., & Wake, G. (2007). Black boxes in workplace mathematics. *Educational Studies in Mathematics, 64*(3), 317-343.

Wolfe, E., & Smith, E. (2006a). Instrument development tools and activities for measure validation using Rasch models: Part I-instrument development tools. *Journal of Applied Measurement*, *8*(1), 97-123.

Wolfe, E., & Smith, E. (2006b). Instrument development tools and activities for measure validation using Rasch models: Part II-validation activities. *Journal of Applied Measurement*, *8*(2), 204-234.

Wood, L., Petocz, P., & Reid, A. (2012). *Becoming a mathematician: An international perspective.* Dordrecht: Springer.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago, IL: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Wu, M., & Adams, R. (2012). Properties of Rasch residual fit statistics. *Journal of Applied Measurement*, *14*(4), 339-355.

Using the Rasch Model to Measure the Extent to which Students Work Conceptually with Mathematics

Eivind Kaspersen
Sør-Trøndelag University College

Differences between working conceptually and procedurally with mathematics are well documented. In short, working procedurally can be characterized as learning and applying 'rules without reason.' Working conceptually, in contrast, means creating and applying a web of knowledge. To continue this line of research, an instrument that is able to measure the level of conceptual work, and that is based on the basic requirements of measurement, is desireable. As such, this paper presents a Rasch calibrated instrument that measures the extent to which students work conceptually with mathematics. From a sample of 133 student teachers and 185 Civil Engineering students, 20 items are concluded as being productive for measurement.

Requests for reprints should be sent to Eivind Kaspersen, Sør-Trøndelag University College, HiST ALT, N-7004,Trondheim, Norway, e-mail: eivind.kaspersen@hist.no.

Introduction

The field of mathematical education has been aware of differences between conceptual and procedural knowledge for more than four decades. Although academics use different words to express these ideas, such as relational vs. instrumental understanding (Skemp, 1989), conceptual vs. procedural knowledge (Hiebert, 2013), or deep versus surface approach (e.g., Booth, 2008), there seems to be some degree of consensus when it comes to qualitative differences between what I will refer to in this paper as conceptual and procedural knowledge. Procedural knowledge can be characterized as easier to 'understand,' immediately rewarding, fast and reliable (Skemp, 1989, p. 9) and focused on understanding syntactic convention and rules for solving problems (Hiebert, 2013, p. 7). In contrast, conceptual understanding is more adaptable, easier to remember, effective as a goal in itself, organic in nature (Skemp, 1989, pp. 9-11) and rich in relationships (Hiebert, 2013, pp. 3-5).

Subsequently, the practices typically associated with conceptual and procedural knowledge have been a frequent topic of discussion. Both Hiebert (2013) and Skemp (1989) highlighted the differences between 'meaningful' and 'rote' learning, whereby what is considered to be meaningful learning is associated with conceptual understanding. Procedures, by contrast, can be learned with or without meaning (Hiebert, 2013, p. 8). Furthermore, Carpenter and Lehrer (1999) highlighted five mental activities that promote understanding: constructing relationships (e.g., relating new ideas to existing knowledge), extending and applying mathematical knowledge (e.g., creating rich and integrated knowledge structures), reflecting about experiences (e.g., conscious examination of one's own actions), articulating what one knows (e.g., communicating one's ideas), and making mathematical knowledge one's own (e.g., developing personal investments in building knowledge) (pp. 20-23).

To expand upon these theories, this study aimed to develop an instrument for measuring the extent to which students work conceptually with mathematics. The rationale behind the study is that such an instrument can (statistically) verify existing knowledge (e.g., test the hypothesis that people who work conceptually with mathematics remember content more easily than those working procedurally) and highlight anomalous results for further research. Additionally, the instrument can be used as a tool for extending theory. For instance, the instrument can be used as a screening tool for studies examining qualitative differences between people working conceptually and those procedurally.

As the purpose is to measure (as opposed to model) how people work with mathematics conceptually, this study is situated within the Rasch paradigm (Rasch, 1960). The Rasch measurement is based on the basic requirements for measurement, such as unidimensionality, additivity, and invariance (Andrich, 1989). A key feature of the Rasch model is that persons and items are not discriminated, which means that they can be measured on the same scale (Wright and Stone, 1979). In the simplest dichotomous version, the probability that a person agrees with, or answers correctly, to an item, is a function of the distance between person 'ability' and item 'difficulty'. The greater the person measure in relation to item measure, the closer to 1 the probability gets. Conversely, the greater the item measure relative to person measure, the closer to 0 the probability gets. As such, the Rasch model consists of only two parameters, β_n and δ_i:

$$\Pr\{x_{ni}=1|\beta_n,\delta_i\}=\frac{\exp[\beta_n-\delta_i]}{1+\exp[\beta_n-\delta_i]},$$

where,

β_n is the measure of person n,

δ_i is the measure of item i, and

$\Pr\{x_{ni}=1|\beta_n,\delta_i\}$ is the probability of person n answering correctly (or agreeing) to item i (Wright and Stone, 1979, p. 15).

One of the consequences of situating the study in the Rasch paradigm is epistemological by nature; how the researcher positions herself in relation to what is being researched. Kuhn (1979) described a typical relationship between theory and measurement:

> Quantitative facts cease to seem simply the "given." They must be fought for and with, and in this fight the theory with which they are to be compared proves the most potent weapon. Often scientists cannot get numbers that compare well with theory until they know what numbers they should be making nature yield. (p. 193)

Unlike traditional statistical modelling, it is not the theory that needs to be fixed when there are discrepancies between the data and the model. Rather, the researcher needs to take a closer look at the data (e.g., the test items) to see if they fulfil the requirements for measurement.

Methodology

Data Collection

The instrument development was heavily informed by theory. Specifically, information regarding the trait to be measured was gathered from three different sources: (a) from existing literature; (b) from persons with expected knowledge about the trait; and (c) from other related instruments. In many cases, there were agreements between two (or all three) sources of information.

(a) The organic nature of relational schemes provides one example of how existing literature guided item selection. Skemp (1989) asserted that when people get satisfaction from relational understanding, they are likely to seek out new material, like a tree growing or an animal seeking out new territory (p. 11). This point led to two items:

- *I take the initiative to learn more about math than what is required at school/work.*
- *Math ideas that I hear or learn about help me inspire new trains of thoughts.*

(b) To get information from persons with expected knowledge about the trait, e-mails were sent to academics teaching mathematics to prospective teachers, as they were expected to possess knowledge about both mathematics and mathematical didactics. The e-mails asked for typical characteristics of persons working 'deeply' with mathematics and characteristics of persons working 'on the surface'. Similar e-mails were also sent to STEM (science, technology, engineering and mathematics) students and newly qualified students from these fields. Accordingly, 13 different persons contributed information. The items below are two examples of items that resulted from this phase:

- *When I try to use a method that doesn't work, I use time to find out why it didn't work.*
- *If I forget a math formula or method, I try to derive it myself.*

(c) Some of the items were influenced by items from an existing instrument: 'The approaches and study skills inventory for students (ASSIST)' (Entwistle, 1997). However, the ASSIST- instrument is not situated in the Rasch-paradigm. Nor is it specifically directed towards mathematics. Consequently, relevant items were translated to yield mathematical work. The following two items are examples of items influenced by the ASSIST-instrument:

- *I struggle with putting math problems aside.*
- *When I work with a problem/assignment, I pause along the way to reflect on what I am doing.*

In order to find items on the 'easy' end of the variable, some of the possible characteristics of people working on the surface were reversed. For instance, one characteristic that emerged was that 'giving up easily' is typical for people working on the surface. As such, this was translated into a possible item at the 'easy' end of the scale:

- *If I do not immediately understand what to do, I keep trying.*

One item (item 9) was reversely coded due to negative point-measure correlation. Therefore, the meaning of this item was reversed: *When I learn a new method I [don't] like to be told exactly what to do.*

After a pool of 50 items had been collected, a Norwegian translation of the items was tested on 88 student teachers. To avoid fatigue, each person responded to a random sample of 36 items only so that each item was responded to an equal number of times (±1). The analysis of this first pilot will not be discussed in detail in this paper, but in short, obvious ill-behaving items were removed or rephrased. Additionally, the person-item map was inspected to search for locations along the variable that needed more items to increase person reliability. Specifically, more items at the 'extremely deep' end were needed. As a consequence, 9 new items intended to belong to this end were formulated, for instance:

- *When I learn a new method, I try to think of situations when it wouldn't work.*
- *I take the initiative to learn more about math than what is required at school/work.*

From this, the revised instrument was administered to yet another 45 student teachers, and, based on item fit and location along the variable, a final version of 30 items was chosen. Subsequently, the instrument was administered to a convenience sample of 185 STEM students at a Norwegian university. To discuss invariance between student teachers and STEM students, the full sample ($N = 318$) will be used in this paper - although 2 items in the final instrument were not administered to the first group of student teachers. In this case, responses to these items were coded as 'missing data.' Each item was then hypothesized to be on the 'surface,' 'middle,' or 'deep' end of the variable, and discussed with three teacher educators, for the purpose of substantive validity later in the analysis. In the analysis, the WINSTEPS software (Linacre, 2006) was used.

Choosing Categories

The choice of rating scales has long been a subject for debate. Generally, more categories provide more precise person estimates due to more statistical information, as long as the respondents can distinguish between the categories (Bond and Fox, 2007, p. 109). In this study the categories *never/almost never*, *sometimes*, *often*, and *always/almost always*, were chosen. However, it was hypothesized that some of the items were cognitively demanding, and, therefore, the option of responding *I don't know* was included. This category is not unproblematic and has been widely discussed. Usually, the *I don't know* - category do not belong in the hierarchy (e.g., Lopez, 1996), and therefore responses to this category were treated as missing data.

Choice of Model

When items have more than two options, two parameterizations of the polytomous Rasch model, the partial credit model (PCM) (Masters, 1982; Wright and Masters, 1982) and the Andrich rating scale model (RSM) (Andrich, 1978), are typically used. The only difference between PCM and RSM is the number of parameters that are being estimated. In the PCM, the structure is unique to all items, whereas the items (or groups of items) in the RSM share rating scale structure. Guidelines presented by (Linacre, 2000, p. 768) were considered to decide which model to use, and the RSM was chosen for three reasons:

1. All items in the instrument were *intended* to share the same scoring structure.
2. If the PCM had been chosen, some of the items would have had less than 10 responses to some of the categories, which is considered to be problematic.
3. The correlation between person measures, when the different approaches were tested, was close to 1—leaving little room for argument as to whether to reject the RSM.

The rating scale model is a generalisation of the dichotomous Rasch model. With m thresholds, $\tau_1, \tau_2, \ldots, \tau_m$ it takes the form:

$$\Pr\{X_{ni} = x|\Omega'\} = \frac{\left[\exp\left(x(\beta_n - \delta_i) - \sum_{k=0}^{x} \tau_k\right)\right]}{\gamma_{ni}},$$

where,

β_n is the measure of person n,

δ_i is the measure of item i

τ_k is the k^{th} threshold location,

$\Pr\{X_{ni} = x|\Omega'\}$ is the probability of person n responding in category x on item i and

$$\gamma_{ni} = \sum_{x=0}^{m_i} \left[\exp\left(x(\beta_n - \delta_i) - \sum_{k=0}^{x} \tau_k\right)\right],$$

is the normalizing factor ensuring that the sum of probabilities is 1 (Andrich, 1978).

Since a response to an item with more than two categories is considered to be a series of independent dichotomous judgments, i.e., deciding whether one should pass each threshold or not, the outcome space, Ω, is reduced to a restricted outcome space, Ω' that follows the Guttman structure. As such, a person that is responding in the k^{th} category to an item is considered to 'succeed' the first $k - 1$ thresholds and 'fail' the rest (e.g., Andrich, 2010).

Validity

To ensure validity, the study is reliant upon guidelines presented by Wolfe and Smith (2006b) which extends Messick's (1995) validation framework with two aspects of evidence put forth in The Medical Outcomes Trust (MOT). This framework has been used in other studies measuring alternative learning outcome (e.g., Pampaka et al., 2013). To summarize, validity is viewed as a unified concept. That is, there are not different kinds of validity, rather, different kinds of evidence that support validity. Accordingly, Messick (1995) presented six different aspects of validity where evidence can be found: the *content, substantive, structural, generalizability, external,* and *consequential* aspects. Furthermore, the MOT presents two aspects not mentioned by Messick: *Responsiveness* and *interpretability*. In short, the *content* aspect of validity refers to the relevance, representativeness, and technical quality of the items. The *substantive* aspect is the degree to which a theoretical foundation can be related to the items. The *structural* aspect concerns how the scoring structure can be related to the construct of measure. The *generalizability* aspect is the degree to which the score and interpretations can be generalized across sample and context. The *external* aspect includes convergent and discriminant evidence in addition to criterion relevance and applicability of the measures. The *consequential* aspect refers the extent to which society benefits from using the test. *Responsiveness* refers to the capacity of detecting change, and, finally, *interpretability* is the degree to which qualitative meaning can be drawn based on the quantitative measures (Wolfe and Smith, 2006a).

Analysis

To ensure *content* validity, INFIT and OUTFIT MNSQ and ZSTD were used to assess data to model fit. OUTFIT is based on the mean of squared standardized residuals, and INFIT is an information-weighted sum (Bond and Fox, 2003, p. 238). Different critical values for INFIT MNSQ and OUTFIT MNSQ are reported in the literature, e.g., 0.5 to 1.5 (Linacre, 2002). Smith, Schumacker, and Busch (1998), however, asserted that critical values are not symmetrical about 1 and warned against the use of a single critical value, since it is affected by the type of the mean square and the number of persons. Since the role of these statistics is merely to 'flag' potential problematic items for closer inspection (Bohlig, Fisher, Masters, and Bond, 1998, p. 607), this study used the more conservative cut values. Specifically, critical values for INFIT- and OUTFIT MNSQ were calculated from the formulas suggested in Smith et al. (1998, p. 78) to be 1.1 (INFIT MNSQ) and 1.3 (OUTFIT MNSQ). Critical values for |ZSTD| were set to 2.0. Consequently, all items above these values were flagged and inspected more closely. No items, however, were removed based on these values alone. In addition, the item-measure correlation was assessed with .40 as the suggested value for flagging.

