The PISA Mathematics Self-Efficacy Scale: Questions of Dimensionality and a Latent Class Concerning Algebra

Boris Girnat

Abstract In 2003 and in 2012, the PISA assessment framework used a scale to measure mathematics self-efficacy. In 2015, this scale was reused in a pretest of an upcoming Swiss assessment of basic mathematical competencies in grade 9. The pretest shows three remarkable results: (1) The scale cannot be seen as unidimensional; moreover, the assumption of unidimensionality disguises some important facts, e.g. concerning gender differences. (2) The items are not worded carefully and do not seem to represent all the relevant content of mathematics adequately; concrete enhancements are suggested. (3) There are latent classes observable within the response patterns to the items, enabling to identify a latent class of "self-proclaimed algebra experts" with interesting connections to other scales measuring beliefs on mathematics.

Measuring Mathematics Self-Assessment

There are different methods for measuring pupils' self-assessment in mathematics. In general, they can be classified into two approaches: The first one is related to a person's so-called mathematical self-concept and is measured by general statements on his mathematical ability like "I have always believed that mathematics is one of my best subjects" (cf. Marsh, 1990). The second approach is called self-efficacy and is based on a suggestion of Bandura to measure a person's self-assessment not by his responses to general statements but by the level of confidence about feeling

B. Girnat (✉)
University of Applied Sciences and Arts Northwestern Switzerland School of Teacher Education, Basel, Switzerland
e-mail: boris.girnat@fhnw.ch

able to solve specific problems that are relevant to the mathematical subdomains of interest (cf. Bandura, 1977, 1986). More explicitly, he defined self-efficacy beliefs as "people's judgments of their capabilities to organize and execute courses of action required to attain designated types of performances" (Bandura, 1986, p. 391).

Research has confirmed a correlation between mathematics self-concept, self-efficacy, and pupils' performance (cf. Multon et al., 1991), but it has been found that the first two concepts are not equivalent and that task-specific mathematics self-efficacy was even a better predictor of career choice than self-concept and test performance (Hackett & Betz, 1989). Insofar mathematics self-efficacy can be seen as a crucial part of a person's mathematical belief system (cf. Philipp, 2007, for the general concept of beliefs and Törner, 2015, for current developments). In the light of these results, scales on mathematics self-concept and self-efficacy have become essential parts of context questionnaires accompanying mathematics performance tests.

The Scales Used in PISA 2003 and 2012

The PISA studies in 2003 and in 2012 measured both the pupils' mathematics self-concept and self-efficacy using the same scales in both of these studies with one minor change (cf. OECD, 2005, pp. 291–294, & OECD, 2014, pp. 322–323). Since the mathematics self-concept is not the main focus of this article, only the eight items of the self-efficacy scale are reported here (cf. OECD, 2014, pp. 322):

Introduction: How confident do you feel about having to do the following mathematics tasks?

1. Using a train timetable to work out how long it would take to get from one place to another
2. Calculating how much cheaper a TV would be after a 30% discount
3. Calculating how many square metres of tiles you need to cover a floor
4. Understanding graphs presented in newspapers
5. Solving an equation like $3x + 5 = 17$
6. Finding the actual distance between two places on a map with a 1:10,000 scale
7. Solving an equation like $2(x + 3) = (x + 3)(x - 3)$
8. Calculating the petrol consumption rate of a car

There were four response categories, labelled with "strongly agree" (coded as 4), "agree" (3), "disagree" (2), and "strongly disagree" (1).

The items are supposed to form a unidimensional scale. Both in PISA 2003 and 2012, the Cronbach's alpha is reported. In 2003, the OECD median of Cronbach's alpha was 0.82; the Swiss value was exactly the same (OECD, 2005, p. 294). In 2012, the OECD median of Cronbach's alpha was 0.85 and 0.83 in Switzerland (OECD, 2014, p. 320). According to the usual standards of interpreting Cronbach's alpha, these values can be seen as good characteristics (cf. Cronbach, 1951).

