International Journal of Educational Research 103 (2020) 101620

Contents lists available at ScienceDirect

International Journal of Educational Research

journal homepage: www.elsevier.com/locate/ijedures

Measuring mathematical identity in lower secondary school☆

Eivind Kaspersen*, Bjørn Owe Ytterhaug

Department of Teacher Education, Faculty of Social and Educational Sciences, Norwegian University of Science and Technology, Trondheim, Norway

ARTICLE INFO

Keywords:
Mathematical identity
Measurement
Rasch
Context-dependency

ABSTRACT

The authors examined whether we can use the same instrument for measuring and comparing the mathematical identities of lower secondary school students and those of university students in science, technology, engineering, and mathematics (STEM). Specifically, Rasch measurement techniques were used on items from an instrument that was earlier validated for measuring mathematical identities in STEM contexts to assess the psychometric properties of the instrument in lower secondary school. Moreover, data from the two contexts were merged to assess the invariance of the instrument. The results indicate that the same instrument can measure mathematical identity in STEM contexts and lower secondary school contexts. Also, evidence is provided that the instrument is practically invariant. Implications and suggestions for further research are provided.

1. Introduction

In mathematics education, one of the dominant constructs during the last two decades has been "identity." Mathematics related identities have been used to examine aspects of power, access, equity, career choice, interactions between individuals in mathematical activities, socio-political issues, and persons' relationships with mathematics (Darragh, 2016, p. 19–20). However, few instruments exist for measuring what we refer to in this paper as mathematical identity (MI). A review of MI studies showed that 45 of 47 reviewed papers reported on eight or fewer learners (Graven & Heyd-Metzuyanim, 2019) and that most studies (76 %) used interview data to understand identities in mathematics education.

One reason why qualitative data dominate studies on identity in mathematics education might be theoretical; it could be the case that principles of identity and principles of measurement, for instance, those that were proposed by Thurstone (e.g., 1959), are incompatible. Although this claim might hold for some definitions of identity, we maintain that the argument is not true in general, and we base our reasoning on the observation that measurement is a frequently applied method for studying identities in social and educational psychology (Abdelal et al., 2009). For example, instruments exist for measuring racial and ethnical identities (e.g., Cross & Vandiver, 2001; Phinney & Chavira, 1992), ego identities (Tan, Kendis, Porac, & Fine, 1977), vocational identities (Holland, Johnston, & Asama, 1993), creative role identities (Huang, Lee, & Yang, 2019), and gender identities (e.g., Henley, Meng, O'Brien, McCarthy, & Sockloskie, 1998), to mention but a few.

Instruments for measuring identity in psychology have been used for understanding characteristics of individuals who identify strongly with particular identities (e.g., Cross & Vandiver, 2001), how characteristics vary between contexts (e.g., Cokley, Komarraju, King, Cunningham, & Muhammad, 2003), how identities develop over time (e.g., Perron, Vondracek, Skorikov, Tremblay, & Corbiere, 1998), and how identities relate to other variables such as teaching practices (e.g., Huang et al., 2019) and racism (e.g.,

☆ This article is based on data published in Kaspersen (2018) and Ytterhaug (2019)

* Corresponding author at: Department of Teacher Education, NTNU, NO-7491, Trondheim, Norway.
E-mail address: eivind.kaspersen@ntnu.no (E. Kaspersen).

https://doi.org/10.1016/j.ijer.2020.101620
Received 20 April 2020; Received in revised form 22 May 2020; Accepted 25 May 2020
Available online 22 June 2020

Swim, Hyers, Cohen, Fitzgerald, & Bylsma, 2003).

A motivation for developing valid and reliable instruments for measuring MI is that such instruments might contribute information for answering similar questions, for instance, questions about characteristics of individuals who identify strongly with mathematics, how characteristics of MI vary between contexts, how MI relates to other variables such as grades and career choice, and how MI develops over time.

Kaspersen, Pepin, and Sikko (2017) developed an instrument for measuring MI, but this instrument was validated in university contexts only, including students in science, technology, engineering, and mathematics (STEM). It is an open question, therefore, whether MI can be measured in other contexts, and if so, how context-dependent the instrument is.

In this study, we asked whether the instrument that measures STEM students' MIs can also be applied to measure lower secondary students' (LS) MIs. Moreover, if the answer to the above question were positive, we questioned whether the instrument is invariant between STEM contexts and LS contexts, that is, whether it makes sense to compare a measure of "LS MI" with a measure of "STEM MI."

With this background, the research questions we answer in this paper are the following:

1 What are the psychometric properties of an instrument for measuring MI in lower secondary school?
2 What is the level of invariance when the instrument is calibrated in STEM contexts and when it is calibrated in LS contexts?

In essence, our procedure was this: First, we administered the "STEM MI instrument" (Kaspersen et al., 2017; Kaspersen, 2018) to lower secondary students in Norway ($n = 332$). We then used a Rasch measurement framework (e.g., Rasch, 1980; Wright & Masters, 1982) for assessing the psychometric quality of the instrument, before we compared the invariance of the instrument when it was administered to STEM students and when it was administered to LS students. Finally, we compared LS students' measured MIs with STEM students' measured MIs.

2. Theoretical framework

Many frameworks exist for studying identity, each offering a constrained abstract world in which empirical research can occur. Commonly applied frameworks in education include identities in cultural worlds (e.g., Holland, Lachicotte, Skinner, & Cain, 2001), ego identities (e.g., Erikson, 1959; Marcia, Waterman, Matteson, Archer, & Orlofsky, 2012), identities in communities of practice (e.g., Wenger, 1998), discursive identities (e.g., Gee, 2000), and narrated identities (e.g., Sfard & Prusak, 2005).

Since we in this study aimed to validate and compare instruments for measuring MI, we adopted a view on MI that, we maintain, is compatible with principles of measurement (Thurstone, 1959). Specifically, we have interpreted MI as a relation between *social mathematical identities* and *personal mathematical identities*. This interpretation is inspired by Deaux (1993), who defined *social identities* as "those roles *or* membership categories that a person claims as representative." *Personal identities*, for Deaux (1993), were "those traits and behaviors that the person finds self-descriptive, characteristics that are typically linked to one or more of the identity categories" (p.6). That is, personal identities are meaningful only relative to social identities, which, in turn, are negotiated by individuals.