Moreover, a rating scale analysis was conducted to find evidence for *substantive* validity. Linacre (2002) presented eight aspects to find evidence for well-functioning rating scales: 1) each rating scale category should contain more than 10 observations; 2) the shape of each rating scale distribution should be smooth and unimodal; 3) the average respondent measure associated with each category should increase with the values of the categories; 4) the category OUTFIT MNSQ fit statistics should be less than 2.0; 5) Step calibrations should advance; 6) Ratings should imply measures, and measures should imply ratings; 7) Step difficulties should advance by at least 1.4 logits; and 8) Step difficulties should advance by less than 5.0 logits. Additionally, Wolfe and

Smith (2006b, p. 210) argued that researchers also should consider several post hoc categorizations. Zhu (2002) indicated that one should consider person reliability and fit of the data to the model when deciding which categorization that optimizes the rating scale structure.

Another source of *substantive* validity is analysis of person fit (Wolfe and Smith, 2006, p. 211). As such, persons with INFIT MNSQ larger than 2.0 were flagged and examined more closely.

Finally, qualitative judgements about the items' placement on the scale were made. That is, the items' theoretically expected location on the variable were compared with their empirical location. Accordingly, items that deviated from their theoretical location were inspected more closely to find explanations for this deviation.

To find evidence for the *generalizability* aspect of validity, analysis on invariance was conducted through differential item functioning (DIF): *the loss of invariance of item estimates across testing occasions* (Bond and Fox, 2003, p. 309). That is, comparison between two items should be independent of which persons are being used in the calibration (Rasch, 1961, pp. 331-332). DIF analysis was conducted to test the hypothesis that item measures remained invariant between institutions (teacher education and STEM). This study used the Rasch-Welch *t*-test, which means that each item is calibrated for each class, holding everything else constant. As proposed by Linacre (2015, p. 542), a threshold of .64 for DIF contrast and .05 for *p*-value was chosen to flag potentially problematic items. These, however, were not used as definite cut-values. Since iterative procedures that aim at identifying DIF free subsets do not distinguish between real and artificial DIF (Andrich and Hagquist, 2012), further analysis of potentially problematic items were conducted. First, qualitative judgements were made to explain the DIF. Second, a subset of items with no DIF (DIF contrast below .2) was used as a reference group, and potentially problematic items were then included in the subset, one at a time. If DIF was reduced to an acceptable value (i.e., providing some evidence that a substantial part of the total DIF was artificial), and qualitative judgements could not explain the invariance, it was decided to keep these items. Finally, to detect non-uniform DIF, expected and empirical ICC curves were inspected and compared between both institutions.

To assess the dimensionality of the items (*structural* validity), principal component analysis of the residuals was conducted (e.g., Bond and Fox, 2007). A threshold of 2.0 in Eigenvalue units was chosen for possible, substantial, dimensions (Linacre, 2015, p. 391). In addition, person measures based on items from different sub-dimensions were correlated to see if the dimensions should belong to separate instruments or not. The degree of local dependency between items was also assessed to see if responses to some items affected the responses to others.

Results

Based on person/item fit statistics and person/item separation, misfitting items and persons were removed until the instrument was no longer improved, leaving 20 items (see Appendix A) and 310 persons. The final results will be discussed in terms of *content, substantive, structural, generalizability* and *responsiveness* validity.

Summary Statistics

Separation is the ratio of 'true' variance to error variance (Linacre, 2015, p. 330). The analysis shows that the real item separation index is 10.68 and that the real person separation index is 2.45. This means that the instrument can distinguish between those who are working conceptually with mathematics, and those who are not.

Item Fit

In Table 2, the items are ordered by measure. All point-measure correlations are above .40. From INFIT and OUTFIT statistics, items 4, 9, 10 and 17 are the most under-fitting, and items 13 and 18 are the most over-fitting. However, deletion of any of these items would reduce person separation. In addition, the misfit in all of these items could be explained by only a few anomalous responses. Removing the three most unexpected results to item 4, for instance, reduced the item INFIT MNSQ to 1.09 and the OUTFIT MNSQ to 1.03.

Furthermore, the practical consequences of keeping the most misfitting items were tested. Specifically, a 'mindless' reduction of misfitting items was conducted (Table 3). Item measures in this analysis were anchored, and a new analysis

Table 1

Summary statistics: 310 measured persons

	Total score	Count	Measure	Model *SE*	Infit		Outfit	
					Mnsq	*Zstd*	*Mnsq*	*Zstd*
Mean	38.5	17.1	−.58	.39	1.00	−.1	1.00	−.1
P.SD	11.7	3.9	1.14	.06	.44	1.3	.44	1.3

Real RMSE .43	True *SD* 1.05	Separation 2.45	Person reliability .86
Model RMSE .40	True *SD* 1.07	Separation 2.67	Person reliability .88

SE of Person mean = .06

Summary statistics: 20 measured items

	Total score	Count	Measure	Model *SE*	Infit		Outfit	
					Mnsq	*Zstd*	*Mnsq*	*Zstd*
Mean	596.7	264.8	.00	.10	1.00	.0	1.00	−.1
P.SD	143.6	25.1	1.09	.01	.15	1.9	.44	1.3

Real RMSE .10	True *SD* 1.08	Separation 10.68	Item reliability .99
Model RMSE .10	True *SD* 1.08	Separation 11.01	Item reliability .99

SE of Item mean = .25

Table 2

Item statistics: Measure order

	INFIT		OUTFIT		
Measure	*Mnsq*	*Zstd*	*Mnsq*	*Zstd*	Item
1.71	.93	−.7	.86	−1.3	1. Takes the initiative to learn more.
1.69	.89	−1.3	.86	−1.5	2. Takes time to find better methods.
1.42	.98	−.2	.93	−.7	3. Thinks of times when methods don't work.
1.10	**1.23**	**2.7**	**1.21**	**2.4**	4. Struggles with putting problems aside.
.58	1.09	1.2	1.11	1.4	5. Derives formulas.
.55	1.12	1.4	1.12	1.3	6. Likes to discuss math.
.54	.99	−.1	.99	−.1	7. Makes his/her own problems.
.45	.91	−1.0	.89	−1.3	8. New ideas lead to trains of thoughts.
.19	**1.25**	**2.9**	**1.25**	**2.9**	9. (x) Likes to be told exactly what to do.
.08	**1.23**	**2.7**	**1.20**	**2.4**	10. Finds out why methods wouldn't work.
−.06	.89	−1.4	.88	−1.5	11. Finds out why formulas/algoritms work.
−.21	.94	−.7	.94	−.7	12. Studies proofs until they make sense.
−.24	.71	−4.1	.72	−3.9	13. Considers different possible solutions.
−.27	.84	−2.0	.86	−1.8	14. Moves back and forth between strategies.
−.35	1.08	1.0	1.08	1.0	15. Wants to learn more things.
−.51	.96	−.4	.99	−.1	16. Pauses and reflects.
−.82	**1.22**	**2.3**	**1.22**	**2.3**	17. Visualizes problems.
−1.20	.75	−3.4	.78	−3.0	18. Can explain why solutions are correct.
−2.27	.98	−.2	1.02	.3	19. Connects new and existing knowledge.
−2.38	1.06	.7	1.04	.5	20. Keeps trying.

was conducted including all twenty items. A cross-plot between person measures, with or without the most misfitting items, did not prove a significant difference (Figure 1). This suggested that the misfitting items should remain in the instrument, though they may point to areas of possible improvement in the future.

Another key source of information about item fit is the item characteristic curves (e.g., Schumacker, 2015). Figure 2 provides three examples, from item 20 (which seemed to be well-behaving), item 13 and 17. These last two items show a general impression; the empirical curves tended to deviate from the model when high scores were expected. This will be discussed more closely later in the paper.

Differential Item Functioning

DIF analysis between institutions (student teachers and STEM) pointed to three possible problematic items: item 2, item 19 (both being relatively easier for the teacher students to agree

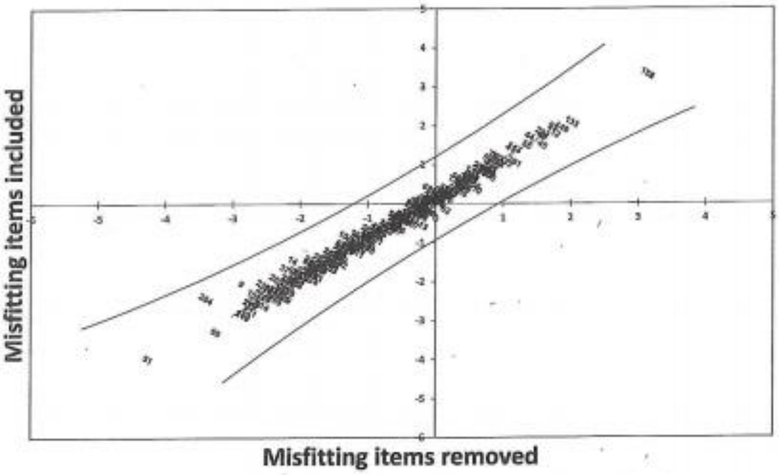


Figure 1. Cross-plot between person measures with or without misfitting items

Table 3

Item statistics: Misfitting items removed

	INFIT		OUTFIT		
Measure	*Mnsq*	*Zstd*	*Mnsq*	*Zstd*	Item
1.84	.96	−.4	.88	−1.1	1. Takes the initiative to learn more.
1.82	.94	−.6	.91	−.9	2. Takes time to find better methods.
1.54	1.01	.1	.95	−.5	3. Thinks of times when methods don't'work.
.64	1.15	1.8	1.17	2.1	5. Derives formulas.
.61	1.04	.5	1.03	.4	7. Makes his/her own problems.
.60	1.14	1.6	1.13	1.5	6. Likes to discuss math.
.51	.92	−1.0	.89	−1.3	8. New ideas lead to trains of thoughts.
−.03	.92	−1.0	.91	−1.1	11. Finds out why formulas/algoritms work.
−.21	.95	−.6	.94	−.7	12. Studies proofs until they make sense.
−.26	.89	−1.4	.92	−1.0	14. Moves back and forth between strategies.
−.34	1.12	1.4	1.12	1.5	15. Wants to learn more things.
−.52	1.00	.0	1.03	.3	16. Pauses and reflects.
−1.26	.81	−2.6	.85	−1.9	18. Can explain why solutions are correct.
−2.41	1.02	.3	1.09	1.0	19. Connects new and existing knowledge.
−2.53	1.15	1.8	1.15	1.6	20. Keeps trying.

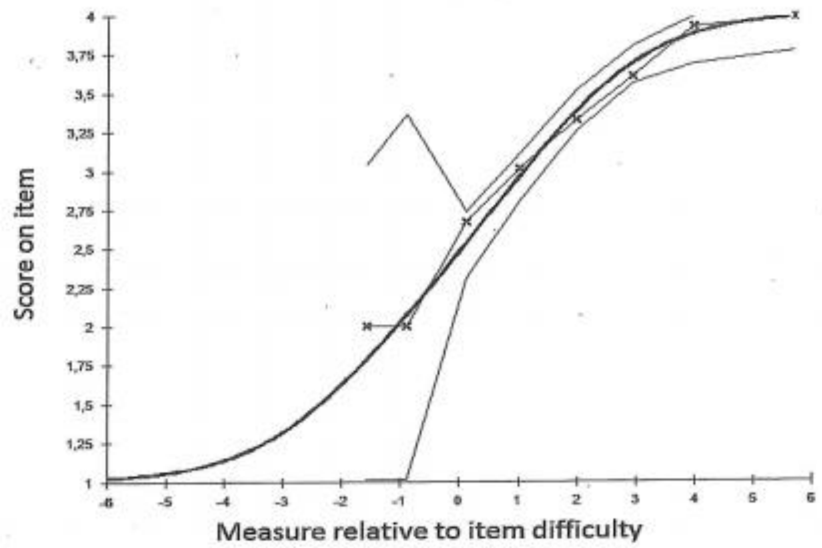


Figure 2a. ICC, item 20

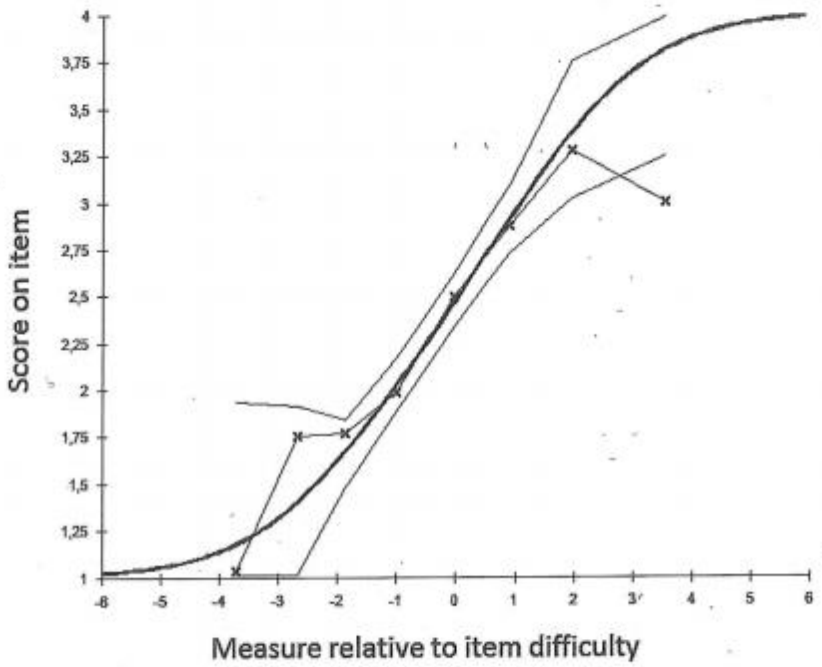


Figure 2b. ICC, item 13

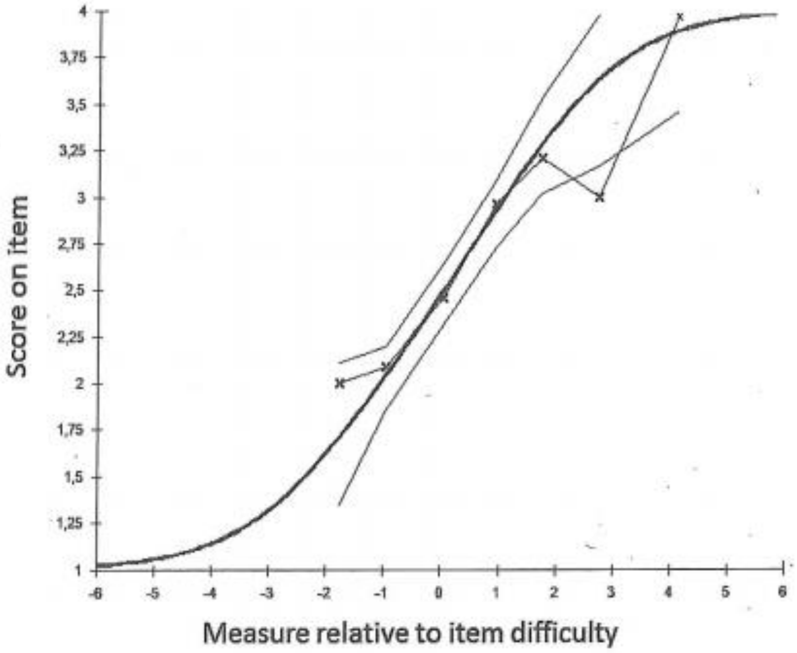


Figure 2c. ICC, item 17

on, with DIF contrasts −.73, and −.72 respectively) and item 15 (being relatively easier for the STEM students to agree on, with DIF contrast .68). The DIF contrast on item 15, however, was reduced to .60 when tested together with the DIF-free subset. A closer look at the empirical ICCs to items 2 and 19 suggested that the DIFs were uniform and could not be explained by a few unexpected results (Figure 3). Instead, they could be explained by qualitative judgements. In the case of item 2: when students learn a new method, it seems relatively easier for a student teacher to search for a better method. This can be explained by the curriculum of teacher education, which emphasizes the search for and comparison between methods. Additionally, the tasks at the technical university seemed to emphasize specific methods to a greater extent than teacher education (e.g., 'Use Newton's method to find the only real root of the equation $x^3 - x - 1 = 0$'), making it naturally harder for STEM students to search for other methods.

