The transition from higher education to the world of work: Measuring student teachers' beliefs and practices for purposeful sample selection

Eivind Kaspersen, Birgit Pepin, Svein Arne Sikko

The transition from higher education to the world of work: Measuring student teachers' beliefs and practices for purposeful sample selection

Eivind Kaspersen[1], Birgit Pepin[2] and Svein Arne Sikko[1]

1 Sør-Trøndelag University College, Trondheim, Norway, eivind.kaspersen@hist.no

2 Technische Universiteit Eindhoven, Eindhoven, The Netherlands

This paper reports on methodological results from an on-going project investigating student transition from teacher education to the world of work. We argue that research benefits from purposeful sampling, and we present two Rasch-calibrated instruments that aim at finding participants with particular characteristics. With items from existing instruments, one practice instrument is calibrated on Norwegian student teachers. Furthermore, these items are rephrased to fit a second instrument measuring beliefs about teaching mathematics. Finally, 'virtual equating' is used to align items so that measures can be compared across instruments.

Keywords: Student-centredness, measurement, Rasch modeling, sample selection.

INTRODUCTION

A considerable amount of research exists that sheds light on 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, many 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 in school is not adequately linked to their future practices as teachers (e.g., Liston, Whitcomb, & Borko, 2006). Other studies describe new teachers' first period in 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 & Day, 2006) and a period where there is a gap between new teachers' actual identities and the 'designated identities' shaped by the workplace communities (Haggarty, Postlethwaite, Diment, & Ellins, 2011). In most studies, however, the sample selections do not seem to be based on pre-determined criteria (or these are not made explicit).

To build further on existing research, in this paper, we argue that Rasch-calibrated instruments can facilitate the selection of persons with certain characteristics. For instance, some studies would benefit from selecting persons who identify with reform-minded practices, traditional practices, or persons who follow certain trajectories in the transition from education to work (e.g. resisting change; complying with traditional practices; coming to identify with reform-minded practices) etc. Thus, our research question in this paper is: *how can Rasch-calibrated instruments inform the sample selection in studies on student teachers' transition from higher education to the world of work/school?*

In the paper, we report on results from an on-going study where the overall aim is to understand how identities are negotiated in the transition from higher education to the world of work, including those of mathematics teachers in schools. The study follows on from previous studies of the TransMaths project (www.transmaths.org) (e.g., Pampaka et al., 2012; Pepin, Lysø, & Sikko, 2012), and we will present two Rasch-calibrated instruments that measure persons' practices and beliefs about ideal practice. Finally, we discuss how these instruments can inform sample selection.

THEORETICAL BACKGROUND

The literature suggests different ways of characterising teachers. As a frame for our instrument, we build on the notion of student-centredness, defined by Stephan (2014) 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).

The notion of student-centred teaching, both in practice and research, has grown in many different directions, and it is impossible to describe one single approach. Nevertheless, Stephan (2014) listed five characteristics of student-centred teaching: problem-solving; collaboration; mathematical discourse; tools/manipulations; and classroom environment (pp. 340–342). Regarding classroom environment, Stephan (2014) emphasized four social norms, documented by Yackel and Cobb (1996), which are supporting student-centred teaching. That is, students are expected to: explain/justify solutions; attempt to make sense of others' explanations; indicate agreement/disagreement; and, ask clarifying questions (p. 340).

The rationale for choosing student/teacher-centredness as a frame is that it is uni-dimensional, and thus meets one basic requirement for Rasch analysis. The construct was used by Pampaka and colleagues (2012) earlier in the TransMaths project, when they used Rasch analysis to construct an instrument to capture teachers' self-reported pedagogies. Their instrument was based upon Swan's (2006) practice-questionnaire, which in turn was based on three teacher orientations: *transmission*, *discovery*, and *connectionist* (Askew, Rhodes, Brown, Johnson, & Wiliam, 1997).