To explain how we have operationalized the term MI, we describe below the outcome of a previous study for measuring STEM students' mathematical identities (Kaspersen, 2018). Without going into methodical details at this point, a set of characteristics of being mathematical was validated as representing one dimension of *social mathematical identities* (we assume that there exist other characteristics, and we believe that there are different dimensions of MI, but they are not the focus in this paper).

Characteristics of being mathematical in the STEM context include (but they are not restricted to) those that are described on the right-hand side in Fig. 1: "struggling with putting mathematical problems aside," "liking to discuss mathematics," "studying proofs until they make sense," and so forth. Each characteristic was associated with a measure using Rasch measurement techniques, which we describe in the next section. Roughly, characteristics with low measures are endorsed by most respondents; characteristics with high measures are endorsed by students with high MI only. For instance, when students learn something new, most students, except for those with extremely low MI measures, reported that they try to "connect new and existing knowledge." By contrast, only persons with extremely strong MI measures reported that, when they learn a new method, they try to "think of times when this method would *not* work." When different sub-samples within the STEM context were used in separate analyses—when the analysis was conducted using females only, males only, persons with strong MIs only, persons with weak MIs only, and so forth—the measures of the characteristics remained practically invariant; this property is a requirement of measurement (Kaspersen, 2018).

In the rest of this paper, this is what we mean when we say *social mathematical identity*: (1) a set of characteristics and (2) how they are structured (3) regardless of which subgroup within the studied context that is used in the analysis.

Individuals are represented on the left side in Fig. 1. They are measured on the same variable as the characteristics. A rough interpretation is this: Persons with a specific measure, say -1.5 units, will most likely respond positively to characteristics below their measure (in our example, they will most likely "keep trying when getting stuck" and "connect new and existing knowledge when learning something new") and negatively to characteristics above their measure (in our example, the rest of the characteristics). Some deviation from this pattern is expected; too much deviation, however, will show up as an anomaly (misfit) using the methods we explain in the next section.

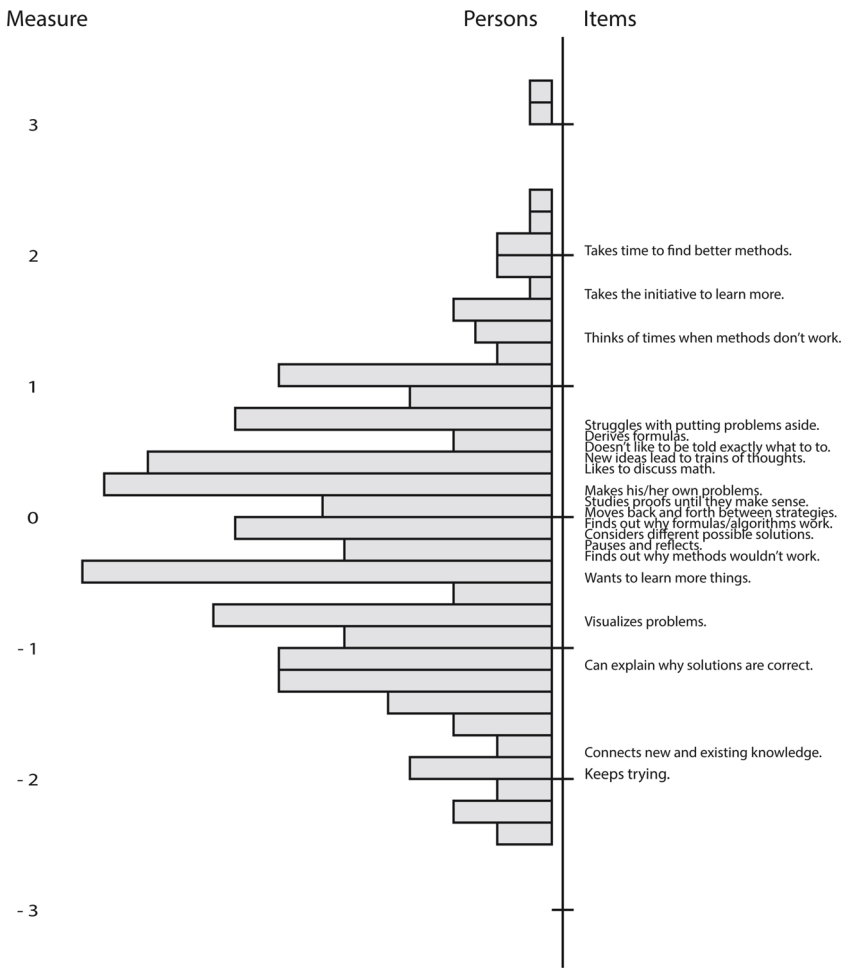

Fig. 1. Mathematical identity in STEM.
Note: The locations of some characteristics in this figure have been shifted slightly to avoid overlapping text.

In the rest of this paper, this is what we mean when we say *personal mathematical identity*: (1) a person's position (2) relative a social mathematical identity. Accordingly, two persons are concluded as having approximately the same MI only when they have the same measure relative to the same social structure.

The reason why we have chosen identity as opposed to alternative terms, such as attitude or personality, is the relational view we have taken, namely, that identity consists of a relation between individuals and social structure, both of which, we argue, can be measured. We appreciate, however, that the choice comes with baggage. In particular, since there exist many conceptualizations of identity, for instance, those that we have mentioned in this section, and since many conceptualizations of identity are incompatible, the instrument we have validated in this study is not an instrument for measuring *every* mathematical identity. A more thorough discussion of the relation between conceptualizations of identity and principles of measurement is provided in Kaspersen (2018).

3. Methods

With the definitions we proposed in the previous section, this study had two objectives: (1) to assess whether the psychometric properties of the characteristics listed in Fig. 1 are sufficiently good for measuring LS students' personal mathematical identities, and (2) to assess whether the structures of the characteristics (i.e., the social mathematical identities) in the two contexts are similar enough for comparisons of personal mathematical identities to make sense.

In this section, we present, first, the methods we applied to assess the psychometric properties of the instrument for measuring LS students' MIs. Then, we describe how we assessed the invariance between the STEM context and the LS context.

Table 1
Participants in the study.

