Malaysian Journal of Mathematical Sciences 19(4): 1351–1375 (2025)
https://doi.org/10.47836/mjms.19.4.10

Malaysian Journal of Mathematical Sciences

Journal homepage: https://mjms.upm.edu.my

Rasch Calibration on The "Beliefs About Statistics" Scale for Prospective Mathematics Teachers

Sutrisno *[1,2], Manuharawati [1], and Masriyah [1]

[1] *Universitas Negeri Surabaya, Surabaya, Indonesia*
[2] *Universitas PGRI Semarang, Semarang, Indonesia*

E-mail: sutrisno.21035@mhs.unesa.ac.id
**Corresponding author*

Received: 28 October 2024
Accepted: 28 May 2025

Abstract

Mathematics teachers must have different beliefs when teaching statistics because statistics have different characteristics from mathematics. Forming these beliefs takes a relatively long time, so, they must be prepared early in the prospective teacher preparation program. Unfortunately, until now, no instrument has been available to measure this construct. Therefore, this research aims to use Rasch analysis to develop a "beliefs about statistics" scale for prospective mathematics teachers containing 60 items. The initial sample was obtained from 380 prospective mathematics teachers spread across various colleges in Indonesia. After repeated calibration, leaving three items and two persons misfits, a dataset containing 57 items and 378 persons was obtained. The research results show that this scale is of good quality and has adequate psychometric properties. This scale successfully identified various problems of beliefs about statistics that prospective mathematics teachers have. Nevertheless, this research has provided a "beliefs about statistics" scale that is much needed for prospective mathematics teacher preparation programs.

Keywords: beliefs about teaching and learning statistics; beliefs about the discipline of statistics; instrument calibration; Rasch model; prospective mathematics teachers.

1 Introduction

In an era where information is widely available, people must be wise and critical when using it [49, 5]. This change in the situation requires educational institutions to update their curriculum frameworks, especially school mathematics, because statistical content is entrusted to them [15, 22]. Statistics is closely related to data and decision-making, so, today's society needs it to face the onslaught of information [64, 35]. However, interestingly, statistics has different characteristics from the main subjects that contain it.

Numerous scholars have articulated the distinctions between mathematics and statistics [4, 9]. Mathematics and statistics differ in their approach to numerical data [29]. Numbers, their operations, generalizations, and "abstractions" are the primary focus of mathematics. For instance, statistical reasoning and mathematical modeling necessitate "data related to context" [29, 32]. To continue the analysis and draw conclusions, it is necessary to understand the nature of the data, its origin, and the methods by which it was generated. In contrast, mathematics is motivated by context in the classroom or as a source of research problems; however, the objective is to generalize, identify patterns, and abstract. To comprehend the model or structure, it is necessary to disregard the context. In conclusion, statistics employs more inductive reasoning than mathematics, which relies on deductive reasoning. Statistics emphasizes interpretation in context, whereas mathematics encourages abstraction [57, 60]. The logical implication is that statistical thinking is distinct from mathematical thinking and that a strong foundation in mathematics does not inherently translate into statistical thinking [29, 32].

Proficiency in mathematical theory and statistical concepts is essential for effectively instructing statistics [32]. Both of these abilities must be integrated into the concepts of variability and uncertainty of conclusions because there are fundamental differences between statistics and mathematics in how they are viewed [29, 55]. In mathematics, results are usually reached through deduction, logical proof, or mathematical induction, and there is usually only one correct answer. Statistics, however, uses inductive reasoning, and conclusions are always uncertain. This is mainly due to interpreting the context and methods surrounding data collection and analysis. This also stems from the nature of the variability of the problem. For example, "How old are the teachers in your school?" is a statistical question that expects variability in age. One needs to decide where to get the data (teachers), measure (age), and choose the appropriate statistic (a measure of central tendency or variation) and graphical display to answer the question. In contrast, giving a set of data points on the ages of teachers and asking students to find the mean of the data set is not a statistical question because the answer must be a single number found using an algorithm. Another example of bivariate data is fitting a linear function between height and weight. In mathematics, students are often asked to see a (deterministic) function over a set of points. In contrast, statistical questions focus on the degree of certainty one can make when using a "best fit" function to predict one variable based on another. In particular, one considers how far such extrapolations can be made given the context and how much error is associated with such predictions. Teaching statistics also calls for knowledge of the environments that support and inspire statistical procedures so that teachers can guide students in developing their statistical thinking [10, 36]. This requires a typical mathematics teacher to believe in teaching statistical material, which is called beliefs about statistics [52]. Pfannkuch [50] states, "To be a statistics teacher is to realize that one does not teach a branch of mathematics" and that "statistical thinking, reasoning, or literacy needs to be recognized as a primary learning goal for all students". This is a problem and challenge for mathematics teachers [29]. Depending on their experiences with statistics, prospective mathematics teachers' opinions on statistics will affect their practices and attitudes toward statistics instruction [25, 26].

Rolka and Bulmer [56] state that one can clarify ideas concerning statistics by applying the same structure as those connected to mathematics. Still, beliefs about statistics also bring in other aspects, namely statistical thinking [50]. Wolfe [71] introduced beliefs about statistics, but this topic did not experience much development until Gal and Ginsburg [28] echoed it again. The subject of beliefs related to statistics has been widely researched [1, 38]. In the beliefs structure, beliefs about statistics are domain-specific [67]. Domain-specific beliefs are related to a specific field or domain of mathematics, such as calculus, geometry, or statistics [56]. The field of statistics is characterized by a variety of beliefs, such as those regarding the discipline, the relationship between mathematics and statistics, learning statistics, the placement of statistics in the curriculum, teaching statistics, the relationship between mathematics and statistics, students and their needs, teaching statistics, beliefs regarding statistical content knowledge, the relevance of statistics, the efficiency of teaching practices, the social order shaping the classroom, and beliefs about technology [18, 24]. This research focuses on two beliefs that have received the most attention in statistics education research: beliefs about the discipline of statistics and beliefs about teaching and learning statistics [52, 34].

In contrast to beliefs about mathematics, which have been researched for a long time [43], beliefs about statistics are new ideas that continue to develop. Currently, many instruments measure beliefs about mathematics [54, 14]. However, instruments to measure beliefs about statistics are few and are only intended for school students [52]. Several studies related to statistics have highlighted aspects of "attitude" that some people consider similar to "beliefs". Several academics have created tools to identify and assess students' attitudes about statistics [44, 62]. Nolan et al. [44] conducted a thorough review of 15 instruments measuring attitudes toward statistics, while Carmichael et al. [16] present a theoretical model and explain the development of the Statistical Literacy Interest Measure (SLIM), which can be used to assess students' interest in statistical literacy. Nolan et al. [44] examined the Survey of Attitudes Toward Statistics (SATS). SATS is a comprehensive tool that has been widely used to evaluate students' perspectives on statistics in many settings [33, 59]. Initially, SATS was designed to consist of 28 items (SATS-28) and then developed into 36 items (SATS-36). SATS was developed based on six main attitude components: affection, cognitive competence, values, difficulties, interests, and effort.

However, experts distinguish between the "attitude" and "beliefs". Eagly and Chaiken [23] call attitude "an individual's tendency to evaluate a particular entity with a certain degree of favor or disfavor". Student attitudes in statistics classes can be described as a mix of positive, neutral, and adverse reactions to various people, places, and things that contribute to their statistical education [21]. Unlike attitudes, Philipp [51] defines beliefs as "psychologically held understandings, premises, or propositions about the world that are considered true". Beliefs are considered cognitive (and so are 'known' in a sense). Beliefs are held to different extents. Beliefs are enduring and relatively impervious to alteration, characterized by a greater emphasis on cognitive aspects and a lesser degree of emotional intensity than attitudes [52]. Pierce and Chick [52] contend that beliefs have exerted a significant and enduring impact on teaching and learning processes. Teachers' opinions regarding the teaching of statistics encompass their beliefs about the subject of statistics itself and its role within the curriculum. Teachers' statistical beliefs will impact their attitudes and practices in teaching statistics, and these ideas will be shaped by their previous experiences with the subject. Therefore, studying teachers' beliefs in teaching statistics is very important.

The difference between attitudes and beliefs and the importance of beliefs in teaching statistics prompted researchers to develop a scale to measure beliefs about statistics for college students as prospective mathematics teachers. Researchers want to see the root of the problems in statistics learning through the beliefs about statistics of prospective mathematics teachers. Beliefs take time to develop, and cultural factors play an essential role in their development [27]. The culture in mathematics teacher preparation programs fosters prospective teachers' beliefs about statistics, which will carry over when they become mathematics teachers. Beliefs are stable and reasonably resistant to change, with a more significant cognitive component and less emotional intensity than attitudes [52, 40]. Through this research, testing was carried out to see whether the "beliefs about statistics" scale for prospective mathematics teachers developed by researchers had adequate psychometric properties. The existence of this instrument will be beneficial for mathematics teacher preparation programs and for producing prospective mathematics teachers who are ready to teach statistics.

2 Methods

This quantitative research uses a "beliefs about statistics" scale in data collection. In this study, beliefs about statistics are defined as the ideas that prospective mathematics teachers individually hold about the discipline of statistics, about themselves as learners of statistics, and about the social context of statistics learning that together provide the context for the statistics experience. Based on this definition, this study focuses on two areas of beliefs that have received the most attention in statistics education research: beliefs about the discipline of statistics and beliefs about teaching and learning statistics [52, 34]. First, beliefs about the discipline of statistics include three parts, namely beliefs about the nature of statistics, beliefs about the relationship between mathematics and statistics, and beliefs about the placement of statistics in the curriculum. Second, beliefs about teaching and learning statistics include beliefs about the importance of statistics for students to learn and beliefs about teaching and learning statistics.

