Journal of Pedagogical Research
Onlinefirst Article, 2024
https://doi.org/10.33902/JPR.202427097

Research Article

Development and calibration of an instrument measuring attitudes toward statistics using classical and modern test theory

Ezi Apino[1], Edi Istiyono[2], Heri Retnawati[3], Widihastuti Widihastuti[4] and Kana Hidayati[5]

[1]*Educational Research and Evaluation, Graduate School, Universitas Negeri Yogyakarta, Indonesia (ORCID: 0000-0001-9711-2807)*
[2]*Department of Physics Education, Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta, Indonesia (ORCID: 0000-0001-6034-142X)*
[3]*Department of Mathematics Education, Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta, Indonesia (ORCID: 0000-0002-1792-5873)*
[4]*Department of Fashion and Food Technology Education, Faculty of Engineering, Universitas Negeri Yogyakarta, Indonesia (ORCID: 0000-0001-8242-658X)*
[5]*Department of Mathematics Education, Faculty of Mathematics and Natural Sciences, Universitas Negeri Yogyakarta, Indonesia (ORCID: 0000-0002-9226-8500)*

Assessment of attitudes towards statistics [ATS] is needed to support the success of statistics education in tertiary institutions, so measuring instruments with high accuracy is required. However, existing instruments to measure ATS have not considered the use of technology as an essential variable affecting success in statistics education. The current study sought to fill this gap by developing a standardized instrument to measure ATS and considering aspects of technology use as a necessity for statistics education in the modern era. The study involved 367 students from various study programs spread across several universities in Indonesia as participants. To examine the quality of the instrument, we performed factor analysis, reliability estimation, and item calibration. We calibrated items based on classical test theory [CTT] and item response theory [IRT] using the graded response model [GRM]. Exploratory factor analysis [EFA] indicated three main factors (i.e., interest, difficulty, and value) for measuring attitudes toward statistics. Factor loading of each factor component > 0.45, indicating that all items contributed to the main factor. Cronbach's alpha coefficient of the three factors ranged from 0.784 to 0.929, indicating that the instrument was reliable. Item calibration based on CTT and IRT-GRM indicated that item performance was satisfactory regarding item endorsement and discrimination. In addition, the information function indicated that the instrument accurately measures attitudes from very low to very high levels. Overall, the psychometric properties of the instrument indicated that the instrument was valid, reliable, and feasible for use in practice and research in the field of education.

Keywords: Attitude toward statistics; Instrument calibration; Classical test theory; Item response theory; Statistics education

Article History: Submitted 2 February 2024; Revised 19 May 2024; Published online 7 June 2024

1. Introduction

One of the successes in statistics education is influenced by attitudes toward statistics [ATS] (Cladera et al., 2019; Fayomi et al., 2022; Hommik & Luik, 2017; Peiró-Signes et al., 2020; Soe et al., 2021). Through the ATS assessment, educators can plan appropriate statistics learning strategies (Saidi & Siew, 2019; Soe et al., 2021; Vanhoof et al., 2011). It illustrates that the ATS assessment must provide accurate information regarding how students perceive statistics. Accurate

Address of Corresponding Author

Ezi Apino, Colombo Street No. 1, Special Region of Yogyakarta, Indonesia.

✉ eziapino.2021@student.uny.ac.id

How to cite: Apino, E., Istiyono, E., Retnawati, H., Widihastuti, W., & Hidayati, K. (2024). Development and calibration of an instrument measuring attitudes toward statistics using classical and modern test theory. *Journal of Pedagogical Research.* Advance online publication. https://doi.org/10.33902/JPR.202427097

assessment results require accurate measurement tools (Vanhoof et al., 2011). In education, this measuring tool is known as the instrument. Thus, the availability of instruments to accurately measure ATS is needed to support the success of statistics education.

Currently, many researchers have developed instruments for measuring ATS. For instance, Statistics Attitude Survey (Roberts & Bilderback, 1980), Attitudes toward Statistics (Wise, 1985), Statistical Anxiety Rating Scale (Cruise et al., 1985), Multifactorial Scale of Attitudes toward Statistics (Auzmendi, 1991), Statistics Anxiety Inventory (Zeidner, 1991), Survey of Attitudes toward Statistics [SATS-28] (Schau et al., 1995), Survey of Attitudes toward Statistics [SATS-36] (Schau, 2003) are among these tools. Although these instruments have been widely used in research and statistics course practices in various disciplines, there is a need to examine whether these instruments are still relevant to the conditions of statistics education in the modern era. The main reason is that existing instruments have been developed over the past decade while statistics education progresses rapidly.

One of the crucial issues in statistics education in the modern era is the use of technology as a tool to make it easier for students to learn statistics. Kinds of literature have reported that using technology positively affects learning outcomes in statistics courses (e.g., Benková et al., 2022; Christmann, 2017; Koparan, 2018; Koparan & Rodríguez-Alveal, 2022; Larwin & Larwin, 2011; Sosa et al., 2011). The massive use of technology in statistics learning is strongly suspected of affecting students' perspectives on statistics (Brezavšček et al., 2016; Counsell et al., 2022; Counsell & Cribbie, 2020; Jatnika, 2015). However, previous instruments to measure ATS have not considered elements of technology use as an essential aspect that affects students' perceptions of statistics. For example, one of the most popular instruments for measuring ATS, the SATS-28 (Schau et al., 1995), only focuses on four aspects: (a) *affect*––describing positive and negative feelings related to statistics; (b) *cognitive competence*––describing attitudes about knowledge and intellectual skills applied to statistics; (c) *value*––describing attitudes about the usefulness, relevance, and value of statistics in personal and professional life; and (d) *difficulty*––describing attitudes about the difficulty of statistics as a subject. Schau (2003) added two new aspects to the SATS-28: *interest*––describing students' self-reported level of individual interest in statistics, and *effort*––describing the effort students put into learning statistics. Although Schau has added two new aspects to SATS-28, later known as SATS-36, the items developed have not considered the impact of technology use on students' perceptions of statistics. Therefore, developing a new instrument to measure ATS that considers the aspects of technology use and the aspects that have existed in previous instruments is necessary. Through the current study, we seek to fill this gap by obtaining an instrument to measure ATS relevant to statistics education in the modern era. The new instrument can be used to obtain accurate information regarding students' ATS profiles, which is helpful for statistics educators to plan and design effective statistics learning.

In order to produce high-quality and reliable measurement instruments, a calibration process for the instrument's psychometric properties is required. However, efforts made to evaluate the psychometric properties of existing ATS instruments, even when developing new instruments, are still dominated by calibrations based on Classical Test Theory [CTT]. For example, Hommik and Luik (2017) adapted the SATS-36 instrument for Estonian secondary school students. To test the quality of the instrument, they used Cronbach's alpha to estimate reliability and confirmatory factor analysis [CFA] to obtain validity evidence. In their study, item quality was evaluated based on factor loading values. Saidi and Siew (2019) adapted the SATS-36 instrument for rural secondary school students in Malaysia. They also used CFA and Cronbach's alpha to evaluate the validity and reliability of the instrument. Sharma and Srivastav (2021) also used CTT to evaluate the psychometric properties of an instrument to measure ATS for business school students in India, which they adapted from SATS-36. They used exploratory factor analysis [EFA] to obtain validity evidence and Cronbach's alpha to evaluate the reliability of the instrument they developed. Koparan (2015) developed an instrument to measure the ATS of students in middle school in Turkey. Although the items developed by Koparan were not adapted from previous

instruments, the validity and reliability of the instrument were still evaluated using CTT. Koparan used EFA to prove validity, while the instrument's reliability was estimated using Cronbach's alpha. Since studies investigating the structure of instruments to measure ATS mostly used CTT, Akour (2022) applied the Rasch model to evaluate the psychometric proportions of SATS-36. Although Akour's (2022) study showed an advance in evaluating the instrument's psychometric properties to measure ATS, studies using different approaches to examine the structure and psychometric properties of the instrument are still needed. It is necessary to overcome the weaknesses of CTT, including the Rasch model, in evaluating the psychometric properties of measurement instruments.