Table 1 PISA 2003 fit statistics of a CFA model including the self-efficacy scale

Country	RMSEA	CFI	Correlation
OECD median	0.077	0.91	0.62
Switzerland	0.085	0.89	0.61

However, it is worth noting that Cronbach's alpha is only a measurement for the internal consistency of a scale, and it is not an indicator for unidimensionality. Also a scale containing (positively correlated) subscales can achieve a high internal consistency though being not unidimensional and not measuring exactly one psychological construct. An indication for the fact that exactly this situation might be instantiated by the self-efficacy scale is given in the technical report of PISA 2003. In contrast to PISA 2012, the previous documentation did not only report Cronbach's alpha but also a confirmatory factor analysis of the self-efficacy scale combined with other scales, namely, the self-concept scale and the anxiety scale. Table 1 contains the fit indices of this model and the latent correlation between self-efficacy and self-concept (cf. OECD, 2005, p. 293, & Beaujean, 2014, pp. 153–166, for interpreting the fit indices; a short summary: the RMSEA should be less than 0.06 and the CFI should be greater than 0.95, but definitely not below 0.90).

Although the PISA group states that the "model fit is satisfactory for the pooled international sample and for most country sub-samples" (OECD, 2005, p. 294), the fit indices of this model are at least on the borderline. However, since the model whose fit values are reported by the PISA group contains not only the self-efficacy scale, it is undecidable if this scale is the reason for the misfit or if one of the two other scales is responsible for the poor fit indices.

The Swiss Pretest

The Swiss Conference of Cantonal Ministers of Education initiated a nationwide assessment of basic competencies in mathematics in grade 9 (cf. EDK, 2015). This assessment is intended to take place in spring 2016. The School of Teacher Education Northwestern Switzerland is responsible for the performance test and is additionally engaged in developing a part of the context questionnaire. This questionnaire is designed in a way to be connectable with existing research. Insofar, several scales of TIMSS and PISA were integrated, and the scale of mathematics self-efficacy was of special interest. However, the results of the two PISA studies reported above give evidence for the fact that some statistical problems might occur. Since it is unclear what the reasons of these problems could be, I decided to check the PISA scale in a pretest without any changes and to revise the scale afterwards, if necessary. The pretest took place in spring 2015. It was a representative and nationwide test with 956 participants. The items of PISA 2012 were integrated into the Swiss test by using the official German, French, and Italian translations of the PISA group. In the following, I will report the results of this pretest, discussing

Table 2 Three-factor EFA of the self-efficacy scale (item 4 has been withdrawn)

Item	Factor 1	Factor 2	Factor 3	Communality	Complexity
1		0.53		0.66	1.6
2		0.68		0.63	1.1
3		0.72		0.64	1.0
5	0.83			0.90	1.3
6			0.64	0.59	1.1
7	0.94			0.87	1.1
8			0.70	0.65	1.1

what problems appeared and what I would suggest to revise this scale. After analysing the scale, I will present an interesting finding that is not based on quantitative statistics but on a latent class analysis concerning response patterns linked to the items of the scale.

Measuring Mathematics Self-Assessment

In the Swiss pretest, Cronbach's alpha was even higher than in the PISA studies having the value 0.87 with a confidence interval of [0.85, 0.89] on 95% level.

Questions of Dimensionality

As stated above, a good Cronbach's alpha does not discharge from testing the dimensionality of the scale. A parallel analysis according to Horn was performed to determine the optimal number of factors to extract (cf. Horn & Engstrom, 1979). I used the psych package (Revelle, 2015) with R (R Core Team, 2014) to process the parallel analysis and the following exploratory factor analysis (EFA). The parallel analyses suggested four factors, but the EFA showed that the fourth item "Understanding graphs presented in newspapers" caused a problem: It had a high complexity and poor and multiple loadings. That might be an evidence for the fact that the wording of this item could be misleading, e.g. it could be unclear what level of "understanding" is desired or how complex the graphs might be. After removing this item, the parallel analyses suggested three factors, and the EFA led to the following clear and simple structure (factor loadings below 0.2 are suppressed) (Table 2):

The three resulting factors can be interpreted as follows: Factor 1 is definitely the "algebra factor", whereas factor 2 and 3 can be seen as factors of applied mathematics. The difference between the latter could be located in the fact that factor 2 contains rather "easy applications", whereas factor 3 aggregates "hard applications", e.g. its items refer to tasks that require "demanding" calculations to gain the results. To confirm the explanatory outcome, the EFA was complemented by a confirmatory

Table 3 Comparison of a unidimensional and a three-dimensional CFA model

Criterion	One-factor model (14 df)	Three-factor model (11 df)
P value (χ^2)	0.000	0.474
CFI	0.971	1.000
RMSEA	0.086	0.000
SRMR	0.068	0.021

