International Journal of Science and Mathematics Education (2020) 18:771–790
https://doi.org/10.1007/s10763-019-09985-0

Diagnosing Preservice Teachers' Understanding of Statistics and Probability: Developing a Test for Cognitive Assessment

Muhammet Arican[1] · Okan Kuzu[1]

Received: 26 February 2018 / Accepted: 8 May 2019 / Published online: 10 June 2019

Abstract

This study investigates preservice middle school mathematics teachers' understanding of statistics and probability and provides a cognitive diagnostic assessment of their strengths and weaknesses on these subjects. A statistical reasoning test that included 15 multiple-choice and 5 open-ended items was developed from the perspective of the log-linear cognitive diagnosis model, which is a general form of the cognitive diagnosis models. The statistical reasoning test was applied to 456 preservice teachers from 4 universities in 3 different regions of Turkey. The collected data were analyzed using the Mplus 6.12 software, and diagnostic feedback on the preservice teachers' responses was provided based on the findings. The analysis suggested that although many preservice teachers were able to represent and interpret the given data, most experienced difficulty in drawing inferences about populations based on samples, selecting and using appropriate statistical methods, and understanding and applying the basic concepts of probability. In addition, preservice teachers had difficulty answering open-ended items. Implications for teaching are also discussed.

Keywords Cognitive assessment · Diagnostic classification models · Preservice teacher education · Statistics and probability · Test design

Introduction

Statistics and probability have been a primary focus of mathematics education for some decades (Franklin, Kader, Mewborn, Moreno, Peck, Perry, & Scheaffer, 2007). While

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10763-019-09985-0) contains supplementary material, which is available to authorized users.

✉ Muhammet Arican
muhammetarican@gmail.com

[1] Department of Mathematics and Science Education, Kırşehir Ahi Evran University, Kırşehir, Turkey

the primary focus of statistics is the handling of data using different data collection and analysis methods, probability is mainly concerned with studying the likelihood of an event's occurrence. Although these two subjects are related to different areas of mathematics, they can also be considered to go hand in hand. Statistics and probability have become a focus of interest for many countries and have been included in mathematics teaching programs and in the learning standards of leading education organizations (e.g. the National Council of Teachers of Mathematics [NCTM], the National Assessment of Educational Progress [NAEP], the College Board) because of their practical application and usefulness in real life, as well as being utilized as tools in other disciplines (e.g. economics, consumer science, physical education) (Batanero & Díaz, 2010; Franklin et al., 2007; Jones, 2005; Makar & Rubin, 2009; Shaughnessy, 2007; Watson, 2006).

Although statistics and probability are considered to be very important in real life and in various disciplines, some problems have been observed in the literature with regard to the learning and teaching of these two subjects (Batanero & Díaz, 2012). Many teachers have difficulty constructing statistical knowledge of their students since they themselves have not had the opportunity to develop accurate knowledge of the principles and concepts underlying the practices of data analysis (Franklin et al., 2007). Therefore, as Batanero and Díaz (2010) stated, students may graduate from secondary school education with little understanding of the basic principles underlying data analysis which explains the problems that preservice teachers (PSTs) encounter in undergraduate statistics courses. In addition, some researchers (e.g. Batanero & Díaz, 2012; Batanero, Godino, & Roa, 2004; Franklin & Mewborn, 2006; Stohl, 2005) have drawn attention to the inefficacy of university programs in training PSTs in the areas of statistics and probability. Thus, there is a significant need to investigate PSTs' understanding of statistics and probability in order to eventually help educators evaluate their ability to teach these subjects.

Determining what teachers need to know in order to teach school mathematics has been an important area of mathematics education research (Bradshaw, Izsák, Templin, & Jacobson, 2014). Although researchers (e.g. Ball, Lubienski, & Mewborn, 2001; Hill, Schilling, & Ball, 2004) have conducted studies to understand teachers' knowledge for teaching mathematics, only a few recent studies (e.g. Baumert et al., 2010; Hill, Rowan, & Ball, 2005) have attempted to measure the link between teachers' mathematical knowledge and student achievement (Bradshaw et al., 2014). However, the measures used in these studies have relied upon unidimensional item response theory (IRT) models, which are not that useful in identifying the multidimensional characteristics of a research topic (Bradshaw et al., 2014). Therefore, inspired by the recent developments in measuring teachers' knowledge, the goal of the current study is to develop a multidimensional test to examine knowledge that PSTs have of statistics and probability contents.