The initial design of the scale contained 60 items, consisting of 37 items related to beliefs about the discipline of statistics and 23 items related to beliefs about learning and teaching statistics. Viewed from the type of statement, there are 27 favorable and 33 unfavorable items (refer to Appendix A). Each item uses a 4-point Likert rating, with a range of 1 (strongly disagree), 2 (disagree), 3 (agree), and 4 (strongly agree). Reverse scoring is done for unfavorable items. Favorable items are items that have characteristics that support scale-measuring attributes, whereas unfavorable items are items that have characteristics that do not support scale-measuring attributes. It is hoped that the existence of items that have a reverse direction (unfavorable items) will condition respondents to read each statement item more carefully and not be careless in responding [70, 68]. The homepage of the online survey states that the scale is anonymous and voluntary in addressing ethical issues. So, by completing the survey, respondents agree to participate in this research.

A total of 380 respondents participated in this research. The respondents came from 24 colleges in Central Java Province, Indonesia. They had a Bachelor of Mathematics Education or Mathematics Tadris (a prospective teacher preparation program designed for Islamic schools) study program. These respondents are currently studying in semester 6 or 8. In general, in the curriculum for the preparation program for prospective mathematics teachers in Indonesia, essential advanced statistics courses and mathematics pedagogy have been taken at least by semester 6. Data collected through a scale of beliefs about statistics were tabulated using Microsoft Excel software and evaluated using the Rasch model, especially the Rating Scale Model (RSM), assisted by WINSTEP software version 5.2.0.0. The sample size in Rasch modeling is adequate for stable item calibration within ± 0.5 logit, and 95% confidence level ranges from $64 - 144$ respondents [46, 58]. Therefore, the sample size in this study is very adequate for Rasch model analysis. The Rasch model is a measurement technique based on a model that has gradually become famous for creating scales.

Rasch models generally provide excellent and reliable results [47, 53]. The analysis results contain two outputs, namely person-output and item-output. Item and person fit statistics show the extent to which the data obtained is suitable and reliable, follows the fundamental measure, and provides information about the quality of the measurement. The person table shows whether the respondent is statistically fit, while the item table describes whether the items used in the instrument are fit or not. When evaluating item match and person match, the primary focus is on identifying the item and person using an outfit match statistic. Specifically, the outfit means square (MNSQ) of item and person. The outfit MNSQ value for the item must be in the interval $0.5 - 1.5$, indicating that the data fits reasonably with the model so that the data is productive for item measurement [2, 13]. Meanwhile, there is a tolerance for outfit MNSQ values for the person in the interval $0.0 - 2.0$ [2]. Outfit statistics were chosen because these statistics are more sensitive to outliers and have more familiar calculations. The sensitivity of outfit statistics to outliers also makes it easier to identify and correct suitability problems [13].

There are several steps for fit analysis [13, 12]. First, the analysis begins by evaluating the outfit MNSQ item. Second, if the items are misfits, specific responses from individuals that might cause the items to misfit are examined. The experiment removed unexpected responses (z-residual less than -2 or more than 2), and the analysis was rerun [13]. Next, the misfit items are rechecked to see whether they are within the acceptable limits of the MNSQ. Deletion of a response does not mean the researcher forgot what was done, and deletion of a response does not mean that the researcher did not learn from the response - such responses obscure information for measurement purposes, distorting the researcher's actions [13]. Third, if these measures have not brought the item within the acceptable MNSQ range, the z-standardized outcome (ZSTD) value is checked to see whether it falls between intervals -1.9 and 1.9 [13]. The ZSTD value interval shows that the data has reasonable predictability. As long as the MNSQ value is within the acceptable fit range, the ZSTD value is ignored because ZSTD is greatly influenced by sample size [66, 6]. ZSTD has insufficient power for sample sizes less than 30, which causes ZSTD to be too insensitive, i.e., "all fit". Meanwhile, ZSTD has excessive power for sample sizes over 300, which causes ZSTD to be too sensitive, i.e., "everything is not appropriate" [66, 65]. Considering the sample size in this study, which was more than 300, the ZSTD value was ignored because it could be misleading in interpretation. Furthermore, the Point Measure Correlation (PT-Measure Corr.) value is confirmed to be not negative [2]. Negative values on correlations usually indicate responses to items that contradict the direction of the latent variable [6, 17]. Analysis of item and person fit was also done by looking at standard error measurement (SEM), measurement reliability, separation index, and strata separation [2]. Not limited to the results of this analysis, this research is also equipped with an analysis of rating scales, dimensionality, ceiling and floor effects, differential item functioning (DIF), Cronbach Alpha reliability, information function, and Wright map to provide a comprehensive picture regarding the "beliefs about statistics" scale.

3 Results

Data calibration was carried out repeatedly using the Rasch model to eliminate misfit items and persons [13, 12]. The most critical first step is to investigate the fit of items and then the fit of persons [13]. After cleaning misfit items and persons, other testing criteria were carried out, such as standard error measurement (SEM), measurement reliability, separation index, and separation strata. It does not stop here, and this research is also equipped with an analysis of rating scales, dimensionality, ceiling and floor effects, differential item functioning (DIF), Cronbach Alpha reliability, information function, and Wright map to provide a comprehensive picture regarding the "beliefs about statistics" scale. The final dataset consisted of 57 items and 378 persons. All items and persons in the final dataset are declared fit, meaning that the data fits reasonably with the model and is productive for item measurement. Respondent demographic data is presented in

Table 1, and a summary of the results of the dataset analysis is presented in Table 2. The model fit section (item and person) explains eliminating misfit items and persons.

Table 1: Profile of respondents.

Attribute	Category	Code	Person (n)	Proportion (%)
Gender	Male	M	65	17.20
	Female	F	313	82.80
Type of college	State	S	196	51.85
	Private	P	182	48.15
Study program	Bachelor of Mathematics Education	E	270	71.43
	Bachelor of Mathematics Tadris*	T	108	28.57
Semester	Semester 6	6	248	65.61
	Semester 8	8	130	34.39

*Mathematics Tadris is a unique teacher preparation program for Islamic schools.

3.1 Model fit: Item summary

Of the 60 items on the confidence scale about statistics calibrated using the Rasch model repeatedly, three misfit items were removed, leaving 57 items. Item fit statistics were utilized to measure the validity of the scale. Misfit items in the scale are indicated by negative values on point measurement correlations or outfit MNSQ values outside the $0.50 - 1.50$ interval [2]. The calibration results show that all items have outfit MNSQ values within this interval. However, from the point measurement correlation values, three items were found to have negative values, namely item i15 (PTMEASUR-AL CORR = -0.12), i20 (PTMEASUR- AL CORR = -0.13), and i37 (PTMEASUR-AL CORR = -0.19). Therefore, these three items need to be deleted from this scale because they indicate responses to the items that contradict the direction of the latent variable [69, 11]. This is possible because most respondents do not understand these things. This item deletion also considers the distribution of items on the scale grid so that all indicators are represented in the remaining scale items. These three misfit items are part of beliefs about discipline statistics with favorable statement types. Finally, this "beliefs about statistics" scale consists of 34 items related to beliefs about the discipline of statistics and 23 items related to beliefs about learning and teaching statistics. Viewed from the type of statement, there are 24 favorable and 33 unfavorable items. The calibration results also show that the items in the scale have excellent measurement accuracy, as indicated by the SEM value of 0.28 (less than 0.5) [69, 61]. To give you a rough idea, if the mean of a measure is less than one measurement error from the center of the item hierarchy, then the instrument is on target. This description explains that the items in the "beliefs about statistics" scale are said to be valid (see Table 2).

On the other hand, the item measurement reliability value was 1.00, which was a perfect category (above 0.94). This reliability is not equivalent to classical test theory. This reliability report shows how reproducible the order of item difficulty levels for this set of items is for this sample of persons. This reliability value indicates that the sample is large enough to confirm the difficulty level of the items in the "beliefs about statistics" scale (i.e., construct validity) [2, 63]. This scale has varying levels of agreement difficulty, from items that are easiest to agree on to the most difficult to decide on. This statement is supported by the item separation index, namely 18.65, approximately 19 levels of difficulty of agreement with the appropriate category (more than 3), and item strata separation, namely 25.2, approximately 25 with difficulty with the perfect category (more than 5). This description shows that the items in this scale are said to be reliable, refer to Table 2.

Table 2: Summary of the Rasch measurement model on the "beliefs about statistics" scale.