Table 4 Correlations between the three factors

	Algebra	Easy applications	Hard applications
Algebra	1		
Easy applications	0.690***	1	
Hard applications	0.552***	0.820***	1

factor analysis (CFA), using the R package lavaan (Rosseel, 2012). The three-dimensional model of the EFA was tested against the unidimensional one. In both cases, a DWLS estimator was used due to the ordinal nature of the responses (diagonally weighted least squares estimator with robust standard errors and a mean- and variance-adjusted test statistic, cf. Beaujean, 2014, pp. 92–113):

A χ^2 test indicates a significant improvement by using three factors, and the fit indices mentioned in Table 3 give also strong evidence to prefer the three-factor solution.

Correlations Between the Three Factors

In addition to the statistical values, the latent correlation between the three factors were estimated, supporting the decision in favour of the three-factor model, since especially the correlations between the algebra factor and the two application factors are too low to perceive the three factors as measuring a single psychological construct. The asterisks here and in the following denote the usual significance levels (Table 4):

Gender Differences: An Example of Practical Relevance

The discussion about dimensionality and model fits might be regarded as "purely academic", since Cronbach's alpha gives support for the operational capability of the unidimensional scale. A look on gender differences is used as an example to stress the practical relevance of these questions and to underline the warning that a questionable unidimensional scale can disguise empirical facts.

To calculate gender differences, all the latent variables are standardised, and the female group is set to be the reference group. Therefore, the female group always has zero as its mean, and the mean of the male group directly expresses the difference to the mean of the female group. Since the latent variables are standardised, the differences can be interpreted as effect sizes using the thumb rule that 0.2 indicates a small, 0.5 a medium, and 0.8 a strong effect (cf. Cohen, 1988). The gender differences are firstly calculated using the unidimensional scale (without the fourth item) and then for each of the three factors of the three-dimensional solution separately.

In case of the unidimensional scale, the group difference is 0.347** in favour of the male group. This is a small to medium effect. This finding is not unusual, but also not very remarkable (cf. Pajares, 2005). If you consider the three factors separately, the situation will change drastically: In case of the algebra factor, the group difference has a value of −0.024. This difference is not significant and practically non-existent. The difference concerning easy applications (factor 2) is small having a value of 0.276*, but the difference linked to hard applications (factor 3) rises up nearly to a strong effect of 0.766***. Insofar the unidimensional scale masks the fact that gender differences in mathematics self-efficacy is no "monolithic" issue, but it is distributed quite diversely with respect to different subdomains of mathematics.

A Proposal for Further Developments

The observation that mathematics self-efficacy can be organised in three factors leads to the question if three factors are enough to represent the mathematical content of secondary school education adequately. At least in the case of Switzerland, two domains of the traditional curriculum are not represented at all: geometry and probability. This circumstance is taken into account to reorganise the self-efficacy scale for the Swiss main test in the following manner: (1) The algebra factor and the hard applications are maintained, but each of these factors is extended to four items to broaden the possibilities of statistical investigations; (2) the easy applications are partly omitted to keep the number of items in an acceptable range; and (3) four items concerning geometry and four items concerning probability are added to represent all the relevant parts of the Swiss curriculum. The new set of items will look as follows, including as many items of the PISA scale as possible (the original PISA items are marked with an apostrophe):

1') Calculating how much cheaper a TV would be after a 30% discount
2') Calculating how many square metres of tiles you need to cover a floor
3') Calculating the petrol consumption rate of a car
4') Finding the actual distance between two places on a map with a 1:10,000 scale
5') Solving an equation like $3x + 5 = 17$
6') Solving an equation like $2(x + 3) = (x + 3)(x − 3)$
7) Developing and simplifying an algebraic expression like $2a(5a − 3b)^2$

8) Solving an equation like $2x - 3 = 4x + 5$
9) Applying the Pythagorean theorem to calculate the length of one side of a triangle
10) Constructing a perpendicular bisector using a compass and ruler.
11) Calculating the area of a parallelogram.
12) Constructing the focus of a triangle.
13) Calculating the probability of throwing a dice twice in succession to achieve two sixes.
14) Calculating the probability of getting the first prize in the lottery.
15) Calculating how likely it is to take two sweets of the same colour from a sweet jar.
16) Calculating how likely it is that two pupils in a class have the same birthday.