STEM		
	n = 372	
		48 Pre-calculus
		72 Calculus 2
		115 Calculus 3
		12 Cryptography
		125 (Norm.) final year students in a variety of courses
LS		
	n = 332	
		128 8th grade (year 12–13)
		117 9th grade (year 13–14)
		87 10th grade (year 14–15)
Total	N = 704	

3.1. Participants and data collection

The data were collected from a convenient sample of 332 LS students in Norway: 8th graders (n = 128), 9th graders (n = 117), and 10th graders (n = 87) (Table 1). When we studied invariance between the LS context and the STEM context, we merged the collected data with data from Kaspersen (2018), which includes responses from STEM students (n = 372).

All LS students responded to a Norwegian version of the STEM MI questionnaire (English translation in Appendix A). The instrument was administered in a pseudo-random order. Two changes were made from the original instrument. First, the wordings were slightly adjusted to fit LS students. For instance, the word "method" (*metode* in Norwegian) was specified to "method for calculation" (*regnemetode* in Norwegian) when the instrument was administred to LS students. Second, the response categories were adjusted from *never/almost never*, *sometimes*, *often*, *always/almost always*, or *don't know* as they appeared in the original instrument to *never*, *sometimes*, *often*, *always*, or *don't know*. The response categories were adjusted due to results from rating-scale analyses (Kaspersen, 2018) which indicated some misfit in the extreme categories. However, the adjustment seemed to have a negligible effect on person and item measures, and hence, on the results we report in this paper. That is, the correlation between item measures when the category thresholds were anchored as they emerged in the LS analysis and the STEM analysis was practically perfect (r = 1.00) (Table 1).

3.2. Assessing the psychometric properties of the MI instrument in lower secondary school

In the analyses, we used the Andrich (1978) rating scale model (RSM). The RSM (Eq. 1) is an expression of the likelihood 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. By convention, $\sum_{k=0}^{m} \tau_k = 0$ (Andrich, 1978). In essence, the model expresses the following relationship: When the measure, β, of some person increases, the likelihood that the person will respond in the lowest category on some item, with measure δ, decreases, and the likelihood that the person will respond in the highest category increases.

$$\mathrm{P(X{=}}x) = \frac{\exp(x(\beta - \delta) - \sum_{k=0}^{x} \tau_k)}{\sum_{n=0}^{m} \exp(n(\beta - \delta) - \sum_{k=0}^{n} \tau_k)} \quad (1)$$

The joint maximum likelihood estimation (JMLE) algorithm was used for estimating the parameters (i.e., the person measures, the item measures, and the threshold measures) in the model. Wright and Stone (1979) include a technical description of JMLE. An intuitive explanation of the algorithm is that it searches a combination of measures that maximizes the likelihood of the observed data. The measures are reported in *logits* which is an arbitrary unit of measurement with the zero-point located at the mean measure of the items.

The JMLE algorithm assumes that basic requirements for measurement that were proposed by Thurstone (1959)—additivity, invariance, and unidimensionality—are met. However, whether this assumption holds is an empirical question. Thus, to test the psychometric quality of the instrument in the LS context, we applied the framework for validity presented by Wolfe and Smith (2007a, b).

To ensure *content validity*, Infit Mnsq and Outfit Mnsq were used to assess data-to-model fit. Outfit Mnsq is a statistic based on the mean of squared standardized residuals; Infit Mnsq is an information-weighted version of the same statistic, one that is less affected by outliers (Bond & Fox, 2003, p. 238). The cut-score of 1.4 was set for flagging items with a possible misfit. Typically, items that are interpreted differently by the respondents, or items that belong to dimensions other than what most of the others do, would show high Outfit Mnsq and Infit Mnsq values. Accordingly, with this analysis we examined the level of unexpected responses to items.

To find evidence for *substantive validity*, we considered aspects for well-functioning rating scales (Linacre, 2002), except for one aspect (ratings should imply measures, and measures should imply ratings), which has been proven inaccurate (Kaspersen, 2019).

With this analysis, we assessed whether the content of the categories (e.g., what *sometimes* means) is interpreted "fairly similar" by the respondents.

Differential item functioning (DIF)—the loss of invariance—was assessed using the Rasch-Welch *t*-test to find evidence for the *generalizability* aspect of validity. Loss of invariance was assessed between gender and between 8th-, 9th-, and 10th-graders. For this analysis, we followed Linacre's (2015) suggestion of 0.64 logits as a cut-value for meaningful differences. Moreover, we set the critical *p*-value to be .05. With this analysis, we assessed the internal stability of the instrument within the LS context.

Structural validity was sought from examinations of the dimensionalities of the items by principal components analysis (PCA) of standardized residuals. A critical threshold of 2.0 in eigenvalue units was chosen for possible multidimensionality (Linacre, 2015, p. 391). With this analysis, we examined whether the items aligned (approximately) on one dimension, or whether multiple instruments would be more appropriate to capture underlying facets of MI.

3.3. Assessing the invariance between the STEM context and the LS context

To address the second research question, we merged the data from the LS students' responses with data from the STEM students' responses. Subsequently, we conducted a DIF analysis of LS responses and STEM responses. This analysis revealed which items were forming similar structures in the two contexts and which items indicated structural differences between the contexts.

When we observed that one set of items structured similarly and that another set of items structured differently between the observed contexts, a natural question was: Is the instrument invariant *enough* for comparing MIs in LS with MIs in STEM? For example, in a longitudinal study, does it make sense to compare STEM students' MIs with their MIs when they were LS students? To answer this question, we conducted the following analysis: First, the LS students were measured relative to the instrument as it was calibrated on LS students only. Then, they were measured relative to the instrument as it was calibrated on STEM students only. Finally, LS students' measures using the LS instrument were correlated with LS students' measures using the STEM instrument. A strong correlation here would indicate that the instrument is practically invariant between the contexts.

4. Results

In this chapter, we answer the first research question when we report on the psychometric properties of the MI instrument when it was administered to students in LS. Subsequently, we answer the second research question when we describe the level of invariance between the MI instrument in the STEM and the LS contexts.

4.1. Psychometric properties of the MI instrument in the LS context

4.1.1. Summary statistics

The meaning of reliability of psychometric measures is analogous to the meaning of reliability of physical measures (e.g., Boone, Staver, & Yale, 2013). In physical measurement (e.g., using a ruler to measure height), reliability is mostly affected by the number of marks on the ruler and how they are positioned relative to the things or the persons being measured. That is, more marks provide more accurate measures than do fewer marks; also, marks that are distributed around what is being measured provide more accurate measures than do marks that are distant from the measured objects. The same is true with psychometric instruments. Reliability is affected mostly by the number of items and how they are positioned relative to person measures.