Analysis	Parameter (with quality criteria)[a]	Value	Category

Model fit: Item summary[b]	Measure item
	(criteria: poor SEM > 2;
	fair SEM $1 - 2$; good SEM $1 - 0.5$;
	very good SEM $0.5 - 0.25$;
	perfect SEM < 0.25)
		M: 0.00;
		SD: 2.12;
		SEM: 0.28
			Very good

	Reliability of item measurement (criteria: poor < 0.67; fair $0.67 - 0.80$; good $0.81 - 0.90$; very good $0.91 - 0.94$; perfect > 0.94)	1.00	Perfect
	Item separation index (criteria: > 3)	18.65	Suitable
	Strata item separation $= [(4 \times \text{item separation index}) + 1]/3$ (criteria: fair $2 - 3$; good $3 - 4$; very good $4 - 5$; perfect > 5)	25.2	Perfect
	Outfit MNSQ item (criteria: $0.50 - 1.50$)	$0.50 - 1.50$	Suitable
	Point measurement correlation (criteria: ≥ 0.00)	$0.00 - 0.61$	Suitable

Model fit: Person summary	Measure person
	(criteria: poor SEM > 2;
	fair SEM $1 - 2$; good SEM $1 - 0.5$;
	very good SEM $0.5 - 0.25$;
	perfect SEM < 0.25)
		M: 0.79;
		SD: 0.92;
		SEM: 0.05
			Perfect

	Person measurement reliability (criteria: poor < 0.67; fair $0.67 - 0.80$; good $0.81 - 0.90$; very good $0.91 - 0.94$; perfect > 0.94)	0.88	Good
	Person separation index (criteria: > 2)	2.77	Suitable
	Strata person separation $= [(4 \times \text{person separation index}) + 1]/3$ (criteria: fair $2 - 3$; good $3 - 4$; very good $4 - 5$; perfect > 5)	4.03	Very good
	Outfit MNSQ item (criteria: $0.00 - 2.00$)	$0.21 - 2.00$	Suitable
	Point measurement correlation (criteria: ≥ 0)	$0.39 - 0.97$	Suitable
Rating scale analysis	Responses per category (criteria: ≥ 10)	1: 952; 2: $6,903$; 3: $10,123$; 4: $2,575$	Suitable
	Measurement means (criteria: increases monotonically	1: -2.47; 2: -1.17;	Suitable

	between rating scales)	3: 1.81; 4: 3.41	

	Outfit MNSQ
	(criteria: $0.00 - 2.00$ logits)
		1: 1.30;
		2: 0.93;
		3: 0.90;
		4: 0.96
			Suitable

	Andrich threshold
	(criteria: increases monotonically between rating scales)
		1: NONE;
		2: -4.04;
		3: 0.03;
		4: 4.01
			Suitable

	Adjusted threshold distance (criteria $1.4 - 5.0$)	Category $1 - 2$: -5.15 to $-2.01 = 3.14$; Category $2 - 3$: -2.01 to $2.02 = 4.03$; Category $3 - 4$: 2.02 to $5.12 = 3.10$	Suitable
Dimensionality[c]	Raw variance in data explained by measurements (criteria: poor $< 20\%$; fair $20\% - 40\%$; good $40\% - 60\%$; excellent $> 60\%$)	57.0%	Good
	Unexplained variance in the 1st-5th contrast of residual PCA (criteria: poor $> 15.0\%$; fair $10.0\% - 15.0\%$; good $5.0\% - 10.0\%$; very good $3.0\% - 5.0\%$; excellent $< 3.0\%$)	1st contrast = 5.0%; 2nd contrast = 3.3%; 3rd contrast = 1.4%; 4th contrast = 1.3%; 5th contrast = 1.2%	Very good
Effect	Ceiling effects (criteria: significant $> 15\%$; moderate $10\% - 15\%$; minor $5\% - 10\%$; negligible $< 5\%$ person with a maximum score of 4)	1: 4.63%; 2: 33.59%; 3: 49.25%; 4: 12.53%	Moderate
	Floor effects (criteria: significant $> 15\%$; moderate $10\% - 15\%$; minor $5\% - 10\%$; negligible $< 5\%$ person with minimum score 1)	1: 4.63%; 2: 33.59%; 3: 49.25%; 4: 12.53%	Negligible

DIF	Items with significant DIF
	(criteria: $p < 0.05$)
		Gender,
		i2: 0.0029;
		i22: 0.0478;
		i34: 0.0034
			3 Items

		Type of college,
		i11: 0.0144;
		i24: 0.0024;
		i25: 0.0176;
		i28: 0.0153;
		i36: 0.0100;
		i41: 0.0152;
			7 Items

		i42: 0.0003	

		Study program,
		i29: 0.0230;
		i36: 0.0000;
		i39: 0.0020;
		i51: 0.0200;
		i58: 0.0310;
		i60: 0.0326
			6 Items

		Semester, i36: 0.0419; i49: 0.0392	2 Items
Reliability	Cronbach Alpha (criteria: bad < 0.50; poor $0.50 - 0.60$; fair $0.60 - 0.70$; good $0.70 - 0.80$, excellent > 0.80)	0.97	Excellent

a: Rating scale quality criteria.
b: All items achieved an acceptable fit because outfit MNSQ values were within the $0.5 - 1.5$ logits, and correlations were not negative.
c: PCA of standardized residuals is used to assess the dimensions of whether a group of items measures the underlying construct.
Note: M: mean; SD: standard deviation; SEM: standard error of measurement; MNSQ: mean-square; PCA: principal component analysis; DIF: differential item functioning.

3.2 Model fit: Person summary

Two misfit persons were repeatedly removed from the 380 respondents calibrated using the Rasch model, leaving 378 persons as the final dataset. Person 018FPE8 was removed from the dataset because it was considered a misfit by the Rasch model (Outfit MNSQ $= 2.67$). Meanwhile, respondent 053MST6 was removed because he responded "Strongly Agree" to all items despite being unfavorable [2]. After the misfit person was eliminated, a research sample size of 378 respondents was obtained, consisting of 65 male and 313 female prospective mathematics teachers (see Table 1). In this study, the respondent's identity was given a code shown in Table 1 (e.g., 053MST6). The first three characters indicate the respondent's serial number, and then the 4th to 7th characters indicate the respondent's attributes, namely gender, type of college, study program, and semester. Respondent attributes will be beneficial during DIF analysis. The calibration results also show that the persons in the sample have a perfect level of measurement accuracy, which is indicated by the SEM value of 0.05 (less than 0.25) [69, 61]. As a rough idea, the sample is on target if the average size is less than one measurement error from the center of the person hierarchy. This description explains that the respondents in this study are said to be valid (see Table 2).

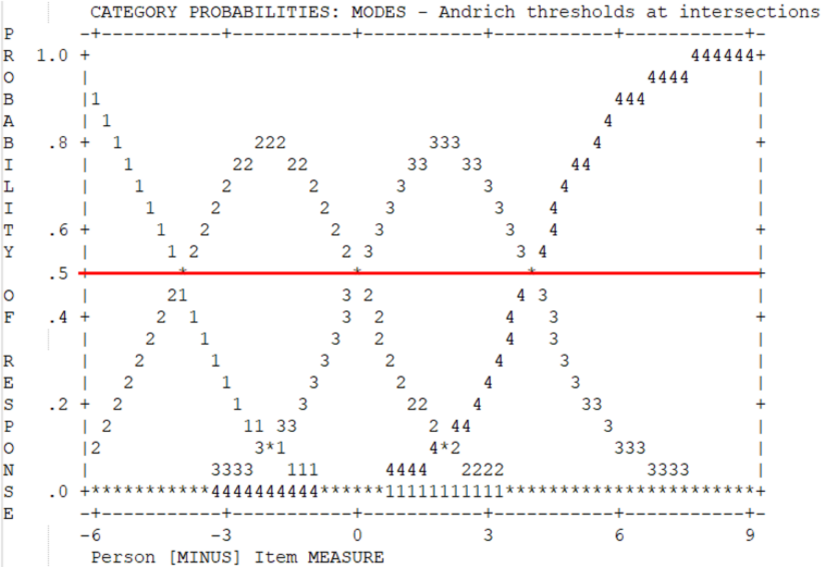
On the other hand, the reliability value obtained for the person measurement was 0.88, which is a suitable category ($0.81-0.90$). This reliability is equivalent to the reliability in classical test theory [13]. This reliability report shows how reproducible a person's measurement sequence is from this sample of persons for this set of items. This reliability value also indicates that many items in the scale are sensitive and adequate to differentiate between different levels of a person's abilities [2, 63]. This research sample has a broad scope, meaning that the sample is spread from persons with the highest beliefs about statistics to those with the lowest. The person separation index supports this statement, namely 2.77, approximately three levels of respondent confidence in the appropriate category (more than 2), and person strata separation, namely 4.03, approximately four levels of respondent confidence in the excellent category $(4-5)$. This description shows that the respondents in this study are said to be reliable (see Table 2).

3.3 Rating scale analysis

The information in Table 2 helps investigate the quality of the rating scale, whether the categories fit the model sufficiently, and whether the thresholds show a hierarchical pattern on the rating scale. An essential examination of the assessment scale shows that each category has provided sufficient observations to estimate a stable threshold value: category 1 is 952 (4.63%), category 2 is 6, 903 (33.59%), category 3 is 10, 123 (49.25%), and category 4 is 2, 575 (12.53%). The recommended minimum number of responses per category is 10 [2].

Next, based on the average measurement and Andrich threshold calibration, all categories are ranked and increase monotonically [41, 37]. The average measurement starting from category 1 is -2.47, category 2 is -1.17, category 3 is 1.81, and category 4 is 3.41. Meanwhile, in the Andrich calibration threshold, starting from category 1, the value is NONE, category 2 is -4.04, category 3 is 0.03, and category 4 is 4.01. Apart from that, the adjusted threshold distance is also known. Category $1-2$ ranges from -5.15 to -2.01 with a gap of 3.14; category $2-3$ starts from -2.01 to 2.02 with a distance of 4.03, and category $3-4$ starts from 2.02 to 5.12 with a distance of 3.10. It can be seen that the adjusted threshold distance meets the criteria, namely in the interval $1.4-5.0$.