Item response theory [IRT] can be used to overcome weaknesses in the CTT calibration process (Aybek & Gulleroglu, 2021; Hambleton et al., 1991; Pardede et al., 2023; Zanon et al., 2016). In educational measurement, IRT was chosen as a superior alternative to CTT (DeVellis & Thorpe, 2022; Kyriazos & Stalikas, 2018). IRT is modeling responses to items of an educational and psychological measurement scale along with latent properties that determine how individuals respond to those items (Foster et al., 2017; Immekus et al., 2019). IRT has some attractive features for investigating the psychometric properties of an instrument, including (1) item characteristics are independent of the examinees; (2) scores describing test takers' abilities are independent of the test; (3) the model emphasizes item-level rather than test-level; (4) the model does not strictly require that scales be parallel to estimate reliability; and (5) the model describes a decision measure for each ability score, i.e. there is a functional relationship between the examinees and their ability levels (Hambleton et al., 1991; Hambleton & Swaminathan, 1985). However, despite these attractive features, the application of IRT in the context of developing and calibrating instruments to measure ATS is still rare. On the other hand, much of the literature recommends that a combination of CTT and IRT be conducted to prove the psychometric properties of a measurement instrument (e.g., DeVellis, 2017; Irwing & Hughes, 2018; Kyriazos & Stalikas, 2018). This research also seeks to fill this gap so that our findings can provide readers with insights into developing high-quality ATS instruments. Therefore, this study aims to produce an instrument to measure ATS that is valid, reliable, and calibrated based on CTT and IRT.

2. Materials and Methods

2.1. Questionnaire Development

The instrument for measuring ATS that we have developed was a questionnaire using a Likert-type scale with seven response categories (1 = Strongly disagree to 7 = strongly agree). Many kinds of literature have suggested that a Likert-type scale with seven response categories effectively obtained a reliable instrument (Comrey & Montag, 1982; Joshi et al., 2015). In this study, we adapted items from the Survey of Attitudes Towards Statistics [SATS] developed and copyrighted by Candace Schau. The first version of the SATS was developed in 1995 (Schau et al., 1995) and then updated in 2003 by adding two new subscales (Schau, 2003). The first version of the SATS consisted of 28 items that measured four components: affect, cognitive competence, difficulty, and value. This version is known as SATS-28. The second version is known as SATS-36 because it has 36 items. In this version, Schau adds two new components: interest and effort. Complete information on both versions of SATS can be seen at https://www.evaluationandstatistics.com/.

Apart from translating the items into Indonesian, we also adjusted the contextual aspects of several items. Because the ATS instrument developed by Schau has not considered aspects of the use of technology in statistics courses, and we believed that this aspect contributed significantly to students' attitudes toward statistics, we added six new items. The six items are as follows: "*I do not like statistical applications/software*"; "*Statistical applications/software are important in statistics courses*"; "*Statistics applications/software make computations/calculations in statistics easier*"; "*I like studying statistics using applications/software statistics*"; "*I find it easier to understand statistics using statistical applications/software*"; and "*statistical applications/software add to my burden in studying statistics*".

Finally, 42 items were generated in the questionnaire development phase. Of the 42 items, there are 21 unfavorable items. Sample unfavorable items, for example, "*I feel insecure when I have to solve statistics questions*"; "*I find it difficult to understand statistical concepts*"; "*Statistics is a complicated subject*". The favorable item samples, for example, "*I like statistics*"; "*I like taking statistics courses*"; "*Statistical formulas are easy to understand*". Furthermore, these items are assembled to get feedback from experts.

2.2. Pilot Study

After compiling all the items, we sent 42 items to experts for feedback. This study involved seven experts: three in statistics education and four in measurement. Each expert was asked to assess the relevance of each item to measure students' attitudes toward statistics. Three categories of assessments could be selected by experts, namely "relevant", "useful but less relevant," and "irrelevant". Experts were also allowed to provide qualitative input on each item being assessed. Quantitative data from expert assessment results were analyzed using Aiken's formula (V) (Aiken, 1980) to obtain evidence of content validity. Meanwhile, qualitative input was considered to improve the substance and grammar of the items. The V index of the 42 items ranged from 0.643 to 1 with a mean V = 0.908. It proved that all items were relevant for measuring attitudes toward statistics. In addition, we also made minor revisions based on expert feedback, especially for items with the lowest V index (items 10 and 13). After revising, we finally got 42 items that were ready to be used for the pilot study.

The pilot study administered the questionnaire online via Google Form (questionnaire link: https://cutt.ly/Sikap_Terhadap_Statistika) for three weeks in May 2023. There was no time limit for participants to complete the questionnaire, but under normal conditions, it was estimated that participants would only need 5–10 minutes to complete the questionnaire. Participants were only allowed to complete the questionnaire once. Participants could fill out the survey via their computer/laptop or smartphone. For anonymity reasons, participants may include nicknames, but respondents must fill in their gender, semester, study program, and university origin. We informed them that filling out the questionnaire was voluntary, so participants could withdraw if they were unwilling to complete it. We also informed them that we guaranteed the confidentiality of all data and participant identities so that nothing would affect their study.

2.3. Participants

In this study, participants were recruited from various universities in various regions of Indonesia. Respondents were a convenient and volunteer sample, so there were no special requirements for recruiting participants. We first sent invitations to participate in this study to our colleagues (lecturers) at various universities. We asked for their consent to distribute the questionnaires to their students at their respective universities. If they agreed, we provided two alternatives for administering the questionnaire: they asked students to complete the questionnaire directly during lectures, or they distributed the questionnaire link in online classes, WhatsApp Groups, or email. Finally, the number of participants who accessed and completed the questionnaire was 367 students from various academic levels and study programs across Indonesia. The complete demographics of the participants in this study are presented in Table 1.

2.4. Data Analysis

We conducted data analysis in five stages. First, we reversed the scores for the unfavorable items. For unfavorable items, score 1 (strongly disagree) is changed to 7, score 2 (agree) is changed to 6, and so on. Second, we performed a factor analysis using the Exploratory Factor Analysis [EFA] procedure. However, before performing the EFA procedure, we performed a Kaiser-Meyer-Olkin [KMO] measure of sampling adequacy and Bartlett's test of Sphericity to examine whether the data were suitable for factor analysis. The initial EFA procedure was carried out on data consisting of 42 items. Following the suggestion of Maskey et al. (2018), we used the varimax rotation and set a minimum factor loading of 0.45, indicating that the items significantly contributed to the factor.