In Fig. 2, we illustrate the distribution of person measures and item measures along the same variable. Most items are positioned around most persons. Accordingly, the Cronbach's alpha ($\alpha = .85$) is relatively high. However, more items could have been included to improve reliability further, particularly on the lower end of the variable.

4.1.2. Item fit statistics

Most items showed good data-model fit (Infit Mnsq/Outfit Mnsq < 1.4) except for Item 9 (Infit Mnsq $= 1.56$, Outfit Mnsq $= 1.89$). Also, the ICC of Item 9, illustrated in Fig. 3, indicated a poor fit between the empirical data and the Rasch model. Together, these results suggest that Item 9 is inappropriate for measuring MI in the LS context, and therefore, we excluded the item in the subsequent analyses. When Item 9 was excluded, the fit-statistics of the remaining items were good, as indicated in Table 2. Also, the ICCs indicated a good data-model fit. A selection of four ICCs (i.e., the ICCs of Items 1, 2, 3, and 4) are provided in Fig. 4.

4.1.3. Invariance

The DIF analyses showed that Item 2 ("takes time to find better methods") had a significant variance ($p < .001$) between 8th- and 9th-graders. All else being equal, the item was relatively easier to agree with for 8th-graders ($\delta_{2,\,8th} = 0.32$) than it was for 9th-graders ($\delta_{2,\,9th} = 1.19$). In our study, the level of invariance had little practical consequence. That is, the correlation between persons' measures when the item was included and their measures when the item was excluded was close to perfect ($r = 1.00$); also, the average absolute difference between persons' measures in the two trials was 0.08 logits. Nevertheless, the observed DIF is an indication that students understand in different ways what it means to "take time to find better methods."

All other items were concluded as invariant between grades. Between gender, all items, including Item 2, were concluded as

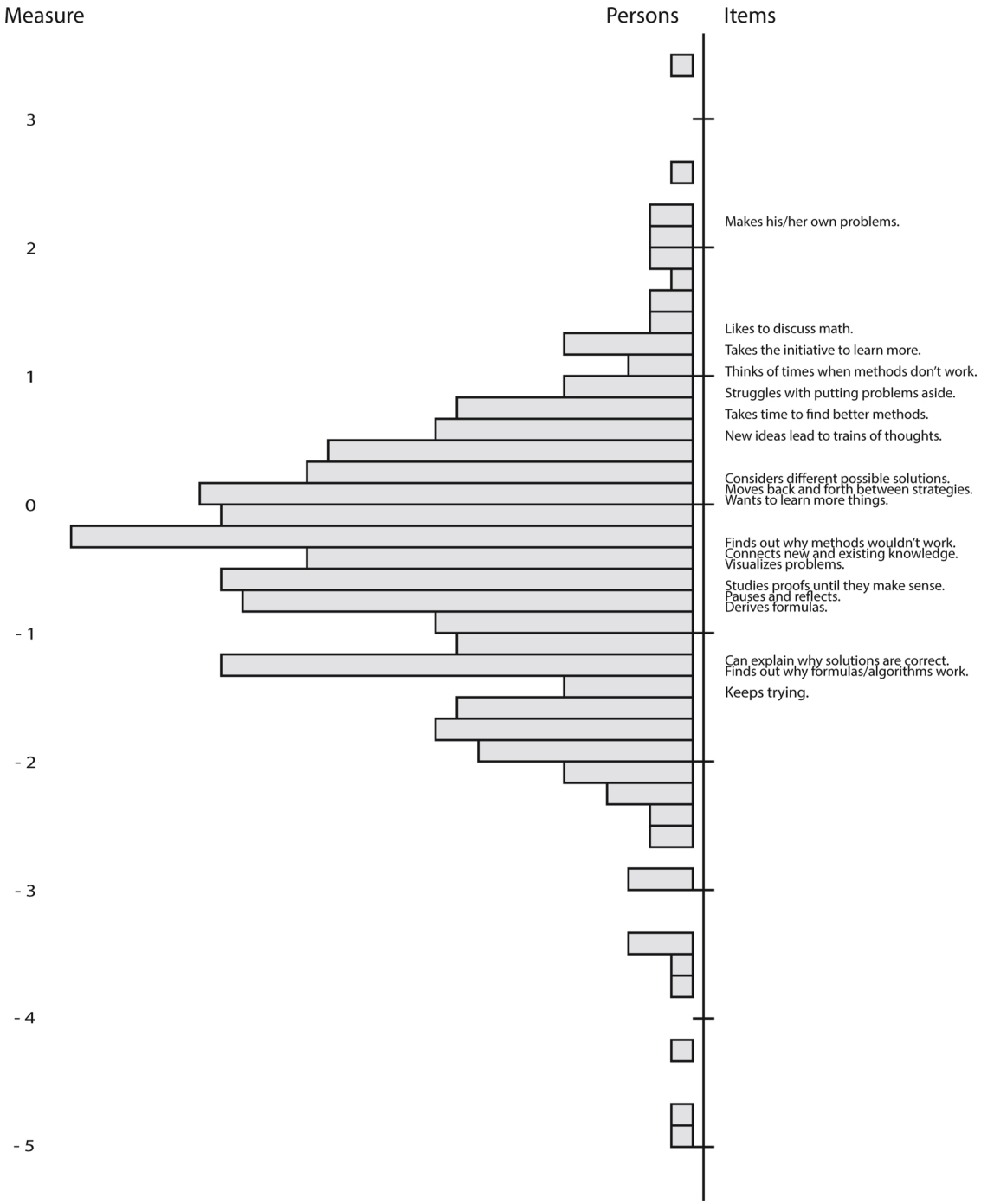

Fig. 2. Person-item map, LS context.
Note: Measures are in logit units.

invariant. In sum, we conclude that, within the observed LS context, the instrument is sufficiently invariant for meaningful measurement.

4.1.4. Unidimensionality and local dependency

In the PCA analysis of the standardized residuals, an unexplained variance of 1.7 (in eigenvalue units) was found in a second contrast; this is considered as sufficiently unidimensional for meaningful measurements (Linacre, 2015, p. 391). Also, the items were concluded to be relatively independent; the largest standardized residual correlation was between Item 13 ("considers different possible solutions") and Item 14 ("moves back and forth between strategies") ($r = .20$).