To further support this, observations based on the outfit MNSQ for each category show that the suitability of each rating scale category to the Rasch model meets the outfit MNSQ statistical criteria of less than 2.0 [2]. The outfit MNSQ value in category 1 is 1.30, category 2 is 0.93, category 3 is 0.90, and category 4 is 0.96. In addition, category probability curves can also be used to strengthen arguments in rating scale analysis. The category probability curve shows differences in each response category, and each curve peak is higher than 0.5 logit, as presented in Figure 1. This analysis indicates that the quality rating scale with its categories reasonably follows the model, and the stable threshold value shows a pattern hierarchy on a ranking scale (see Table 2).



```
        CATEGORY PROBABILITIES: MODES - Andrich thresholds at intersections
P      -+-----------+-----------+-----------+-----------+-----------+-
R  1.0 +                                                       444444+
O      |                                                   4444      |
B      |1                                              444          |
A      | 1                                             4             |
B   .8 +  1            222             333            4              +
I      |   1         22   22         33   33        44               |
L      |    1       2       2       3       3      4                 |
I      |     1     2         2     3         3    4                  |
T   .6 +      1   2           2   3           3   4                  +
Y      |       1 2             2 3             3 4                   |
    .5 +        *               *               *                    +
O      |       21              3 2             4 3                   |
F   .4 +      2  1             3  2           4   3                  +
       |     2    1           3   2           4   3                  |
R      |    2      1         3     2         4     3                 |
E      |   2        1       3       2       4       3                |
S   .2 +  2          1     3         22    4         33              +
P      | 2            11 33            2 44            3             |
O      |2              3*1             4*2              333          |
N      |           3333   111      4444   2222             3333      |
S   .0 +***********4444444444******11111111111***********************+
E      -+-----------+-----------+-----------+-----------+-----------+-
       -6          -3           0           3           6           9
        Person [MINUS] Item MEASURE
```

Figure 1: Category probability curve.

3.4 Dimensionality

Dimensional analysis investigates whether a scale is unidimensional. Unidimensionality was analyzed using Principal Component Analysis of Rasch measures and residuals. The instrument has excellent unidimensionality (see Table 2). The raw variance value of 57.0% is located in the interval $40\% - 60\%$ (good). It is supported by the unexplained variance values of 1st contrast $= 5.0\%$, 2nd contrast $= 3.3\%$, 3rd contrast $= 1.4\%$, 4th contrast $= 1.3\%$, and 5th contrast $= 1.2\%$, which all lie in the $3.0\% - 5.0\%$ interval (perfect). This indicates that the instrument can effectively measure the beliefs about statistics of prospective mathematics teachers (see Table 2). As expected, all items in the scale regarding statistics measure one dimension, so the unidimensional requirement has been met [17, 37].

3.5 Ceiling and floor effects

The words "ceiling effect" and "floor effect" are opposed but refer to the same phenomenon: the clustering of a person's answers at specified numbers [2, 39]. More specifically, the ceiling effect happens when many participants receive the maximum score, whereas the floor effect occurs when many people receive the lowest score. This can be seen when an instrument is too simple (ceiling effect) or challenging (floor effect). As a result, researchers cannot use the instrument to rank individuals on either end of the scale.

The analysis results show that the "beliefs about statistics" scale has a moderate ceiling effect but is insignificant (see Table 2). This is demonstrated by the response to a maximum score of 4 of 12.53% $(10\% - 15\%)$. On the other hand, the floor effect of the instrument can be ignored because the response to a minimum score of 1 was 4.63% $(< 5\%)$. Based on this effects analysis, it can be concluded that the "beliefs about statistics" scale is of good quality because the ceiling and floor effects are not significant.

3.6 Differential item functioning

Differential Item Functioning (DIF) investigates whether instrument items function differently for different groups of respondents [3, 72]. In other words, DIF examines the items in a test, one by one, to look for signs of interaction with sample characteristics, such as gender, ethnicity, and socio-economic status. This study has several respondent attributes, such as gender, type of college, study program, and semester, which can be used to investigate the DIF of instrument items. An item is said to have significant DIF (item bias) if the probability value is less than 0.05. Interpretation of DIF must consider the risk of a "pseudo-DIF" phenomenon if fewer than 200 respondents are in each attribute category. There is debate around sample size for DIF analysis as this depends on many factors, such as the method of DIF analysis, distribution of item responses, scale length, and ceiling and floor effects [39].

Three items on the "beliefs about statistics" scale have DIF on the gender attribute, namely items i2 $(p = 0.0029)$, i22 $(p = 0.0478)$, and i34 $(p = 0.0034)$. This means that these three items function differently for men and women. Seven items on the "beliefs about statistics" scale have DIF on type of college attributes, namely items i11 $(p = 0.0144)$, i24 $(p = 0.0024)$, i25 $(p = 0.0176)$, i28 $(p = 0.0153)$, i36 $(p = 0.0100)$, i41 $(p = 0.0152)$, and i42 $(p = 0.0003)$. This means that the seven items function differently for public and private colleges. Six items on the "beliefs about statistics" scale have DIF on study program attributes, namely items i29 $(p = 0.0230)$, i36 $(p = 0.0000)$, i39 ($p = 0.0020$), i51 ($p = 0.0200$), i58 ($p = 0.0310$), and i60 ($p = 0.0326$). This means the six items function differently for mathematics education and mathematics education study programs. Two items on the "beliefs about statistics" scale have DIF on the semester attribute: items i36 ($p = 0.0419$) and i49 ($p = 0.0392$). This means the two items function differently for students in semesters 6 and 8. Item i36, which reads, "based on my experience as a student at school, I have enough time studying statistics in mathematics subjects", has a DIF in the type of college, study program, and semester attribute. This item needs more attention and consideration for future improvements (see Table 2).

The DIF items in this study inform future DIF research to develop a sampling strategy of more than 200 respondents for each category on the attributes of gender, type of college, study program, and semester. Deleting or modifying DIF items is not recommended, as it may affect the precision of the matched variables. Therefore, more researchers are advocating qualitative methods to complement DIF analysis for an in-depth assessment of whether a particular DIF effect is of sufficient practical importance [39].

3.7 Cronbach alpha reliability

Cronbach Alpha explains the internal consistency reliability of the scale, and if the value is close to 1, the internal measurement consistency is excellent [61]. This confidence scale about statistics has a Cronbach Alpha value of 0.97, which is in the excellent category (> 0.80). This reflects that the interaction between 378 persons and 57 items is reliable. In other words, the results show the suitability between the person and the item used (see Table 2) [2].

3.8 Information function

Fisher information refers to the quantity of information data offered on a parameter [72]. For item information, this refers to the amount of information the item response provides concerning the person parameter. For instrument information, this is the sum of the information provided by all items a person encounters regarding the person's parameters. The item information functions are added together to generate the instrument information function. Figure 2 depicts the Fisher information for the instrument (collection of items) at different points along the latent variable. This function reports "statistical information" in data relating to each score or metric on the instrument [3, 72].

Simply put, Figure 2 shows the measurement information obtained from the "beliefs about statistics" scale. The $x-$axis shows prospective mathematics teachers' confidence level in working on a given scale, while the $y-$axis shows the value of the information function. In practice, the functional values of information are usually overlooked, with just its form being evaluated. We often want the information function to peak at (1) the most critical cut point (criteria-referenced instruments) or (2) the sample mode (norm-referenced instruments) [3, 72].

Based on Figure 2, it can be seen that the curve has two peaks. The peak of the information function is at 1.875 logits (high level of ability) with information of 14.1581; apart from that, there is also another peak that is slightly lower, namely at -1.3375 logits (low level of ability) with details of 14.1479. The information obtained by the measurement is very high at high and low levels of confidence, meaning that this scale is suitable or optimal for use with prospective mathematics teachers in both groups. Furthermore, this information function ranges from -12 to 13 logits, which means that this scale of beliefs about statistics has an extensive and effective measurement range to reach various levels of respondent confidence [3, 72].