Table 1
Participant demographics (n = 367)

Aspect demographics	*n*	*(%)*
Gender		
Male	89	24.3
Female	278	75.7
Year in program		
1st-year	137	37.3
2nd-year	184	50.2
3rd-year	32	8.7
4th-year	14	3.8
Academic level		
D3/D4 – Diploma	1	0.3
S1 – Undergraduate	258	70.3
S2 – Master	79	21.5
S3 – Doctoral	29	7.9
Study program		
Mathematics education	90	24.6
Psychology	66	18.0
Early childhood teacher education	58	15.8
Educational research and evaluation	56	15.3
Educational management	42	11.4
Primary teacher education	21	5.7
Sports science	9	2.5
Vocational education	9	2.5
Public administration science	6	1.6
Language education	3	0.8
Sufism	3	0.8
Culinary art	2	0.5
Science education	2	0.5

Following this criterion, items with a factor loading of less than 0.45 were excluded from the factor analysis, and the EFA procedure was repeated. The EFA procedure was iteratively continued to obtain a factor loading for all items of at least 0.45. The results of the final CFA analysis were used for further analytical procedures. Third, after the factors and their components have been formed, we estimated the reliability of each factor using Cronbach's alpha formula.

Fourth, we calibrated the items based on the Classical Test Theory [CTT] to determine each item's parameters of item endorsement and item discrimination. Fifth, we calibrated the items based on the Item Response Theory [IRT] using the Graded Response Model [GRM] (Samejima, 1969). Literature has suggested that GRM was suitable for data types using a Likert scale (Aybek & Gulleroglu, 2021; Aybek & Toraman, 2022; Zanon et al., 2016). IRT calibration focused on knowing each item's slope and location parameters. In addition, IRT calibration was also used to examine the fit of items with the measurement model (item fit) and obtain information functions. Because IRT assumed that the latent trait measured was unidimensional and there was no correlation between factors, IRT calibration was performed on each factor formed (Zanon et al., 2016).

All data analysis procedures used the help of RStudio software version 2023.3.1.446 (Posit Team, 2023). For the purposes of this study, we used several R packages which were available for free. For the EFA procedure, we used the 'psych' package (Revelle, 2023). We estimated reliability and calibration based on CTT using the 'CTT' package (Willse, 2018). Finally, for item calibration based on IRT, we used the 'mirt' package (Chalmers, 2012).

3. Results

3.1. Validity Evidence

Before carrying out the EFA procedure, we examined the sample's adequacy and whether the correlation matrix was not an identity matrix. The Kaiser-Meyer-Olkin [KMO] measure of sampling adequacy, which was 0.932, indicates that the sample size was sufficient for factor analysis. Bartlett's test of Sphericity was significant (χ^2 = 7292.883, *df* = 561, p < .01), indicating that the correlation matrix was not an identity matrix. Therefore, the data was suitable for factor analysis.

In the EFA procedure, we first analyzed 42 questionnaire items. The initial EFA found that several items had a factor loading < 0.45. The items were excluded, and the EFA procedure was repeated. We removed eight items through an iterative EFA process and generated 34 items for the final factor analysis. The final factor analysis produced three factors based on the eigenvalue criteria more than 1 (see Figure 1). It indicated three main factors formed by the 34 questionnaire items developed. These three factors accounted for 48% of the total variance (factor 1 = 22.3%, factor 2 = 18.8%, and factor 3 = 6.9%).

Factor 1, consisting of 16 items (see Table 2), has a factor loading ranging from 0.515 to 0.724. Item 16 contributed the highest factor loading (0.724), while item 17 contributed the lowest factor loading to factor 1. We named factor 1 "interest" because it represented students' interest in statistics and its learning. Factor 2 consisted of 14 items (see Table 2) with a loading factor ranging from 0.502 to 0.733. Item 12 contributed the highest factor loading (0.733), while item 30 contributed the lowest factor loading (0.502). We named this factor "difficulty", representing students' difficulties and feelings when taking statistics courses. Factor 3, consisting of four items (see Table 2), has a loading factor ranging from 0.472 to 0.668. Item 24 contributed the highest factor loading (0.668), while item 14 contributed the lowest factor loading (0.472). We named this factor "value", representing students' views on the usefulness of statistics.

Figure 1
Scree plot of the factor structure

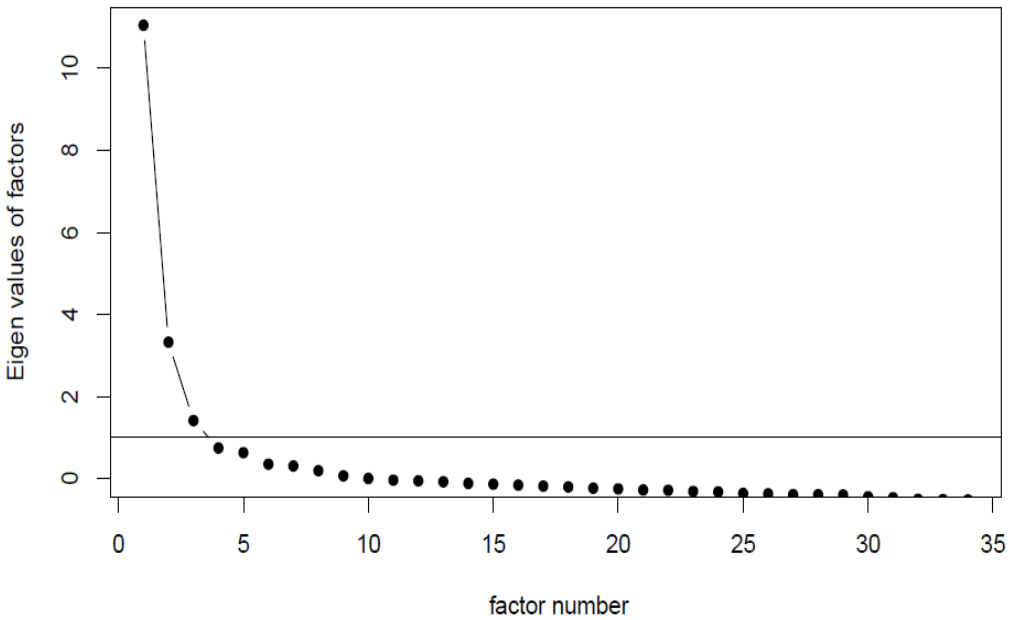


Table 2
Exploratory analysis of the factors in the attitude toward statistics questionnaire

No. item	*Questionnaire item*	*1*	*2*	*3*
	Interest – Feelings and interest in statistics and its learning			
16	I am interested in using statistics.	0.724		
19	I am interested in understanding statistical information.	0.724		
18	I try to study hard for every statistics examination.	0.715		
15	I am interested in studying statistics in more depth.	0.668		
20	Statistical applications/software are important in statistics courses*	0.663		
32	I enjoy taking statistics classes.	0.659		
9	Statistics skills will make me more employable.	0.648		
2	I worked hard in the statistics course.	0.642		
11	I am interested in being able to communicate statistical information to others.	0.629		
8	Statistics should be a part of my professional development.	0.622		
1	I am trying to complete all my statistics assignments.	0.618		
23	I like learning statistics using statistical applications/software*	0.602		
26	Statistical applications/software simplify computations in statistics*	0.585		
29	It is easier for me to understand statistics using statistical applications/software*	0.576		
3	I like statistics.	0.555		
17	I try to attend every statistics class.	0.515		
	Difficulty – Difficulties and feelings during statistics class			
12	I find it challenging to understand the concept of statistics.		0.733	
22	I feel frustrated when completing the statistics test in class.		0.732	
25	I feel pressured during statistics class.		0.705	
7	Statistics is a complicated subject.		0.660	
34	I am afraid of statistics.		0.654	
31	I made a lot of computational errors in statistics.		0.604	
5	I have difficulty understanding statistics because of my way of thinking.		0.598	
27	I can learn statistics easily.		0.582	
10	I have no idea what to do in a statistics course.		0.576	
4	I feel insecure when I have to complete statistical problems.		0.563	
28	Statistics is a subject that most people easily understand.		0.527	
13	Statistics are too technical for me.		0.520	
6	Statistical formulas are easy to understand.		0.504	
30	I understand statistical equations/formulas.		0.502	
	Value – Views on the usefulness of statistics			
24	I will not use statistics in my work.			0.668
21	Statistics is not useful for the profession in general.			0.655
33	Statistics are irrelevant to my life.			0.609
14	Statistical thinking is useless in my life outside of my job.			0.472