The purpose of these items is not only seen in representing the Swiss curriculum adequately but is also motivated by statistical reasons: The set of items contains four subsets, each of them consisting of four items. This "4x4 arrangement" is the ideal situation not only to model four independent factors but also to estimate a bi-factor model (cf. Beaujean, 2014, pp. 145–152): The entire items load on one common factor, and, additionally, each subset of four items load on one specific factor concerning applications, algebra, geometry, and probability. The common factor can be interpreted as the representation of "general" mathematics self-efficacy; the four specific factors express differences in self-efficacy related to particular subdomains of mathematics. The bi-factor model (if it will work) could fulfil two demands: primarily, the wish to measure mathematics self-efficacy in general and, secondly, the insight to take the observation seriously that it is not advisable to bundle the entire items into one unidimensional scale.

A Latent Class Analysis

A latent class analysis (LCA) is located in the qualitative or nominal part of item response theory. It uses the response patterns to items to classify the probands with a certain probability into different classes characterised by a pattern of conditional probabilities that indicate the chance that their responses to the items take on certain values (cf. Bartulucci, Bacci, & Gnaldi, 2016, pp. 81–82 and 140–141). According to the BIC criterion, the optimal number of latent classes with respect to the items of the self-efficacy scale is seven (all LCA calculations are performed by using R with the poLCA package, cf. Linzer & Lewis, 2011). Figure 1 gives a graphical overview on the probabilities of the response patterns: For each group and for each item, the red column represents the probability that a member of the respective group chooses one of the four response categories linked to this item.

The most interesting class is class 1, since its members have an extraordinarily high probability to choose the highest response option "strongly agree" with respect to the two algebra items (items 5 and 7), whereas their response probabilities to the
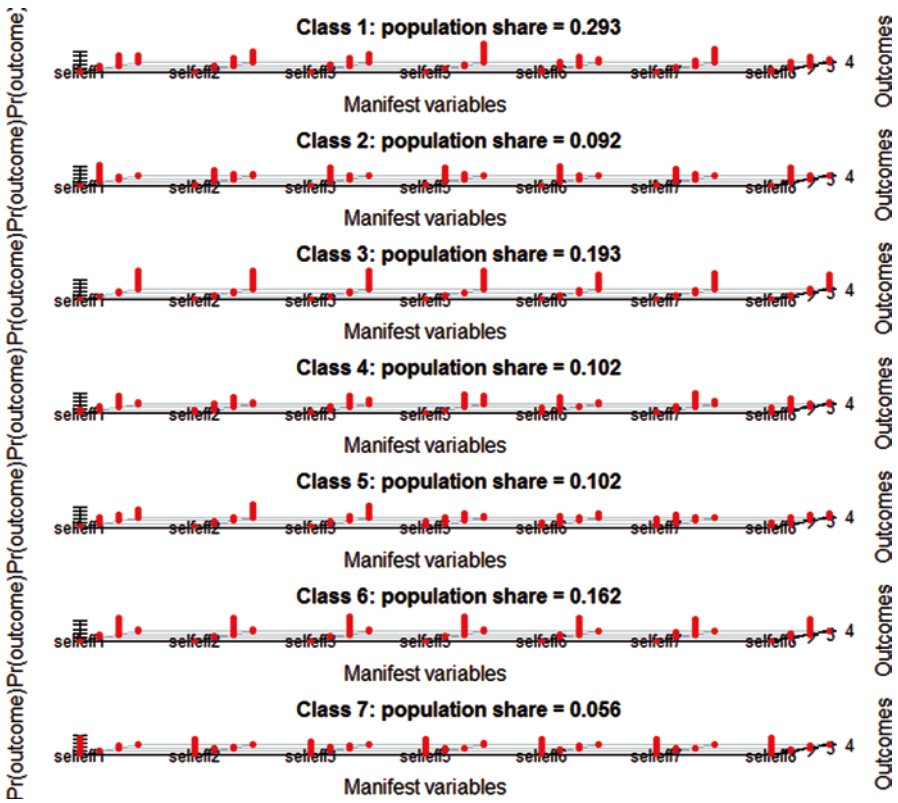
Fig. 1 Latent classes of the self-efficacy scale

other items are rather normally distributed. Insofar, this class can be labelled as the group of "self-proclaimed algebra experts".