In order to measure the relationship between practice and belief about ideal practice, we have extended the practice instrument (Pampaka et al., 2012) with a 'belief-dimension'. We argue that there are three reasons for including a belief-dimension to the original instrument. First, several studies suggest that teachers' practices are influenced by workplace norms (e.g., Haggarty et al., 2011). Gresalfi and Cobb (2011) distinguish between three ways in which teachers can relate to these workplace norms: *consent and identify*; *consent and comply*; or, *resist*. As such, knowledge about teachers' practices alone does not distinguish between those who identify with their own practice and those who merely comply with a workplace norm (or are constrained by other contextual influences). Second, knowledge about teachers' beliefs alone does not inform the researcher about how central those beliefs are. That is, two persons that express the same beliefs about teaching mathematics can hold those beliefs with different strengths. Green (1971) identified 'the degree of conviction' as one of the three dimensions in belief-systems. That is, beliefs can be central or peripheral, where central beliefs are more strongly held than peripheral beliefs. If the researcher wants to locate persons who hold certain beliefs strongly, we argue that persons who can relate those espoused beliefs to actual practice are more likely to meet this criteria than persons with different espoused beliefs and practices. Third, in longitudinal studies, knowledge about both practice and belief can provide information about participants' trajectories in terms of the three categories presented by Gresalfi and Cobb (2011).

In sum, we claim that knowledge about persons' beliefs and practices can help researchers in making well-targeted sample selections, and that this is pertinent to studies conducted in the context of higher education.

DATA COLLECTION AND ANALYSIS

In our study items in the instruments were influenced by the original items in the 'practice instrument' (Swan, 2006). For the belief instrument, these practice items were translated into belief items. For instance, the practice item: "Students work with tasks with a clear answer" was translated into a belief item: "Students should work with tasks with a clear answer." The items and the response categories were then discussed at a Ph.D. seminar in mathematics education. From this discussion, 27 practice items and 27 belief items were piloted on 42 Norwegian student teachers in their second and third year of education, in addition to 9 teacher educators at the same institution. The items were discussed briefly with all participants. The items were then revised and piloted again on 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. 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').

In the analysis, the Rasch-Andrich Rating-Scale Model and the WINSTEPS software were used to construct

one practice scale and *one belief scale*. The Rasch model turns categorical data into interval measurements. Moreover, the model assumes an underlying trait (e.g., teacher-centredness) and is based upon the idea that persons with high measures (e.g., highly teacher-centred) are more likely to agree with the items that define the trait than persons with low measures (e.g. highly student-centred). Similarly, each person is more likely to agree on items with low measures than on items with high measures. A key feature of the Rasch model is that persons and items are not discriminated, which means that they can be measured on the same scale (Wright & Stone, 1979).

Since the purpose of the instruments was to detect persons with particular characteristics (e.g., persons that identify with their practice, or persons with a more teacher-centred practice than belief), we pursued equally scaled instruments. That is, if a person identified with her practice, then her practice-measure should, ideally, be equal to her belief-measure. Moreover, if her practice was more teacher-centred than her belief, then her practice-measure should be larger than her belief-measure. Thus, we conducted four steps for "virtual equating" (Luppescu, 2005): 1) identified pairs of items with, possibly, similar 'difficulties'; 2) cross-plotted the pairs of items; 3) removed pairs of items that were not close (within a .95 confidence bound); and, 4) rescaled the measures on the beliefs test to compensate for different item spacing. In rescaling the belief-instrument, we used two raw scores at each end of the practice-scale (20 and 45), and their corresponding measures, and computed new UIMEAN (mean of the item difficulty) and USCALE (the user-scale value for 1 logit) for the belief-instrument, so that the measures of the raw scores 20 and 45 were equal in both instruments. This is in principle similar to rescaling a Fahrenheit-instrument to a Celsius-instrument by defining two values (e.g. points of freezing and boiling).

To ensure further validity, we used guidelines presented by Wolfe and Smith Jr (2006) which extends Messick's (1995) validation framework with two aspects of evidence put forth in The Medical Outcomes Trust (MOT). To summarize, validity is viewed as a unified concept. That is, there are not different kinds of validity, rather, different kinds of evidence that support validity. Accordingly, Messick (1995) presents six different aspects of validity where evidence can be found: the *content, substantive, structural, generalizability, external,* and *consequential* aspects. Furthermore, the MOT presents two aspects not mentioned by Messick: *Responsiveness* and *interpretability*.