Note. Factor loadings smaller than 0.45 were removed; **new item.*

3.2. Reliability Estimation

We used Cronbach's alpha formula to estimate the reliability of the statistical attitude instrument. The Cronbach's alpha coefficient of the instrument (all items) was 0.938 (see Table 3), indicating that the developed instrument was reliable. The Cronbach's alpha coefficient for each subscale ranged from 0.784 to 0.929. It indicated that each sub-scale was considered a reliable measurement

scale. Subscale 1 (interest) has the highest reliability coefficient compared to subscale 2 (difficulty) and subscale 3 (value).

Table 3
The reliability of the instrument and the three subscales of ATS

	No. of items	*n*	*M*	*SD*	*Cronbach's alpha*
All items	34	367	166.169	28.131	0.938
Subscale 1 (interest)	16	367	87.297	14.628	0.929
Subscale 2 (difficulty)	14	367	57.714	14.619	0.905
Subscale 3 (value)	4	367	21.158	4.696	0.784

3.3. Item Statistics using Classical Test Theory

The item endorsement index for the 34 items of the attitude towards statistics instrument ranged from 0.515 to 0.897 (see Table 4). The item endorsement index ranged from 0.3 to 0.7, indicating respondents' acceptance or support for the item was in the "moderate" category. Meanwhile, the item endorsement index of more than 0.7 indicated that the respondent "easily" supports or accepts an item. The item with the highest endorsement index was item 17 ("I try to attend every statistics course"). In contrast, the item with the lowest endorsement index was item 28 ("Statistics is a subject that is easy for most people to understand"). Of the 34 items, 18 have an endorsement index in the "easy" category, while 16 other items have an endorsement index in the "moderate" category. Based on these results, it could be assumed that all items satisfactorily performed in item endorsement.

The item discrimination index of the 34 items of the attitude instrument towards statistics ranged from 0.384 to 0.778 (see Table 4). The discrimination index of more than 0.7 indicated that the item's performance to differentiate the respondent's attitude (high vs. low) level is "high". While the discrimination index ranged from 0.3 to 0.7, indicating that the item's performance to differentiate the respondents' attitude level was "medium". In this study, item 16 ("I am interested in using statistics") has the highest discrimination index. In contrast, item 21 ("Statistics are not useful for the profession/work in general") has the lowest discrimination index. Of the 34 items, only five had a discrimination index in the "high" category, while the other 29 had a "medium" category. Based on these results, it could be assumed that the performance of all items in differentiating levels of student attitudes toward statistics was satisfactory.

Table 4
Item endorsement and discrimination based on CTT

No. item	*Item Endorsement*	*Item Discrimination*	*No. item*	*Item Endorsement*	*Item Discrimination*
1	0.854	0.504	18	0.816	0.549
2	0.826	0.515	19	0.752	0.712
3	0.664	0.712	20	0.891	0.398
4	0.547	0.509	21	0.777	0.384
5	0.565	0.457	22	0.561	0.634
6	0.608	0.631	23	0.760	0.594
7	0.534	0.570	24	0.757	0.495
8	0.746	0.578	25	0.664	0.659
9	0.807	0.591	26	0.865	0.465
10	0.647	0.602	27	0.621	0.681
11	0.700	0.695	28	0.515	0.519
12	0.552	0.574	29	0.740	0.522
13	0.582	0.488	30	0.629	0.658
14	0.762	0.549	31	0.547	0.423
15	0.712	0.745	32	0.717	0.764
16	0.724	0.778	33	0.727	0.516
17	0.897	0.416	34	0.675	0.625

3.3. Item Parameter using Item Response Theory

Calibration using IRT for each subscale was carried out separately. It fulfills the IRT assumption that the measured construct must be unidimensional. In addition, this study also found no significant correlation between the subscales. The results of item calibration with the GRM model for the interest, difficulty, and values subscales were presented in Table 5, Table 6 and Table 7. Parameter slope (*a*) was an item discrimination that represents the ability of the item to distinguish between students with high and low levels of attitude. Items with a slope parameter between 0.65 and 1.34 indicated item discrimination was "moderate", between 1.35 and 1.69 indicated item discrimination was "high", and more than 1.70 indicated item discrimination was "very high" (Aybek & Gulleroglu, 2021; Baker, 2001). The location parameter (*b*) was an endorsement item representing how difficult or easy it was for students to accept or support an item. The location parameter between –2 to 2 indicated the endorsement item was "medium", more than 2 indicated the endorsement item was "high", and less than –2 indicated the endorsement item was "low" (Hambleton et al., 1991). Parameters b_1 to b_6 indicated endorsement items to select for each response category. When $b_1 < b_2 < ... < b_6$ indicated that each response category functioned appropriately.

On the interest subscale, the slope parameter ranged from 1.208 to 3.73, which indicated that the scale has a discrimination parameter in the "moderate" to "very high" category. The location parameter ranged from –2.326 to –0.185, which indicated that the subscale has an endorsement parameter in the "easy" to "moderate" category. Parameters b_1 to b_6 indicated that each response category for each item was functioning correctly. However, item 2 indicated that the participant did not select one of the response categories. The item fit test indicated that most items (15 out of 16) fit the GRM measurement model. Based on these results, it could be assumed that the item parameters on the interest subscale were satisfactory and feasible.

Table 5
Parameters slope (a) and location (b) of the "interest" subscale based on IRT