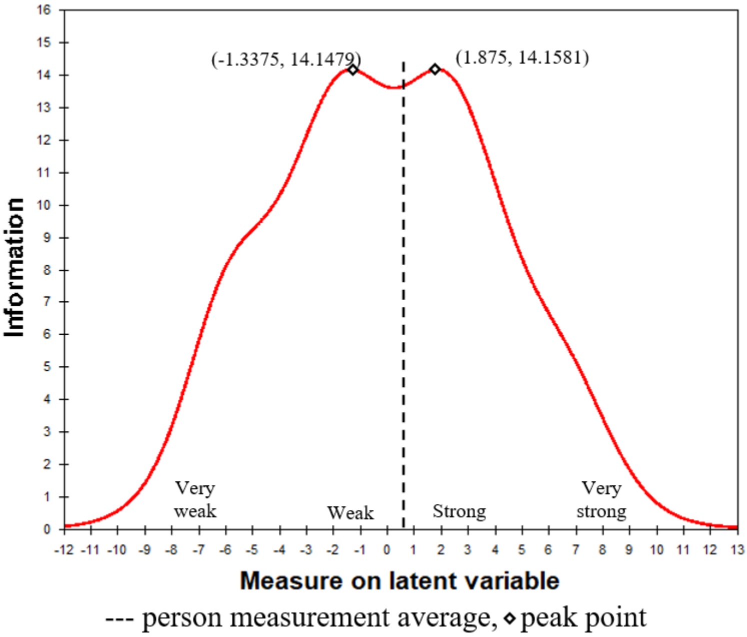
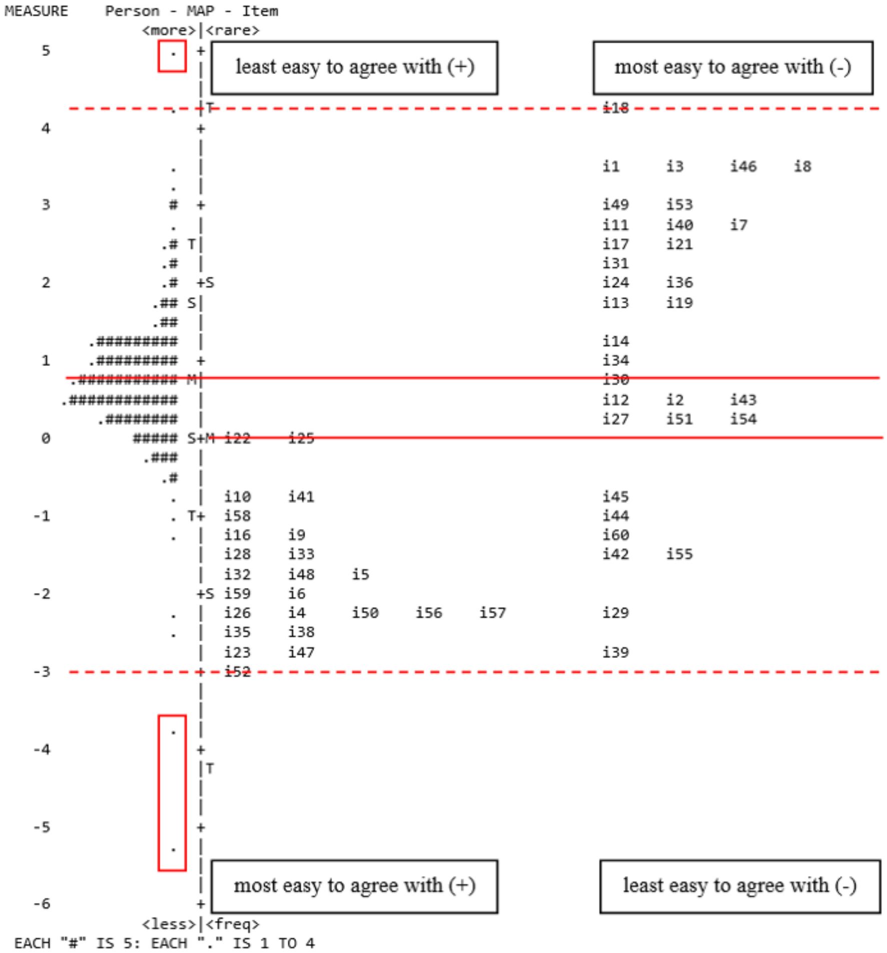


Figure 2: Test information function.

3.9 Wright map

Figure 3 displays the Wright map of the research data. Generally, the person's ability (M= 0.79, SD = 0.92) on the left side is higher than the item difficulty (M = 0.00, SD = 2.12) on the right side. This shows that, in general, prospective mathematics teachers have sufficient beliefs about statistics. However, if examined further, it is revealed that only 10.58% of prospective mathematics teachers have beliefs about statistics at high and very high levels (see Table 2). This indicates a problem with beliefs about statistics among prospective mathematics teachers. Furthermore, based on the Wright map, the items on the scale of beliefs about statistics function well and can separate each respondent, meaning they have good discrimination abilities.

The person with the highest beliefs about statistics is 173FPE6 (4.89 logit); conversely, the person with the lowest beliefs about statistics is 137FST6 (-5.21 logit). Person 173FPE6 (4.89 logit) has a higher logit than any item on the scale (-2.90 to 4.23). This means the respondent's beliefs about statistics exceed the scale's measurement capability. This is in contrast to person 137FST6 (-5.21 logit) and 281FPE6 (-3.74 logit), who have lower logits than any item on the scale (-2.90 to 4.23). This means that both respondents have beliefs about statistics below the scale's measurement capability.



```
MEASURE     Person - MAP - Item
                <more>|<rare>
   5                . +
                      |
                      |
                    . +T                                          i18
   4                  +
                      |
                    . |                                           i1     i3     i46    i8
                    . |
   3                # +                                           i49    i53
                    . |                                           i11    i40    i7
                  .# T|                                           i17    i21
                   .# |                                           i31
   2               .# +S                                          i24    i36
                  .## S|                                          i13    i19
                  .## |
            .######### |                                          i14
   1        .######### +                                          i34
          .########### M|                                         i30
        .############ |                                           i12    i2     i43
            .######## |                                           i27    i51    i54
   0          ##### S+M i22    i25
                 .### |
                   .# |
                    . | i10    i41                                i45
  -1                . T+ i58                                      i44
                    . | i16    i9                                 i60
                      | i28    i33                                i42    i55
                      | i32    i48    i5
  -2                  +S i59    i6
                    . | i26    i4     i50    i56    i57           i29
                    . | i35    i38
                      | i23    i47                                i39
  -3                  + i52
                      |
                      |
                    . |
  -4                  +
                      |T
                      |
                      |
  -5                  +
                    . |
                      |
                      |
  -6                  +
                <less>|<freq>
 EACH "#" IS 5: EACH "." IS 1 TO 4
```

(+) item favorable; (-) item unfavorable

M: mean, S: 1 time standard deviation, T: 2 times the standard deviation

Figure 3: Wright map.

The item with the highest difficulty level is i18 (4.32 logit), and the item with the lowest is i52 (-2.90 logit). The items on the scale that need more attention are the items that have logits above the person's measurement average. In other words, items with a higher difficulty level than the person's average ability ($\mathrm{M} = 0.79$). This has the potential to be explored further so that problems in the beliefs about statistics of prospective mathematics teachers can be identified. There are 20 items with such criteria, namely items i18, i46, i1, i3, i8, i53, i49, i11, i40, i7, i17, i21, i31, i24, i36, i19, i13, i14, i34, and i30 (sorted from the highest item measurement), which coincidentally are all unfavorable items. Interestingly, even though the items are unfavorable, many respondents agree with them. As a result, these respondents get low scores on these items on this scale.

4 Discussion

The calibration results show that the "beliefs about statistics" scale containing 57 items has adequate psychometric properties, such as category function, item properties, item bias, reliability, unidimensionality, ceiling and floor effects, information function, and Wright map. This "beliefs about statistics" scale consists of 34 items related to the discipline of statistics and 23 items related to learning and teaching statistics [52, 34]. Viewed from the type of statement, there are 24 favorable and 33 unfavorable items. Favorable items are items that have characteristics that support scale-measuring attributes, whereas unfavorable items are items that have characteristics that do not support scale-measuring attributes. It is hoped that the existence of items that have a reverse direction (unfavorable items) will condition respondents to read each statement item more carefully and not be careless in responding [70, 68].

Using 4−point Likert ratings on each item in the scale worked well and was fully understood by respondents [41, 37]. The items in the scale are declared valid and have excellent measurement accuracy [69, 61]. Items in the scale represent all indicators of the construct of beliefs about statistics. This scale has varying levels of agreement difficulty, from the easiest to agree on to the most difficult to decide on [13, 63]. The insignificant ceiling and floor effects on this scale support this. This means that the scale is not too easy to cause a ceiling effect and not too difficult to cause a floor effect, so researchers can use this scale to determine respondents' ranking at both ends [2, 39]. This item's difficulty level has been confirmed through a research sample of adequate size, namely 378 respondents [19, 58]. The responses given by respondents were declared valid and had a perfect level of measurement accuracy [69, 61]. Many of the items in the scale are sensitive and adequate to distinguish between different levels of a person's abilities, from persons with the highest beliefs about statistics to those with the lowest [2, 63]. In other words, the items on the "beliefs about statistics" scale have good discrimination ability. This description emphasizes that the items and respondents in this research dataset are valid and reliable. Hence, the dataset fits the Rasch model and is productive for item measurement [2, 13].

Based on Wright's map, this study also found something interesting, namely that many respondents agreed with unfavorable items. This research identified various problems in the beliefs about statistics of prospective mathematics teachers. Many prospective mathematics teachers agree that statistics is part of mathematics (i18). This finding aligns with the opinion of Begg and Edwards [8], who stated that most teachers and prospective teachers believe that statistics is part of mathematics. This is not surprising because they know that statistics is taught in school mathematics. Therefore, they believe that learning statistics is the complete responsibility of mathematics teachers (i34) and a person's mathematical ability is a requirement for success in learning statistics (i24). This finding is confirmed by Chick and Pierce [20], who stated that prospective teachers surveyed believed they had to be good at mathematics to understand basic statistical concepts. Many prospective mathematics teachers believe mathematics and statistics have the same characteristics (i7) related to numbers and mathematical calculations (i1). It is believed that all numbers in mathematics and statistics can be operated mathematically (i8). This aligns with the opinion of Pierce and Chick [52], who stated that mathematics teachers may not consider themselves teaching "statistics" but instead applying calculations. Such beliefs conflict with Pfankuch's opinion [50], who states that being a statistics teacher means realizing that one does not teach a branch of mathematics. The solution and truth in statistical problems are always single (i13 and i14). These findings confirm that prospective mathematics teachers with such beliefs are identified as seeing statistics as part of mathematics, so that all mathematical properties are transferred to statistics. Although closely related to mathematics, statistics is a separate discipline and did not originate from mathematics [42, 30]. Statistics uses mathematics to solve problems, such as algorithms and formulas, theoretical probability models, and several forms of graphical representation [31, 45]. Mathematics is a tool to help investigate statistical questions, but it is not the sole end of statistics. Many of the core ideas of statistics are not mathematical. For example, statistical inquiry is based on context and depends on data, unlike mathematical investigations, which are independent of context [9, 32]. Mathematics and statistics differ in essential defining characteristics, which include the role of context, reasoning methods, precision, data, and data collection

[7].