RESULTS

We now present the final instruments and corresponding validity arguments. Both instruments consisted of 12 items, whereas item 3 (marked with an [x]) was reversely coded due to negative point measure correlation earlier in the analysis. To find evidence for the *content* aspect of validity, the technical quality of items has been evaluated. Mean squared fit statistics are chi-squared statistics divided by their degrees of freedom (and hence, have an expected value of 1). OUTFIT is outlier sensitive fit, and INFIT is information-weighted fit. Linacre (2002) suggests values between .5 and 1.5 as productive for measurement, and all items in both tests were within this interval, with belief item 3 having the largest misfit (Figure 2). Furthermore, person reliability values, analogous to Cronbach's alpha, were .87 (practice) and .78 (belief),

Item STATISTICS: PRACTICE MEASURE

ENTRY NUMBER	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASURE-A CORR.	PTMEASURE-A EXP.	Item
10	4.35	.20	1.19	1.2	1.16	1.0	.69	.71	I make sure that all use the same methods
11	2.41	.20	.77	−1.5	.76	−1.6	.67	.68	I teach each subject separately
2	2.13	.20	.82	−1.2	.80	−1.3	.79	.69	Students work with tasks with a clear answer
3	2.04	.20	1.18	1.1	1.32	1.9	.51	.68	[x]Students work with mathematics in small groups
8	1.95	.20	.86	−.9	.85	−.9	.80	.69	I explain for the students how to solve different tasks
4	1.79	.20	1.00	.1	.99	.0	.62	.68	Students start with easy tasks
1	1.58	.20	1.10	.7	1.12	.8	.67	.67	Students work with tasks from the textbook
9	1.53	.20	.96	−.2	.94	−.3	.71	.67	Students work with tasks individually
5	.95	.21	.98	−.1	.94	−.2	.69	.65	I explain methods for the whole class
7	.42	.21	.92	−.4	.88	−.6	.65	.63	Students work with tasks where they use methods they have learned
13	−.86	.23	.99	.0	.99	.1	.54	.56	Most students work with the same subject
6	−1.97	.28	1.25	1.3	1.02	.2	.38	.47	I know before class which subjects the students will work on
MEAN	1.36	.21	1.00	.0	.98	−.1			
S.D.	1.55	.02	.15	.9	.15	.9			

Person: REAL SEP.: 2.57 RELIABILITY: .87 Item: REAL SEP.: 6.96 RELIABILITY: .98

Figure 1: Final practice items

Item STATISTICS: BELIEF MEASURE

ENTRY NUMBER	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASURE-A CORR.	PTMEASURE-A EXP.	Item
10	3.82	.20	.99	.0	1.06	.4	.55	.51	The teacher should make sure that all use the same methods
2	2.53	.19	.76	−1.5	.74	−1.5	.61	.56	Students should work with tasks with a clear answer
11	2.44	.19	1.00	.1	1.01	.1	.58	.56	The teacher should teach each subject separately
3	2.34	.19	1.35	1.9	1.39	1.9	.49	.54	[x]Students should work with mathematics in small groups
8	2.10	.19	.80	−1.1	.83	−.9	.67	.57	The teacher should explain for the students how to solve different tasks
1	2.06	.19	.72	−1.7	.72	−1.6	.62	.57	Students should work with tasks from the textbook
9	1.24	.17	.90	−.6	.88	−.7	.56	.58	Students should work with tasks individually
4	1.22	.18	1.15	.9	1.10	.6	.59	.57	Students should start with easy tasks
5	1.07	.17	1.13	.9	1.13	.8	.56	.59	The teacher should explain methods for the whole class
7	.13	.16	.86	−1.0	.81	−1.3	.65	.60	Students should work with tasks where they use methods they have learned
13	−.30	.16	1.11	.8	1.09	.6	.52	.60	Most students should work with the same subject
6	−1.44	.17	1.18	1.2	1.20	1.0	.51	.57	The teacher should know before class which subjects the students will work on
MEAN	1.43	.18	1.00	.0	1.00	.0			
S.D.	1.38	.01	.19	1.1	.20	1.1			

Person: REAL SEP.: 1.90 RELIABILITY: .78 Item: REAL SEP.: 7.33 RELIABILITY: .98

Figure 2: Final belief items

and item reliability was .98 on both tests, an indicator that the sample was big enough to provide information for item calibration.