No. of item	*a*	*b*	b_1	b_2	b_3	b_4	b_5	b_6	*S_X2*	*df.S_X2*	*p*
1	1.359	−1.842	−5.037	−3.638	−2.736	−1.994	−0.991	0.416	84.587	68	.084
2	1.378	−1.153	−3.479	−2.712	−1.798	−0.668	0.849	–	87.909	75	.146
3	1.725	−0.232	−2.591	−1.852	−1.168	−0.332	0.698	1.922	75.059	80	.635
8	1.953	−0.713	−3.171	−2.257	−1.578	−0.728	0.026	1.070	100.324	81	.072
9	1.896	−1.032	−3.201	−2.743	−1.994	−1.289	−0.388	0.725	71.442	66	.302
11	2.383	−0.325	−2.494	−1.838	−1.251	−0.438	0.280	1.401	64.691	67	.557
15	3.162	−0.230	−2.171	−1.662	−1.090	−0.420	0.205	0.920	71.329	59	.130
16	3.730	−0.185	−2.192	−1.630	−1.215	−0.525	0.130	1.090	62.514	47	.064
17	1.208	−2.326	−4.908	−4.305	−3.572	−2.619	−1.738	−0.114	55.479	55	.457
18	1.871	−1.269	−4.018	−2.937	−2.358	−1.377	−0.527	0.806	89.539	67	.034*
19	3.434	−0.374	−2.348	−1.903	−1.407	−0.676	0.026	0.874	71.524	57	.093
20	1.491	−1.904	−4.147	−3.634	−3.109	−2.167	−1.421	−0.025	55.467	52	.345
23	1.865	−0.830	−3.387	−2.272	−1.774	−0.915	0.029	0.963	84.108	76	.245
26	1.455	−1.738	−3.998	−3.468	−3.056	−1.998	−1.136	0.309	67.614	60	.233
29	1.245	−1.015	−4.039	−3.175	−2.025	−1.008	0.188	1.602	110.225	92	.095
32	2.543	−0.390	−2.594	−1.888	−1.232	−0.526	0.230	1.125	77.338	71	.284

Note. * $p < .05$, the item did not fit with the measurement model.

On the difficulty subscale, the slope parameter ranged from 1.314 to 2.399, which indicated that the scale has a discrimination parameter in the "moderate" to "very high" category. The location parameter ranged from –0.242 to 0.637, which indicated that the subscale has an endorsement parameter in the "moderate" category. Parameters b_1 to b_6 indicated that each response category for each item was functioning properly. The item fit test indicated that most items (13 out of 14) fit the GRM measurement model. Based on these results, it could be assumed that the item parameters on the difficulty subscale were satisfactory and feasible to use.

On the value subscale, the slope parameter ranged from 1.859 to 2.589, which indicated that the scale has a discrimination parameter in the "very high" category. The location parameter ranged from –0.714 to –0.520, which indicated that the sub-scale has an endorsement parameter in the "moderate" category. Parameters b_1 to b_6 indicated that each response category for each item was functioning properly. The item fit test indicated that most items (3 out of 4) fit the GRM measurement model. Based on these results, it could be assumed that the item parameters on the difficult subscale were satisfactory and feasible to use.

Table 6

Slope (a) and location (b) parameters of the "difficulty" subscale based on IRT-GRM

No. of item	*a*	*b*	b_1	b_2	b_3	b_4	b_5	b_6	*S_X2*	*df.S_X2*	*p*
4	1.603	0.271	−2.310	−0.952	−0.136	0.495	1.099	2.197	91.272	106	.845
5	1.597	0.259	−2.235	−1.120	−0.333	0.346	1.049	2.502	125.507	110	.148
6	1.502	0.098	−2.574	−1.956	−0.834	0.257	1.215	2.846	73.971	86	.819
7	1.878	0.381	−1.955	−0.875	−0.084	0.576	1.195	2.072	99.975	98	.426
10	1.721	−0.242	−3.149	−1.775	−0.813	−0.018	0.696	1.811	83.152	91	.709
12	2.399	0.432	−1.971	−1.041	−0.248	0.400	1.056	2.573	97.089	82	.122
13	1.331	0.127	−2.801	−1.587	−0.760	0.543	1.336	2.756	111.095	105	.323
22	2.247	0.324	−1.604	−0.955	−0.248	0.352	0.890	1.797	89.125	96	.677
25	2.273	−0.110	−2.032	−1.405	−0.751	−0.168	0.394	1.235	104.487	88	.111
27	1.766	0.094	−2.187	−1.567	−0.906	0.093	1.089	2.333	103.985	86	.091
28	1.314	0.637	−2.283	−1.059	−0.133	0.979	1.889	3.523	103.465	106	.552
30	1.530	−0.063	−3.058	−1.945	−1.178	0.105	1.153	2.850	106.191	81	.032*
31	1.455	0.438	−2.504	−1.312	−0.413	0.661	1.662	3.325	119.580	98	.068
34	1.987	−0.211	−2.149	−1.532	−0.811	−0.201	0.317	1.167	89.456	100	.766

Note. * $p < .05$, the item did not fit with the measurement model.

Table 7

Slope (a) and location (b) parameters of the "value" subscale based on IRT-GRM

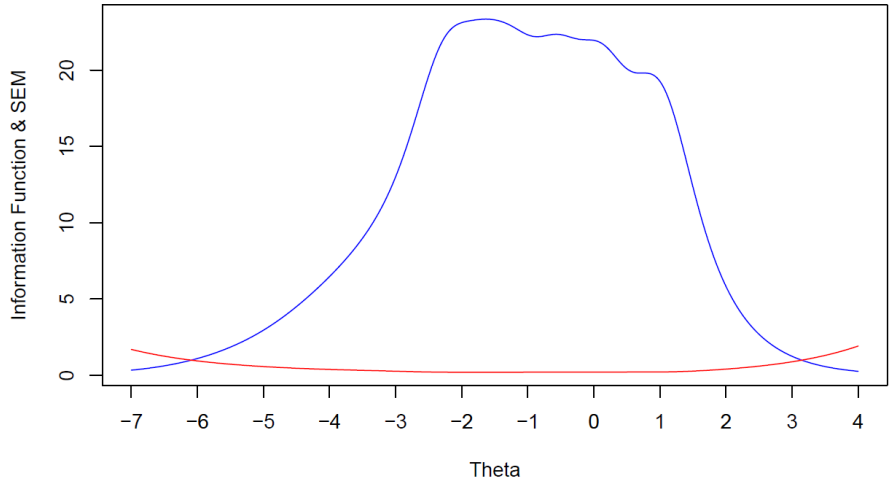
No. of item	*a*	*b*	b_1	b_2	b_3	b_4	b_5	b_6	*S_X2*	*df.S_X2*	*p*
14	1.859	−0.714	−2.723	−2.064	−1.465	−0.862	−0.283	0.843	46.791	33	.056
21	2.021	−0.700	−2.442	−2.134	−1.592	−0.902	−0.292	0.742	63.543	32	.001*
24	2.589	−0.564	−2.539	−2.068	−1.596	−0.649	−0.051	0.832	33.385	23	.075
33	2.151	−0.520	−2.653	−1.988	−1.515	−0.487	−0.091	1.115	37.783	28	.103

Note. * $p < 0.05$, the item did not fit with the measurement model.

Furthermore, the information functions for the interest, difficulty, and value subscales are presented in Figures 1, 2, and 3, respectively.

Figure 1

Information function of the "interest" subscale



The intersection of the information function graph (blue) and the standard error (red) on the "interest" subscale provided information that the subscale of this scale would be accurate when measuring students with attitude scores ranging from –6.1 to 3.1 (on a logit scale) (see Figure 1). When a student's attitude score is less than –6.1 (on a logit scale) or more than 3.1 (on a logit scale), the measurement error will be higher than the information provided by the "interest" subscale. Therefore, the "interest" subscale is appropriate for students with very low to high attitudes toward statistics.

