Measuring student teachers’ practices and beliefs about teaching mathematics using the Rasch model

Several attempts have been made to measure and categorise beliefs and practices of mathematics teachers (e.g. Swan 2006). One of the reasons for measuring both beliefs and practices is to characterise teachers in terms of the relationship between their beliefs and practices. However, the practice-instruments and the beliefs-instruments discussed in the literature are predominately based on differing constructs. Consequently, it is challenging to compare the relationship between beliefs and practices solely based on these instruments. As such, we argue that practice- and beliefs-instruments based on the same construct would be desirable when both beliefs and practices are at stake. This paper presents two Rasch calibrated instruments that measure the level of teacher-centredness with respect to student teachers’ practices and beliefs about teaching mathematics. From a sample of 160 student teachers, 15 items have been established as being beneficial for measurement, both when the items are translated as practice items and when translated as beliefs items. By studying the invariance between both translations, we conclude that measures from both instruments can be directly compared.

**Keywords**: beliefs, practice, measurement, Rasch modelling.

Much has been said about student teachers’ transition from higher education to the world of work. Most studies on novice teachers’ experiences show that the transition from teacher education to work as a teacher is problematic. For instance, selected studies describe a gap between higher education and work, e.g. weak relationships between courses and field experiences (Feiman-Nemser 2001), and that what students learn at university is not appropriately linked to their future practice as teachers (e.g. Liston, Whitcomb, and Borko 2006). Other studies describe new teachers’ first period at work as a time where the workplace communities expect new employees to be able to teach like experienced teachers (Worthy 2005), a period with high emotional intensity (Flores and Day 2006) and one where there is a gap between new teachers’ actual identities and the ‘designated identities’ shaped by the workplace communities (Haggarty, Postlethwaite, Diment, and Ellins 2011). In many studies, however, the sample selections seem to be based on contextual rather than personal criteria, e.g. selecting teachers from certain schools, or teachers who have undertaken a particular course or education (e.g. Flores and Day 2006).

This study builds on previous studies of the TransMaths-project (www.transmaths.org) (e.g. Pampaka et al. 2012; Pepin, Lysø, and Sikko 2012) and seeks instruments that can facilitate the selection of student teachers with certain characteristics for researching the transition from teacher education to the world of work. The aim is to select persons with different relationships between beliefs and practices: persons who identify with their reform-minded practices, traditional practices, or persons with conflicting beliefs and practices (e.g. persons with ‘traditional’ practices and ‘reform-minded’ beliefs), etc.

Rasch measurements (Rasch 1960) have previously been used to measure practices in terms of teacher-centredness (Pampaka et al. 2012). The aim of this study is to validate instruments that measure both practices and beliefs, and to equate the instruments on the same scale, in order for measures to be directly comparable. Thus, the research question that guides this study is:

How can the Rasch model be used to measure student teachers’ practices and beliefs about the teaching of mathematics?

# Theoretical background

Different conceptualisations of teachers’ pedagogic practice with respect to learning environments are presented in the literature. Carpenter and Lehrer (1999), for instance, highlighted five mental activities that promote understanding in mathematics classrooms: 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). More recently, Schoenfeld (2014) discussed five dimensions of mathematically powerful classrooms: the mathematics (e.g. making connections between procedures, concepts, and contexts are addressed); cognitive demand (e.g. creating an environment of productive intellectual challenge); access to mathematical content (e.g. active involvement of all the students); agency, authority, and identity (e.g. making opportunities for students to conjecture, explain, make mathematical arguments, and build on one another’s ideas); and uses of assessment (e.g. building on productive beginnings or addressing emerging misunderstandings) (Schoenfeld 2014, 407). Without going into detail (for each dimension), it is clear that learning environments differ and that not all ‘ingredients’ are visible in each.

Turning to teachers’ pedagogic practice (in those environments), a common idea (and categorisation) in the literature is to regard teacher practice as a continuum, between teacher-centredness and student (learner)-centredness. Stephan (2014) defined learner-centred approaches as

...an approach to mathematics instruction that places heavy emphasis on the students taking responsibility for problem solving and inquiry. The teacher is viewed as a facilitator by posing problems and guiding students as they work with partners toward creating a solution (p. 338).

Stephan (2014) listed five characteristics of student-centred teaching: problem-solving; collaboration; mathematical discourse; tools/manipulations; and classroom environment (p. 340–342). With regard to classroom environment, Stephan (2014) emphasised four social norms, documented by Yackel and Cobb (1996), which will support student-centred teaching: students are expected to (1) explain/justify solutions; (2) attempt to make sense of others’ explanations; (3) indicate agreement/disagreement; and (4) ask clarifying questions (p. 340).

In the affect domain, research on beliefs about the nature of mathematics and the teaching and learning of mathematics has also received much attention over past decades (e.g. De Corte, Op 't Eynde, and Verschaffel 2002; Ernest 1991; Thompson 1992). In particular, the relationship between beliefs and practices has been a topic of considerable debate, and there seems to be a disagreement in the literature, in which ways beliefs might be related to practice. While some report on a high degree of consistency (e.g. Thompson 1984), others report on a disjunctive relationship between teachers’ practice and intentions of practice (e.g. Skott 2001).

In connection with these studies, several attempts have been made to measure or categorise the practices and beliefs of mathematics teachers. Swan (2006), for instance, described two research instruments: a beliefs questionnaire; and a practice questionnaire. The beliefs questionnaire further developed characterisations based on early research by Ernest (1991) and Askew et al. (1997). In short, the proposed categories/orientations are: the *transmission* orientation (viewing mathematics as a series of rules and truths); the *connectionist* orientation (viewing mathematics as a network of ideas); and the *discovery* orientation (viewing mathematics as a human creation) (Swan 2006, 59). The practice questionnaire, however, consisted of items categorised as being either teacher- or student-centred. Although there are obvious benefits to categorising teachers in terms of both their practices and beliefs, it should however be noted that the belief and practice items presented by Swan (2006) consist of different ‘units’: (transmission/connectionist/discovery)-orientations and (student/teacher)-centrism.