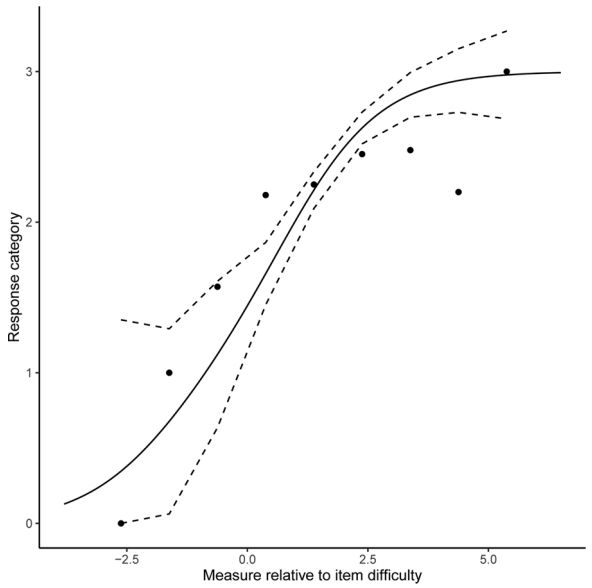

Fig. 3. Item Characteristic Curve, Item 9.
Note: Empirical responses (each dot represents the mean response for persons within a 1 logit-interval); expected responses (solid); 95 % confidence interval (dashed).

Table 2
Item statistics.

Measure	Infit Mnsq	Outfit Mnsq	Item
1.26	1.00	0.94	1. Takes the initiative to learn more.
0.70	1.18	1.14	2. Takes time to find better methods.
0.97	1.26	1.38	3. Thinks of times when methods don't work.
0.83	1.34	1.37	4. Struggles with putting problems aside.
−0.82	1.01	1.00	5. Derives formulas.
1.27	1.11	0.98	6. Likes to discuss math.
2.10	0.97	0.90	7. Makes his/her own problems.
0.56	0.85	0.84	8. New ideas lead to trains of thoughts.
−0.47	1.01	0.99	10. Finds out why methods wouldn't work.
−1.43	1.22	1.19	11. Finds out why formulas/algorithms work.
−0.71	0.91	0.92	12. Studies proofs until they make sense.
0.19	0.71	0.70	13. Considers different possible solutions.
0.07	0.77	0.82	14. Moves back and forth between strategies.
0.01	0.98	0.97	15. Wants to learn more things.
−0.75	1.03	1.05	16. Pauses and reflects.
−0.56	1.07	1.09	17. Visualizes problems.
−1.32	1.00	1.00	18. Can explain why solutions are correct.
−0.50	0.86	0.86	19. Connects new and existing knowledge.
−1.40	0.90	0.92	20. Keeps trying.

4.1.5. Rating scale analysis

The rating scale analysis showed that all response categories functioned appropriately: (1) each response category had more than 10 responses (423 responses in the fourth category being the least); (2) the shape of each rating scale was smooth and unimodal; (3) the average respondent measure associated with each category increased with the values of the categories (the average measures of persons who responded in categories 1, 2, 3, and 4 were −2.14, −0.74, 0.33, and 1.05 logits, respectively); (4) the category Outfit Mnsq was less than 2.0 (1.28 in the fourth category being the most); (5) Andrich thresholds advanced (−1.96, 0.25, 1.71); (6) Andrich thresholds advanced by at least 1.4 logits (the minimum advance being 1.96 logits between the second and third threshold);