When practice items were cross-plotted against belief items (Figure 3), most items were within a .95 confidence bound. An exception was one border-line item (item 4), but the effect on person measures, when this item was removed, was negligible. Thus we have decided to keep the item in the analysis.

Ideally, we would want each pair of items to have equal measures. The DTF-analysis, however, showed that measures were close but for most items not equal (Figure 3). The next step, then, was to see if the measures were 'close enough'. And indeed, when person measures were compared with the ideal situation (anchoring practice items to be equal to the belief items), only small differences in person measures could be found (.26 logits at the most with r^2=1.00).

Moreover, rating scale analysis has been conducted for both instruments, to find evidence for *substantive* validity. None of the four guidelines suggested by Wolfe and Smith Jr (2006, p. 210) were violated: 1) each rating scale category contained more than 10 observations (65 observations in the first category in the practice instrument, being the least); 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; and, 4) the unweighted mean-squared fit statistics were all less than 2.0 (1.25 at the most).

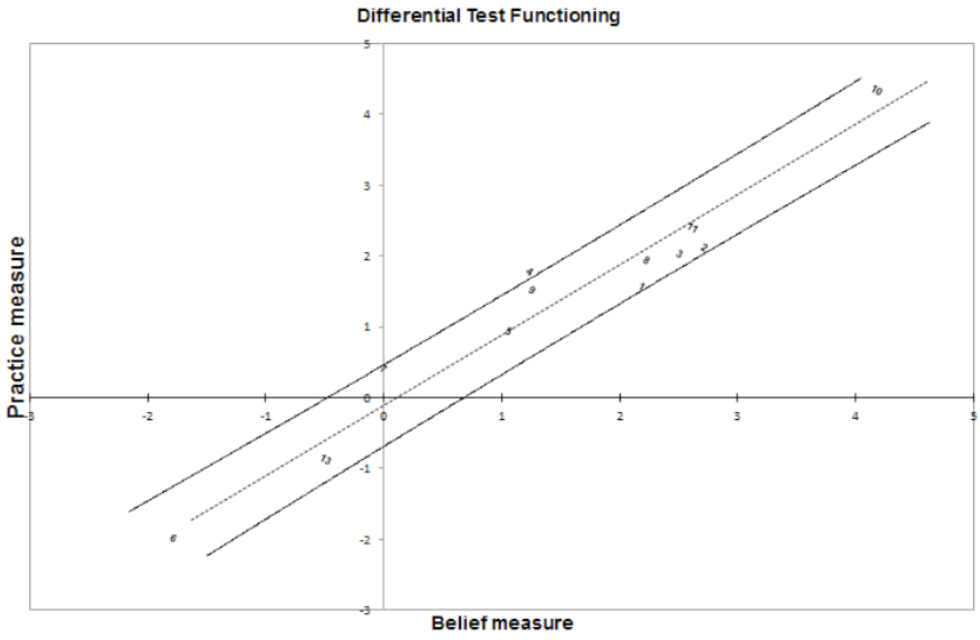
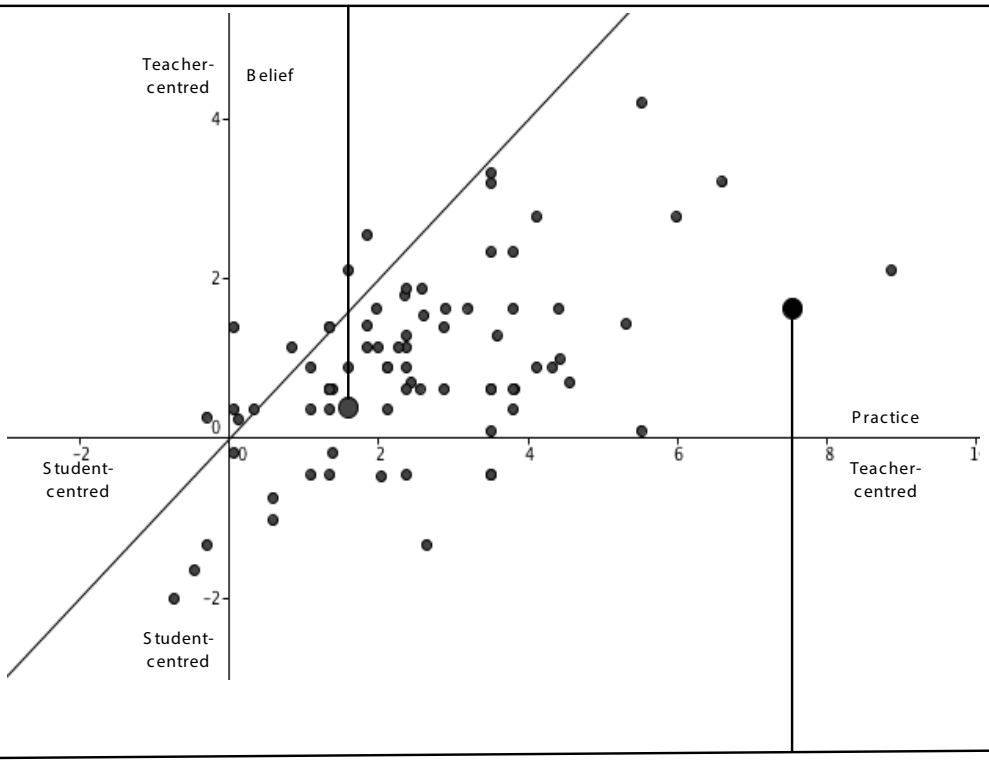