This research identified biased items for each respondent's attributes, such as gender, type of college, study program, and semester. Three items on the gender attribute, namely items i2, i22, and i34, show that these three items function differently for males and females. The seven items on type of college attributes, namely items i11, i24, i25, i28, i36, i41, and i42, show that these items function differently for public and private colleges. Six items in the study program attributes, namely items i29, i36, i39, i51, i58, and i60, show that these items function differently for mathematics education and mathematics education study programs. Two items in the semester attribute, namely items i36 and i49, show that these two items function differently for students in semester 6 and semester 8. Item i36, which reads, "based on my experience as a student at school, I have enough time studying statistics in mathematics subjects", has DIF on all three attributes: type of college, study program, and semester. This item needs more attention and consideration for future improvements. DIF research is recommended to have a sample of more than 200 respondents in each attribute category. Because the sample size is insufficient for DIF analysis in this study, it is not recommended that these items be deleted [39]. As an alternative, many researchers advocate qualitative methods to complement DIF analysis for an in-depth assessment of whether a particular DIF effect is of practical importance.

Even though it has several items indicating DIF, this confidence scale about statistics has excellent reliability, namely 0.97, which reflects the correspondence between the person and the items in the research dataset [2, 61]. This unidimensional scale, where all items measure one dimension, namely beliefs about statistics. This scale can effectively measure the beliefs about statistics of prospective mathematics teachers [48, 37]. Furthermore, the information function strengthens this argument. The information obtained by the measurement is very high at high and low levels of confidence, meaning that this scale is suitable or optimal for use with prospective mathematics teachers in both groups. This scale has an extensive, effective measurement range to reach various levels of respondent beliefs [3, 72].

One of the limitations of this research is that participants were recruited using an online survey and were limited to students in Central Java Province, Indonesia. The internet has reached most areas in the province, and the education sector is increasing. However, Indonesian people in rural areas may interpret the scale content differently than expected. This must be studied empirically by expanding the research population. This population expansion can also sharpen DIF analysis because DIF research is recommended for sample sizes of more than 200 in each category. Therefore, future research needs to develop a sampling strategy. A second possible limitation is that this study focused only on the internal psychometric characteristics of the scale. Although these findings are promising, external validity measures can indicate how much the measure aligns with theoretically related constructs. Therefore, it is necessary to test external validity using this instrument in different populations. Despite these limitations, the procedures of this study are considered sufficient to illustrate the instrument's reliability.

5 Conclusions

Based on Rasch's analysis, this research dataset has good reliability and validity, and matches the expected model. The "beliefs about statistics" scale is generally of good quality. It has adequate psychometric properties, especially regarding category function, item properties, item bias, reliability, unidimensionality, ceiling and floor effects, information function, and Wright map. This scale successfully identified various problems of beliefs about statistics that prospective mathematics teachers have. On this scale, it is indicated that several items are biased, but because the sample size is not sufficient for DIF analysis, it is not recommended to delete these items. Further testing and refinement of the instrument need to be carried out to increase test precision. Nevertheless, this research has contributed to providing a much-needed scale of beliefs about statistics for prospective mathematics teacher preparation programs.

Acknowledgement The authors express their gratitude to the Center for Higher Education Funding and Assessment (*Pusat Pembiayaan dan Asesmen Pendidikan Tinggi*) and the Indonesian Endowment Fund for Education (*Lembaga Pengelola Dana Pendidikan*) from the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia for providing Indonesian Education Scholarship [ID Number: 202101121527] and funding this research.

Conflicts of Interest The authors declare no conflict of interest.

References

[1] S. E. Aguilera & E. Martinez (2023). Average or outlier? Introductory statistics adjunct instructors' beliefs, practices, and experiences. *The Qualitative Report*, *28*(6), 1769–1786. https://doi.org/10.46743/2160-3715/2023.5971.

[2] H. Akhtar & B. Sumintono (2023). A Rasch analysis of the international personality item pool big five markers questionnaire: Is longer better? *Primenjena Psihologija*, *16*(1), 3–28. https://doi.org/10.19090/pp.v16i1.2401.

[3] N. Azizah, M. Suseno & B. Hayat (2021). Item analysis of the Rasch model items in the final semester exam Indonesian language lesson. *World Journal of English Language*, *12*(1), 15–26. https://doi.org/10.5430/wjel.v12n1p15.

[4] N. Baharun & A. Porter (2016). Use of CAOS test in introductory statistics subject. In *7th International Conference on University Learning and Teaching (InCULT 2014) Proceedings: Educate to Innovate*, pp. 61–74. Singapore. Springer. https://doi.org/10.1007/978-981-287-664-5_6.

[5] J. Bailey, B. Cowie & B. Cooper (2020). "Maths outside of maths": Pre-service teachers' awareness of mathematical and statistical thinking across teachers' professional work. *Australian Journal of Teacher Education*, *45*(1), 1–18. https://doi.org/10.14221/ajte.2020v45n1.1.

[6] O. Barbarin, A. Hitti & J. Brown (2020). Assessing the severity of concerns about preschool children's self-regulation of attention, behavior, and emotions using the ABLE universal screener: A Rasch analysis. *Journal of Emotional and Behavioral Disorders*, *28*(3), 167–179. https://doi.org/10.1177/1063426619864932.

[7] C. Batanero, G. Burrill & C. Reading (2011). *Teaching Statistics in Challenges for Teaching and teacher education: A joint ICMI/IASE study: The 18th ICMI study*. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-1131-0.

[8] A. Begg & R. Edwards (1999). Teachers' ideas about teaching statistics. In *Annual Meeting of the Australian Association for Research in Education and the New Zealand Association for Research in Education. Melbourne, Australia, December 1-4, 1999*, pp. 1–11. Australian Association for Research in Education, Canberra, Australia. https://www.aare.edu.au/data/publications/1999/beg99082.pdf.

[9] D. Ben-Zvi & J. Garfield (2004). *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*, volume 66, chapter Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges, pp. 3–15. Springer, Dordrecht. https://doi.org/10.1007/1-4020-2278-6_1.

[10] R. Biehler, D. Frischemeier & S. Podworny (2018). Elementary preservice teachers' reasoning about statistical modeling in a civic statistics context. *ZDM - Mathematics Education*, *50*(7), 1237–1251. https://doi.org/10.1007/s11858-018-1001-x.

[11] T. G. Bond & C. M. Fox (2015). *Applying the Rasch Model: Fundamental in Measurement the Human Sciences*. Routledge, New York. https://doi.org/10.4324/9781315814698.

[12] W. J. Boone (2016). Rasch analysis for instrument development: Why, when, and how? *CBE–Life Sciences Education*, *15*(4), 1–7. https://doi.org/10.1187/cbe.16-04-0148.

[13] W. J. Boone, J. R. Staver & M. S. Yale (2014). *Rasch Analysis in the Human Sciences*. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-6857-4.

[14] G. T. L. Brown, S. K. F. Hui, W. M. Flora & K. J. Kennedy (2011). Teachers' conceptions of assessment in Chinese contexts: A tripartite model of accountability, improvement, and irrelevance. *International Journal of Educational Research*, *50*(5–6), 307–320. https://doi.org/10.1016/j.ijer.2011.10.003.

[15] R. Callingham, J. Watson & G. Oates (2021). Learning progressions and the Australian curriculum mathematics: The case of statistics and probability. *Australian Journal of Education*, *65*(3), 329–342. https://doi.org/10.1177/00049441211036521.

[16] C. Carmichael, R. Callingham, I. Hay & J. Watson (2010). Measuring middle school students' interest in statistical literacy. *Mathematics Education Research Journal*, *22*(3), 9–39. https://doi.org/10.1007/bf03219776.

[17] S. W. Chan, C. K. Looi & B. Sumintono (2021). Assessing computational thinking abilities among Singapore secondary students: A Rasch model measurement analysis. *Journal of Computers in Education*, *8*(2), 213–236. https://doi.org/10.1007/s40692-020-00177-2.

[18] C. H. Chen (2008). Why do teachers not practice what they believe regarding technology integration? *The Journal of Educational Research*, *102*(1), 65–75. https://doi.org/10.3200/JOER.102.1.65-75.

[19] W. H. Chen, W. Lenderking, Y. Jin, K. W. Wyrwich, H. Gelhorn & D. A. Revicki (2014). Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Quality of Life Research*, *23*, 485–493. https://doi.org/10.1007/s11136-013-0487-5.

[20] H. L. Chick & R. U. Pierce (2008). Teaching statistics at the primary school level: Beliefs, affordances, and pedagogical content knowledge. In *Joint ICMI/IASE Study: Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education. Proceedings of the ICMI Study 18 and 2008 IASE Round Table Conference, Mexico, 2008*, pp. Article ID: T2P3. ICMI/IASE, Mexico. https://doi.org/10.52041/SRAP.08303.

[21] F. Chiesi & C. Primi (2010). Cognitive and non-cognitive factors related to students' statistics achievement. *Statistics Education Research Journal*, *9*(1), 6–26. https://doi.org/10.52041/serj.v9i1.385.

[22] S. Dingman, D. Teuscher, J. A. Newton & L. Kasmer (2013). Common mathematics standards in the United States: A comparison of K–8 state and common core standards. *The Elementary School Journal*, *113*(4), 541–564. https://doi.org/10.1086/669939.