Figure 2
Information function of the "difficulty" subscale

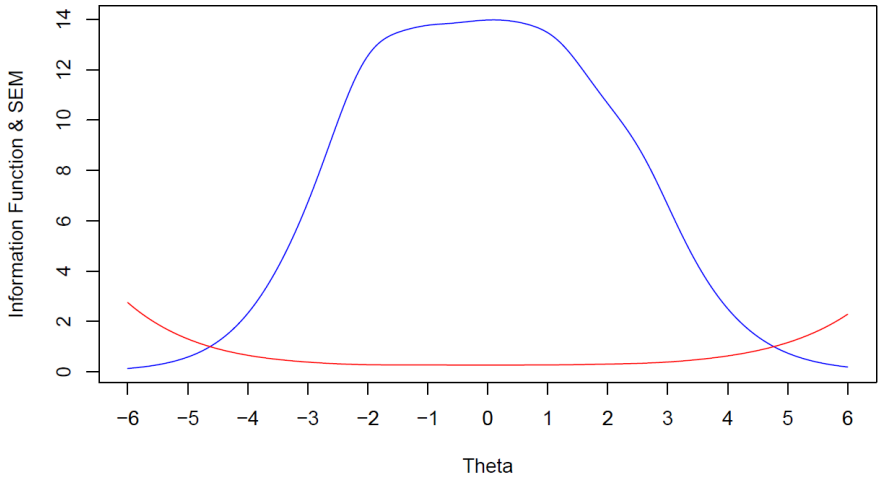
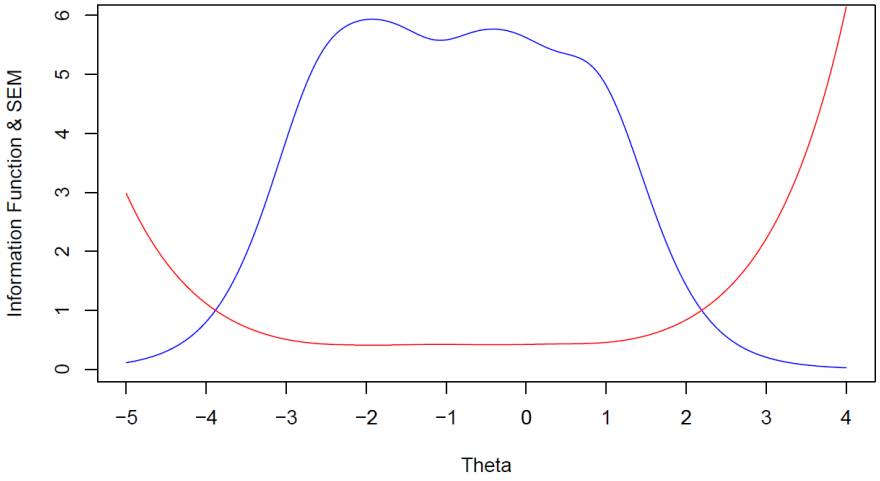


Figure 2 shows that the intersection of the information function graph (blue) and the standard error (red) on the "difficulty" subscale ranged from –4.7 to 4.8. It suggests that this subscale will be accurate when measuring students with attitude scores ranging from –4.7 to 4.8 (on a logit scale). When a student's attitude score is less than –4.7 (on a logit scale) or more than 4.8 (on a logit scale), the measurement error will be higher than the information provided by the "difficulty" subscale. Thus, the "interest" subscale is appropriate for students with very low to very high attitudes toward statistics.

Figure 3
Information function of the "value" subscale



Furthermore, Figure 3 shows that the intersection of the information function graph (blue) and the standard error (red) on the "value" subscale ranged from –3.9 to 2.1. It indicates that this subscale is accurate when measuring students with attitude scores ranging from –3.9 to 2.2 (on a logit scale). When a student's attitude score is less than –3.9 (on a logit scale) or more than 2.2 (on a logit scale), the measurement error will be higher than the information provided by the "value"

subscale. Thus, the "value" subscale is appropriate for students with very low to high levels of attitudes toward statistics.

4. Discussion

This study found that the instrument for measuring ATS measured three factors only: interest, difficulty, and value. It indicates that the structure of the instrument we developed differs from the original ATS instrument construct [SATS-36] developed by Schau (2003). This finding is unsurprising, considering that previous studies have also found inconsistent factors in the ATS instrument construct. For example, Koparan (2015) found only four factors that can be used to measure the ATS construct. Escalera-Chávez et al. (2014) found only five factors to measure the ATS construct. However, on the other hand, several studies have confirmed the construct on SATS-36, where six factors were found to be valid for measuring ATS (Judi et al., 2011; Saidi & Siew, 2019). Therefore, our findings make it clear that the literature regarding the ATS construct is debatable.

The fewer factors we found compared to other studies are also not surprising. Vanhoof et al. (2011) examined the SATS-36 structure between those using six and four factors. Their study found that several factors have a high correlation (affect, cognitive competence, and difficulty), so they can be combined into one factor. This merge does not cause much loss of information. The reason behind the merger is a high correlation between the three factors (affect, cognitive competence, and difficulty). Consistent with our findings, several items to measure affect, cognitive competence, and difficulty have a strong factor loading on one of the factors (factor 2, we named this factor "difficulty"). Thus, even though the ATS construct that we tested only involved three factors, it did not reduce the substance of the measurement. This finding further confirms that the construct of the ATS instrument needs to be readjusted by considering the performance of the items in predicting ATS.

One of the differences between the ATS instrument that we developed and the SATS-36 is the addition of items to measure perceptions regarding the use of statistical technology. In fact, these items do not form factors on their own. It indicates that perceptions of the use of technology cannot be separated from other aspects, such as interest in statistics and its learning. Factor loading for these items is also convincing (ranging from 0.576 to 0.663). It indicates that perceptions of the use of technology in statistics learning will influence attitudes toward statistics. This finding is consistent with the findings of previous research that the massive use of technology in statistics learning is strongly suspected of influencing students' perspectives on statistics (Brezavšček et al., 2016; Counsell et al., 2022; Counsell & Cribbie, 2020; Jatnika, 2015).

In this study, we not only focus on uncovering the quality of the instrument from the aspect of validity and reliability but also carry out analysis at the item level (item calibration). The calibration that we do is not only CTT-based but also IRT-based. As stated in the results section, overall, CTT-based calibration shows that item quality is satisfactory for measuring ATS. The results of IRT-based calibration also strengthen it. It provides evidence that the developed instrument is valid, reliable, and supported by standardized items. Therefore, there is a guarantee that the ATS instrument that we have developed is suitable for use in both research and statistical educational practice.

Calibration using IRT to examine the psychometric properties of instruments to measure ATS is a strength of our study. This procedure has rarely been used before, as most researchers prefer CTT to test the psychometric properties of ATS instruments (see Homik & Luik, 2017; Koparan, 2015; Saidi & Siew, 2019). The combination of CTT and IRT applied in this study is novel and innovative in the context of developing instruments to measure ATS. Consistent with experts' recommendations (i.e., DeVellis, 2017; Irwing & Hughes, 2018; Kyriazos & Stalikas, 2018), combining CTT and IRT is very beneficial to obtain high-quality instruments. By applying CTT and IRT as in the current study, we seek to provide best practices for obtaining standardized

instruments to measure ATS. We hope that our best practices will also inspire the development of other psychological measurement instruments.