Figure 3: Differential Test Functioning (DTF)

[…] So, we used envelopes and paperclips. Eh, and that X was the envelope, and the number of paperclips in the envelope was the value of X. So, by taking away and adding on both sides of a line, they could solve equations. Eh, and it worked very well.

[…] A good task. Hmm… […] that there is, kind of, a possibility for individual interpretation and exploration. Eh, that it is, it is open. That it is possible to do it in your own way. That it is not constrained by, by the method you are supposed to use, for instance.



[…] it was jus a recap of, they were having a test. And, from one chapter in the book. And then, I read in the book and made my own summary. Formula booklet summary. And then I went to the classroom. And then I went through all I had written down in advance.

[…] that I want, I don't want it to be loose and floating, and, now you are going outside to think about this, and now you are going to play with these bricks. I want some structure, and I want it to be like, ok, now we are going to do this, why do we do it? And I want some rules and stuff. […] But, I think it is also important that you should explore and work with things you know. Perhaps go outside, in mathematics in particular, I think it is important that mathematics is not only abstract, but that you have to work with concretes, real life situations.

Figure 4: The relationship between Norwegian teacher students' practice and espoused beliefs about ideal practice

Uni-dimensionality is a basic assumption in Rasch analysis (Bond & Fox, 2003, p. 32). Thus, we have evaluated dimensionality (*structural* validity) using principal component analysis of the standardized residuals after the Rasch dimension was extracted. Among the practice items, a second dimension could explain 1.9 (in Eigenvalue units) (6.5 %) of the unexplained variance, and among the belief items, a second dimension could explain 1.9 (6.8 %) of the unexplained variance. By default, WINSTEPS stratifies items in three clusters for each contrast. The dis-attenuated correlations between person measures in these clusters were close to 1 on the belief instrument, and .85 in the practice instrument. Thus, we treated the second dimensions as strands (like addition and subtraction on a mathematics test), and not as dimensions that needed separate instruments.

To find evidence for the *generalizability* aspect of validity, we have used Differential Item Functioning (DIF): *the loss of invariance of item estimates across testing occasions* (Bond & Fox, 2003, p. 309). Item calibrations have been compared between genders, classes, and high/low-measured persons. The DIF (Rasch-Welch) t-value was less than 2.0 in all cases, where the belief item 5 had the most misfit between males and females (DIF-contrast = .78 with p = .07).