[23] A. H. Eagly & S. Chaiken (2007). The advantages of an inclusive definition of attitude. *Social Cognition*, *25*(5), 582–602. https://doi.org/10.1521/soco.2007.25.5.582.

[24] A. Eichler (2007). Individual curricula: Teachers' beliefs concerning stochastic instructions. *International Electronic Journal of Mathematics Education*, 2(3), 208–226. https://doi.org/10.29333/iejme/184.

[25] A. Eichler (2008). Teachers' classroom practice and students' learning. In *Joint ICMI/IASE Study: Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education. Proceedings of the ICMI Study 18 and 2008 IASE Round Table Conference, Mexico, 2008*, pp. Article ID: T2P5. ICME/IASE, Mexico. https://doi.org/10.52041/SRAP.08305.

[26] A. Eichler (2011). *Teaching Statistics in School Mathematics-Challenges for Teaching and Teacher Education: A Joint ICMI/IASE Study: The 18th ICMI Study*, volume 14 of *New ICMI Study Series*, chapter Statistics Teachers and Classroom Practices, pp. 175–186. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-1131-0_19.

[27] I. Gal (1997). Assessing students' interpretation of data. In *IASE Papers on Statistical Education ICME-8. International Congress on Mathematics Education-8. Seville, Spain, July 14-21, 1996*, pp. 49–57. Swinburne Press, Hawthorn, Australlia. http://iase-web.org/documents/papers/icme8/Gal.pdf.

[28] I. Gal & L. Ginsburg (1994). The role of beliefs and attitudes in learning statistics: Towards an assessment framework. *Journal of Statistics Education*, *2*(2), 1–16. https://doi.org/10.1080/10691898.1994.11910471.

[29] L. Gattuso & M. G. Ottaviani (2011). *Teaching Statistics in School Mathematics-Challenges for Teaching and Teacher Education: A Joint ICMI/IASE Study: The 18th ICMI Study*, volume 14 of *New ICMI Study Series*, chapter Complementing mathematical thinking and statistical thinking in school mathematics, pp. 121–132. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-1131-0_15.

[30] J. L. Green, W. M. Smith, A. T. Kerby, E. E. Blankenship, K. K. Schmid & M. A. Carlson (2018). Introductory statistics: Preparing in-service middle-level mathematics teachers for classroom research. *Statistics Education Research Journal*, *17*(2), 216–238. https://doi.org/10.52041/serj.v17i2.167.

[31] R. E. Groth (2007). Research commentary: Toward a conceptualization of statistical knowledge for teaching. *Journal for Research in Mathematics Education*, *38*(5), 427–437. https://www.jstor.org/stable/30034960.

[32] A. Hannigan, O. Gill & A. M. Leavy (2013). An investigation of prospective secondary mathematics teachers' conceptual knowledge of and attitudes towards statistics. *Journal of Mathematics Teacher Education*, *16*, 427–449. https://doi.org/10.1007/s10857-013-9246-3.

[33] A. Hannigan, A. C. Hegarty & D. McGrath (2014). Attitudes towards statistics of graduate entry medical students: The role of prior learning experiences. *BMC Medical Education*, *14*, Article ID: 70. https://doi.org/10.1186/1472-6920-14-70.

[34] T. R. Harrison (2020). *Decision-Making of Secondary Statistics Teachers*. Phd thesis North Carolina State University, United States. https://repository.lib.ncsu.edu/handle/1840.20/38370.

[35] B. Jahn, S. Friedrich, J. Behnke, J. Engel, U. Garczarek, R. Münnich, M. Pauly, A. Wilhelm, O. Wolkenhauer, M. Zwick et al. (2022). On the role of data, statistics and decisions in a pandemic. *AStA Advances in Statistical Analysis*, *106*(3), 349–382. https://doi.org/10.1007/s10182-022-00439-7.

[36] F. Kien-Kheng, N. Azlan, S. N. D. Ahmad, N. L. H. Leong & I. Mohamed (2016). Relationship between cognitive factors and performance in an introductory statistics course: A Malaysian case study. *Malaysian Journal of Mathematical Sciences*, *10*(3), 269–282.

[37] L. A. R. Laliyo, B. Sumintono & C. Panigoro (2022). Measuring changes in hydrolysis concept of students taught by inquiry model: Stacking and racking analysis techniques in Rasch model. *Heliyon*, *8*(3), Article ID: e09126. https://doi.org/10.1016/j.heliyon.2022.e09126.

[38] H. S. Lee, G. F. Mojica & J. N. Lovett (2020). Examining how online professional development impacts teachers' beliefs about teaching statistics. *Online Learning*, *24*(1), 5–27. https://doi.org/10.24059/olj.v24i1.1992.

[39] W. Ling Lee, K. Chinna & B. Sumintono (2021). Psychometrics assessment of HeartQoL questionnaire: A Rasch analysis. *European Journal of Preventive Cardiology*, *28*(12), e1–e5. https://doi.org/10.1177/2047487320902322.

[40] D. B. McLeod (1992). *Handbook of Research on Mathematics Teaching and Learning: A project of the National Council of Teachers of Mathematics.*, chapter Research on affect in mathematics education: A reconceptualization, pp. 575–596. Macmillan Publishing Co Inc, New York. https://peterliljedahl.com/wp-content/uploads/Affect-McLeod.pdf.

[41] Z. Mohd Zabidi, B. Sumintono & Z. Abdullah (2022). Enhancing analytic rigor in qualitative analysis: Developing and testing code scheme using Many Facet Rasch model. *Quality & Quantity*, *56*, 713–727. https://doi.org/10.1007/s11135-021-01152-4.

[42] D. S. Moore & G. W. Cobb (2000). Statistics and mathematics: Tension and cooperation. *The American Mathematical Monthly*, *107*(7), 615–630. https://doi.org/10.2307/2589117.

[43] M. Muhtarom, S. Sutrisno, N. Nizaruddin & Y. H. Murtianto (2024). Research on mathematical beliefs: Systematic literature review. *International Journal of Evaluation and Research in Education*, *13*(2), 693–704. https://doi.org/10.11591/ijere.v13i2.25968.

[44] M. M. Nolan, T. Beran & K. G. Hecker (2012). Surveys assessing students' attitudes toward statistics: A systematic review of validity and reliability. *Statistics Education Research Journal*, *11*(2), 103–123. https://doi.org/10.52041/serj.v11i2.333.

[45] J. Noll (2011). Graduate teaching assistants' statistical content knowledge of sampling. *Statistics Education Research Journal*, *10*(2), 48–74. https://doi.org/10.52041/serj.v10i2.347.

[46] T. R. O'Neill, J. L. Gregg & M. R. Peabody (2020). Effect of sample size on common item equating using the dichotomous Rasch model. *Applied Measurement in Education*, *33*(1), 10–23. https://doi.org/10.1080/08957347.2019.1674309.

[47] H. Othman, I. Asshaari, H. Bahaludin, Z. M. Nopiah & N. A. Ismail (2012). Application of Rasch measurement model in reliability and quality evaluation of examination paper for engineering mathematics courses. *Procedia-Social and Behavioral Sciences*, *60*, 163–171. https://doi.org/10.1016/j.sbspro.2012.09.363.

[48] Y. K. Ow-Yeong, I. H. Yeter & F. Ali (2023). Learning data science in elementary school mathematics: A comparative curriculum analysis. *International Journal of STEM Education*, *10*(1), Article ID: 8. https://doi.org/10.1186/s40594-023-00397-9.

[49] L. Pangrazio, A. L. Godhe & A. G. L. Ledesma (2020). What is digital literacy? A comparative review of publications across three language contexts. *E-learning and Digital Media*, *17*(6), 442–459. https://doi.org/10.1177/2042753020946291.

[50] M. Pfannkuch (2008). Training teachers to develop statistical thinking. In *Joint ICMI/IASE Study: Teaching Statistics in School Mathematics. Challenges for Teaching and Teacher Education. Proceedings of the ICMI Study 18 and 2008 IASE Round Table Conference, Mexico, 2008*, pp. Article ID: T4P2. ICME/IASE, Mexico. https://doi.org/10.52041/SRAP.08505.

[51] R. A. Philipp (2007). Mathematics teachers' beliefs and affect. In *Second handbook of research on mathematics teaching and learning*, pp. 257–315. Information Age Publishing, Charlotte.

[52] R. Pierce & H. Chick (2011). *Teaching Statistics in School Mathematics-Challenges for Teaching and Teacher Education: A Joint ICMI/IASE Study: The 18th ICMI Study*, volume 14 of *New ICMI Study Series*, chapter Teachers' Beliefs About Statistics Education, pp. 151–162. Springer, Dordrecht, Netherlands. https://doi.org/10.1007/978-94-007-1131-0_17.

[53] M. Planinic, W. J. Boone, A. Susac & L. Ivanjek (2019). Rasch analysis in physics education research: Why measurement matters. *Physical Review Physics Education Research*, *15*(2), Article ID: 020111. https://doi.org/10.1103/PhysRevPhysEducRes.15.020111.

[54] Y. W. Purnomo (2017). A scale for measuring teachers' mathematics-related beliefs: A validity and reliability study. *International Journal of Instruction*, *10*(2), 23–38. https://doi.org/10.12973/iji.2017.10120a.

[55] A. H. M. Rahmatullah Imon & K. Das (2014). How to tell the truth with statistics. *Malaysian Journal of Mathematical Sciences*, *8*(2), 289–309.

[56] K. Rolka & M. Bulmer (2005). Picturing student beliefs in statistics. *ZDM - Mathematics Education*, *37*(5), 412–417. https://doi.org/10.1007/s11858-005-0030-4.