A similar judgement can explain the DIF in item 19. First, the emphasis teacher education puts on connecting new and existing knowledge could be a plausible explanation for why it is relatively easier for the student teachers to do so. Furthermore, the new content STEM students face (such as linear algebra) seems further away from high school mathematics than the new content student teachers face. To illustrate, it makes sense that a student learning about Eigenvalues would be having more difficulty connecting this to existing knowledge than a student learning about different division algorithms.

In conclusion, there seemed to be a real DIF existing in item 2 and item 19. However, when separate analyses were conducted, both items worked well in both groups, with INFIT MNSQ in the range of .8 - 1.03 and OUTFIT MNSQ in the range of .79 - .99. As such, the items were included in the instrument.

Targeting

In order to be able to detect change (*responsiveness* validity), it is important that the items in the instrument not only fit well with the

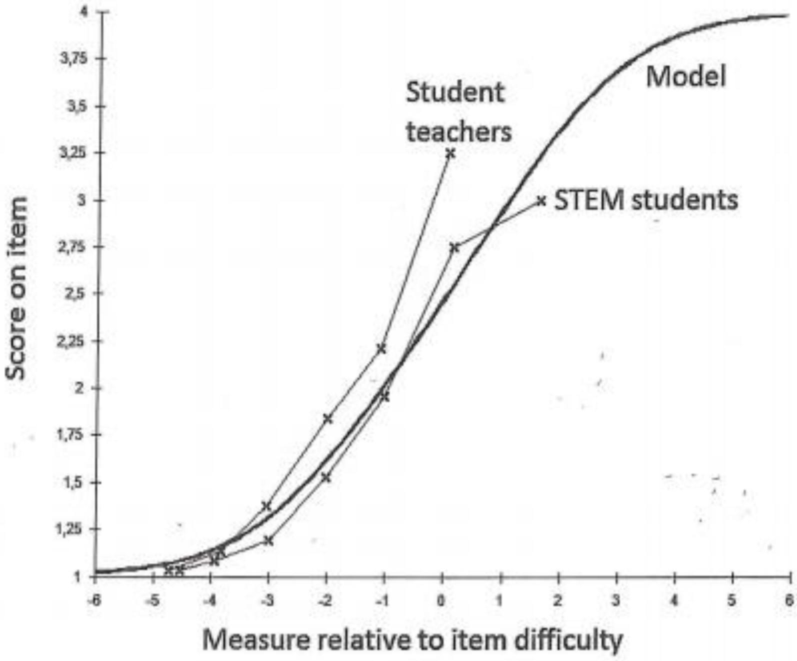


Figure 3. ICC curves for item 2

sample abilities but also possible abilities in the future (Wolfe and Smith, 2006b, p. 222). Figure 5 puts persons and items on the same scale. Accordingly, future improvements can be found at three distinct places: items in the range of –2 to –1 logits, and items and persons on both ends of the scale. Moreover, a general observation is that the items are slightly 'difficult' for this sample. This also suggests more items at the 'surface' end of the scale.

Unidimensionality and Response Dependency

From principal component analysis of the residuals (*structural* validity), a 1.7 unexplained variance (in Eigenvalue units) was found in a second contrast. This is below the general guideline. Additionally, the disattenuated correlation between person measures on the cluster of items with highly positive loadings, and the cluster of items with highly negative loadings on this second dimension, was close to 1. Thus, this dimension was treated as a strand, and not a dimension that needed a separate instrument

Moreover, the largest standardized residual correlation between items was .16 (between item 2 and 16). Although there seems to be a qualitative similarity between these items (as they both express autonomy), the relatively small residual correlation did not lead to further action.

Theory-data Fit

Based on their (theoretically) predicted placement on the variable, two items were flagged with theory-data misfit. Item 10 (item measure = .08) was expected to have a substantially higher measure, and item 9 (.19) was expected to have a lower measure. As items 9 and 10 were among the items with the greatest INFIT MNSQ, the theory-data misfit strengthened the impression that these items are areas for consideration if the instrument is to be improved. However, comparing analyses with and without these items did not show a significant difference in person measures, only a reduction of person separation when the items were removed. As such, the items should be included until better alternatives are presented.

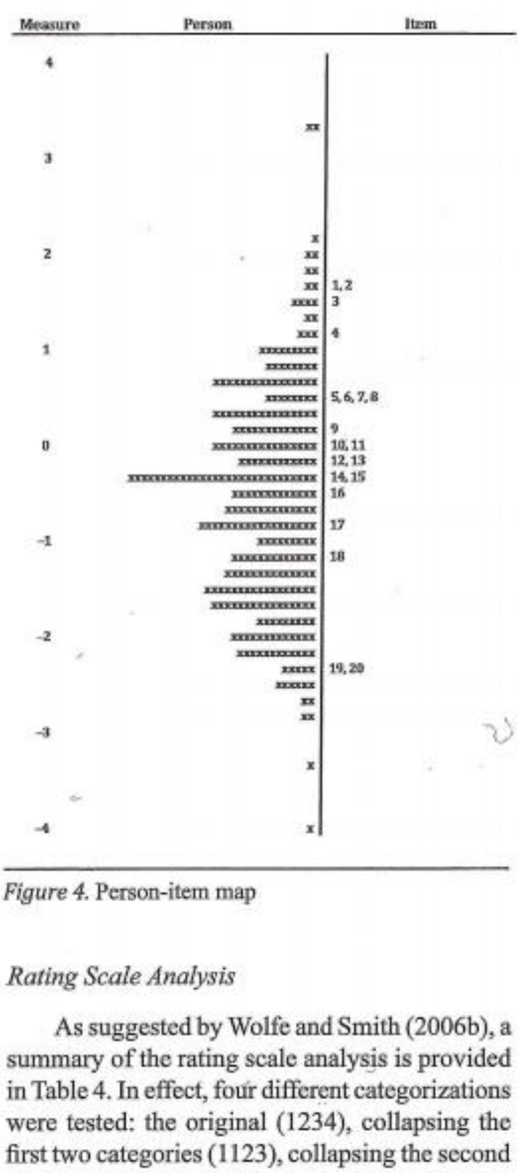


Figure 4. Person-item map

Rating Scale Analysis

As suggested by Wolfe and Smith (2006b), a summary of the rating scale analysis is provided in Table 4. In effect, four different categorizations were tested: the original (1234), collapsing the first two categories (1123), collapsing the second and third category (1223), and collapsing the last two categories (1233). In the original categorization, one violation to Linacre's criteria, the category to measure coherence, was found. That is, only 28% of responses in the fourth category belonged to persons that were expected. Linacre (2002) suggested at least 40%. This problem was also evident in the ICCs (e.g., Figure 2b and 2c). Collapsing the last two categories (1233) seemed, however, to resolve the problem, as all of Linacre's criteria were satisfied. Accordingly, a thorough analysis of the collapsed version was conducted and compared with the original version.

In DIF and dimensionality analysis, only small differences were detected between the original and collapsed version, and thus not discussed further.

ICC plots proved an improvement in the collapsed version. This comes as no surprise, as the major problems with the ICCs in the original version were deviations when high scores were expected. Two examples of the improved ICCs can be seen in Figures 5a and 5b. The original ICCs to these items were discussed earlier (Figure 2).

Item fit statistics was also slightly improved in the collapsed version. The misfit in items 4 and 10 that was discussed earlier was resolved. The new INFIT/OUTFIT MNSQ for these items were 1.10/1.08 and 1.07/ 1.13 respectively. Only item 9 (1.34/1.35) and item 17 (1.23/1.18) remained misfitting.

Nevertheless, the correlation between item measures (r^2 = 1.00) and person measures (r^2= .98) between the original and collapsed versions indicated that the practical consequences of choosing either version are small.

Conclusions

The 20 items presented in this article are productive for measuring the extent to which students work conceptually with mathematics. Two versions of the instrument are discussed, and evidence can be found that supports the use of both of these. If the collapsed version is favoured, however, a cross-validation on new data is suggested by Wolfe and Smith (2006b, p. 210). Items 2 and 19 seem to have real DIF between teacher—and STEM students in both versions. Moreover, the analysis shows that future developments should consider item 9 for improvement as it shows model-data and theory-data misfit in both versions. Further improvements should also consider filling the gaps on the variable, in the range –2 to –1 logits and at both extremes.

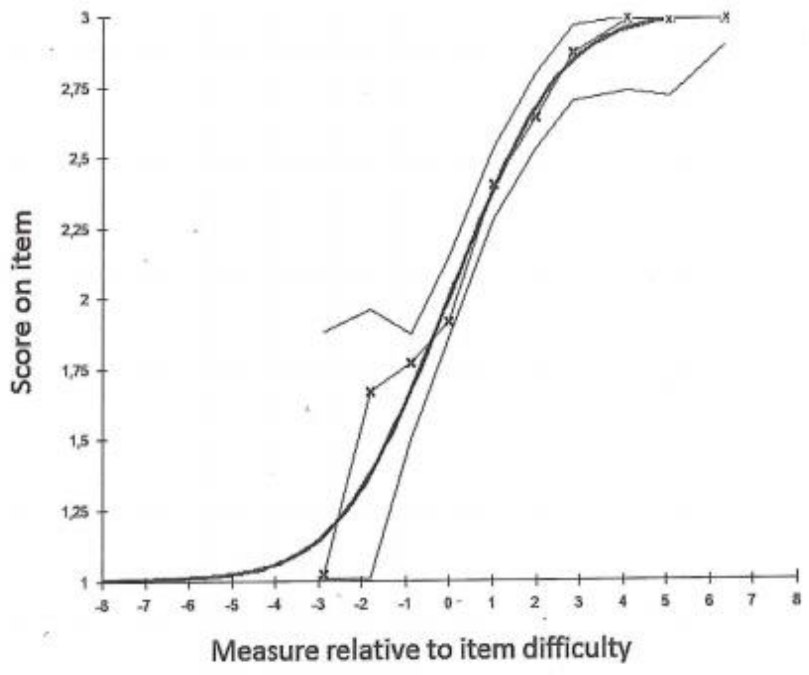


Figure 5a. ICC, item 13

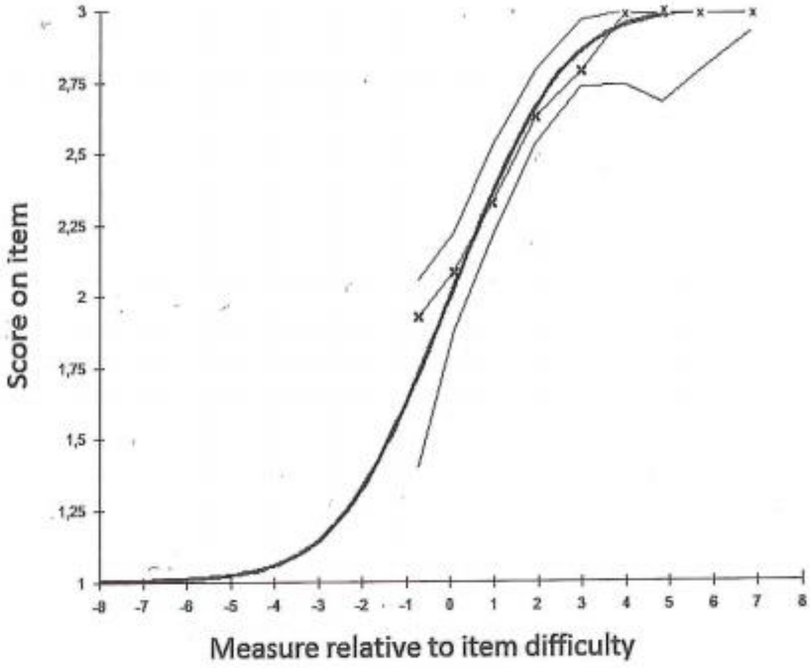


Figure 5b. ICC, item 17

Table 4

Rating scale analysis: Summary table

	Rating scale			
Original Recoded	1234 1234	1234 1123	1234 1223	1234 1233
Person separation	2.45	1.78	2.03	2.30
OUTFIT MNSQ/SD				
Person	1.00/.44	.98/.61	.99/.60	1.00/.43
Item	1.00/.15	.97/.19	.98/.20	1.00/.14
LINACRE'S CRITERIA				
N	☑	☑	☑	☑
Unimodal	☑	☑	☑	☑
M(q)	☑	☑	☑	☑
OUTFIT MNSQ	☑	☑	☑	☑
τs increase	☑	☑	☑	☑
M→C	☑	☑	☑	☑
C→M	☒	☒	☒	☑
τs distance	☑	☑	☒	☑

Moreover, further improvements should consider changes in the rating categories. In effect, more categories are preferred as long as respondents are able to discriminate between them, because they provide more information than fewer categories. As such, it is suggested to rephrase before collapsing categories to check whether this can fix possible ambiguity. Specifically, rephrasing the last category from *always/almost always* to *always* may be productive if the category-to-measure problem is due to a shaded difference between *often* and *almost always*.