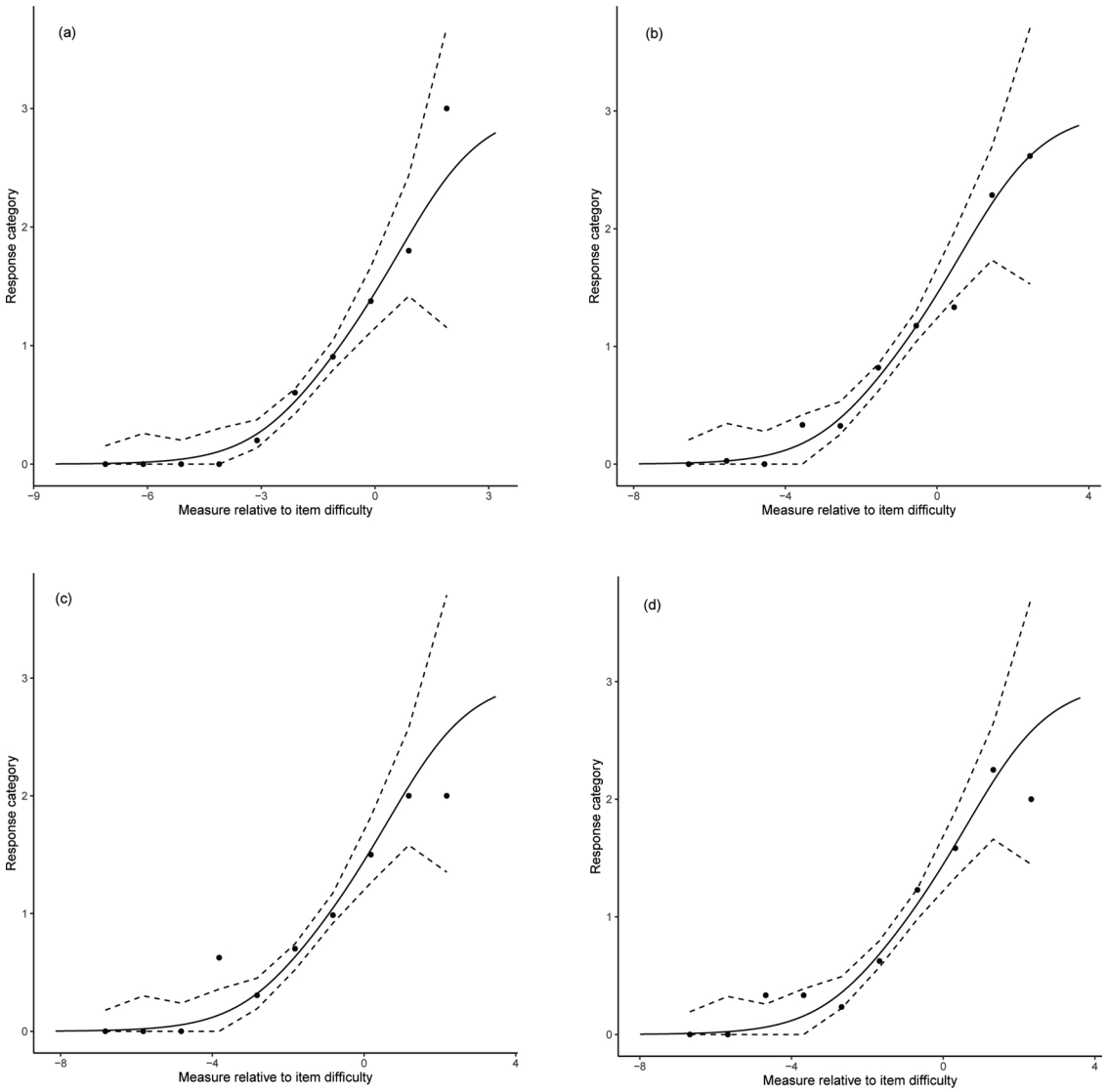

Fig. 4. A selection of four ICCs.
Note: (a) ICC of Item 1. (b) ICC of item 2. (c) ICC of item 3. (d) ICC of item 4.

and (7) Andrich thresholds advanced no more than 5.0 logits (the maximum advance being 2.21 logits between the first and the second threshold). In Fig. 5, we illustrate the category probability curves, which shows that each response category at some interval along the variable was the most probable.

In conclusion, except for one item, the psychometric quality of the instrument was sufficiently good for measuring MI in LS. Accordingly, if we exclude one item, we can use the same instrument for measuring MIs in STEM contexts and LS context. In the next section, we report on an analysis of invariance between the STEM context and the LS context.

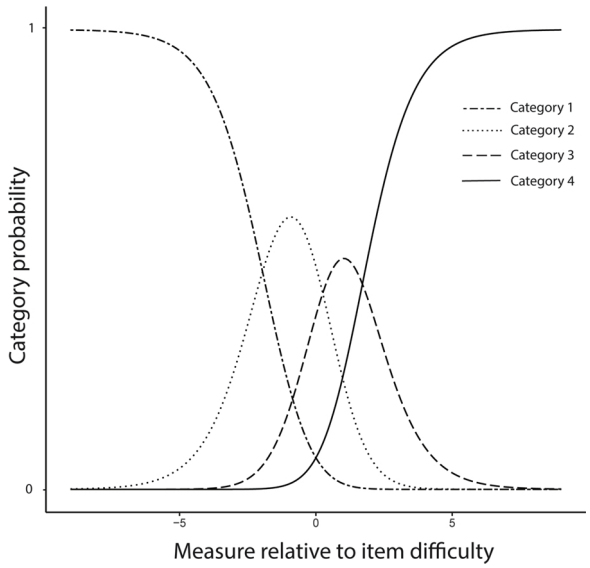

Fig. 5. Category probability curves.

4.2. Invariance of MI between STEM and LS

When Item 9 was excluded from the instrument, and data from STEM students and LS students were merged, the analysis showed significant DIF (DIF-contrast larger than 0.64 logits, $p < .05$) in six items (Items 2, 5, 6, 7, 11, and 19). The remaining 13 items were

Table 3
Items with significant DIF between LS and STEM.

Item	Sample	Measure	S.E.	Sample	Measure	S.E.	Contrast	*t*	*p*
2	LS	0.63	0.10	STEM	1.91	0.09	−1.28	−9.5	.000
5	LS	−0.61	0.11	STEM	0.53	0.08	−1.14	−8.5	.000
6	LS	1.16	0.11	STEM	0.34	0.08	0.82	6.3	.000
7	LS	1.93	0.12	STEM	0.22	0.07	1.71	11.9	.000
11	LS	−1.27	0.09	STEM	−0.08	0.07	−1.20	−10.1	.000
19	LS	−0.48	0.16	STEM	−1.76	0.08	1.28	7.25	.000

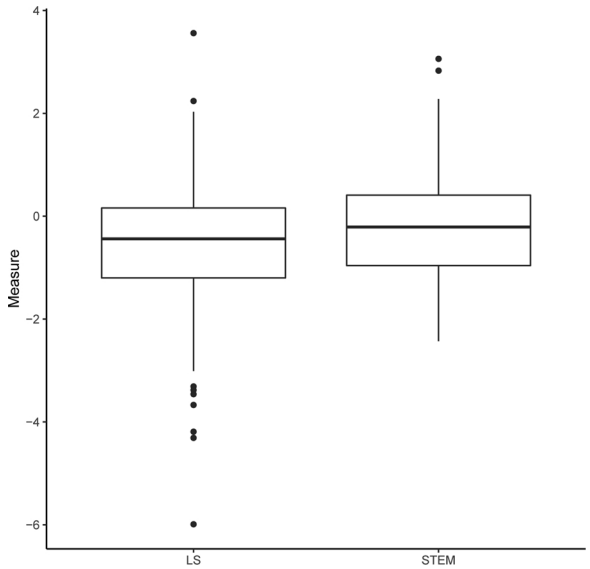

Fig. 6. MI measures of LS students and STEM students.

concluded to be relatively invariant (i.e., within the proposed limits). A list of the items with significant DIF is provided in Table 3.

Item 2 showed significant DIF-contrasts also within the LS context. Together, these results suggest that, in general, Item 2 is sensitive to contextual change. The rest of the items in Table 3, however, had significant DIF-contrasts only *between* the contexts. Although we do not have qualitative evidence to explain these results, one hypothesis is that the LS students and the STEM students make different interpretations of specific words. For instance, it is generally agreed upon that people who engage with mathematical activities interpret "understanding" differently (e.g., Skemp, 1987). Hence, it might be the case that STEM students, more often than LS students, interpret "finding out why algorithms work" to mean a conceptual, rather than a procedural, understanding of algorithms.

Another reason that could explain some of the DIF-contrasts is that some students, more than others, *alienate* themselves (or: are alienated) from mathematics (e.g., Radford, 2016; Solomon & Croft, 2016; Williams, 2016). Consequently, when students perceive the products of mathematical activities as belonging to others—e.g., when students perceive homework as something done *by* the students *for* the teachers—it might be relatively difficult for these students to come up with new tasks or to discuss mathematics in their spare time. By contrast, when students perceive products of mathematical activities as belonging to them—e.g., when students perceive homework as something done *by* the students *for* themselves or a community in which they are central agents—it might be relatively easy for these students to come up with new tasks or to discuss mathematics in their spare time.