The different ‘units’ of Swan’s (2006) beliefs-instrument and practice-instrument do not constitute a problem per se. In some cases, however, it is desirable to compare beliefs and practices more directly (i.e. being able to say that one is more than, less than, or about the same as the other). One such example is when the relationship between beliefs and practices is being studied. In those cases research would benefit from practice- and beliefs-instruments that are based on the same ‘unit’, and also on basic requirements of measurement, e.g. those formulated by Thurstone (1959).

Indeed, anchored in Swan’s (2006) original data (*N*=120), Pampaka et al. (2012) used Rasch analysis to find that the practice items showed acceptable fit to a uni-dimensional scale, and a further revision resulted in a 28-item instrument that could measure teacher practice on a single scale; from extreme teacher-centred to extreme student-centred. However, Pampaka et al. (2012) did not discuss a similar beliefs-instrument.

To expand upon the developments of Swan’s (2006) and Pampaka et al.’s (2012) instruments, this study has set out to measure self-reported practices and beliefs about the teaching of mathematics, and we have chosen the construct of ‘teacher-centredness’ for two reasons. First, the methodological justification is that teacher-centredness can, intuitively, be thought of as uni-dimensional which is a basic requirement of measurement. Second, the theoretical justification is that connections between teacher/student-centredness and transmission/connectionist/discovery-orientations have already been established. Both Swan (2006, 62) and Pampaka et al. (2012, 485) explicitly referred to the transmission-orientation as being teacher-centred, and the constructivist-orientation as being student-centred. As such, even if we, for the sake of consistency, have chosen the construct of ‘teacher-centredness’ throughout this paper, the reader might prefer to think of it as ‘degree of transmissionism’.

In short, the study was influenced by the practice-instrument outlined by Pampaka et al. (2012) and used the same methodology to validate a similar beliefs-instrument based on items directly translated from the practice-instrument. Such adaption of practice to belief-items (or belief characteristics) is common in the literature. What can be said to be transmission/teacher-centred items or characteristics such as ‘Students should be expected to use only those methods that their text or teacher uses’ (Collier 1972, 157); ‘The teacher instructs solely from the textbook’ (Raymond 1997, 559); and ‘Teaching is (…) giving verbal explanations and checking that these have been understood through practice questions’ (Swan 2006, 60) are examples of items and characteristics found in the literature that are based on teacher practice, and these items are closely related to items used in our study.

Calibrating practice items and beliefs items on the same scale is a prerequisite for assessing the relationship between beliefs and practices, but does not say anything a priori about how they are related. How beliefs and practices are aligned is something to be discussed in this paper.

# Methodology

## Rasch analysis

Since the purpose was to construct instruments for measuring, as opposed to model, the beliefs and practices of mathematics student teachers, the study was situated within the Rasch paradigm (Rasch 1960). Rasch (1960) based his model on basic requirements of measurement such as unidimensionality, additivity, and invariance (Andrich 1989). There are several reasons why Rasch measurement was favoured over classical test theory (CTT). One reason was that CTT is person- and item-dependent. For example, including a set of difficult questions to a test would most likely increase the differences between persons. In a similar way, including a set of persons with high measures would most likely increase the difference between items. Measures obtained by Rasch modelling, in contrast, are not dependent upon who takes the test, and which items are chosen (e.g. Bond and Fox 2003).

Like physical measurements, Rasch measurement does not discriminate between persons and items, which means that they can be measured on the same scale. Moreover, when the data fit the model, the estimation of person measures does not depend on the item distribution and vice versa (Wright and Stone 1979). As such, the dichotomous Rasch model consists of two parameters only: person ‘ability’; and item ‘difficulty’ (where ‘ability’ and ‘difficulty’ should be interpreted in the broadest sense). In this study, ‘ability’ refers to the degree of teacher-centrism, and ‘difficulty’ refers to being hard to endorse. More specifically, people with high measures are highly teacher-centred, and items with high measures are the ones that only highly teacher-centred persons find ‘easy’ to agree with.

Two parameterisations of the Polytomous Rasch Model, the Partial Credit Model (PCM) (Masters 1982; Masters and Wright 1997) and the Andrich Rating Scale Model (RSM) (Andrich 1978), are typically used when items have more than two options, and the only difference between PCM and RSM is the number of parameters estimated. In the PCM, the structure is unique to all items, whilst the items (or groups of items) in the RSM share a rating scale structure. This study considered guidelines presented by Linacre (2000, 768), in order 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. Some of the items would have had less than 10 responses to some of the categories if the PCM had been chosen, which is considered to be problematic.
3. The correlation between person measures was 1.00 when the different approaches were tested, leaving little room for argument as to whether to reject the RSM.

The RSM is a generalisation of the dichotomous Rasch Model. With *m* thresholds, it takes the form:

Where:

 = 0

is the measure of person ,

is the measure of item

 is the outcome space,

is the probability of person responding in category *x* on item

and

is the normalising 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, , is reduced to a restricted outcome space, that follows the Guttman structure; i.e. the possibility of passing the *k-*th threshold when the (*k*-1)-th threshold is not passed, is non-existing. 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).

Intuitively, the model represents the probability of a person with a certain measure, responding in a certain category, on a certain item. The likelihood of a person responding in one of the higher categories increases as the person measure increases for any given item. Conversely, the likelihood of a person responding in one of the lower categories increases as the person measure decreases. Moreover, the model is a logistic model, and the denominator, , ensures that the likelihood for any case is restricted between 0 and 1.

## Data collection

In our study items of the instruments were influenced by the practice-instrument of Pampaka et al. (2012), which in turn was influenced by Swan’s (2006) items. These items were then translated into beliefs items. For instance, the practice item ‘Students work with tasks with a clear answer’ was translated into a beliefs item ‘Students should work with tasks with a clear answer’. The final items, which have been established to be productive for measurement, are presented in Appendix A, and Appendix B.