Finally, further research should consider two aspects of validity that has not been discussed in this paper: *external* validity and *interpretability* validity. Evidence for *external* validity will be found when measures are related to external measures (Wolfe and Smith, 2006b, p. 221). Specifically, this involves testing well-established hypothesis such as: 'persons working with mathematics conceptually are more likely to adapt knowledge to new tasks,' 'persons working conceptually remember the content more easily' (Skemp, 1989, pp. 9-10), or 'persons working conceptually are more likely to monitor procedural outcomes for erroneous results' (Hiebert, 2013, p. 13) to mention but a few.

Evidence for *interpretability* validity can be found when meanings of the measure are clear to the audience. Accordingly, mixed-method studies that combine individual measures and qualitative data may provide qualitative references to different parts of the instrument.

Acknowledgement

The author would like to thank Professor Birgit Pepin, Professor Pauline Vos, Dr. Svein Arne Sikko, and Dr. Maria Pampaka for invaluable advice.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Andrich, D. (1989). Distinctions between assumptions and requirements in measurement in the social sciences. In J. A. Keats, R. Taft, R. A. Heath, and S. H. Lovobond (Eds.), *Mathematical and theoretical systems* (pp. 7-16). North Holland, NY: Elsevier Science Publications.

Andrich, D. (2010). Understanding the response structure and process in the polytomous Rasch model. In M. L. Nering and R. Ostini (Eds.), *Handbook of polytomous item response* *theory models* (pp. 123-152). New York, NY: Routledge.

Andrich, D., and Hagquist, C. (2012). Real and artificial differential item functioning. *Journal of Educational and Behavioral Statistics, 37*, 387-416.

Bohlig, M., Fisher, W., Masters, G., and Bond, T. (1998). Content validity and misfitting items. *Rasch Measurement Transactions, 12*, 607.

Bond, T. G., and Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Oxford, UK: Lawrence Erlbaum.

Booth, S. (2008). Learning and teaching engineering mathematics for the knowledge society. *European Journal of Engineering Education, 33*, 381-389.

Carpenter, T. P., and Lehrer, R. (1999). Teaching and learning mathematics with understanding. In E. Fennema and T. A. Romberg (Eds.), *Mathematics classrooms that promote understanding* (pp. 19-32). Mahwah, NJ: Lawrence Erlbaum.

Entwistle, N. (1997). *The approaches and study skills inventory for students (ASSIST)*. Edinburgh, Scotland: Centre for Research on Learning and Instruction, University of Edinburgh.

Hiebert, J. (2013). *Conceptual and procedural knowledge: The case of mathematics.* Hillsdale, NJ: Lawrence Erlbaum.

Kuhn, T. S. (1977). *The essential tension: Selected studies in scientific tradition and change.* Chicago, IL: University of Chicago Press.

Linacre, J. M. (2000). Comparing "Partial Credit Models" (PCM) and "Rating Scale Models" (RSM). *Rasch Measurement Transactions, 14*, 768.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*, 85-106.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*, 878.

Linacre, J. M. (2006). WINSTEPS: Rasch measurement computer program [Computer software]. Chicago, IL: Winsteps.com.

Linacre, J. M. (2015). *A user's guide to WINSTEPS/MINISTEP: Rasch-model computer programs.* Beaverton, OR: Winsteps.com.

Lopez, W. (1996). Communication validity and rating scales. *Rasch Measurement Transactions, 10*, 482-483.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Medical Outcomes Trust Scientific Advisory Committee. (1995). Instrument review criteria. *Medical Outcomes Trust Bulletin*, 1-4.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Pampaka, M., Williams, J., Hutcheson, G., Black, L., Davis, P., Hernandez-Martinez, P., and Wake, G. (2013). Measuring alternative learning outcomes: Dispositions to study in higher education. *Journal of Applied Measurement, 14*, 197-218.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago, IL: University of Chicago Press.)

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In proceedings of the *Fourth Berkley Symposium on Mathematical Statistics and Probability (Vol. 4)* (pp. 321-334). Berkley, CA: University of California Press.

Schumacker, R. E. (2015). Detecting measurement disturbance effects: The graphical display of item characteristics. *Journal of Applied Measurement, 16*, 76-81.

Skemp, R. R. (1989). *Mathematics in the primary school.* London, UK: Psychology Press.

Smith, R. M., Schumacker, R. E., and Busch, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, *2*, 66-78.

Wolfe, E., and Smith, E. (2006a). Instrument development tools and activities for measure validation using Rasch models: Part I-instrument development tools. *Journal of Applied Measurement*, *8*, 97-123.

Wolfe, E., and Smith, E. (2006b). Instrument development tools and activities for measure validation using Rasch models: Part II-validation activities. *Journal of Applied Measurement*, *8*, 204-234.

Wright, B. D., and Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Wright, B. D., and Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Zhu, W. (2002). A confirmation study of Rasch-based optimal categorization of a rating scale. *Journal of Applied Measurement*, *3*, 1-15.

Appendix A

The Mathematical Depth Instrument (English version)

Never/almost never (1), Sometimes (2), Often (3), Always/almost always (4), Don't know (9)

1.	I take the initiative to learn more about math than what is required at school/work.	①	②	③	④	⑨
2.	When I learn a new method, I take time to find out if I can find a better method.	①	②	③	④	⑨
3.	When I learn a new method, I try to think of situations when it wouldn't work.	①	②	③	④	⑨
4.	I struggle with putting math problems aside.	①	②	③	④	⑨
5.	If I forget a formula or method, I try to derive it myself.	①	②	③	④	⑨
6.	I get engaged when someone starts a mathematical discussion.	①	②	③	④	⑨
7.	When I learn something new, I make my own problems.	①	②	③	④	⑨
8.	Math ideas that I hear or learn about help me inspire new trains of thoughts.	①	②	③	④	⑨
9.	When I learn a new method, I like to be told exactly what to do.	①	②	③	④	⑨
10.	When I try to use a method that doesn't work, I spend time to find out why it didn't work.	①	②	③	④	⑨
11.	When I learn a new formula/algorithm, I try to understand why it works.	①	②	③	④	⑨
12.	When I face a proof, I study it until it becomes meaningful.	①	②	③	④	⑨
13.	When I face a math problem, I consider different possible ways I can solve it.	①	②	③	④	⑨
14.	When I work with a math problem, I move back and forth between various strategies.	①	②	③	④	⑨
15.	When I learn something new, it makes me want to learn more things.	①	②	③	④	⑨
16.	When I work with a problem, I pause along the way to reflect on what I am doing.	①	②	③	④	⑨
17.	If I get stuck on a problem, I try to visualize it.	①	②	③	④	⑨
18.	I can explain why my solutions are correct.	①	②	③	④	⑨
19.	I try to connect new things I learn to what I already know.	①	②	③	④	⑨
20.	If I immediately do not understand what to do, I keep trying.	①	②	③	④	⑨

Educ Stud Math (2017) 95:163–179
DOI 10.1007/s10649-016-9742-3

Measuring STEM students' mathematical identities

Eivind Kaspersen[1] · Birgit Pepin[2] · Svein Arne Sikko[1]

Published online: 14 December 2016

Abstract Studies on identity in general and mathematical identity in particular have gained much interest over the last decades. However, although measurements have been proven to be potent tools in many scientific fields, a lack of consensus on ontological, epistemological, and methodological issues has complicated measurements of mathematical identities. Specifically, most studies conceptualise mathematical identity as something multidimensional and situated, which obviously complicates measurement, since these aspects violate basic requirements of measurement. However, most concepts that are measured in scientific work are both multidimensional and situated, even in physics. In effect, these concepts are being conceptualised as sufficiently uni-dimensional and invariant for measures to be meaningful. We assert that if the same judgements were to be made regarding mathematical identity, that is, whether identity can be measured with one instrument alone, whether one needs multiple instruments, or whether measurement is meaningless, it would be necessary to know how much of the multidimensionality can be captured by one measure and how situated mathematical identity is. Accordingly, this paper proposes a theoretical perspective on mathematical identity that is consistent with basic requirements of measurement. Moreover, characteristics of students' mathematical identities are presented and the problem of "situatedness" is discussed.

Keywords Mathematical identity · CHAT · Rasch measurement

✉ Eivind Kaspersen
eivind.kaspersen@ntnu.no

Birgit Pepin
b.e.u.pepin@tue.nl

Svein Arne Sikko
svein.a.sikko@ntnu.no

[1] Faculty of Teacher and Interpreter Education, Norwegian University of Science and Technology (NTNU), NO-7491 Trondheim, Norway

[2] Eindhoven School of Education, Technische Universiteit Eindhoven, Eindhoven, The Netherlands

1 Introduction

Much has been said about what it means to be a successful mathematics student. The early writings of Skemp (1987, 1989) and Hiebert (1986) emphasised understanding and knowledge, whereby a combination of conceptual knowledge/relational understanding and procedural knowledge/instrumental understanding was perceived to be a characteristic of successful mathematics students. Subsequently, research in mathematics education has had a tradition of focusing on the teaching and learning for relational understanding (e.g. Carpenter & Lehrer, 1999).

Another line of research goes beyond the cognitive aspects of learning and focuses on learning as a process of identification with certain practices or activities (e.g., Holland, Lachicotte, Skinner, & Cain, 2001; Wenger, 1998). From this perspective, being a successful mathematics student is not determined by cognitive aspects alone, but also by how persons are identifying with mathematical practices (Boaler, Wiliam, & Zevenbergen, 2000; Sfard & Prusak, 2005).

Although the mainstream in identity research seems to follow a qualitative tradition (e.g., Black et al., 2010; Hernandez-Martinez et al., 2011; Hossain, Mendick, & Adler, 2013; Solomon, 2007), some quantitative studies on mathematical identities, and related concepts, such as students' views on themselves as learners, exist (e.g., Alexander, 2015; Roesken, Hannula, & Pehkonen, 2011). It appears problematic, however, to measure mathematical identity based on basic requirements of measurement, for instance, as formulated by Thurstone (1959), and we will in this paper address three particular problems of measuring mathematical identity.

The *first problem* is that there is no consensus on philosophical aspects of identity in general, and mathematical identity in particular. As such, it seems impossible to construct a measure of mathematical identity that is consistent with all conceptualisations of identity. The *second problem* is that selected quantitative studies have suggested multiple dimensions to mathematical identity, like knowledge, ability, motivation, and anxiety (e.g., Axelsson, 2009). However, the nature of any measure is uni-dimensional, that is, one can measure only one dimension at a time (Thurstone, 1959). The *third problem* is that measurements are required to be invariant across contexts (Thurstone, 1959). That is, in the case of mathematical identity, "what it means to be mathematical"–what we will refer to as the social structure of being mathematical–has to be identical (ideally) or similar enough (practically) across contexts for cross-contextual comparisons of measurements to make sense. However, a great proportion of studies conceptualise mathematical identity as something more or less situated, which obviously violates the requirement of invariance. Accordingly, although measurements have been proven to be extremely potent in the history of science (Kuhn, 1977), serious threats to meaningful measurements can be found in a lack of theoretical consensus, in the requirement of uni-dimensionality, and in the requirement of invariance.

The problem with a lack of consensus is not a new one. Rather, the state of identity research seems to have many resemblances to what typically is found in pre-paradigmatic phases. Kuhn (1970) described these phases as filled with confusion and a lack of consensus, but not incompatible with significant discoveries and inventions. Thus, the negotiations, confusions, and disagreements of ontological, epistemological, and methodological issues seem inevitable, if identity research is perceived as an emerging paradigm. Moreover, the uni-dimensional and invariance issues are also known problems in the history of science. That is, uni-dimensional and invariant constructs are only theoretical ideas, just like a line in Euclidean geometry. In practice, every construct is multidimensional, and every *dimension* is situated. Thus, when studies conclude that a construct, like identity, consists of a finite number of invariant dimensions, none of these will be truly uni-dimensional nor completely invariant.

Accordingly, in this paper, we claim that questions about how many dimensions mathematical identity comprises of, and how situated mathematical identity is, are pragmatic ones. When instruments are used to measure constructs such as motivation, ability, and anxiety, researchers have decided that those instruments capture "enough" of what one tries to measure, and that these constructs are invariant enough for practical purposes, whilst appreciating that the measures that are produced are only approximations. We claim that the same decisions must be made when mathematical identity is measured.

As such, mathematical identity (as a unit) can be measured, if it is perceived to capture "enough" of what one wants to study. Moreover, measures across contexts can be compared, if the social structure of being mathematical is perceived similar "enough" in these contexts. However, to make such decisions, one needs to know how much information one main mathematical identity dimension covers, and how much the structure differs (if measures are to be compared across contexts). Furthermore, from a measurement perspective, this kind of information can be obtained only from a well-defined and, most importantly, measurable perspective on mathematical identity.

Although some of our arguments in this paper are general, we will focus specifically on Science, Technology, Engineering, and Mathematics (STEM) students' mathematical identities. Moreover, when the problem of invariance is discussed, a second sample consisting of student teachers has been chosen for illustrative purposes. Thus, the questions that guide this research, all of which are related to the three problems discussed above, are:

(1) Which theoretical perspective on mathematical identity is consistent with basic requirements of measurement?
(2) What are the characteristics of STEM students' measured mathematical identities?
(3) How can mathematical identity measures provide information on how much the social structure of being mathematical differs across the STEM context and the student teacher context?