To assess the practical significance of the observed DIF, LS students were measured twice: first, relative to the LS instrument, subsequently, relative to the STEM instrument (i.e., anchoring the item measures as they appeared when only STEM students were measured). The correlation between the students' measures in the two trials was $r = .98$.

From these results, we make the following conclusions. Between the studied contexts, there are some observable differences in six items. These items are, undoubtedly, compelling cases that one might pursue qualitatively to understand social differences in the two contexts better. Regarding the measurement of personal MIs, however, the structural similarities outweigh the structural differences. We conclude, therefore, that the instrument we have studied is practically invariant; even if STEM mathematical activities are different from LS mathematical activities, measured MIs of LS students can be compared with measured MIs of STEM students.

When personal MIs were compared between the contexts, the 372 STEM students ($M = -0.22$, $SD = .99$), compared to the 332 LS students ($M = -0.51$, $SD = 1.11$), demonstrated significantly stronger personal MIs, $t(669.8) = -3.72$, $p < .001$ (see Fig. 6). However, if we return to the social MI (e.g., Fig. 2), we observe that the substantial difference between the means is relatively small. (Roughly, a STEM student with average MI is more likely to respond positively to items in the range -0.22 to -0.51 logits than an LS student with an average MI). Also, within both contexts, personal MIs covered a wide range (from extremely low to extremely strong MI). The Levene's test did not indicate unequal variance between the LS and the STEM context ($F = .07$, $p = .79$).

5. Discussion

In this paper, we have provided evidence for reliability and validity of an instrument for measuring LS students' MIs, and we have shown that measures of MIs in LS can be compared with measures of MIs in STEM. Moreover, the results showed little difference between LS students' personal MIs and STEM students' personal MIs, regarding both mean and variance.

Although we in this paper have focused on MI, we do not assume that MI exists in isolation or that this identity has a special role relative to other identities. On the contrary, we accept the thesis that people have multiple identities and that learning involves a negotiation of such identities (Black et al., 2010). Accordingly, questions about how MI relates to alternative identities is an empirical question. Instruments for measuring different kinds of identities might help to answer this question.

A didactic implication of this study is that it provides examples of things that characterize students who identify strongly with mathematics. That is, if we look at the items with the highest measures in Fig. 2, we see that students who identify most strongly with mathematics often make mathematical problems for themselves, often like to discuss mathematics, and often take the initiative to learn more than what is required in school. In other words, it seems that students who identify most strongly with mathematics tend to perceive mathematics as more than a school subject.

The results we have presented in this paper are relevant also for researchers who study MI. For instance, the instrument for measuring MI could be used to assess how MI relates to variables proven to be productive in mathematics education such as conceptual understanding, procedural fluency, strategic competence, adaptive reasoning, productive disposition (Kilpatrick, Swafford, & Findell, 2001), the ability to ask and answer questions in and with mathematics, and the ability to deal with mathematical language and tools (Niss, 2004). Also, future research could use the instrument for assessing how various teaching and learning practices affect the development of MI.

Although the instrument might be relevant mostly for researchers in mathematics education, we believe that researchers in other fields who study identities quantitatively could adapt the combination of theory and method we have applied in this study. Specifically, we argue that all theories of measurable identities must be compatible with theories of measurement. From theories of measurement (e.g., Thurstone, 1959), we know that (1) for all measurements, there exists a body of reference and (2) within the measured context, the body of reference must be perfectly invariant (in theory) or approximately invariant (in practice). Hence, it follows (1) that every theory of measurable identities must include something that has the function of a body of reference and (2) that the body of reference must be defined in a way that fulfils the requirement of invariance. We maintain that Rasch measurement theory together with the definitions of personal and social identities we have presented in this paper are compatible with these requirements.

In this study, we examined the extent to which social mathematical identities are context-dependent between LS and STEM. A limitation of this study is that we do not know the reasons for the observed differences, although we have suggested some

hypothetical causes. Based on previous research (Radford, 2016; Solomon & Croft, 2016; Williams, 2016), we proposed that alienation might cause a change in the structure of MI. In future studies, the relationship between alienation and MI can be examined using the theoretical perspective and method we have presented in this paper.

In this study, we did not examine the extent to which *personal* MIs are context-dependent between LS and STEM. It is true that we compared personal MIs in LS with personal MIs in STEM, but the results from this analysis do not answer the fundamental question: To what extent do students change their MIs when they move from one context to the other? We believe that answers to this question require longitudinal data, for instance, multiple measures of persons' MIs as they transfer between contexts.

However, when data from different contexts are merged using multiple responses from the same sample, there might be a violation of the requirement of local independence (e.g., Marais & Andrich, 2008). Regarding items, a data set that includes subsets of relatively similar items (in the extreme case: a test where some questions are raised multiple times) violates the requirement of local independence. Symmetrically, a data set that includes dependent responses (in the extreme case: a test where responses from some persons are duplicated) violates the requirement of local independence.

One solution to this problem is to use different items when persons are measured multiple times, keeping a set of items for equating (Wright & Stone, 1979). However, this procedure would require a larger item bank than what we have presented, and therefore, we suggest that future studies expand the set of characteristics that can be used for measuring MI.

Since the Rasch model allows the comparisons of measures obtained from different instruments, insofar as there exist some common items that can be used for equating (Wright & Stone, 1979), it is possible to reconsider the content of the items we used in the LS context without compromising the ability to compare measures in LS with measures in other contexts (e.g., STEM). In our study, the instrument that was validated for STEM students was administered to LS students with minor adjustments. However, there might exit characteristics of MI that are more relevant for LS students than they are for STEM students. In future studies, the inclusion of such items would add content validity to the instrument we have presented in this paper.

We make two suggestions for researchers who choose to exchange some of the items in the instrument. First, we suggest that the researchers attempt to fill the gaps in the instrument (see Fig. 2). Specifically, items at the lower end of the variable that are also particularly relevant for LS students would add both content validity and reliability. Second, we suggest that researchers should be careful about which items they exclude from the instrument. If researchers remove too many of the items that in our study proved to be relatively invariant, it might be challenging to compare measures of LS students' MIs with MIs in other contexts.