The items and the response categories were subsequently discussed with five teacher educators. From this discussion, 27 practice items and 27 beliefs items were piloted with 42 Norwegian student teachers in their second and third years of teacher education, and, also, with nine teacher educators at the same institution. The items were also briefly discussed with all participants. The items were subsequently revised, and again piloted with 36 student teachers in their second year of education.

After a final revision, 32 items were assigned to a convenience sample of 83 student teachers in their fourth (and for many, final) year of education. In the final analysis, all responses were used, although the pilot group did not respond to two of the items. In this case, responses were coded as ‘missing data’, which is considered to be non-problematic in the Rasch paradigm, except for some loss of statistical information of course. Indeed, it is a specific requirement of meaningful measurements that ‘missing data must not matter’ (Wright 1997). As the original practice-instrument identified some problems with the use of five response categories, four response categories were chosen in our study (‘in none/almost none of the lessons’; ‘in some of the lessons’; ‘in most of the lessons’; and ‘in all/almost all of the lessons’).

The items were administered in random order, but all participants answered the practice-instrument before answering the beliefs-instrument. Moreover, it was made clear that these were two separate instruments. However, a legitimate concern is that a successful validation of the beliefs-instrument might be due to the close resemblance between the beliefs-items and the practice-items. As such, two actions were taken to assess this concern. First, correlations between standard residuals across instruments were tested. Standard residuals were favoured over raw scores since any standardised residual, in a theoretically perfect instrument, is considered to be random noise with an expected value of 0 (and SD = 1), regardless of person- and item-measure. As such, a high correlation between residuals of, say, practice item 7 and beliefs item 7, means that when persons respond, for example, higher than expected to practice-item 7 they also respond higher than expected to beliefs-item 7. A weak correlation between residuals means that, for example, unexpectedly high responses to the practice-item do not affect how they respond to the beliefs-item. In the Rasch paradigm, only items with the correlation *r* > .7 are considered to be problematic when a local dependency is tested within an instrument (Linacre 2015, 401). However, since response dependency had been tested between instruments, a more conservative cut-value of .3 was chosen. Consequently, any pair of items with *r* > .3 were considered to be potentially ‘problematic’ items. Moreover, the impact of the response dependency was tested by comparing beliefs measures with potentially ‘problematic’ items included in the test versus beliefs measures with those items omitted.

Second, from examining the correlation between students’ beliefs and practices (not residuals this time), evidence for external validity was sought to clarify whether responses to the beliefs-items seemed to be significantly affected by responses to the practice-items. That is, a review on existing literature show that the majority of studies report on a weak relationship between beliefs and practices (e.g. Skott 2001; Liljedahl 2009), even if we did not find any studies that used equated instruments based on requirements of measurement. If students’ responses to the beliefs-instrument were significantly affected by their responses to the practice-instrument one would, however, expect practices and beliefs to be highly correlated, i.e., inconsistent with these studies. Hence, we assert that a weak correlation between students’ beliefs and practices would be consistent with existing literature and strengthen the external validity, as opposed to a high correlation that would be inconsistent with a majority of existing literature and weaken, not only the external validity, but also the internal validity since this result would suggest possibly highly dependent responses.

In addition to the quantitative data, nine student teachers in their final year of education were interviewed about their practice. Before the interviews, the interviewees shared documents relevant to their practice of teaching mathematics that they had stored on their computers. These documents mostly consisted of essays, tasks, and planning documents. The interviews were audio-recorded and lasted for 60 – 90 minutes. During the interviews the students were asked about their experiences with mathematics, their time at University College, and their experiences and practice as mathematics teachers. Moreover, the students were asked to comment upon two planning documents, or papers that described teaching sequences related to mathematics. Finally, the students were asked to complete the practice- and beliefs-questionnaires. In this paper, excerpts from the interviewed cases are presented for *interpretability* validity, to provide the reader with some examples of the kinds of statements one can expect along the variable.

## Validity

To find evidence of validity, we used guidelines presented by Wolfe and Smith Jr. (2006), which extend Messick’s (1995) validation framework by two aspects of evidence suggested by The Medical Outcomes Trust (MOT). To summarise, Messick (1995) viewed validity as a unified concept, in the sense that he claims that there do not exist different kinds of validity, but rather different kinds of evidence that support validity. Accordingly, Messick (1995) presents six different aspects of validity where evidence can be found: the *content* (e.g. relevance, representativeness, and technical quality)*; substantive* (e.g. theoretical foundation)*; structural* (e.g. evidence of uni-dimensionality)*; generalisability* (e.g. generalisation across sample and context)*; external* (e.g. convergent and discriminant evidence)*;* and *consequential* (e.g. the extent to which society benefits from using the test)aspects. Furthermore, the MOT presents two aspects not proposed by Messick: *responsiveness* (e.g. the capacity to detect change); and *interpretability* (e.g. how quantitative measures are related to qualitative meaning).

## Simulation studies

As a tool for interpreting some of the results in the analysis, 100 simulation studies were conducted for both instruments, which comprised of the empirical measures used to generate Rasch-fitting data.

## Analysis

In the analysis, the WINSTEPS (Linacre 2014) software was used. To find evidence for *content* validity, infit and outfit mean-squared and ZSTD (*t*-tests that test the hypothesis of whether the data fits the model) were used to assess data to model fit. Outfit is based on the mean of squared standardised residuals, whilst infit is an information-weighted sum (Bond and Fox 2003, 238). Although different critical values for infit and outfit mean-squared can be found in the literature, e.g. 0.5 to 1.5 (Linacre 2002b) or above 1.2 - 1.3 (Smith, Schumacker, and Busch 1998, 67), we have chosen to use the more conservative cut-values, since the role of these statistics was merely to ‘flag’ potential problematic items for closer inspection (Bohlig et al. 1998, 607). Thus, critical values for infit and outfit mean-squared were calculated, from the formulas suggested in Smith et al. (1998, 78), to be 1.1 (infit mean-squared) and 1.4 (outfit mean-squared). The simulation studies confirmed that these values seemed reasonable (e.g. the mean of the largest outfit in the beliefs-simulations was 1.25 with SD= .16). Values for |ZSTD| were set to 2.0. Positive values indicate lack of predictability, whilst negative values indicate ‘too predictable’ items (Linacre 2002b). Moreover, the item-measure correlation was assessed with .40 as the suggested value for flagging. No items, however, were removed based upon these criteria alone.