This study seeks to update the ATS instrument to be more relevant to the needs of statistics education in the modern era. Much literature has suggested that the use of technology has a positive effect on learning outcomes in statistics learning (Christmann, 2017; Koparan, 2018; Larwin & Larwin, 2011; Tishkovskaya & Lancaster, 2012), but previous instruments to measure ATS have not accommodated this. We are trying to improve through this study so that the ATS instrument we have developed can measure students' perceptions of statistics more accurately and comprehensively. Statistics educators may use our instrument to investigate students' initial attitudes toward statistics. The strength of our instrument is that it allows statistics educators to know students' perceptions of the use of technology in statistics courses. This information is very beneficial for statistics educators to decide what and how this technology may be applied in statistics courses. It is essential because the use of technology may be a solution to overcome students' difficulties in learning statistics (Carver et al., 2016; Christmann, 2017; Koparan, 2018; Larwin & Larwin, 2011). Therefore, statistics educators can adequately plan and design their statistics courses. More broadly, other aspects, such as how students are interested in statistics, their potential difficulties in learning statistics, and the value of statistics to them, can be identified more accurately. More accurate and comprehensive measurement results will help statistics education stakeholders improve the quality of statistics education in the future.

5. Conclusion

This study found that the developed ATS instrument consisted of three factors, namely interest in statistics and learning (interest), difficulties and feelings of participating in statistics courses (difficulty), and views on the usefulness of statistics (value). All items (34 items) significantly contributed to measuring each factor. Reliability estimation provided evidence that the developed ATS instrument was reliable. The results of item calibration based on CTT and IRT-GRM provided empirical evidence that the quality of the items was satisfactory. The instrument information function provided additional information that the ATS instrument accurately measured student attitudes at very low to very high levels. All psychometric properties indicate that the developed ATS instrument is valid, reliable, and supported by high-quality items. We hope this instrument can contribute to improving the quality of statistics education in the future, especially in learning and teaching statistics in many fields. Considering the limitations of this study, we encourage that the psychometric properties of the instrument we developed can be re-examined in the future through longitudinal studies and considering cultural aspects.

6. Limitations and Future Directions

One of the limitations of this study is that the number of participants involved is still limited. Even though we have tried to recruit participants from various study programs, the participants we used have not been able to represent all the characteristics of students taking statistics courses. Therefore, future studies are needed to re-examine the structure and quality of the ATS instrument that we developed by involving a more significant number of participants representing various demographic characteristics. It helps improve the ATS instruments that we have developed and provides another perspective regarding their quality.

In addition, we have not conducted Confirmatory Factor Analysis [CFA] procedures in this study to obtain evidence of construct validity. Additional validity evidence, such as concurrent and predictive validity, has not been applied in this study. This is due to the limited access to the sample and the study duration. In future research, we hope the ATS constructs we found through EFA procedures can be empirically re-examined using CFA procedures. Also, future studies can investigate the concurrent validity or predictive validity of the instruments we developed to obtain complete validity evidence.

Finally, the cultural differences between countries raise concerns about the instrument's

applicability across different cultures. We encourage the instrument we developed to be adapted by researchers in other countries, according to their culture, when they are interested in using this instrument. Just as many other researchers have done (e.g., Akour, 2022; Homik & Luik, 2017; Koparan, 2015; Saidi & Siew, 2019), adapting the instrument to measure ATS according to the characteristics of the sample in their country is expected to reduce bias caused by differences in cultural background. The contributions from future studies are expected to enrich further the literature regarding measuring attitudes towards statistics.

Author contributions: All the authors contributed significantly to the conceptualization, analysis, and writing of this paper.

Availability of data and materials: The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Declaration of interest: No conflict of interest is declared by authors.

Ethics declaration: All participants gave informed consent to participate in this study. Data were analyzed in anonymized form only, and personal information that could lead to the identification of participants was removed. No additional ethical approval was required.

Funding: No funding source is reported for this study.

References

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement*, *40*(4), 955–959. https://doi.org/10.1177/001316448004000419

Akour, M. M. (2022). Rasch rating scale analysis of the survey of attitudes toward statistics. *Eurasia Journal of Mathematics, Science and Technology Education*, *18*(12), em2190. https://doi.org/10.29333/ejmste/12646

Auzmendi, E. (1991). *Factors related to statistics: A study with a Spanish sample* (ED333049). ERIC.

Aybek, E. C., & Gulleroglu, H. D. (2021). Attitudes toward pirated content: A scale development study based on graded response model. *Eurasian Journal of Educational Research*, *2021*(91), 127–144. https://doi.org/10.14689/ejer.2021.91.7

Aybek, E. C., & Toraman, C. (2022). How many response categories are sufficient for Likert type scales? An empirical study based on the item response theory. *International Journal of Assessment Tools in Education*, *9*(2), 534–547. https://doi.org/10.21449/ijate.1132931

Baker, F. B. (2001). *The basics of item response theory* (2nd ed). ERIC Clearinghouse on Assessment and Evaluation.

Benková, M., Bednárová, D., Bogdanovská, G., & Pavlíčková, M. (2022). Redesign of the statistics course to improve graduates' skills. *Mathematics*, *10*(15), 1–26. https://doi.org/10.3390/math10152569

Brezavšček, A., Šparl, P., & Žnidaršič, A. (2016). Factors influencing the behavioural intention to use statistical software: The perspective of the Slovenian students of social sciences. *EURASIA Journal of Mathematics, Science and Technology Education*, *13*(3), 953–986. https://doi.org/10.12973/eurasia.2017.00652a

Carver, R., Everson, M., Gabrosek, J., Horton, N. J., Lock, R., Mocko, M., Rossman, A., Rowell, G. H., Velleman, P., Witmer, J., & Wood, B. (2016). *Guidelines for assessment and instruction in statistics education (GAISE) college report 2016*. American Statistical Association.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Christmann, E. P. (2017). A comparison of the achievement of statistics students enrolled in online and face-to-face settings. *E-Learning and Digital Media*, *14*(6), 323–330. https://doi.org/10.1177/2042753017752925

Cladera, M., Rejón-Guardia, F., Vich-i-Martorell, G. À., & Juaneda, C. (2019). Tourism students' attitudes toward statistics. *Journal of Hospitality, Leisure, Sport & Tourism Education, 24*, 202-210. https://doi.org/10.1016/j.jhlste.2019.03.002

Comrey, A. L., & Montag, I. (1982). Comparison of factor analytic results with two-choice and seven-choice personality item formats. *Applied Psychological Measurement*, *6*(3), 285–289. https://doi.org/10.1177/014662168200600304

Counsell, A., & Cribbie, R. A. (2020). Students' attitudes toward learning statistics with R. *Psychology Teaching Review*, *26*(2), 36–56. https://doi.org/10.53841/bpsptr.2020.26.2.36

Counsell, A., Rovetti, J., & Buchanan, E. (2022). Psychometric evaluation of the students' attitudes toward statistics and technology scale (SASTSc). *Statistics Education Research Journal*, *21*(3), 1–18. https://doi.org/10.52041/serj.v21i3.29

Cruise, R. J., Cash, R. W., & Bolton, D. L. (1985). Development and validation of an instrument to measure statistical anxiety. In E. Team (Eds.), *American Statistical Association 1985 proceedings of the section on statistical education* (pp. 92–97). American Statistical Association.

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Sage Publication.

DeVellis, R. F., & Thorpe, C. T. (2022). *Scale development: Theory and applications* (5th ed.). Sage Publication.