The study is influenced by the TransMaths project (TransMaths.org), and builds on data from a previous study that validated a Rasch-calibrated instrument for measuring the extent to which Norwegian STEM students and student teachers are working conceptually with mathematics (Kaspersen, 2015). Although these measures initially were conceptualised as static traits, subsequent analyses suggested that the measures could only be understood in relation to the context in which the observations were made. As such, we will outline in this paper a theoretical perspective on identity that, we argue, is both contextual and measurable. Thus, concepts from cultural-historical activity theory (CHAT) are adopted to account for the structure of the activity. Specifically, "the social structure of being mathematical" is outlined as the invariant background in which individuals are measured. Subsequent quantitative and qualitative analysis will exemplify how mathematical identities are characterised in the STEM context and discuss the invariance of identities across the STEM context and the student teacher context.

2 Theoretical background

The field of identity research has been claimed to lack a consensus on the meaning of identity (e.g., Beijaard, Meijer, & Verloop, 2004). To illustrate, Gee (2000) defined identity as "being recognized as a 'certain kind of person', in a given context (…)" (p. 99), whereas Holland et al. (2001) conceptualised identity as narrated/authored, and "identities-in-practice". Sfard and Prusak (2005) criticised these definitions, as they both emphasised "who one is", as if identity is independent of one's actions. Instead, Sfard and Prusak (2005, p. 14) conceptualised identity as "discourse" constituted as stories and, thus, defined identity to be "a set of reifying, significant, endorsable stories about a person".

However, although little consensus has been found regarding critical issues on identity, some similarities are described in the literature. In particular, the locus of identity, or the "structure-agency" debate, distinguishes those who see identity as primarily context-free (i.e., within the individual) from those who see it as primarily individual-free (i.e., identity can only be understood in context) (Cote & Levine, 2014). Accordingly, identity research can be said to be in a pre-paradigmatic phase. One characteristic of research in pre-paradigmatic fields is the relationship between data and theory (Kuhn, 1970). Researchers that are working in well-established paradigms often try to "force" the data to fit some commonly accepted theory. The situation is quite the opposite in pre-paradigmatic fields, where the data at hand often guide the choice of theory. Indeed, the choice of theoretical perspective in this study has not been pre-determined—we have not adopted any ready-made interpretation of identity (e.g., those mentioned above). Rather, we have started with a few basic assumptions of identity and, based on the argument to follow, chosen a perspective that fits our purpose of how identity can be measured.

So, our fundamental assumption about identity in general, and mathematical identity in particular, is that *identity is relational by nature* (unless proven otherwise). This assumption is based on the fact that most concepts, even physical ones, are relational (the speed of light would be an exception). Thus, identity, like physical concepts, such as speed and weight, can be assessed (e.g., measured) only relative to a context. Accordingly, theoretical concepts are needed to account for the context, and we will ground these concepts in CHAT. It is important to note, however, that we do not assume that the context/activity itself is static—it most likely is not. Just like the speed of a car is only a measure of how fast the car is moving relative to the earth, and does not take into account that the earth itself is moving, so we claim that any measure of identity would only be a measure relative to the activity in which it is situated.

2.1 Cultural—historical activity theory

The origin of CHAT can be traced back to Vygotsky (e.g., 1978) who built a psychology on Marxist ideas. In doing so, individuals and the social context were incorporated into a unifying framework for understanding the human psyche. As a response to theories of behaviour as a direct relationship between the subject and the object, Vygotsky (1978) replaced stimulus-response processes with complex mediated acts, where signs and tools served as mediating links. Moreover, Vygotsky (1978) distinguished between *externally* oriented tools and *internally* oriented signs. Externally oriented tools lead to real changes on the object, much like Marx described the way man uses tools as "forces that affect other objects in order to fulfill his personal goal" (p. 54). Internally oriented signs, by contrast, change nothing on the external object, but work as mental auxiliaries.

In what has been called the second generation of CHAT (Engeström, 2001), Leont'ev (1978) continued the work of Vygotsky by taking up a more collective view on activity. At a general level, *activity* is directed towards a collectively motivated object. *Actions*, in contrast, are directed toward goals. From Leont'ev's (1978) point of view, one cannot talk about individual activity, but one may speak of the activity of the individual. Only actions and operations are individual. Moreover, as the smallest unit of analysis, one cannot analyse activity without actions, and vice versa. This is because activity is a sum of actions, and the same action can bear different meanings depending on the activity into which it is incorporated. As such, one cannot study personal characteristics related to working with mathematics without taking into account what any particular characteristic means in the activity in which it is observed.

Later, Engeström (1987) formulated Leont'ev's ideas in a framework that represents the activity system in a triangle (Fig. 1). The mediating triangle (Vygotsky, 1978) on the top is extended with *rules*, *community*, and *division of labour*. As such, Engeström's (1987) framework takes a less personal position than the original works of Vygotsky and puts more emphasis on the structure of the activity. Moreover, in what is called the third-generation CHAT, Engeström provided conceptual tools to analyse multiple activity systems and their interactions, for instance, though linking boundary persons/objects.

2.2 Mathematical identity

CHAT is not a unified framework, and one of the arguments is related to the notion of identity. Instead of going into detail about different interpretations of identity in CHAT and related theoretical frameworks, such as the theory of communities of practice (Wenger, 1998), we will address these controversies at a more general level, in what has been referred to as the structure-agency debate (e.g., Cote & Levine, 2014).

In contrast to more cognitive traditions, cultural–historical approaches often view identity as mainly a social phenomenon. Yet, there seem to be disagreements regarding the locus of identity: how context-bound it is. Leont'ev (1978), for instance, asserted that the chief task is to understand consciousness as a subjective product of activity, whereas Stetsenko and Arievitch (2004) suggested a middle ground and asserted the self as a leading activity.

We have already assumed that identity is relational by nature, but to address this issue, we need two more assumptions: (a) we assume that the locus of identity question is not a dichotomous one, which means that identity does not have to be either *completely* context-free, or *completely* context-bound (though it might be located closer to one end than the other); and (b) the locus of identity does not have to be static (we believe that it probably is not),

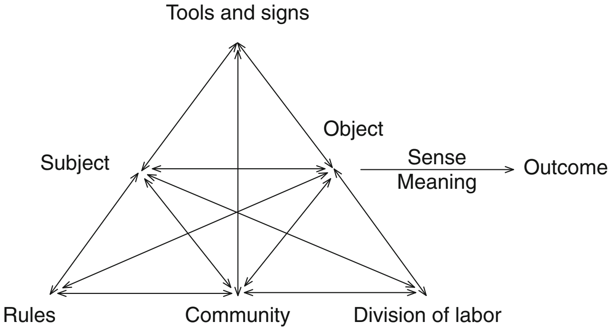


Fig. 1 The structure of an activity system as presented in Engeström (1987, p.78)

which means that there is not one universal answer to the question on how context-bound/free mathematical identity is. As such, we assert that any general claim that attempts to locate identity at any "magical" position between context-free and context-bound is a distraction from the debate. Rather, the locus of identity question should be empirical, not theoretical.

Accordingly, we suggest that researchers need information about at least two different activities before making any judgements about whether mathematical identity is regarded as relatively situated or relatively context-free, though these activities do not have to be completely isolated from each other (they could, for example, be connected by some boundary objects or boundary agents). From this, one problem immediately emerges. If identity can be measured only in context, and we need different contexts to get access to how situated mathematical identity is, then it seems that we have lost the common point of reference that makes comparisons of cross-contextual measurements possible. As such, although it initially looked promising, the activities per se cannot be used as the common point of reference that identity is measured against if we want to compare measures across activities.

To resolve this problem, we need a common point of reference, one common background that can, at least hypothetically, be similar across different activities, and we propose *the social structure of being mathematical* as such a background. That is, we define mathematical identity to be *where persons position themselves relative to the social structure of being mathematical within the activity in which they participate and contribute*.

According to fundamental requirements of measurement (Thurstone, 1959), it is essential that the background of the persons acting stays invariant. Consequently, the social structure of being mathematical in any given activity is required not to change if persons are added or removed from the analysis. An important note is that this property of the social structure of being mathematical is a *requirement* as opposed to an *assumption*. Thus, we appreciate that there exists multiple characteristics of being mathematical within any activity, some of which are variant and others that are not. However, the social structure of being mathematical is required to contain the invariant characteristics only, without assuming that others do not exist. If no such structure can be established, then measures cannot be compared.

Following from this, the social structure of being mathematical is, in theory, the person-independent (i.e., invariant) subset of the complete set of possible positions a person can occupy within a given activity. That is, the social structure of being mathematical is the subset of positions designated by the activity *alone*, whereby any two positions are distinguished by at least one invariant characteristic of being mathematical. Accordingly, we argue that in any activity, at any given time, there is a set of positions, and their associated characteristics, that exists independently from individuals within or without the activity (including researchers). Surely, any person can hypothesise characteristics that belong to the social structure of being mathematical, but some empirical evidence of person-independence is needed before such characteristics are included in the social structure.

To say that a position is person-independent, however, is not the same as saying that it is person-free (clearly, any activity consists of persons). Rather, it means that any position in the social structure exists regardless of whether individuals occupy this position or not. Therefore, the results of an analysis of the social structure of being mathematical should, in theory, depend on which activity is being observed only, and not on which positions persons happen to take in this structure. However, in practice, the theoretical social structure of being mathematical can never be completely observed, nor is it entirely person-independent. Thus, we define the empirical social structure of being mathematical to be a fragmented and approximately person-independent representation of the theoretical invariant structure.

Furthermore, the requirement of invariance is not a requirement for the structure to be static. Indeed, a change of the level of invariance over time is an indicator of where the structure is being negotiated. Likewise, the requirement of invariance is not a requirement of the structure to dictate human action. The structure of being mathematical is an invariant abstraction and not a description of the complete set of possibilities a person has. Any person has agency and can move freely between each position. Moreover, any person can choose to take positions not included in the social structure, and there are many reasons for them to do so (or not). One reason is to negotiate the social structure. Another reason is that persons contribute in many activities, and therefore, what seems to be an outlying position in one activity might, in fact, be a common/firm position in another.

To sum up, considering a simple case with only two activities, if the social structure of being mathematical in activity A is about the same as the social structure of being mathematical in activity B, then mathematical identity is more or less context-free between activity A and activity B—here it does make sense to compare measures between these activities. As the structure across these contexts diverges, mathematical identity becomes more situated, until a hypothetical example where the social structure of being mathematical is completely different between the activities in which we would conclude that mathematical identity is completely situated between activity A and B—here it would no longer make sense to compare mathematical identities across these activities.

3 Methodology

The starting-point of this study is a Rasch-calibrated instrument (Appendix 1) that measures the extent to which students are working conceptually with mathematics (Kaspersen, 2015). Rasch measurement (Rasch, 1961, 1980) is based on fundamental requirements for measurement as formulated by Thurstone (Andrich, 1989), and a key feature of the Rasch model is that persons and items are not discriminated, which means that they can be measured on the same scale (Wright & Stone, 1979). That is, a Rasch measure is a relational measure, which makes it suitable for relational concepts such as identity. In the simplest dichotomous version, the probability that a person agrees with an item is a function of the distance between the person and the item. The greater the person measure in relation to item measure, the closer to 1 the probability gets. Conversely, the greater the item measure relative to person measure, the closer to 0 the probability gets (Wright & Stone, 1979).

Validation of the instrument has been extensively discussed elsewhere (Kaspersen, 2015), and therefore it will only be summarised in this paper. As a means for finding evidence of validity, the study relies upon a framework formulated by Wolfe and Smith (2006), who perceived validity not as a unified concept, but rather as a collection of evidence on different aspects of validity. From a sample of 133 student teachers and 185 STEM students, a group of 20 items was in Kaspersen (2015) concluded to be productive for measurement (Table 1).

Invariance, additivity, and uni-dimensionality have been formulated as the basic requirements of quantitative measurement (Thurstone, 1959). The last requirement, that is uni-dimensionality, means that one could measure only one dimension at a time, and thus, mathematical identity as a unit could be measured only if it was considered to be sufficiently uni-dimensional. If, on the contrary, mathematical identity was perceived to be multidimensional, multiple instruments would be needed.

Table 1 Item statistics: measure order

Measure	INFIT		OUTFIT		Item
	MNSQ	ZSTD	MNSQ	ZSTD	
1.71	.93	–.7	.86	−1.3	1. Takes the initiative to learn more.
1.69	.89	−1.3	.86	−1.5	2. Takes time to find better methods.
1.42	.98	–.2	.93	–.7	3. Thinks of times when methods do not work.
1.10	1.23	2.7	1.21	2.4	4. Struggles with putting problems aside.
.58	1.09	1.2	1.11	1.4	5. Derives formulas.
.55	1.12	1.4	1.12	1.3	6. Likes to discuss math.
.54	.99	–.1	.99	–.1	7. Makes his/her own problems.
.45	.91	−1.0	.89	−1.3	8. New ideas lead to trains of thoughts.
.19	1.25	2.9	1.25	2.9	9. (x) Likes to be told exactly what to do.
.08	1.23	2.7	1.20	2.4	10. Finds out why methods would not work.
–.06	.89	−1.4	.88	−1.5	11. Finds out why formulas/algorithms work.
–.21	.94	–.7	.94	–.7	12. Studies proofs until they make sense.
–.24	.71	−4.1	.72	−3.9	13. Considers different possible solutions.
–.27	.84	−2.0	.86	−1.8	14. Moves back and forth between strategies.
–.35	1.08	1.0	1.08	1.0	15. Wants to learn more things.
–.51	.96	–.4	.99	–.1	16. Pauses and reflects.
–.82	1.22	2.3	1.22	2.3	17. Visualises problems.
−1.20	.75	−3.4	.78	−3.0	18. Can explain why solutions are correct.
−2.27	.98	–.2	1.02	.3	19. Connects new and existing knowledge.
−2.38	1.06	.7	1.04	.5	20. Keeps trying.

Item 9 negatively coded

From a theoretically uni-dimensional case, any deviation from the model is considered to be random noise. Therefore, a principal component analysis (PCA) of the standardised residuals is one way of finding multidimensionality in the data. Since residuals, in a theoretically uni-dimensional case, are random noise, unexpected responses should not correlate across items.

However, any instrument, psychometrical or physical, is in practice multidimensional. Consequently, the question is not really about the instrument being uni-dimensional or not, but how multidimensional it is, and how much it matters. This is an empirical question and depends on the context in which the study is conducted. What in one case is sufficiently uni-dimensional might not be in another. A general guideline on the Rasch paradigm, however, is to use 2.0 in Eigenvalue units as a threshold on second dimensions when PCA on standardised residuals is used to assess multidimensionality. Eigenvalues below 2.0 indicate that, although there are multiple dimensions in the data, the sub-dimensions are in most purposes measuring the same thing (Linacre, 2015, p. 391).