To find evidence for *substantive* validity, a rating scale analysis was conducted. Linacre (2002a) presented eight aspects whereby evidence for well-functioning rating scales can be found: (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 outfit mean-squared fit statistics should be less than 2.0; (5) step calibrations should advance; (6) Step difficulties should advance by at least 1.4 logits; (7) step difficulties should advance by less than 5.0 logits; and (8) ratings should imply measures (C → M), and measures should imply ratings (M → C). Additionally, Wolfe and Smith Jr. (2006, 210), referring to Zhu (2002), stated that researchers should also consider several post-hoc categorisations, and to use person reliability and fit of the data to the model when deciding upon which categorisation optimises the rating scale structure.

Another source of *substantive* validity is the analysis of person-fit (Wolfe and Smith Jr. 2006, 211). As such, persons with infit mean-squared larger than 2.0 were flagged and examined more closely. Specifically, qualitative meanings of the misfit were sought, in addition to the consequences of removing obviously ‘misfitting’ persons from the data (e.g. in terms of changes in reliability).

To find evidence for the *generalisability* aspect of validity, *the loss of invariance of item estimates across testing occasions* (Bond and Fox 2003, 309) was conducted through Differential Item Functioning (DIF): the comparison between two items should be independent of the persons who are used for calibration (Rasch 1961, 331–332). Hence, DIF analysis was conducted to test the hypothesis that item measures remained invariant between males/females and between pilot responses/final responses. This study used the Rasch-Welch method, which meant that ‘everything’ (i.e. person and item measures) but the item under investigation was held constant. As proposed by Linacre (2015, 542), a threshold of .64 for DIF contrast and .05 for *p*-value was chosen to flag potentially problematic items. However, when multiple hypothesis tests are conducted (one for each item), it is entirely probable that at least one of these will, by chance, show significant DIF. Indeed, amongst the simulation studies (that were simulated to show no DIF), 56% had at least one item with DIF. Therefore, it was decided to use a Bonferroni correction, which accounts for multiple hypothesis-tests (Bland and Altman 1995), leaving a corrected threshold for *p*-value = .0033. Finally, to detect non-uniform DIF, expected and empirical item characteristic curves (ICC) were inspected.

To find evidence of *structural* validity, the dimensionality of the items was assessed by Principal Component Analysis (PCA). Since classical factor analysis on raw scores tends to extract too many dimensions (e.g. items with similar measures tend to be reported as separate factors) (Bond 1994), PCA on standardised residuals was favoured. Based on suggestions presented by Linacre (2015), a threshold of 2.0 in Eigenvalue units was chosen for potential substantial dimensions. Furthermore, qualitative meanings were sought to make sense of potential additional dimensions. In addition, person measures based on items from different sub-dimensions were correlated to examine the impact on those dimensions. The degree of local dependency between items was also assessed, to see if responses to some items affected the responses to others.

## Equating

One of the benefits of measuring practice in terms of teacher-centredness is that one can say that one teacher’s practice is more, or less, teacher-centred than another teacher’s (and also by how much). Subsequently, one of the benefits of measuring both practice and beliefs on the same ‘unit’ is the following: in addition to comparing one person with another, persons can also be compared ‘with themselves’. That is, measurements can be used to decide whether people’s practices are more, or less, teacher-centred than their (measured) beliefs. However, to make such conclusions based on the measures alone is not straightforward. To illustrate, imagine that a teacher’s practice is measured to be, say 1.7, and her beliefs measured to be 2.1. Appreciating that these measures come from separately calibrated instruments, how can one decide whether the qualitative meaning of 2.1 on the beliefs-instrument is actually more, less, or about the same as 1.7 on the practice-instrument?

Since each practice item (e.g. ‘I teach each subject separately’) was related to a pseudo-common item in the beliefs-instrument (‘the teacher should teach each subject separately’), we propose that the instruments can be linked through a stack-analysis (Wright 2003). In effect, each pseudo-common pair of items was treated as identical items, and persons’ belief responses and practice responses were treated as distinct persons, in a pooled analysis. Moreover, DIF analysis between ‘beliefs’-responses and ‘practice’-responses was conducted. Finally, person measures based on the pooled calibrated items versus the original calibrations were correlated, to see if the equating process affected persons’ measures.

In the analysis we will demonstrate how the instruments can be equated, although we emphasise that different samples might yield different results. For instance, in some cases using textbooks might be harder to endorse as a practice item than a beliefs item, if the textbooks provided in the schools are of poor quality. Because of this, we suggest that the instruments should be recalibrated if they are to be used in other contexts.

# RESULTS

Based on the person/item fit statistics and person/item separation, obvious ‘misfitting persons’ (5) and items (6) were removed from the instrument (see Linacre 2010). Since some respondents only responded to one of the instruments, and the teacher educators only responded to the beliefs-instrument, the data-sets consisted of 150 responses to the practice-instrument, and 160 responses to the beliefs-instrument. 148 respondents completed both instruments. The final results will be discussed in terms of *content, substantive, structural, generalisability, interpretability* and *responsiveness* validity.

## Uni-dimensionality and response dependency

Uni-dimensionality is a basic requirement of measurement (Andrich 1989). That is, PCA of the standardised residuals after the part that is explained by the measure should indicate no other dimensions. However, Yamaguchi (1997) indicated that items with negative wording might cause a violation to this requirement. In effect, ‘not being A’ is not always the same as ‘being the opposite of A’. In this study, PCA clearly indicated that the items comprised at least two dimensions, as 3.58 unexplained variance in Eigenvalue units was found in a second contrast. Moreover, the standardised residual plot for the second contrast (Figure 1) shows how positively and negatively worded items were stratified. The negatively-worded items were the ones that were negatively coded; that is the ‘student-centred’ items. The positively worded items were the ‘teacher-centred’ items. Further analysis proved that the dis-attenuated correlation between the ‘student-centred’ items and ‘teacher-centred’ items was .70, that is too low to be ignored. Thus, we concluded that ‘not being teacher-centred’ does not imply ‘being student-centred’. Consequently, all the student-centred items were removed from the analysis, leaving 15 items.