In sum, we suggest that researchers can use instruments for measuring MI for answering two fundamental questions: (1) How is MI affected by learning, and (2) how is MI affected by contextual change? The empirical results in this paper answer neither of these questions, although there is some indication that MI is inconsiderably affected by learning. That is, individuals might change their MIs, either due to learning or contextual change, but the evidence we have provided is that the mean MI change only marginally as persons grow older.

Appendix A

An instrument for measuring mathematical identity, English translation.

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

Note: Item 9 was reversely coded.

References

Abdelal, R., Herrera, Y. M., Johnston, A. I., & McDermott, R. (Eds.). (2009). *Measuring identity: A guide for social scientists*Cambridge University Presshttps://doi.org/10.1017/CBO9780511810909.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–573. https://doi.org/10.1007/BF02293814.

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.* Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2013). *Rasch analysis in the human sciences.* Springerhttps://doi.org/10.1007/978-94-007-6857-4.

Cokley, K., Komarraju, M., King, A., Cunningham, D., & Muhammad, G. (2003). Ethnic differences in the measurement of academic self-concept in a sample of African American and European American college students. *Educational and Psychological Measurement, 63*(4), 707–722. https://doi.org/10.1177/0013164402251055.

Cross, W. E., Jr, & Vandiver, B. J. (2001). Nigrescence theory and measurement: Introducing the Cross Racial Identity Scale (CRIS). In J. G. Ponterotto, J. M. Casas, L. A. Suzuki, & C. M. Alexander (Eds.). *Handbook of multicultural counseling* (pp. 371–393). Sage Publications, Inc.

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

Deaux, K. (1993). Reconstructing social identity. *Personality and Social Psychology Bulletin, 19*(1), 4–12. https://doi.org/10.1177/0146167293191001.

Erikson, E. (1959). Identity and the life cycle. *Psychological Issues, 1*(1), 1–171.

Gee, J. P. (2000). Identity as an analytic lens for research in education. *Review of Research in Education, 25*, 99–125. https://doi.org/10.2307/1167322.

Graven, M., & Heyd-Metzuyanim, E. (2019). Mathematics identity research: The state of the art and future directions. *ZDM Mathematics Education, 51*, 361–377. https://doi.org/10.1007/s11858-019-01050-y.

Henley, N. M., Meng, K., O'Brien, D., McCarthy, W. J., & Sockloskie, R. J. (1998). Developing a scale to measure the diversity of feminist attitudes. *Psychology of Women Quarterly, 22*(3), 317–348. https://doi.org/10.1111/j.1471-6402.1998.tb00158.x.

Holland, D., Lachicotte, W., Skinner, D., & Cain, C. (2001). *Identity and agency in cultural worlds.* Harvard University Press.

Holland, J. L., Johnston, J. A., & Asama, N. F. (1993). The vocational identity scale: A diagnostic and treatment tool. *Journal of Career Assessment, 1*(1), 1–12. https://doi.org/10.1177/106907279300100102.

Huang, X., Lee, J. C. K., & Yang, X. (2019). What really counts? Investigating the effects of creative role identity and self-efficacy on teachers' attitudes towards the implementation of teaching for creativity. *Teaching and Teacher Education, 84*, 57–65. https://doi.org/10.1016/j.tate.2019.04.017.

Kaspersen, E. (2018). *On measuring and theorising mathematical identity.* Doctoral dissertations at the University of Agder.

Kaspersen, E. (2019). Expected values for category-to-measure and measure-to-category statistics: A simulation study. *Journal of Applied Measurement, 20*(2), 146–153.

Kaspersen, E., Pepin, B., & Sikko, S. A. (2017). Measuring STEM students' mathematical identities. *Educational Studies in Mathematics, 95*(2), 163–179. https://doi.org/10.1007/s10649-016-9742-3.

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

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

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

Marais, I., & Andrich, D. (2008). Formalising dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement, 9*(3), 1–16.

Marcia, J. E., Waterman, A. S., Matteson, D. R., Archer, S. L., & Orlofsky, J. L. (2012). *Ego identity: A handbook for psychosocial research.* Springer Science & Business Media.

Niss, M. (2004). The Danish "KOM" project and possible consequences for teacher education. In R. Strässer, G. Brandell, B. Grevholm, & O. Helenius (Eds.). *Educating for the future. Proceedings of an international symposium on mathematics teacher education* (pp. 179–190). The Royal Swedish Academy of Science.

Perron, J., Vondracek, F. W., Skorikov, V. B., Tremblay, C., & Corbiere, M. (1998). A longitudinal study of vocational maturity and ethnic identity development. *Journal of Vocational Behavior, 52*(3), 409–424. https://doi.org/10.1006/jvbe.1997.1638.

Phinney, J. S., & Chavira, V. (1992). Ethnic identity and self-esteem: An exploratory longitudinal study. *Journal of adolescence, 15*(3), 271–281. https://doi.org/10.1016/0140-1971(92)90030-9.

Radford, L. (2016). On alienation in the mathematics classroom. *International Journal of Educational Research, 79*, 258–266.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests: Expanded edition.* University of Chicago Press.

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. https://doi.org/10.3102/0013189X034004014.

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

Solomon, Y., & Croft, T. (2016). Understanding undergraduate disengagement from mathematics: Addressing alienation. *International Journal of Educational Research, 79*, 267–276.

Swim, J. K., Hyers, L. L., Cohen, L. L., Fitzgerald, D. C., & Bylsma, W. H. (2003). African American college students' experiences with everyday racism: Characteristics of and responses to these incidents. *Journal of Black Psychology, 29*(1), 38–67. https://doi.org/10.1177/0095798402239228.

Tan, A. L., Kendis, R. J., Porac, J., & Fine, J. T. (1977). A short measure of Eriksonian ego identity. *Journal of Personality Assessment, 41*(3), 279–284. https://doi.org/10.1207/s15327752jpa4103_9.

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

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge University Presshttps://doi.org/10.1017/CBO9780511803932.

Williams, J. (2016). Alienation in mathematics education: Critique and development of neo-Vygotskian perspectives. *Educational Studies in Mathematics, 92*(1), 59–73.

Wolfe, E., & Smith, E. (2007a). 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. (2007b). 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., & Masters, G. N. (1982). *Rating scale analysis.* MESA Press.

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

Ytterhaug, B. O. (2019). *Matematisk identitet i ungdomsskolen*Master's thesis. Norwegian University of Science and Technology.