Accordingly, a PCA of the residuals was conducted on the original instrument to find sub-dimensions that were likely to be connected. In Kaspersen (2015), 1.7 unexplained variance (in Eigenvalue units) was found in a second contrast. Thus, any sub-dimensions that appear in the PCA analysis are likely to be highly correlated. Yet, although an Eigenvalue of 1.7 indicates highly correlated sub-dimensions, the relationships between the dimensions will be discussed by comparing person measures from the full instrument with person measures when different sub-dimensions are excluded from the instrument. The purpose is to validate that the sub-dimensions can be treated as strands highly connected with the main mathematical identity dimension, instead of as independent dimensions.

In addition to the PCA analysis, the data consist of semi-structured interviews conducted with four purposefully selected STEM-students: two students (A and B) with high measures and two students (Y and Z) with lower measures. These students were asked: how they perceived their time at the university; which project they were currently engaged in; how they worked with mathematical problems; and how they learned new mathematics. Transcripts of the data were subsequently analysed using the item characteristics in addition to concepts described in CHAT: *objective*, *tools*, *rules*, *communities*, and *division of labour*.

Finally, the relationship between data and our ontological perspective on identity needs some attention. That is, we assert that identity is how people respond to the activities in which they engage, and not how they answer to a questionnaire. Yet, our data consist of what people say. The choice of data has some obvious benefits (one can collect much data on a short amount of time; one can "look back in time"; etc.). However, an implicit assumption is that what students say is a reasonable representation of how they respond to the activity. Thus, our data consists of indirect observations of mathematical identities. As such, we claim that there is no best way to represent identity, and future research might use different kinds of data.

4 Results

In this section, a quantitative description of STEM students' mathematical identities is presented and arguments are made that the social structure of being mathematical is mostly affected by the activity, in which the measure is conducted, however practically unaffected by the individuals' identities. Moreover, statistical arguments make the point that three dimensions are discovered in the data, but that these are all highly connected with the main mathematical identity dimension. Finally, qualitative data will provide illustrative examples of characteristics along this variable.

4.1 Students' mathematical identities

STEM students' measured mathematical identities can be found in the first column of Fig. 2. An approximate and fragmented representation of the social structure of being mathematical is displayed on the right-hand side, and individuals' measures are displayed on the left-hand side. Roughly, most persons "agree" with characteristics much lower than their measure and "disagree" with characteristics much higher than their measure. For instance, a person with measure 0.5 would most likely "agree" with most items 10–20 and "disagree" with most items 1–5 (see Wright and Stone (1979) for a more thorough description of the relationship between persons and items).

We have claimed that the social structure of being mathematical is theoretically independent of individuals' mathematical identities. This is illustrated in the second and third column of Fig. 2, where we have conducted artificial changes to the sample distribution. In the second column, we have removed 40 persons with strongest mathematical identities

	Full STEM sample		Top 40 removed		Bottom 40 removed		Student teacher sample	
Measure	Person	Item	Person	Item	Person	Item	Person	Item
	XX				XX			
3								
	X			2	X			
2	X	2		1	X	2	X	
	X				X		XXX	
		1				1	XX	1, 3
	X			3	X		X	4
		3				3	XXX	2
	X				X		X	
1	XXXXXXXXX	4		4	XXXXXXXXX	4	X	6
	XXXXXXX	5		5	XXXXXXX	5	XXX	
	XXXXXXXXX			8	XXXXXXXXX		XXXX	7
	XXXX	6, 7, 8		6, 7	XXXX	6, 7	XXXX	
	XXXXXXXXXXXX	9	XXXXXXXX		XXXXXXXXXXXX	8, 9	XXXXXX	5, 8, 10
	XXXXXX		XXXXXX	9	XXXXXX	14	XXXXXXXX	15
0	XXXXXXXXX	10, 11	XXXXXXXXX	10, 11	XXXXXXXXX		XX	9
	XXXXXXX	12, 14	XXXXXXX	12	XXXXXXX	10, 11	XXXXXXX	11, 12, 13
	XXXXXXXXXX	13	XXXXXXXXXX	13, 14	XXXXXXXXXX	12, 13	XXXXXXXXXX	
	XXXXXXXXXXXXXX	16	XXXXXXXXXXXXXX		XXXXXXXXXXXXXX	16	XXXXXXXXXX	14, 16
	XXXXXXXXX	15	XXXXXXXXX	15, 16	XXXXXXXXX	15	XX	
	XXXXXXXXXX	17	XXXXXXXXXX	17	XXXXXXXXXX	17	XXXXX	
−1	XXXXX		XXXXX		XXXXX		XXXXXX	17, 18
	XXXXXXXXXX		XXXXXXXXXX		XXXXXXXXXX	18	XXXX	
	XXXXXXXXXX	18	XXXXXXXXXX		XXXXXXXXXX		XXXXX	
	XXXXXXXXX		XXXXXXXXX	18	XXX		XXXXXXXX	
	XXXXXXX		XXXXXXX				XXXXXX	
	XXX		XXX				XXXXXXX	
−2	XXXXXXX	19	XXXXXXX			19	XXXXXXX	
	XXXXXXXX		XXXXXXXX	19			XX	
	XXX		XXX			20	XXXXX	20
	XXXX	20	XXXX	20			XX	
	X		X				XX	
	X		X					19
−3							X	
							X	
−4								
							X	
	Mean = −0.53						*Mean* = −0.69	
	SD = 1.13						*SD* = 1.20	

Fig. 2 Persons and items measured on the same variable

from the analysis. In the third column, we have removed 40 persons with weakest identities. This illustrates how the social structure of being mathematical, as represented by the items' locations along the variable, is more or less unaffected by which persons we include in the analysis (Item 14, "moving back and forth between strategies", would be an exception where there is a slight person dependency). We can observe how positions on the higher end of the variable are quite accurately described, even when persons that typically occupy these positions are removed from the analysis (column 2), and the same can be said about positions on the lower end (column 3).

In this part of the analysis, we also have included the student teacher sample (column 4) for illustrative purposes. Accordingly, we make two conclusions: First, the social structure of being mathematical is more affected by the change of activity, that is, from the STEM context to the student teacher context, than the change of sample distribution. One example is Item 2 ("searching for better methods"), which is a stronger characteristic in the STEM context than in the student teacher context. This has previously been explained by the fact that teacher education emphasises the search for and comparison between methods, whereas the tasks at the technical university that the STEM students attended seemed to emphasise specific methods to a greater extent than teacher education, making it naturally harder for STEM students to search for other methods (Kaspersen, 2015).

Our second conclusion is that, although the social structure of being mathematical differs between the STEM context and the student teacher context, the structures are close enough for comparisons of measures to make sense. That is, student teachers' measures are almost identical ($r = 1.00$) regardless of whether they are measured relative to the social structure of being mathematical in the student teacher context or relative to the social structure of being mathematical in the STEM context (i.e., forcing the items to be identically calibrated as in the STEM context).

4.2 Dimensionality

The standardised residual plot (Fig. 3) for the second contrast, that is, after the main dimension is accounted for, indicates that the instrument consists of three sub-dimensions. The items that do not load to this second contrast can be said to belong entirely to the main dimension. Furthermore, the items that load positively and negatively to this contrast can be thought of as separate from the main dimension.

One way of seeing whether these sub-dimensions are to be considered as strands highly connected with the overarching mathematical identity dimension or as independent dimensions is to split the items into three dimensions and to examine person measures when different sub-dimensions are omitted from the instrument. This procedure illuminates whether a person's mathematical identity measure changes significantly when one sub-dimension is chosen in favour of another.

In all cases, the dis-attenuated correlation, that is, the correlation between person measures when measurement error is accounted for, is close to 1. This result indicates that, in most cases, choosing items from the *second dimension* or items from the *third dimension* does not have a significant effect on persons' measures. Thus, the practical consequences of perceiving mathematical identity as uni-dimensional or three-dimensional are in most cases insignificant. This means that the three dimensions can be considered to be strands of the same construct, and not independent dimensions that need separate instruments.

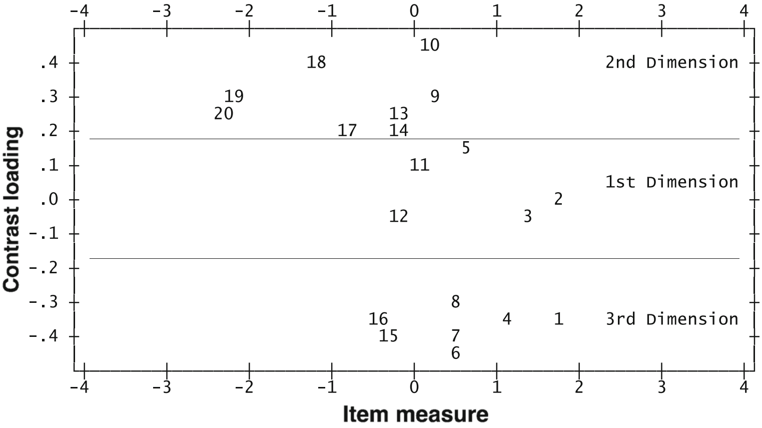


Fig. 3 Standardised residual plot

4.3 STEM students' mathematical identities—a qualitative analysis

In the following section, we provide some qualitative examples of how STEM students' mathematical identities differ across the variable. We have chosen concepts from CHAT in addition to the item characteristics as analytical categories. As such, those characteristics that were found using concepts from CHAT but not included in the items are considered as suggestive for future research.

4.3.1 Characteristics in the lower end

Characteristics on the lower end of mathematical identity are the ones that most STEM students do, except for students with extremely weak mathematical identities, and they seem to be related to working with mathematical problems, such as visualising problems, being able to explain solutions, connecting new and existing knowledge, and so forth. One example is one of the students with lower identity who looked up on his calculus when he faced a more advanced problem:

> Y: Em, when I first started to look at Padè approximation, I was reading an article that was quite mathematical. And I realised, wow, I cannot remember any of this. So, I looked up in some of the books we had in the basic courses, where we had about sequences. Took a quick look. It wasn't very deep, just a recap on the concepts. When I had a basic understanding, I went back on the article.

4.3.2 Medium characteristics

Characteristics around zero are the ones that most persons with medium and high mathematical identities do, in contrast to persons with low identities. These characteristics include making his/her own problems, disliking to be told exactly what to do, finding out why methods would not work, studying proofs until they make sense, deriving formulas, and so forth. One example is one of the students with lower identity who expressed that he had problems working on proofs:

> Z: I never did a proof until I had been here for 2-3 years. I was…I didn't get it. I didn't have… I didn't have a chance. I always ended up with going to an assistant, and they helped me solving the proofs-tasks we had on the assignments.

Another example is how one of the students with stronger mathematical identity derived some formulas instead of remembering them:

> B: One typical thing is integration by parts, which I have derived a million times in my life. You can learn the formula by heart, but I never bothered. So, it's like, we have U and we have V, and… [writes]… there you go. And I often do this with other problems too.

4.3.3 Characteristics on the higher end

Characteristics on the higher end are what mostly students with high mathematical identities do—it is what distinguishes those with high identities from those with medium mathematical identities, and includes struggling with putting problems aside, thinking of times when methods would not work, taking time to find better methods, and so forth. Some of these items seem to be related with extending mathematics beyond the institutionalised community. One example is one of the students with stronger mathematical identity who repeatedly talked about how he was struggling to stop thinking about mathematical problems:

> A: I have had the problem [with a simulation algorithm] for a while, three weeks maybe. I think about the problem everywhere. I think about it when I am at home, on my spare time, and when I relax. […] Yesterday, I was at a café with my girlfriend. […] We were talking about something, and then I just started to think about it.

4.3.4 Suggestive characteristics

When categories from the CHAT framework were used to analyse the data, some additional characteristics of mathematical identity in the STEM context were found. One of those was "being mathematically entertained", that is, enjoying mathematically entertaining problems, results, or stories about other mathematicians, and was, for the most part, addressed outside the institutional setting. One of the students with weaker mathematical identity talked about how he had been interested in Cesàro sums at a social gathering:

> Y: When I was at a 'waffel night', one in the group had read about Cesàro sums. It is one of those classic proofs where you can prove that […] if you have an infinite sum of one minus one plus one minus one, you will get a half. And if you assume that this is right, you can do some other funny proofs, so, you can prove that the sum of all natural numbers equals minus one-twelfth. And we ended up discussing that. […] It was quite exciting.

In this study, this characteristic was discovered in a form hypothesised to belong on the lower end of the variable, that is, it is likely to discriminate students with low identities from students with extremely low identities in the STEM context (if this characteristic aligns with mathematical identity). All the interviewed students—even the two students with low measures—talked about some aspect of mathematics as being entertaining or amusing.

Another characteristic is "using internalised version of complex tools", which has been described in Vygotsky (1978). This characteristic was addressed by one of the students with strong mathematical identity when he was describing a friend who was "more clever than him". As such, this is a characteristic that we hypothesise to be on the higher end of the variable:

> B: It was very fascinating to see how he could imagine… spatial understanding. […] It was like, you could give him an equation in three dimensions, and it was like he could imagine the behaviour. What would this look like? Would it look like a vase, would it look like a ball? He was very much like this. I don't know how he did it.

Another characteristic that was found was the role of mathematics in the activity. That is, the person with weakest mathematical identity expressed the role of mathematics as a "rule"—something he had to do to get on with his education. The person with the second weakest identity expressed mathematics as being both something he had to do, but also as a “tool” for solving non-mathematical problems:

> Y: [Learning mathematics] is very targeted to solve a problem. For me, mathematics is more of a tool than something I immerse in. […] It is means to an end. Something I use to solve a different problem.

The two students with strongest identities both expressed mathematics as a “rule”, as a “tool”, and as the “objective”.

5 Conclusion and discussion

In this paper, we have argued that personal mathematical identities can be measured relative to the social structure of being mathematical within the activity in which persons participate and contribute. Moreover, we have discussed characteristics that belong to a one-dimensional mathematical identity in the STEM context. That does not mean that multiple dimensions, or multiple identities, do not exist—they certainly do. But, if meaningful measures are to be conducted, only highly correlated sub-dimensions can be included. Other dimensions of mathematical identity or other identities that researchers find relevant for their study will need separate instruments.