[Figure 1 about here]

After the student-centred items had been removed from the instrument, unexplained variances that could be explained in a second contrast were reduced to 1.9 (practice-instrument) and 1.9 (beliefs-instrument) in Eigenvalue units. Thus, the second dimensions were treated as strands (like addition and subtraction on a mathematics test), and not as dimensions that needed separate instruments.

Moreover, the largest standardised residual correlation between items within each instrument was *r* = .24 (between practice item 12 and 9). Although there seemed to be a qualitative similarity between these items (as they both expressed students working on tasks), the relatively small residual correlation did not lead to further action. Standardised residual correlations between items across instruments showed that item 12 had a correlation *r* = .66 between the practice- and beliefs-item. In addition, 7 items (items 2, 3, 5, 6, 7, 9, and 10) showed modest correlations in the range of .30 and .44. However, the dis-attenuated correlation between persons’ beliefs measures, with these items included versus beliefs measures with these items omitted, was *r* = 1.00, i.e. none of the differences between persons’ measures were found to be significant.

As consistent with the majority of existing literature, the relationship between students’ beliefs and measures was weak (*r2* = .35). One explicit example can be found in one of the students who reported that his practice (6.1 units) was more teacher-centred than his beliefs (3.5 units). These measures were consistent with statements in his interview, such as ‘That the students should be challenged. That they should understand. So, when there are so many different [students]. So, you need to teach differently’ (coded as a belief-statement) and ‘Like, teaching on the blackboard, work with exercises, maybe rehearse, and ask if anything is difficult. And then more exercises’ (coded as a practice-statement). Since the correlation is weak, we conclude that the successful validation of the beliefs-instrument was not caused by the close resemblance to the practice-instrument.

## Summary statistics

Person reliability values, analogous to Cronbach’s alpha, were .85 (practice) and .81 (beliefs), and item reliability was .99 on both tests. Since the simulated data also showed a highly significant difference between person reliability on the two instruments (mean = .89, SD = .01 [practice] against mean = .85, SD = .01 [beliefs]), it seemed that this difference, on the whole, could be explained by the person and item distributions. Thus, although the person reliability was acceptable for both instruments, further improvements should consider more items and samples with greater variance, as person reliability is heavily affected by the test length and person distribution.

## Item fit

To find evidence for the *content* aspect of validity, the technical quality of the items was evaluated. In Table 1, the practice-instrument and beliefs-instrument are presented with items in measure order. In both instruments, all the point-measure correlations were positive, with beliefs-item 12 (.38) being the least. Furthermore, both infit and outfit statistics were in the range of critical values for most of the items. In the practice-instrument, items 12 and 15 were the most ‘underfitting’ with infit/outfit mnsq 1.18/1.21 and 1.20/1.05 respectively. In the beliefs-instrument, item 6 was the most ‘underfitting’ with infit/outfit mnsq 1.19/1.23. Removing any of these items did not, however, have any significant effect on person measures, only a negative effect on person reliability. Moreover, the misfit in all items could be explained by selected anomalous responses. Removing the two most unexpected responses to the practice item 15, for instance, reduced the infit/outfit to 1.07/ .98. Thus, it was decided to keep these items in the instrument.

[Table 1 about here]

Another key source of information about item fit is the ICC (e.g. Schumacker 2015). An inspection of the misfitting items indicated that practice item 12 was the most problematic, although practice item 15 had the highest infit mean-squared values. From the ICC of practice item 12, one can clearly see how the item has been systematically under-discriminating, i.e. the probability of agreeing with this item did not increase as much as expected when person measures increased. In contrast, the ICC of item 15 shows how the empirical ICC is close to the model ICC.

[Figure 2a about here]

[Figure 2b about here]

## Differential item functioning

When DIF analysis was conducted to test the hypothesis that item measures stayed invariant between genders, beliefs item 15 showed some DIF (.88 logits harder to endorse to for males than females). However, the DIF was concluded to be non-significant. Moreover, the correlation between persons’ measures was *r* = 1.00 when item 15 was fixed to be –.78 (female estimation) and .10 (male estimation). Thus, it was decided not to take further action on this item.

The DIF analysis between the pilot respondents (second and third-year students) and the final respondents (fourth-year students) showed that beliefs item 12 had significant DIF. Furthermore, the ICC indicated that the DIF was uniform, i.e. item 12 was relatively easier to be endorsed by fourth-year students for all ‘ability’ levels (e.g. Figure 3). Although removing item 12 had only a negligible effect on the persons’ measures, this item seems to be an area for future improvement.

[Figure 3 about here]

## Targeting

To ascertain *responsiveness* validity, it is important that the items in the instrument not only fit well with the sample abilities, but also potential abilities in the future (Wolfe and Smith Jr. 2006, 222). As such, the person-item maps in Figure 4 were good indicators for areas of potential improvement. In particular for the beliefs-instrument, there seemed to be a ‘missing tail’ in the person distribution. Thus, including persons with measures above 4 logits would likely improve the estimates of the items that are hardest to endorse (e.g. item 1 and item 2), and increase person reliability.

Moreover, the person-item map on the practice-instrument suggested that more items at the extreme teacher-centred end, in particular between 3 and 4 logits and above 5 logits, would improve the estimates of extremely teacher-centred persons, and increase person reliability.

[Figure 4 about here]

## Rating scale analysis

In both instruments, most of Linacre’s criteria were satisfied: (1) each rating scale category contained more than 10 observations (at least 151 observations in the first category in the practice-instrument); (2) the shape of each rating scale distribution was smooth and unimodal; (3) the average respondent measure associated with each category increased with the values of the categories; (4) the ‘unweighted’ mean-squared fit statistics were all less than 2.0 (1.21 at the most); (5) step calibrations advanced; (6) step difficulties advanced with more than 1.4 logits; and (7) step difficulties advanced by less than 5 logits.