[57] A. Rossman, B. Chance & E. Medina (2006). *Thinking and Reasoning with Data and Chance: Sixty-eighth Yearbook*, chapter Some key comparisons between statistics and mathematics and why teachers should care, pp. 323–333. National Council of Teachers of Mathematics, Reston, Virginia.

[58] A. Şahin & D. Anil (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice*, *17*(1), 321–335. https://doi.org/10.12738/estp.2017.1.0270.

[59] E. E. Sarikaya, A. Ok, Y. C. Aydin & C. Schau (2018). Turkish version of the survey of attitudes toward statistics: Factorial structure invariance by gender. *International Journal of Higher Education*, *7*(2), 121–127. https://doi.org/10.5430/ijhe.v7n2p121.

[60] R. L. Scheaffer (2006). Statistics and mathematics: On making a happy marriage. In G. F. Burrill (Ed.), *Thinking and Reasoning with Data and Chance: Sixty-eighth Yearbook*, volume 68 pp. 309–321. National Council of Teachers of Mathematics, Reston, Virginia.

[61] B. Setiawan, M. Panduwangi & B. Sumintono (2018). A Rasch analysis of the community's preference for different attributes of Islamic banks in Indonesia. *International Journal of Social Economics*, *45*(12), 1647–1662. https://doi.org/10.1108/ijse-07-2017-0294.

[62] E. Songsore & B. J. G. White (2018). Students' perceptions of the future relevance of statistics after completing an online introductory statistics course. *Statistics Education Research Journal*, *17*(2), 120–140. https://doi.org/10.52041/serj.v17i2.162.

[63] M. A. P. Souza, W. J. Coster, M. C. Mancini, F. C. M. S. Dutra, J. Kramer & R. F. Sampaio (2017). Rasch analysis of the participation scale (P-scale): Usefulness of the P-scale to a rehabilitation services network. *BMC Public Health*, *17*, Article ID: 934. https://doi.org/10.1186/s12889-017-4945-9.

[64] A. M. Stefan (2022). Statistics for making decisions. *The American Statistician*, *76*(1), 87–88. https://doi.org/10.1080/00031305.2021.2020003.

[65] L. Tesio, A. Caronni, A. Simone, D. Kumbhare & S. Scarano (2024). Interpreting results from Rasch analysis 2. Advanced model applications and the data-model fit assessment. *Disability and Rehabilitation*, *46*(3), 604–617. https://doi.org/10.1080/09638288.2023.2169772.

[66] R. E. Tornabene, E. Lavington & R. H. Nehm (2018). Testing validity inferences for Genetic Drift Inventory scores using Rasch modeling and item order analyses. *Evolution: Education and Outreach*, *11*, Article ID: 6. https://doi.org/10.1186/s12052-018-0082-x.

[67] G. Törner (2002). *Beliefs: A Hidden Variable in Mathematics Education?*, volume 31 of *Mathematics Education Library*, chapter Mathematical beliefs–A search for a common ground: Some theoretical considerations on structuring beliefs, some research questions, and some phenomenological observations, pp. 73–94. Springer, Dordrecht, Netherlands. https://doi.org/10.1007/0-306-47958-3_5.

[68] A. Vigil-Colet, D. Navarro-González & F. Morales-Vives (2020). To reverse or to not reverse Likert-type items: That is the question. *Psicothema*, *32*(1), 108–114. https://doi.org/10.7334/psicothema2019.286.

[69] S. Wei Chan, Z. Ismail & B. Sumintono (2015). The impact of statistical reasoning learning environment: A Rasch analysis. *Advanced Science Letters*, *21*(5), 1211–1215. https://doi.org/10.1166/asl.2015.6077.

[70] B. Weijters, H. Baumgartner & N. Schillewaert (2013). Reversed item bias: An integrative model. *Psychological Methods*, *18*(3), 320–334. https://doi.org/10.1037/a0032121.

[71] M. L. Wolfe (1978). Anxiety and stereotyped beliefs about statistics. *Evaluation & the Health Professions*, *1*(4), 251–260. https://doi.org/10.1177/016327877800100411.

[72] A. E. Wyse & R. Mapuranga (2009). Differential item functioning analysis using Rasch item information functions. *International Journal of Testing*, *9*(4), 333–357. https://doi.org/10.1080/15305050903352040.

Appendix A

"Beliefs about Statistics" Scale for Prospective Mathematics Teachers

Note: (+) is favorable item, (-) is unfavorable item

1. Statistics related to numbers and mathematical calculations. (-)
2. Statistics contains a collection of analysis techniques, but they are used separately and are not related to each other. (-)
3. Statistics is a collection of analysis techniques related to each other.(-)
4. Statistics as data analysis and interpretation. (+)
5. Statistics as a way of understanding real life using various statistical models. (+)
6. Statistics is a tool that can develop one's thinking and create new interpretations of data and life. (+)
7. Statistics has the same characteristics as mathematics. (-)
8. Mathematics and statistics use numbers similarly by operating on them mathematically. (-)
9. Mathematics emphasizes abstraction, but statistics emphasizes interpretation in context. (+)
10. Mathematics exploits deductive reasoning, while statistics uses more inductive reasoning. (+)
11. Mathematics and statistics rely on context as a source of problems, only to motivate students to learn. (-)
12. Problems in statistics are always solved through one method, such as solving mathematical problems in general. (-)
13. Solving problems in statistics always provides a single answer, such as solving mathematical problems in general. (-)
14. The truth of solving cases in statistics is always singular, and there are no other alternative answers. (-)
15. Mathematics is an exact and precise science, while statistics is a science that contains uncertainty and error. (+, drop out)
16. Different representations (tables or graphs) of the same data in statistics are used to identify different aspects. (+)
17. Different representations (tables or graphs) of the same data in statistics are used to show the same relationship. (-)
18. Statistics is a part of mathematics. (-)
19. Mathematics is a part of statistics. (-)
20. Statistics is a separate discipline that is different from mathematics. (+, drop out)
21. Learning activities in statistics have the same characteristics as mathematics, centering on mathematical calculations for all numbers. (-)

22. Statistics learning activities are more suitable for group work than individual work. (+)
23. Technology such as calculators, smartphones, computers, and the internet is appropriate for supporting statistics learning activities. (+)
24. A person's mathematical ability is required to learn statistics. (-)
25. Learning statistics does not require being good at mathematics. (+)
26. Statistics material is very applicable in all fields. (+)
27. Statistics material is taught only through mathematics subjects. (-)
28. Statistics gives meaning to mathematics because statistical material is motivating and fun for students. (+)
29. Statistics is not useful for mathematics lessons because it has no connection to other materials. (-)
30. If I become a teacher later, I will teach statistics using the same method as in other mathematics materials. (-)
31. If I become a teacher later, I will use a learning method oriented towards independent learning activities for students in statistics learning. (-)
32. If I become a teacher later, I will use a group learning method in statistics learning, where students carry out activities together in a group to achieve learning goals. (+)
33. Statistics learning is taught across subjects because of its applicability. (+)
34. Statistics learning is the responsibility of mathematics teachers. (-)
35. Everyone needs to learn statistics because it can be used as a tool to solve problems in everyday life. (+)
36. Based on my experience as a student at school, I have enough time to study statistics in mathematics. (-)
37. Based on my experience as a student at school, statistics should be given more hours per week, even if this means that other mathematics materials get less. (+, drop out)
38. Statistics is useful for solving various problems in the real world. (+)
39. Statistics is only a matter of calculating numbers and is not useful for everyday life. (-)
40. If I become a teacher later, I will prioritize students' procedural skills, such as making graphs and calculating statistical measurements. (-)
41. If I become a teacher, I will prioritize statistical thinking and make it the main goal of statistics learning. (+)
42. Types of data and measurement scales do not need to be taught because they are useless and confusing for students. (-)
43. All types of diagrams or graphs can be used to present all types of data, so there is no need to differentiate their use. (-)
44. Distribution and variability of data are not important in statistics learning, so there is no need to teach them. (-)

45. Population and sample do not need to be differentiated in statistics learning for school students. (-)

46. If I become a teacher, I will provide data for students to work on when teaching statistics. (-)

47. If I become a teacher, I will allow students to explore things around them to be researched and discussed in class. (+)

48. If I become a teacher, I will design statistics learning based on projects and group discussions. (+)

49. If I become a teacher, I will choose quantitative data (metric/continuous) in example questions, assignments, and tests. (-)

50. If I become a teacher, I will emphasize the context from which the data was obtained so that data analysis techniques can be determined and how to interpret them. (+)

51. If I become a teacher, I will teach students procedural skills to solve statistical problems, regardless of the context of the problem from which the data was obtained. (-)

52. If I become a teacher, I will utilize relevant computer or smartphone software to support statistics learning. (+)

53. If I become a teacher, I will avoid using technology to solve statistical problems because it encourages students to depend on technology. (-)

54. If I become a teacher, I will direct students to learn using the same method as when learning other mathematics materials by memorizing the formulas and giving lots of practice questions. (-)

55. If I become a teacher, I will prohibit students from using technology to solve statistical problems because it makes students dependent on technology. (-)

56. If I become a teacher, I will design statistics learning where students learn through data collection activities in their surroundings according to their interests. (+)

57. If I become a teacher, I will get students used to discussing various problem solutions in statistics learning. (+)

58. If I become a teacher, I will get students used to using software to solve statistical problems. (+)

59. If I become a teacher, I will design statistics learning where students learn in groups, work on projects, and present them. (+)

60. Statistics material is not related to other subjects. (-)