We will end with two remarks. The first one is the problem with missing data. Most studies tend to refer to missing data as relative to the empirically complete data set, for instance missing responses in a survey or unclear voices in an interview. This is, however, only a subset of the theoretically complete set of missing data, which includes every question we did not ask, everything the interviewees did not say but could have, and so forth. As such, when we conclude that the social structure of being mathematical is invariant enough between the STEM context and the student teacher context for practical purposes, we do so well knowing that most data are missing. Consequently, future research could deliberately search for characteristics that belong to student teachers’ mathematical identities but not to STEM students’ identities, or indeed, any other activity. We believe that measurements can be valuable tools in this search, since such characteristics would be the ones that align with the identity items in one context but not the other. Additionally, future studies might even find characteristics that are bipolar across activities—characteristics that are considered as mathematical in some context, whilst considered as anti-mathematical in other contexts. Such characteristics would correlate in opposite direction across such contexts.

This leads to our final remark. We have shown how one can measure not only personal identities but also the social structure of being mathematical within an activity. Thus, it seems probable that there exist some trajectories of invariance—clusters of activities where the social structure of being mathematical is quite similar. Moreover, it seems equally probable that there exist trajectories of variance—clusters of activities where the social structure of being mathematical is so different that comparisons do not make sense. Accordingly, we suggest future research to measure and compare the social structure of being mathematical along different trajectories, all starting from school mathematics. If it can be statistically verified that some trajectories are invariant whereas other are not, it seems likely that students who travel through these trajectories of invariance benefit more from participating in school mathematics than those who travel through trajectories of variance.

Appendix 1

The Mathematical Depth Instrument (English version)

Never/almost never (1), Sometimes (2), Often (3), Always/almost always (4), Don't know (9)

1. I take the initiative to learn more about math than what is required at school/work.	1	2	3	4	9
2. When I learn a new method, I take time to find out if I can find a better method.	1	2	3	4	9
3. When I learn a new method, I try to think of situations when it wouldn't work.	1	2	3	4	9
4. I struggle with putting math problems aside.	1	2	3	4	9
5. If I forget a formula or method, I try to derive it myself.	1	2	3	4	9
6. I get engaged when someone starts a mathematical discussion.	1	2	3	4	9
7. When I learn something new, I make my own problems.	1	2	3	4	9
8. Math ideas that I hear or learn about help me inspire new trains of thoughts.	1	2	3	4	9
9. When I learn a new method, I like to be told exactly what to do.	1	2	3	4	9
10. When I try to use a method that doesn't work, I spend time to find out why it didn't work.	1	2	3	4	9
11. When I learn a new formula/algorithm, I try to understand why it works.	1	2	3	4	9
12. When I face a proof, I study it until it becomes meaningful.	1	2	3	4	9
13. When I face a math problem, I consider different possible ways I can solve it.	1	2	3	4	9
14. When I work with a math problem, I move back and forth between various strategies.	1	2	3	4	9
15. When I learn something new, it makes me want to learn more things.	1	2	3	4	9
16. When I work with a problem, I pause along the way to reflect on what I am doing.	1	2	3	4	9
17. If I get stuck on a problem, I try to visualize it.	1	2	3	4	9
18. I can explain why my solutions are correct.	1	2	3	4	9
19. I try to connect new things I learn to what I already know.	1	2	3	4	9
20. If I immediately do not understand what to do, I keep trying.	1	2	3	4	9

References

Alexander, N. N. (2015). *Statistical models of identity and self-efficacy in mathematics on a national sample of black adolescents from HSLS: 09* (Doctoral dissertation). Retrieved from http://academiccommons.columbia.edu/catalog/ac:187103

Andrich, D. (1989). Distinctions between assumptions and requirements in measurement in the social sciences. In J. A. Keats, R. Traft, R. A. Heath, & S. H. Lovibond (Eds.), *Mathematical and theoretical systems. Proceedings of the XXIV international congress of psychology* (pp. 7–16). Amsterdam, Holland: Elsevier Science Ltd.

Axelsson, G. B. M. (2009). Mathematical identity in women: The concept, its components and relationship to educative ability, achievement and family support. *International Journal of Lifelong Education, 28*(3), 383–406.

Beijaard, D., Meijer, P. C., & Verloop, N. (2004). Reconsidering research on teachers' professional identity. *Teaching and Teacher Education, 20*(2), 107–128.

Black, L., Williams, J., Hernandez-Martinez, P., Davis, P., Pampaka, M., & Wake, G. (2010). Developing a 'leading identity': The relationship between students' mathematical identities and their career and higher education aspirations. *Educational Studies in Mathematics, 73*(1), 55–72.

Boaler, J., Wiliam, D., & Zevenbergen, R. (2000). The construction of identity in secondary mathematics education. In J. F. Matos & M. Santos (Eds.), *Proceedings of mathematics education and society conference* (pp. 192–202). Montechoro, Portugal: Centro de Investigaçãem Eduação da Faculdade de Ciências Universidade de Lisboa.

Carpenter, T. P., & Lehrer, R. (1999). Teaching and learning mathematics with understanding. In E. Fennema & T. A. Romberg (Eds.), *Mathematics classrooms that promote understanding* (pp. 19–32). Mahwah, NJ: Lawrence Erlbaum Associates.

Cote, J. E., & Levine, C. G. (2014). *Identity, formation, agency, and culture: A social psychological synthesis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Engeström, Y. (1987). *Learning by expanding*. Helsinki: Orienta-Konsultit.

Engeström, Y. (2001). Expansive learning at work: Toward an activity theoretical reconceptualization. *Journal of Education and Work, 14*(1), 133–156.

Gee, J. P. (2000). Identity as an analytic lens for research in education. *Review of Research in Education, 25*, 99–125.

Hernandez-Martinez, P., Williams, J., Black, L., Davis, P., Pampaka, M., & Wake, G. (2011). Students' views on their transition from school to college mathematics: Rethinking 'transition' as an issue of identity. *Research in Mathematics Education, 13*(2), 119–130.

Hiebert, J. (1986). *Conceptual and procedural knowledge: The case of mathematics*. Hillsdale, NJ: Erlbaum.

Holland, D., Lachicotte, W., Skinner, D., & Cain, C. (2001). *Identity and agency in cultural worlds*. Cambridge, MA: Harvard University Press.

Hossain, S., Mendick, H., & Adler, J. (2013). Troubling "understanding mathematics in-depth": Its role in the identity work of student-teachers in England. *Educational Studies in Mathematics, 84*(1), 35–48.

Kaspersen, E. (2015). Using the Rasch model to measure the extent to which students work conceptually with mathematics. *Journal of Applied Measurement, 16*(4), 336–352.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago, IL: The University of Chicago Press.

Kuhn, T. S. (1977). *The essential tension: Selected studies in scientific tradition and change*. Chicago, IL: The University of Chicago Press.

Leont'ev, A. N. (1978). *Activity, consciousness, and personality*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Linacre, J. M. (2015). *A user's guide to WINSTEPS MINISTEP* [Computer software]. Chicago, IL: Winsteps.com.

Rasch, G. (1961). *On general laws and the meaning of measurement in psychology*. Paper presented at the Proceedings of the fourth Berkeley symposium on mathematical statistics and probability. Retrieved from http://projecteuclid.org/download/pdf_1/euclid.bsmsp/1200512895

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests (Expanded ed.). Chicago, IL: The University of Chicago Press. (Original work published 1960).

Roesken, B., Hannula, M. S., & Pehkonen, E. (2011). Dimensions of students' views of themselves as learners of mathematics. *ZDM – International Journal of Mathematics Education, 43*(4), 497–506.

Sfard, A., & Prusak, A. (2005). Telling identities: In search of an analytic tool for investigating learning as a culturally shaped activity. *Educational Researcher, 34*(4), 14–22.

Skemp, R. R. (1987). *The psychology of learning mathematics*. London: Psychology Press.

Skemp, R. R. (1989). *Mathematics in the primary school*. London: Routledge.

Solomon, Y. (2007). Not belonging? What makes a functional learner identity in the undergraduate mathematics? *Studies in Higher Education, 32*(1), 79–96.

Stetsenko, A., & Arievitch, I. M. (2004). The self in cultural-historical activity theory reclaiming the unity of social and individual dimensions of human development. *Theory & Psychology, 14*(4), 475–503.

Thurstone, L. L. (1959). *The measurement of values*. Chicago, IL: The University of Chicago Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. New York, NY: Cambridge University Press.

Wolfe, E., & Smith, E. (2006). Instrument development tools and activities for measure validation using Rasch models: Part II-validation activities. *Journal of Applied Measurement, 8*(2), 204–234.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: Mesa Press.

The association between engineering students' self-reported mathematical identities and average grades in mathematics courses

Eivind Kaspersen[1], Birgit Pepin[2], and Svein Arne Sikko[1]

[1]Norwegian University of Science and Technology, Faculty of Teacher and Interpreter Education, Trondheim, Norway; eivind.kaspersen@ntnu.no, svein.a.sikko@ntnu.no

[2]Technische Universiteit Eindhoven, Eindhoven, The Netherlands; b.e.u.pepin@tue.nl

Arguments have been made that one purpose of learning mathematics successfully is for students to develop mathematical identities. Thus, since students are frequently evaluated with grades in university mathematics courses, a relevant question is how mathematical identities are associated with average grades. This study has measured engineering students' mathematical identities and compared these measures with grades in university mathematics courses, and a Welch's ANOVA conclude that the mean average grade amongst students with high mathematical identities is significant, and about one grade higher than students with low mathematical identities. Moreover, the variance is greater amongst students with low mathematical identities, which indicates a strong association between mathematical identity and average grade only when mathematical identities are high.

Keywords: Mathematical identity, Rasch, ANOVA.

INTRODUCTION

The transfer of mathematical knowledge from university to the world of work seems problematic. Specifically, evidence has been provided that "attainment" in university mathematics courses is poorly transferred. One example is an experiment that illustrated how 17 students and researchers all failed a mathematics examination they had previously passed, even the students who had recently passed the original exam with an "A" (Rystad, 1993). Moreover, selected studies illustrate how the mathematics is often hidden in "black-boxes" (e.g. Williams & Wake, 2007) in the world of work, and consequently, arguments have been made that the world of work seeks more general mathematical characteristics than what is typically assessed in standard exams (e.g. Hoyles, Wolf, Molyneux-Hodgson, & Kent, 2002). On a general note of education, Wenger (1998) argued that learning is about developing identities in communities of practice. In general, over the last decades, there has been an increased attention towards the construct of identity, and mathematical identity in particular (e.g. Axelsson, 2009; Black et al., 2010; Wenger, 1998). Thus, if the world of work seeks general characteristics of working mathematically, a relevant question is how mathematical attainment in university mathematics courses, as represented by average grades, is associated with mathematical identity. This paper addresses this question.

This study has examined the association between self-reported mathematical identities and average grades in university mathematics courses. From a Rasch calibrated instrument, previously validated in Kaspersen (2015), the students were categorised as having a "low," "medium," or "high" mathematical identity, and the paper will illustrate how the mean average grade of students with high mathematical identities was significant and about one grade higher than students with low mathematical identities. Moreover, the variance amongst students with low mathematical identities was higher than amongst students with high mathematical identities, although the difference was not significant (*p*=0.06). The paper concludes that high mathematical identities are associated with high average grades in university mathematics courses. However, the same conclusion is not true amongst students with lower mathematical identities.

THEORETICAL FRAMEWORK

The construct of identity suffers from a lack of consensus on general philosophical issues (Cote & Levine, 2014). Specifically, identity is defined differently across different studies and paradigms, such as "a certain kind of person" (Gee, 2000, p. 99), "those narratives about individuals that are reifying, endorsable and significant" (Sfard & Prusak, 2005, p. 44), and "self-perceived mathematical knowledge, ability, motivation and anxiety" (Axelsson, 2009, p. 387).

This lack of consensus is typical in pre-paradigmatic fields (Kuhn, 1970). Unlike firm paradigmatic fields where well-established theories tend to guide the analyses, research in pre-paradigmatic areas has a more dialectical relationship between data and theory (Kuhn, 1977). This description is a fair representation of how the theoretical perception in this study was chosen. That is, no ready-made theory was chosen on pure faith. Rather, a definition of identity was established that was consistent with measurement (i.e., consistent to conclude some persons to have stronger mathematical identities than others), yet, influences by fragments of multiple existing theories. The following theoretical perspective and a wider discussion on practical significance has been provided in more detail in Kaspersen, Pepin, and Sikko (2017).

On another note, we do not regard theories as mirrors of some true reality. Thus, we do not believe that some theories *are* true, and that others *are* false. When we propose the following theoretical perspective, therefore, we are not refusing other perspectives, for instance, a narrative view on identity. Rather, we claim that *if* we choose the following perspective, then the practical consequence is that mathematical identity can be measured.

The perspective of mathematical identity relies on two assumptions. First, we assume that identity (originated from the Latin *idem*) is about sameness and distinction. As such, the position in this study juxtaposes perspectives that consider persons to have their unique identity. That is, persons are indeed unique. However, they can be defined as identical with respect to a set of characteristics, just like mathematical objects can be identified by certain characteristics while remaining unique on others. Moreover, since there exists an infinite number of characteristics, identities have a varying degree of complexity. That is, mathematical identity can be binary, linear, or multidimensional, and we argue that there is no ontological limit to the number of dimensions. Consequently, there exists no set of criteria that dictates when researchers have arrived at the final dimension. Hence, the choice of complexity can be nothing but pragmatic, and in this study, we have chosen a one-dimensional perspective on mathematical identity, whereby persons are distinguished on a continuum from having a low to having a high mathematical identity within the engineering education context.

Furthermore, if we accept that persons participate and contribute in multiple activities, a consequence is that each person has multiple identities, a position that is shared by many authors, for example Black and colleagues (2010) who, inspired by Leont'ev (1981), presented the idea of "leading identity." Since there is no limit to how many ways persons can be distinguished, we argue that there exists no limit to the number of identities, although the number of identities that individuals are consciously aware of is likely to be finite. Moreover, in this study, we take no definite position on the relationship between identities. Thus, when we later will conclude that selected persons have (more or less) the same mathematical identity, we do not make claims about how these are related to the multiplicity of identities–for instance, whether they are central/leading or peripheral identities.