One violation, however, was found in both instruments: the category to measure coherence criteria. In the practice-instrument, 32% of the responses in the first category belonged to persons who were ‘expected’ (i.e. 68% of the persons that responded in the first category were expected to respond elsewhere). Linacre (2002a) suggested at least 40%. However, the low C→M could be explained by the observed count distribution. In effect, if we had grouped together all the persons whom we had expected to respond to a category *n*, we should have also expected that a fraction of these persons would make unexpected responses, and then lowering the C→M in the adjacent categories. Since the observed counts in the second category were more than three times the counts in the first category, it seemed reasonable to assume that a fraction of unexpected responses from category two to category one has had more effect on the category one C→M than the same fraction of unexpected responses from category one to category two would have. This conclusion was also supported by the simulation studies where the mean C→M in the first category was 37.2% (SD = 4.8).

A similar result was found in the beliefs-instrument, where the C→M in the fourth category was 33%. Again, the low value could be explained by the count distribution, as there were three times as many responses to category three as to category four, and the beliefs-simulations showed low C→M in the same category (mean = 38.3%, SD = 3.4).

As Wolfe and Smith Jr. (2006, 210) suggested, several post-hoc categorisations were tested, and a summary is provided in Table 2. In addition to the original categorisation (1234), the results from collapsing the first two categories (1123), collapsing the second and third category (1223), and collapsing the last two categories (1233) were compared. In effect, the lack of a category to measure coherence in the original categorisation could be resolved by collapsing the first and the second category in the practice-instrument, and collapsing the third and fourth category in the beliefs-instrument. In both cases, the quality of the instruments (e.g. person reliability) remained largely identical. Moreover, the correlations between measures on the original categorisations and the collapsed categorisations were *r* = .98 (practice-instrument) and *r* = .96 (beliefs-instrument) respectively, providing few arguments for collapsing response categories.

To summarise, the rating scales in both instruments were functioning productively. Although a lack of category to measure coherence could have been resolved if categories were collapsed, this would not have improved the quality of the instruments. Nor would it have had any practical consequences on persons’ measures. Thus, it was decided to leave the lack of C→M coherence as a suggested area for improvement. In particular, we suggest that future studies should consider using a sample with greater variance, as our study indicates that the sample should include more persons with high values on the beliefs-instrument, and more persons with low values on the practice-instrument.

[Table 2a about here]

[Table 2b about here]

## Equating the instruments

In the equating process, a pooled analysis was conducted, treating responses to each pair of pseudo-common item as two different responses to the same item. Subsequently, DIF analysis was used to assess if ‘beliefs-measures’ to items were significantly different from ‘practice-measures’ to pseudo-common items. Measures were favoured instead of raw scores in this analysis, since analysis of differences between raw scores would only be of use if we assumed that persons’ practice distribution was the same as persons’ beliefs distribution. Specifically, since the difference between persons’ practice (mean = 2.85, SD = 1.36) and beliefs (mean = 1.37, SD = 1.24) rejected this assumption (see Figure 4), the expected raw score on each beliefs-item was lower than the expected raw score for the pseudo-common practice-item when data fitted the model. Finally, person measures from the pooled analysis were correlated with measures from the separate practice and beliefs analyses to reveal if the DIFs have had any practical consequences.

In general, the fit statistics of the items in the pooled analysis showed acceptable fit with most items. Pooled item 15 seemed to be the most ‘under-fitting’, and the practical consequences of removing it were negligible. Moreover, the only item with DIF between ‘beliefs-responses’ and ‘practice-responses’ was item 6 (.86 logits easier as a ‘beliefs item’ than a ‘practice item’). However, anchoring this item to 1.34 (beliefs-calibrated) or 2.20 (practice-calibrated) did not have any practical consequences on person measures. As such, it was decided to keep the item. Bearing in mind that item 6 also had some misfit in the beliefs-instrument, this item might, however, be an area for improvement in future developments.

Finally, the correlation between persons’ measures was close to 1 when the pooled calibration was used versus the separate analyses. Figure 5 shows the correlation (*r* = 1.00) between the pooled- and the beliefs-calibrated instruments. Hence, it would have had minor effects on persons’ measures if, say item 9, had been set to 1.68 (beliefs-calibrated), or 1.84 (pooled-calibrated). The same conclusion was drawn when practice-calibrated items were correlated with pooled-calibrated items. Thus, if beliefs and practices are to be compared, we propose the use of pooled calibrated measures, since equal response patterns on both instruments would lead to equal measures. However, we recognise that arguments can be made to avoid pooled calibrated measures (e.g. strictly speaking, a beliefs statement is not the same as a practice statement). In any case, proving little or minor DIF between the practice version and beliefs version should leave little practical consequences of either choice. To illustrate, the mean absolute difference between persons’ beliefs measures using pooled-calibrated measures or beliefs-calibrated measures was .07. As a consequence, we propose that when persons show similar response patterns on both the practice and beliefs-instruments, they will get either equal (in the pooled case) or really close (in the separate case) beliefs and practice measures.

[Table 3 about here]

[Figure 5 about here]

The *interpretability* aspect of validity is the degree to which qualitative meaning can be assigned to quantitative measures (Wolfe and Smith Jr. 2006, 227). Excerpts from interviewed cases are presented in Figure 6. Transcripts of the interviews were coded, based on the items in the instrument. The purpose of presenting these excerpts is merely to provide the reader with some statements that one can expect from interviewing persons with different measures along the variable. Moreover, nine teacher educators from the same institution were asked to respond to the beliefs-test. Other than providing construct validity (as teacher educators, so we expected, were expressing significantly less teacher-centred beliefs than the students), the instrument was rescaled, so that the mean of the teacher educators’ measures was set to zero. The rescaling was used for qualitative interpretation: values close to zero could be regarded as in agreement with the values of the educational institution.