To look for *external* validity, we have compared the results from our study with the literature on the relationship between beliefs and practices. The results in Figure 4 show that there was only a moderate correlation ($r^2 = .28$) between students' practice and espoused beliefs, consistent with the existing literature (Liljedahl, 2009). However, the responses lie heavily on one side of the identity line; inconsistency was more evident for those who held student-centred beliefs. Although it has not been expressed explicitly, we assert that traces of this relationship can also be found in the literature. Even if different notions are being used, inconsistency is mostly described in situations where participants express reform-minded beliefs (e.g., Kesler, 1985; Vacc & Bright, 1999).

Evidence for *responsiveness* validity can be found in the Person-Item Map (Figure 5). Marks on the right represent item measures, and marks on the left represent person measures. From this, we can see that when person measures exceeded -2 to 4 logits, we can expect that measures were being less accurate. However, since we have used the Rasch-Andrich Rating Scale Model, each item covered more of the trait than a dichotomous item. Thus, the 'lowest' and the 'highest' item measures were not to be considered the 'floor' and the 'ceiling'.

The *interpretability* aspect of validity is the degree to which qualitative meaning can be assigned to quantitative measures (Wolfe & Smith Jr, 2006, p. 227). Thus, excerpts from two interviewed cases are presented in Figure 4. In addition, nine teacher educators, at the same institution, were asked to respond to the belief-test. Other than providing construct validity (as teacher educators were expressing significantly more student-centred beliefs than the students), the instrument was rescaled, so that the mean of the teacher educators' measures was set to zero. This was then used for qualitative interpretation: values close to zero could be thought of as in accordance with the values of the educational institution.

CONCLUSION

In this paper, we have argued that research can benefit from purposeful sample selection, and we have presented two instruments for this purpose. From these instruments, certain 'kinds of persons' can be selected: persons who seem to practise and identify with institutional values (measures close to the origin); persons who seem to practise and identify with traditional values (high measures on both instruments); persons who seem to consent and com-

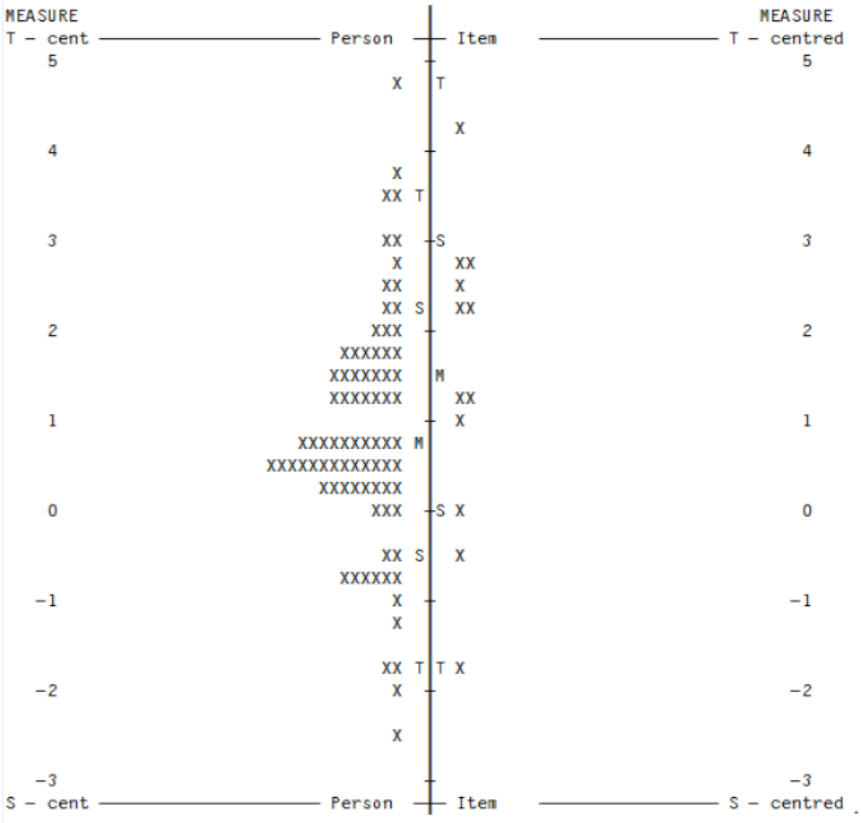


Figure 5: Person-Item Map (belief instrument). M=mean, S=one standard deviation from person or item mean, T=two standard deviations from person or item mean