Second, we assume that identity is relational by nature. That is, persons can be concluded to be identical relative to a set of characteristics, only if the structure of these characteristics is person-independent. Thus, in quantitative studies, we reject the assumption that persons with the same score on some test or questionnaire are identical unless statistical evidence is provided that the items stay invariant across relevant subgroups. Hence, there likely exist contexts that are so different that comparisons of identities across these contexts do not make sense. Consequently, we argue that the methods that are applied to capture identities should also capture the level of invariance.

In conclusion, we define mathematical identity to be *where persons position themselves relative to the social structure of being mathematical within the activity in which they participate and contribute*. From a one-dimensional perspective, "the social structure of being mathematical" is a person-independent set of characteristics and their internal structure (i.e., their relative distance) that distinguishes persons on a continuum from having a "low" to having a "high" mathematical identity. "Where persons position themselves" is persons' positions relative to the social structure.

METHOD

To test the relationship between engineering students' self-reported mathematical identities and average grade in mathematics courses, a convenience sample consisting of Norwegian engineering students (*N*=361) was selected. 47 students attended an "Introductory course in mathematics," 71 students attended a "Calculus 2" course, 113 attended a "Calculus 3" course, 11 a "Cryptography" course, and 119 were students from a variety of courses in their normalised final year of education. The participants responded to a Rasch-calibrated instrument (Rasch, 1960), previously validated in Kaspersen (2015), that measures persons on a continuum from having a low to having a high mathematical identity relative to 20 uni-dimensional characteristics. The items in the instrument were collected from three sources: the literature, other related instruments, and from persons contributing in mathematical activities (e.g., students and lecturers). The validation of the instrument will not be discussed in depth, as details can be found in Kaspersen (2015). The person reliability, analogous to Cronbach's alpha, was 0.87. Moreover, from principal component analysis of residuals, the instrument was found to be sufficiently uni-dimensional for the purpose of measurement with a 1.99 unexplained variance (in Eigenvalue units) in a second contrast. Furthermore, the mean of the squared standardised residuals (outfit mnsq) and the information-weighted version (infit mnsq) (see e.g., Bond & Fox, 2003, p. 238 for a detailed description) indicated a sufficient data-model fit, with Item 6 and Item 15 as the most underfitting items (Table 1).

Rasch measurement requires additivity, uni-dimensionality, and invariance, and the probability of an observation is a function of the difference between a person's measure and a characteristic's measure (e.g. Wright & Stone, 1979). Thus, most response strings follow a Guttman-like structure with most deviations around the measure of the person. Consequently, persons with approximately the same measures, except those with large misfit, have, not only the same measures but also approximately the same combination of self-reported characteristics (and thus concluded to be identical with respect to these characteristics).

After the validation of the instrument, the respondents were categorised as having either low (measures lower than -1), medium (measures between -1 and 1), or high (measures above +1) mathematical identities (all measures are in logit units). The distance from the "low"/"medium" to the "medium"/"high" thresholds was about the same distance as one response category. Consequently, persons with "high" mathematical identities were expected to respond at least one category higher on each characteristic than persons with "low" mathematical identities. Subsequently, a one-way ANOVA was conducted to compare the association between mathematical identity and the self-reported average grade in mathematics courses at the University (from grade F=1 to grade A=15). However, since the Levene's (1960) test barely accepted the null hypothesis of homogeneity of variances (p=0.06), and the sample sizes across categories were unequal, the Welch's ANOVA was chosen since it is more robust to unequal sample size and variance.

Moreover, the assumption of normality was violated, and the grades were ordinal as opposed to interval measures. Since Welch's ANOVA assumes normal and interval measures, 10,000 simulations were made in R (R Core Team, 2015) to assess how these violations affected the robustness of the analysis. To ease this part of the analysis, we considered a transformed data set which had no difference in the mean across groups but was otherwise identical to ours–the assumptions of Welch's ANOVA were violated equally in the empirical study and the simulated studies. This transformation eased the interpretation since we could compare the results with the statistical ideal situation (perfectly normal interval data, equal sample size and variance). If our data set was as good as the ideal situation, we would expect the Welch's ANOVA to show a significant difference in about 5% of the simulations.

Specifically, from the empirical data frame, M, a new data frame, M', was made whereby each grade in the medium and high groups was shifted so that the mean of all three categories in M' were equal (i.e., keeping the sample sizes and distributions, but aligning the means). From M', 10,000 data frames, $M_1 – M_{10,000}$, were randomly sampled whereby the sample sizes in the three groups were equal to the original M. Subsequently, Welch's ANOVA was conducted on each simulated data frame. Since the result showed that 5.2% of the p-values in the simulations were less than .050, it was concluded to ignore violations of Welch's ANOVA's assumptions since they had only a trivial negative effect on the robustness.

RESULT

Mathematical identities

Due to the Guttman-like response strings, a rough interpretation of Table 1 is that most students with low mathematical identities (measures lower than -1) agreed with characteristics much lower than -1, and disagreed with those much higher than -1. That is, students with low mathematical identities often keep trying when they get stuck, but they rarely study proofs until they make sense (to them), they rarely like to discuss mathematics, they rarely derive formulas, etc. Likewise, students with medium mathematical identities (measures between -1 and 1) frequently keep trying, connect new and existing knowledge, and can explain why their solutions are correct, but rarely take the initiative to learn more than expected, rarely take the time to find better methods, etc. Students with high mathematical identities (measures above +1) agree with most characteristics in the instrument. A more thorough discussion is discussed in Kaspersen, Pepin, and Sikko (2017).

Table 1: Characteristics of mathematical identities amongst Norwegian Engineering students

Item statistics: Measure order

Measure	INFIT MNSQ	OTFIT MNSQ	Item
1.91	.81	.83	1. Takes time to find better methods
1.58	1.08	.99	2. Takes the initiative to learn more
1.24	.91	.86	3. Thinks of times when methods don’t work
.55	1.22	1.20	4. Struggles with putting problems aside
.51	1.05	1.07	5. Derives formulas
.45	1.36	1.37	6. (x) Likes to be told exactly what to do
.41	.96	.95	7. New ideas lead to trains of thoughts
.32	1.05	1.05	8. Likes to discuss math
.20	1.07	1.07	9. Makes his/her own problems
.05	.99	.99	10. Studies proofs until they make sense
.04	.86	.88	11. Moves back and forth between strategies
–.10	.87	.86	12. Tries to understand formulas/algorithms
–.20	.72	.74	13. Considers different possible solutions
–.26	1.03	1.05	14. Pauses and reflects
–.38	1.32	1.31	15. Finding out why methods do not work
–.47	.86	.86	16. Wants to learn more things
–.77	1.20	1.20	17. Visualises problems
–1.19	.71	.76	18. Can explain why solutions are correct
–1.83	.83	.88	19. Connects new and existing knowledge
–2.05	1.02	1.06	20. Keeps trying

Note. Item 6 was negatively coded

Items in their entirety in https://www.researchgate.net/publication/309740755_math_identity_questionnaire

Moreover, it is evident from Table 1 how the identities in this study were situated amongst the engineering student context. That is, persons with measures, say, around 0.5 in other contexts would be identical to engineering students with the same measures, only if the same set of characteristics were proven to be invariant (i.e., calibrated to have the same structure) in both contexts.

The relationship between self-reported mathematical identities and average grade

Figure 1 illustrates the relationship between self-reported mathematical identity and average grade in university mathematics courses. The Welch's ANOVA showed that the association between mathematical identity and self-reported average grade was significant, *F*(2, 110.79)=31.966, *p*=0.000. Moreover, the mean of the self-reported average grade amongst students with high mathematical identities was about one grade higher than those with low mathematical identities. The Games-Howell test showed that the difference was significant between all groups with low-medium as the least significant (*p*=0.001).

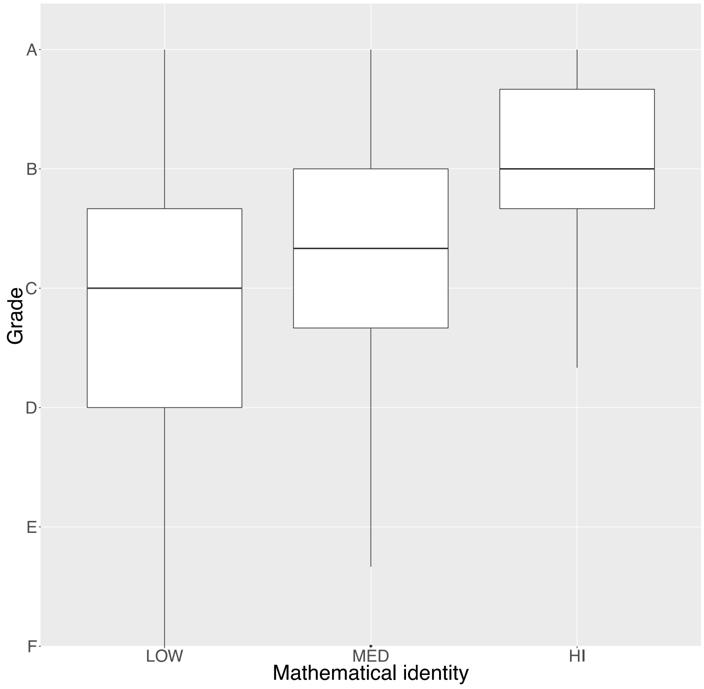


Figure 1: The relationship between self-reported mathematical identity and average grade in university mathematics courses

The unequal variance is also illustrated in Figure 1. Specifically, the variances decreased with the increase of mathematical identity. That is, high mathematical identities are associated with high self-reported average grade. However, there seems to be no limit to how low mathematical identities students can have and still get high grades.

CONCLUSION AND DISCUSSION

In this paper, we have argued that the average grade in university mathematics courses amongst students with high mathematical identities is about one grade higher than amongst students with low mathematical identities, and the difference is significant. Moreover, we have shown that the variance of self-reported average grades amongst students with low mathematical identities is higher than amongst students with high mathematical identities. That is, students with high identities get, for the most, high grades. However, the grades of students with lower identities are more uncertain.

We have in this study examined the association, and not the causal relationship, between self-reported mathematical identities and average grades, and therefore we argue that the significance of the result is that it points the direction for future research. Specifically, we suggest future research to address the following:

First, replicates of this study should seek more precise measures. That is, the precisions of the mathematical identity measures can be improved by including more response categories (as long as they are sufficiently validated) and more items, particularly near the "gaps" (e.g., between 0.5 and 1.2 logits). Moreover, the precision of the average grade would most likely be improved if self-reported average grades were substituted with actual average grades.

Second, future research should seek a more causal relationship between identities and grades. Specifically, this study does not conclude that an increase in mathematical identity infers an increase in attainment.

Third, future research could study the significance of mathematical identity versus the significance of attainment. For instance, students can be categorised as having "low identities and low grades," "low identities and high grades," or "high identities and high grades," and subsequently studied with respect to other variables, for example, in the transition from university to the world of work.

Fourth, we argue that future research can transfer the design of this study to other samples and forms of testing students' attainment. For example, relationships between mathematical identity and measures on international standardised tests, such as PISA and TIMSS, can be tested. Accordingly, we argue that future research can nuance the debate on the significance of these tests. If some districts/countries are "teaching to the test," then one might hypothesise that a relatively great proportion of students in these districts/countries are in the "top left corner"– that is, students with low mathematical identities, yet, high measures of attainment.

REFERENCES

Axelsson, G. B. M. (2009). Mathematical identity in women: The concept, its components and relationship to educative ability, achievement and family support. *International Journal of Lifelong Education, 28*(3), 383–406.

Black, L., Williams, J., Hernandez-Martinez, P., Davis, P., Pampaka, M., & Wake, G. (2010). Developing a 'leading identity': The relationship between students' mathematical identities and their career and higher education aspirations. *Educational Studies in Mathematics, 73*(1), 55–72.

Cote, J. E., & Levine, C. G. (2014). Identity, formation, agency, and culture: A social psychological synthesis. New York, NY: Psychology Press.

Gee, J. P. (2000). Identity as an analytic lens for research in education. *Review of research in education, 25*, 99–125.

Hoyles, C., Wolf, A., Molyneux-Hodgson, S., & Kent, P. (2002). *Mathematical skills in the workplace*. London, UK: The Science, Technology and Mathematics Council.

Kaspersen, E. (2015). Using the Rasch model to measure the extent to which students work conceptually with mathematics. *Journal of Applied Measurement, 16*(4).

Kaspersen, E., Pepin, B., & Sikko, S.A. (2017). Measuring students' mathematical identities. *Educational Studies in Mathematics*, 1–17. Advance online publication. doi: 10.1007/s10649-016-9742-3

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago, IL: The University of Chicago Press.

Kuhn, T. S. (1977). *The essential tension: Selected studies in scientific tradition and change*. Chicago, IL: The University of Chicago Press.

Leont'ev, A. N. (1981). *Problems of the development of mind*. Moscow, RU: Progress.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics* (pp. 278–292). Stanford: Stanford University Press

Nyström, S. (2009). The dynamics of professional identity formation: Graduates' transitions from higher education to working life. *Vocations and learning, 2*(1), 1–18.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980). Chicago, IL: University of Chicago Press.

R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria: URL https://www.R-project.org/

Rystad, J. (1993). Alt glemt på grunn av en ubrukelig eksamensform? En empirisk undersøkelse av Matematikk 2 eksamen ved NTH. *UNIPED* (2–3), 15–29.

Sfard, A., & Prusak, A. (2005). Identity that makes a difference: Substantial learning as closing the gap between actual and designated identities. In H. L. Chick and J. L. Vincent (Eds.), *Proceedings of the 29th Conference of the International Group for the Psychology of Mathematics Education* (pp. 37–52). Melbourne: Psychology of Mathematics Education.

Wenger, E. (1998). *Communities of practice: learning, meaning, and identity*. Cambridge, UK: Cambridge University Press.

Williams, J., & Wake, G. (2007). Black boxes in workplace mathematics. *Educational Studies in Mathematics, 64*(3), 317–343.

Wright, B., & Stone, M. H. (1979). *Best test design*. Chicago, IL: Mesa Press.