[Figure 6 about here]

# CONCLUSION

The two instruments presented in this article were useful and productive for measuring the level of teacher-centredness in student teachers’ practices and beliefs with respect to teaching mathematics. Moreover, in the Norwegian context most items stayed invariant when they were translated from ‘practice’ to ‘beliefs’, which made it possible to assess the relationship between beliefs and practices directly (e.g. higher measures on the practice-instrument than on the beliefs-instrument lead to the conclusion that ‘practice was more teacher-centred than beliefs’). In other contexts, however, the items might not be invariant. For instance, implicit or explicit rules are likely to affect the measure of items. Thus, we propose that DIF analysis should be conducted in each new context before the instruments are used together.

Further improvements to the instrument should be considered, in particular the following: (i) a greater variance in the person distribution; (ii) item 12, as it (in addition to DIF) showed under-fit both in fit statistics and the ICC; and (iii) more items, particularly at the extreme teacher-centred end of the variable. Moreover, since some response dependency has been detected between instruments, research that uses both instruments should consider examining person statistics for indicators of response dependency, in addition to considerations whether it is reasonable to remove cases that are highly biased in regard to response dependency from the data set.

Many instruments exist that categorise or measure teachers’ practices and beliefs, some of which have been discussed in this paper. However, the main argument in this paper has been that measurable differences between practice and beliefs (i.e. to say that one is more/less than the other) can only be tested if both are measured on the same scale. Our conclusion is that, although practice and beliefs are different constructs (like length/width or length/weight), they can be measured on the same scale (like length/width, but unlike length/weight); and the unit that has been proposed is teacher-centredness.

Obviously, the instruments that are presented in this paper can be used separately. For instance, the practice-instrument can be used alone to assess how certain educational-programmes affect the teaching of mathematics, or as a screening tool for selecting particular teachers from a large sample (e.g. teachers with ‘extreme measures’ in comparative studies; or teachers who seem to significantly change over a particular time period). However, since both instruments were based on the same construct of teacher-centredness, we argue that the instruments are also suited for studies that include both practice and beliefs.

Although we have argued that the instruments are tools for sample selection (e.g. for research on student teachers’ transition from higher education to the world of work), the instruments could also be used for other purposes. In particular the relationship between beliefs and practice has long been the subject of considerable debate (e.g. Thompson 1992). Hence, we claim that instruments that measure practice and belief on the same scale contribute to this discussion.

Ideally, the instruments should be recalibrated on each new context. We recognise, however, the practical shortcomings of this. In particular, if the research was mainly qualitative, researchers might hesitate to go through the process of recalibrating the instruments. Nonetheless, research could benefit from measuring beliefs and/or practices. For instance, if researchers wanted to know how some ordinary teachers changed to become less teacher-centred, the instruments presented in this paper could be used as screening tools to search for those teachers that seemed to change the most from a larger sample. Thus, we present a scoring key (from the pooled analysis) that can be used on both instruments in Appendix C.

## Limitations of the study

Although we assert that our and similar instruments can be helpful tools, we acknowledge their limitations. In our case, we have reduced the practice and beliefs of teaching mathematics to one dimension. The reduction was performed due to the statistical benefits, but we emphasise that teaching is clearly multidimensional. Thus, persons with similar measures might, and are likely to have, different practices/beliefs, even when they are measured with reliable instruments. It can only be stated that they probably have some characteristics in common.

Moreover, ontological issues, such as the nature of beliefs, have not been discussed in this paper. The ‘beliefs’ that were measured in this study were self-reported beliefs, just as ‘practices’ were self-reported practices.

# 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 *Proceedings of the XXIVth International Congress of Psychology 4. Mathematical and Theoretical Systems*, edited by J.A. Keats, R. Taft, R.A. Heath, and S.H. Lovibond, 7–16. B.V. North Holland: Elsevier Science Publications.

Andrich, D. 2010. “Understanding the Response Structure and Process in the Polytomous Rasch model.” In *Handbook of Polytomous Item Response Theory Models,* edited by M. L. Nering and R. Ostini, 123–152. New York, NY: Routledge.

Askew, M., V. Rhodes, M. Brown, D. Johnson, and D. Wiliam. 1997. *Effective Teachers of Numeracy: Final Report*. London: King’s College.

Bland, J. M., and D. G. Altman. 1995. “Multiple Significance Tests: The Bonferroni Method.” *Bmj* 310: 170.

Bohlig, M., W. Fisher, G. Masters, and T. Bond. 1998. “Content Validity and Misfitting Items.” *Rasch Measurement Transactions* 12(1): 607.

Bond, T.G. 1994. “Too Many Factors in Factor Analysis?.” *Rasch Measurement Transactions* 8(1): 347.

Bond, T. G., and C. M. Fox. 2003. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences.* 2nd ed. Oxford: Routledge.

Carpenter, T. P., and R. Lehrer. 1999. “Teaching and Learning Mathematics With Understanding.” In *Mathematics Classrooms That Promote Understanding,* edited by E. Fennema and T. A. Romberg, 19–32. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Collier, C. P. 1972. ”Prospective Elementary Teachers' Intensity and Ambivalence of Beliefs About Mathematics and Mathematics Instruction.” *Journal for Research in Mathematics Education* 3 (3): 155–163.

De Corte, E., P. Op 't Eynde, and L. Verschaffel. 2002. "’Knowing What to Believe’: The Relevance of Students' Mathematical Beliefs for Mathematics Education.” In *Personal Epistemology: The Psychology of Beliefs About Knowledge and Knowing,* edited by B. K. Hofer and P. R. Pintrich, 297–320. Mahwah, NJ: L. Erlbaum Associates: Routledge.

Ernest, P. 1991. “The Impact of Beliefs On the Teaching of Mathematics.” In *Mathematics Teaching: The State of the Art,* edited by P. Ernest, 249–254. London: Falmer.

Feiman-Nemser, S. 2001. “From Preparation to Practice: Designing a Continuum to Strengthen and Sustain Teaching.” *The Teachers College Record* 103 (6): 1013–1055.

Flores, M. A., and C. Day. 2006. “Contexts Which Shape and Reshape New Teachers’ Identities: A Multi-perspective Study.” *Teaching and Teacher Education* 22 (2), 219–232.