Now, we will have a brief look on the properties of this group. Firstly, it is remarkable that 55% of the members are female, and, secondly, it is not surprising that the percentage of "algebra experts" increases according to the three school levels of Switzerland: On the lowest level, 16.1% are members of this group, 29.1% on the middle level, and 54.8% on the highest level (the latter schools are called "Gymasien", "Bezirksschulen", or "Kantonsschulen" which could be translated as academic lower secondary schools).

After considering the demographic background, I will address three topics concerning the beliefs of these pupils: The first topic is related to the self-efficacy scale, the second to preferences for teaching methods and mathematical worldviews (cf. Girnat, 2017), and the last to other scales of PISA 2012 used in the Swiss pretest like motivation, interest, and anger (cf. OECD 2014, pp. 48–66). To illustrate what these scales refer to, I will cite one item of each scale. To estimate the group differences, all the other pupils are regarded as the reference group. Insofar, the mean of the "algebra experts" can be directly interpreted as the mean difference to the other pupils.

The first topic is connected to the three subscales of the self-efficacy scale proposed above. Unsurprisingly, the group of "algebra experts" has a much higher mean on the algebra scale as the other ($d = 1.135***$), but there is just a small difference with regard to the "easy applications" ($d = 0.282**$) and remarkably a negative difference with respect to the "hard applications" ($d = -0.128*$). Insofar, the "algebra experts" do not perceive themselves as "good mathematicians" in general but only as "algebra experts", not being confident about solving "hard" mathematical applications. This finding agrees with the means of the self-concept scale ("I have always believed that mathematics is one of my best subjects"), where the difference between these two groups is not significant ($d = 0.034$).

The "algebra experts" have specific beliefs concerning preferences for teaching methods and mathematical worldviews (cf. Girnat, 2017): The system aspect ("It's necessary to understand mathematical methods. It's not enough just to apply them") is predominant ($d = 0.435***$), and also the formal aspect ("In mathematics it's important to use technical terms and conventional notations") is higher with $d = 0.330***$. Concerning the preferences for teaching methods, there is one significant contrast to other pupils: The "algebra experts" value the technique of learning by repetitive exercises higher than the others ($d = 0.464***$, "I think it's useful to solve a lot of similar tasks in order to understand a method correctly").

Finally, I will mention some group difference concerning scales adapted from PISA 2012 (cf. OECD 2014, pp. 48–66). At first, I will have a look where no significant differences could be detected. This occurs in case of the scales on anger ("I'm often that angry about my mathematics lessons that I could leave immediately"), enjoyment ("I do mathematics because I enjoy it"), instrumental motivation ("Making an effort in mathematics is worth it because it will help me in the work that I want to do later on"), and extrinsic motivation ("I want to have good marks in mathematics"). It is quite remarkable that no differences could be detected in these fields, since they might be regarded as essential to "good" performers in mathematics. The only differences that are significant could only be observed concerning two scales: intrinsic motivation ("It's important to me to understand the topics of mathematics") with $d = 0.386***$ and the learn target ("I want to learn something interesting in mathematics") with an effect size of $d = 0.362***$.

To summarise this paragraph, the "self-proclaimed algebra experts" can be seen as a relevant group of about 30% that is characterised by a special self-esteem in algebra, a specific intrinsic motivation for (the abstract and formal part of) mathematics, and a preference for repetitive techniques to learn mathematics.

Final Remarks

This article should have explained two points: Firstly, the PISA self-efficacy scale is an interesting instrument to measure pupils' mathematics self-assessment, but the scale has to be revised and cannot be regarded as unidimensional. At least according to the Swiss curriculum, a concrete enhancement of the scale was proposed, and a bi-factor model was suggested as an alternative to a unidimensional analysis.

Secondly, a latent class analysis was performed leading to the result that an interesting group of "self-proclaimed algebra experts" could be detected having specific properties related to beliefs on mathematical worldviews and the teaching and learning of mathematics: They prefer repetitive techniques to learn mathematics; and they are intrinsically interested in the formal and "theoretical" parts of mathematics, but not in its real-world applications.