ply to a workplace norm (having trajectories moving away from the identity line), etc. Our intention is to use these instruments for studying student teachers' transition from higher education to the world of work. However, these instruments can also be used in other kinds of research, we argue. For instance, research on teaching and learning of mathematics in higher education can benefit from selecting particular kinds of participants (e.g. lecturers with a certain practice).

Although we assert that our and similar instruments can be helpful tools for sample selection, we acknowledge their limitations. In our case, we have reduced the practice (and belief) of teaching mathematics to one dimension. This was done due to the statistical benefits, but we emphasize 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. All we can say is that they probably have some characteristics in common. Nevertheless, we conclude that sample selection is, in many cases, better when it is well-targeted rather than opportunistic.

REFERENCES

Askew, M., Rhodes, V., Brown, M., Johnson, D., & Wiliam, D. (1997). *Effective Teachers of Numeracy: Final Report*. London, UK: King's College.

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

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., & Day, C. (2006). Contexts which shape and reshape new teachers' identities: A multi-perspective study. *Teaching and Teacher Education, 22*(2), 219–232.

Green, T. F. (1971). *The activities of teaching*. New York, NY: McGraw-Hill.

Gresalfi, M. S., & Cobb, P. (2011). Negotiating identities for mathematics teaching in the context of professional development. *Journal for Research in Mathematics Education, 42*(3), 270–304.

Haggarty, L., Postlethwaite, K., Diment, K., & Ellins, J. (2011). Improving the learning of newly qualified teachers in the induction year. *British Educational Research Journal, 37*(6), 935–954.

Kesler, R., Jr. (1985). *Teachers' instructional behaviour related to their conceptions of teaching and mathematics and their level of dogmatism: Four case studies*. Unpublished doctoral dissertation, University of Georgia, Athens, GA.

Liljedahl, P. (2009). Teachers' Insights Into the Relationship Between Beliefs and Practice. In J. Maas & W. Schlöglmann (Eds.), *Beliefs and Attitudes in Mathematics Education*. Rotterdam, The Netherlands: Sense Publishers.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions, 16*(2), 878.

Liston, D., Whitcomb, J., & Borko, H. (2006). Too little or too much teacher preparation and the first years of teaching. *Journal of Teacher Education, 57*(4), 351–358.

Luppescu, S. (2005). Virtual Equating. *Rasch Measurement Transaction, 19*(3).

Pampaka, M., Williams, J., Hutcheson, G., Wake, G., Black, L., Davis, P., & Hernandez Martinez, P. (2012). The association between mathematics pedagogy and learners' dispositions for university study. *British Educational Research Journal, 38*(3), 473–496.

Pepin, B., Lysø, K. O., & Sikko, S. A. (2012). Student educational experiences at transition from upper secondary to higher education mathematics. In F. Rønning, R. Disen, H. Hoveid, & I. Pareliussen (Eds.), *FoU i praksis 2011. Rapport fra konferanse om praksisrettet FoU i lærerutdanning* (pp. 275–285). Trondheim, Norway: Tapir Akademisk Forlag.

Stephan, M. (2014). Learner-Centered Teaching in Mathematics Education. In S. Lerman (Ed.), *Encyclopedia of Mathematics Education* (pp. 338–343). Dordrecht, The Netherlands: 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.

Vacc, N. N., & Bright, G. W. (1999). Elementary preservice teachers' changing beliefs and instructional use of children's mathematical thinking. *Journal for Research in Mathematics Education*, 89–110.

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

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

Yackel, E., & Cobb, P. (1996). Sociomathematical norms, argumentation, and autonomy in mathematics. *Journal for Research in Mathematics Education*, 458–477.