Haggarty, L., K. Postlethwaite, K. Diment, and J. Ellins. 2011. “Improving the Learning of Newly Qualified Teachers in the Induction Year.” *British Educational Research Journal* 37 (6): 935–954.

Liljedahl, P. (2009). ”Teachers' Insights Into the Relationship Between Beliefs and Practice.” In *Beliefs and Attitudes in Mathematics* Education, edited by J. Maas and W. Schlöglmann, 33–43. Rotterdam: Sense Publishers.

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

Linacre, J. M. 2002a. “Optimizing Rating Scale Category Effectiveness.” *Journal of Applied Measurement* 3(1): 85–106.

Linacre, J. M. 2002b. “What Do Infit and Outfit, Mean-square and Standardized Mean.” *Rasch Measurement Transactions* 16(2): 878.

Linacre, J. M. 2010. “When to Stop Removing Items and Persons in Rasch Analysis.” *Rasch Measurement Transactions* 23(4): 1242.

Linacre, J. M. 2014. *WINSTEPS: Rasch Measurement Computer Program (Version 3.8.1)*. Chicago, IL: Winsteps.com.

Linacre, J. M. 2015. *A User’s guide to WINSTEPS/MINISTEP: Rasch-model Computer Programs*. Chicago, IL: Winsteps.com.

Liston, D., J. Whitcomb, and H. Borko. 2006. “Too Little or Too Much Teacher Preparation and the First Years of Teaching.” *Journal of Teacher Education* 57 (4): 351–358.

Masters, G. N. 1982. “A Rasch Model For Partial Credit Scoring.” *Psychometrika* 47(2): 149–174.

Masters, G. N., and B. D. Wright. 1997. “The Partial Credit Model.” In *Handbook of Modern Item Response Theory,* edited by W. J. v. d. Linden and R. K. Hambleton, 101–121. New York: 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.

Pampaka, M., J. Williams, G. Hutcheson, G. Wake, L. Black, P. Davis, and P. Hernandez Martinez. 2012. “The Association Between Mathematics Pedagogy and Learners’ Dispositions for University Study.” *British Educational Research Journal* 38(3): 473–496.

Pepin, B., K. O. Lysø, and S. A. Sikko. 2012. “Student Educational Experiences at Transition From Upper Secondary to Higher Education Mathematics.” In *FoU i Praksis 2011: Rapport Fra Konferanse om Praksisrettet FoU i Lærerutdanning*, edited by F. Rønning, R. Disen, H. Hoveid, and I. Pareliussen, 275-285. Trondheim: Tapir Akademisk Forlag.

Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: 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*, edited by J. Neyman, 321-334. Berkley, CA: University of California Press.

Raymond, A. M. 1997. ”Inconsistency Between a Beginning Elementary School Teacher's Mathematics Beliefs and Teaching Practice.” *Journal for Research in Mathematics Education* 28 (5): 550–576.

Schoenfeld, A. H. 2014. “What Makes for Powerful Classrooms, and How Can We Support Teachers in Creating Them? A Story of Research and Practice, Productively Intertwined.” *Educational Researcher* 43: 404–412.

Schumacker, R. E. 2015. “Detecting Measurement Disturbance Effects: The Graphical Display of Item Characteristics.” *Journal of Applied Measurement* 16 (1): 76–81.

Skott, J. 2001. “The Emerging Pracitces of a Novice Teacher: The Roles of His School Mathematics Images.” *Journal of Mathematics Teacher Education* 4(1): 3–28.

Smith, R. M., R. E. Schumacker, and M. J. Busch. 1998. “Using Item Mean Squares to Evaluate Fit to the Rasch Model.” *Journal of Outcome Measurement* 2 (1): 66–78.

Stephan, M. 2014. “Learner-Centred Teaching in Mathematics Education.” In *Encyclopedia of Mathematics Education,* edited by S. Lerman, 338–343. Dordrecht: Springer.

Swan, M. 2006. “Designing and Using Research Instruments to Describe the Beliefs and Practices of Mathematics Teachers.” *Research in Education 75* (1): 58–70.

Thompson, A. G. 1984. “The Relationship of Teachers' Conceptions of Mathematics and Mathematics Teaching to Instructional Practice.” *Educational Studies in Mathematics* 15(2): 105–127.

Thompson, A. G. 1992. “Teachers’ Beliefs and Conceptions: A Synthesis of the Research.” In *Handbook of Research on Mathematics Teaching and Learning,* edited by D. A. Grouws, 127–146. Reston, VA: National Council of Teachers of Mathematics.

Thurstone, L. L. 1959. *The Measurement of Values*. Chicago, Illinois: The University of Chicago Press.

Wolfe, E., and E. Smith Jr. 2006. “Instrument Development Tools and Activities for Measure Validation Using Rasch Models: Part II - Validation Activities.” *Journal of Applied Measurement* 8 (2): 204–234.

Worthy, J. 2005. “‘It Didn't Have to Be So Hard’: the First Years of Teaching in an Urban School.” *International Journal of Qualitative Studies in Education* 18 (3): 379–398.

Wright, B. 1997. “Fundamental Measurement for Outcome Evaluation.” *Physical Medicine and Rehabilitation: State of the Art Reviews* 11 (2): 261–288.

Wright, B. 2003. “Rack and Stack: Time 1 vs Time 2.” *Rasch Measurement Transactions* 17(1): 905–906.

Wright, B. D., and M. H. Stone. 1979. *Best Test Design*. Chicago: Mesa Press.

Yackel, E., and P. Cobb. 1996. ”Sociomathematical Norms, Argumentation, and Autonomy in Mathematics.” *Journal for Research in Mathematics Education* 27 (4): 458–477.

Yamaguchi, J. 1997. “Positive vs. Negative Wording: PCA of Residuals.” *Rasch Measurement Transactions* 11(2): 567.

Zhu, W. 2002. “A Confirmation Study of Rasch-based Optimal Categorization of a Rating Scale.” *Journal of Applied Measurement* 3(1): 1–15.