The latent class analysis stresses a possibly unintended advantage of Bandura's concept of self-efficacy: The competitive approach of a person's mathematical self-concept based on general statements on mathematics self-assessment would not be suitable to detect different response patterns related to diverse subdomains of mathematics. Insofar, Bandura's concept offers possibilities being interesting both with regard to contents and statistical methods: The latent class analysis and the bi-factor model suggested above allow a more fine-grained investigation of pupils' performance-related beliefs than the mathematical self-concept. But this statement is not to be interpreted as an advice to replace the mathematics self-concept by self-efficacy in general. As shown above, both approaches can complement one another: The "self-proclaimed algebra experts" only have got a higher self-esteem concerning algebra (measured by self-efficacy scales) and not concerning mathematics in general (measured by self-concept scales). Insofar, a combination of both approaches is an opportunity to detect subtle properties of pupils' beliefs related to their own mathematical performance and potential.

References

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84*, 191–215. https://doi.org/10.1037//0033-295x.84.2.191

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, N.J: Prentice Hall. https://doi.org/10.4135/9781446221129.n6

Bartulucci, F., Bacci, S., & Gnaldi, M. (2016). *Statistical analysis of questionnaires – A unified approach based on R and Stata*. Boca Raton, FL: CRC Press.

Beaujean, A. (2014). *Latent variable modeling using R – A step-by-step guide*. New York: Routledge.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334. https://doi.org/10.1007/bf02310555

EDK (Swiss Conference of Cantonal Ministers of Education). (2015). *Faktenblatt Nationale Bildungsziele (Grundkompetenzen): Definition, Funktion, Überprüfung* (fact sheet national goals of education (basic competencies): definition, purpose, assessment) http://www.edudoc.ch/static/web/arbeiten/harmos/grundkomp_faktenblatt_d.pdf

Girnat, B. (2017). Gender differences concerning pupils' beliefs on teaching methods and mathematical worldviews at lower secondary schools. In C. Andrà, D. Brunetto, E. Levenson, & P. Liljedahl (Eds.), *Teaching and learning in maths classrooms: Emerging themes in affect-related research: teachers' beliefs, students' engagement and social interaction (Research in Mathematics Education)* (pp. 253–263). Cham: Springer International Publishing AG, S. https://doi.org/10.1007/978-3-319-49232-2_24)

Hacket, G., & Betz, N. (1989). An exploration of the mathematics efficacy/mathematics performance correspondence. *Journal of Research in Mathematics Education, 20*, 261–273. https://doi.org/10.2307/749515. National Council of Teachers of Mathematics, Reston.

Horn, J. L., & Engstrom, R. (1979). Cattell's scree test in relation to bartlett's chi-square test and other observations on the number of factors problem. *Multivariate Behavioral Research, 14*(3), 283–300. https://doi.org/10.1207/s15327906mbr1403_1

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software, 42*(10), 1–29. URL http://www.jstatsoft.org/v42/i10/. https://doi.org/10.18637/jss.v042.i10

Marsh, H. W. (1990). *Self-description questionnaire (SDQ) II: A theoretical and empirical basis for the measurement of multiple dimensions of adolescent self-concept: An interim test manual and a research monograph*. San Antonio, TX: The Psychological Corporation.

Multon, K. D., Brown, S. D., & Lent, R. W. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counselling Psychology, 38*., American Psychological Association, Washington, D.C, 30–38. https://doi.org/10.1037/0022-0167.38.1.30

OECD. (2005). *PISA 2003 technical report*. Paris, France: OECD.

OECD. (2014). *PISA 2012 technical report*. Paris, France: OECD.

Pajares, F. (2005). Gender differences in mathematics self-efficacy beliefs. In A. M. Gallagher & J. C. Kaufman (Eds.), *Gender differences in mathematics – An integrative psychological approach* (pp. 294–315). Cambridge, UK/New York: Cambridge University Press. https://doi.org/10.1017/cbo9780511614446.015

Philipp, R. A. (2007). Mathematics teachers' beliefs and affect. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 257–315). Charlotte, NC: Information Age Publishing.

R Core Team. (2014). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Revelle, W. (2015). *Psych: Procedures for personality and psychological research*. Evanston, IL: Northwestern University. http://CRAN.R-project.org/package=psych Version = 1.5.8.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. URL http://www.jstatsoft.org/v48/i02/, Version = 0.5-20 from https://cran.r-project.org/package=lavaan. https://doi.org/10.18637/jss.v048.i02

Törner, G. (2015). Beliefs – No longer a hidden variable, but in spite of this: Where are we now and where are we going? In L. Sumpter (Ed.), *Current state of research on mathematical beliefs XX: Proceedings of the MAVI-20 conference September 29–October 1, 2014* (pp. 7–20). Falun, Sweden: Högskolan Dalarna.