Escalera-Chávez, M. E., García-Santillán, A., & Venegas-Martínez, F. (2014). Modeling attitude toward statistics by a structural equation. *Eurasia Journal of Mathematics, Science and Technology Education*, *10*(1), 23–31. https://doi.org/10.12973/eurasia.2014.1019a

Fayomi, A., Mahmud, Z., Algarni, A., & Almarashi, A. M. (2022). Arab and Malay students' attitudes toward statistics and their learning styles: A Rasch measurement approach. *Mathematical Problems in Engineering*, *2022*, Article 4144254. https://doi.org/10.1155/2022/4144254

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Springer. https://doi.org/10.1007/978-94-017-1988-9

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

Hommik, C., & Luik, P. (2017). Adapting the survey of attitudes towards statistics (SATS-36) for Estonian secondary. *Statistics Education Research Journal*, *16*(1), 228–239.

Immekus, J. C., Snyder, K. E., & Ralston, P. A. (2019). Multidimensional item response theory for factor structure assessment in educational psychology research. *Frontiers in Education*, *4*(45), 1–15. https://doi.org/10.3389/feduc.2019.00045

Irwing, P., & Hughes, D. J. (2018). Test development. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing* (pp. 1–47). Wiley. https://doi.org/10.1002/9781118489772.ch1

Jatnika, R. (2015). The effect of SPSS course to students' attitudes toward statistics and achievement in statistics. *International Journal of Information and Education Technology*, *5*(11), 818–821. https://doi.org/10.7763/IJIET.2015.V5.618

Joshi, A., Kale, S., Chandel, S., & Pal, D. (2015). Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, *7*(4), 396–403. https://doi.org/10.9734/bjast/2015/14975

Judi, H. M., Mohamed, H., Ashaari, N. S., & Wook, T. S. M. T. (2011). A structural validity study of students' attitudes towards statistics: An initial empirical investigation. *Journal of Quality Measurement and Analysis*, *7*(2), 163–171. http://www.ukm.my/ppsmfst/jqma/

Koparan, T. & Rodriguez-Alveal, F. (2022). Probabilistic thinking in prospective teachers from the use of TinkerPlots for simulation: Hat problem. *Journal of Pedagogical Research,* *6*(5), 1-16. https://doi.org/10.33902/JPR.202217461

Koparan, T. (2015). Development of an attitude scale towards statistics: A study on reliability and validity. *Karaelmas Journal of Educational Sciences*, *3*(1), 76–86. https://dergipark.org.tr/en/pub/kebd/issue/67216/1049124

Koparan, T. (2018). Examination of the dynamic software-supported learning environment in data analysis. *International Journal of Mathematical Education in Science and Technology*, *50*(2), 277–291. https://doi.org/10.1080/0020739X.2018.1494861

Kyriazos, T. A., & Stalikas, A. (2018). Applied psychometrics: The steps of scale development and standardization process. *Psychology*, *9*(11), 2531–2560. https://doi.org/10.4236/psych.2018.911145

Larwin, K. H., & Larwin, D. (2011). A meta-analysis examining the impact of computer-assisted instruction on postsecondary statistics education: 40 years of research. *Journal of Research on Technology in Education*, *43*(3), 253–278. https://doi.org/10.1080/15391523.2011.10782572

Maskey, R., Fei, J., & Nguyen, H. O. (2018). Use of exploratory factor analysis in maritime research. *Asian Journal of Shipping and Logistics*, *34*(2), 91–111. https://doi.org/10.1016/j.ajsl.2018.06.006

Pardede, T., Santoso, A., Diki, D., Retnawati, H., Rafi, I., Apino, E., & Rosyada, M. N. (2023). Gaining a deeper understanding of the meaning of the carelessness parameter in the 4PL IRT model and strategies for estimating it. *Research and Evaluation in Education*, *9*(1), 86–117. https://doi.org/10.21831/reid.v9i1.63230

Peiró-Signes, Á., Trull, Ó., Segarra-Oña, M., & García-Díaz, J. C. (2020). Attitudes towards statistics in secondary education: Findings from fsQCA. *Mathematics*, *8*(5), 1–17. https://doi.org/10.3390/MATH8050804

Posit Team. (2023). *RStudio: Integrated development environment for R* (2023.3.1.446). Posit Software. http://www.posit.co/

Revelle, W. (2023). *Psych: Procedures for psychological, psychometric, and personality research* (R package version 2.3.3). Cran-r Project. https://cran.r-project.org/package=psych

Roberts, D. M., & Bilderback, E. W. (1980). Reliability and validity of a statistics attitude survey. *Educational and Psychological Measurement*, *40*(1), 235–238. https://doi.org/10.1177/001316448004000138

Saidi, S. S., & Siew, N. M. (2019). Investigating the validity and reliability of survey attitude towards statistics instrument among rural secondary school students. *International Journal of Educational Methodology*, *5*(4), 651–661. https://doi.org/10.12973/ijem.5.4.651

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores (psychometric monograph no. 17). *Psychometrika*, *34*(1), 1–100. https://doi.org/10.1007/BF02290599

Schau, C. (2003). Students attitudes: The "other" important outcome in statistics education. *Joint Statistical Meetings-Section on Statistical Education*, *2003,* 3673–3683.

Schau, C., Stevens, J., Dauphinee, T. L., & Vecchio, A. Del. (1995). The development and validation of the survey of attitudes toward statistics. *Educational and Psychological Measurement*, *55*(5), 868–875. https://doi.org/10.1177/0013164495055005022

Sharma, A. M., & Srivastav, A. (2021). Study to assess attitudes towards statistics of business school students: An application of the SATS-36 in India. *International Journal of Instruction*, *14*(3), 207–222. https://doi.org/10.29333/iji.2021.14312a

Soe, H. H. K., Khobragade, S., Lwin, H., Htay, M. N. N., Than, N. N., Phyu, K. L., & Abas, A. L. (2021). Learning statistics: Interprofessional survey of attitudes toward statistics using SATS-36. *Dentistry and Medical Research*, *9*(2), 121–125. http://doi.org/10.4103/dmr.dmr_68_20

Sosa, G. W., Berger, D. E., Saw, A. T., & Mary, J. C. (2011). Effectiveness of computer-assisted instruction in statistics: A meta-analysis. *Review of Educational Research*, *81*(1), 97–127. https://doi.org/10.3102/0034654310378174

Tishkovskaya, S., & Lancaster, G. A. (2012). Statistical education in the 21st century: A review of challenges, teaching innovations and strategies for reform. *Journal of Statistics Education*, *20*(2), 1–55. https://doi.org/10.1080/10691898.2012.11889641

Vanhoof, S., Kuppens, S., Sotos, A. E. C., Verschaffel, L., & Onghena, P. (2011). Measuring statistics attitudes: Structure of the survey of attitudes toward statistics (SATS-36). *Statistics Education Research Journal*, *10*(1), 35–51. https://doi.org/10.52041/serj.v10i1.354

Willse, J. T. (2018). *CTT: Classical test theory functions* (R package version 2.3.3). Cran-r Project. https://cran.r-project.org/package=CTT

Wise, S. L. (1985). The development and validation of a scale measuring attitudes toward statistics. *Educational and Psychological Measurement*, *45*(2), 401–405. https://doi.org/10.1177/001316448504500226

Zanon, C., Hutz, C. S., Yoo, H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica*, *29*(1), 1–10. https://doi.org/10.1186/s41155-016-0040-x

Zeidner, M. (1991). Statistics and mathematics anxiety in social science students: Some interesting parallels. *British Journal of Educational Psychology*, *61*(3), 319–328. https://doi.org/10.1111/j.2044-8279.1991.tb00989.x