This study investigates Turkish preservice middle school mathematics teachers' mastery of four fundamental cognitive skills, which are also referred to as attributes, that are required in solving middle school statistics and probability problems. These four cognitive skills are determined as follows: representing and interpreting data, drawing inferences about populations based on samples, selecting and using appropriate statistical methods to analyze data, and understanding and applying basic concepts of probability. Furthermore, this study provides diagnostic feedback on the PSTs' particular strengths and weaknesses by analyzing their responses to various statistics and probability problems using the log-linear cognitive diagnosis model (LCDM) (e.g. Henson, Templin, & Willse, 2009).

Background

Literature Review

Students' and PSTs' errors and difficulties in statistics and probability have been reported in the literature. Studies that have focused on the development of students' statistical reasoning and their understanding of statistical concepts have revealed many difficulties even with moderately elementary concepts (Garfield & Ben–Zvi, 2008). Regarding these difficulties, researchers (e.g. delMas, Garfield, Ooms, & Chance, 2007; Shaughnessy, 2007) noted that "ideas of probability and statistics are very difficult for students to learn and often conflict with many of their own beliefs and intuitions about data and chance" (Garfield & Ben–Zvi, 2008, p. 51). For instance, students were reported to have difficulty reading and interpreting data (e.g. Curcio, 1987; delMas, Garfield, & Ooms, 2005; Li & Shen, 1992), understanding the measures of central tendency and measures of spread (e.g. Brown & Silver, 1989; Zawojewski & Heckman, 1997; Zawojewski & Shaughnessy, 2000), and understanding dependent and independent events and calculating the probability of these events (e.g. Dereli, 2009).

Similar to students, both PSTs and in-service teachers experience "many difficulties understanding and teaching core ideas of probability and statistics" (Garfield & Ben–Zvi, 2008, p. 34). In fact, as stated by Garfield and Ben-Zvi (2007), inappropriate reasoning about statistical concepts and ideas is widespread and similar at all age levels, and changing this inappropriate reasoning is quite difficult, even after an instruction to statistics has been provided. Moreover, Garfield and Ahlgren (1988) stated that the majority of university students fail to understand many of the concepts they are studying in introductory statistics courses. For example, on a college-level statistics course, Mathews and Clark (2003) interviewed eight A-grade PSTs and observed the PSTs' lack of understanding when it came to mean (confusing mean with mode) and standard deviation concepts and their heavy reliance on algorithmic procedures. Similarly, O'Connell (1999) observed university students' misjudgments on independent events and their misrepresentation of problem situations when calculating the probability of dependent and independent events. Moreover, Leavy (2010) reported PSTs' difficulties developing pedagogical contexts for advancing their future students' informal inferential reasoning. In addition, while reporting in-service teachers' difficulties, Stohl (2005) noted "conceptual complexity of probability as a major issue for the development of teachers' knowledge" (p. 350).

Besides detecting students' difficulties in statistics and probability, in recent years, there has been significant attention paid to determining teachers' mathematical knowledge for teaching (Fennema & Franke, 1992). Regarding teachers' knowledge in statistics and probability, several researchers (e.g. Burgess, 2007; Groth, 2007; Groth & Bergner, 2006; Lee & Hollebrands, 2011) have attempted to determine the statistical knowledge needed for teaching. Although these studies provided valuable information on the knowledge required for teaching statistics and probability, they did not aim at offering constructive feedback to teachers in terms of understanding which parts of this multifaceted domain they were strong or weak. Thus, this study examines the knowledge that PSTs need for teaching middle school statistics and probability topics and provides diagnostic feedback for their particular strengths and weaknesses.

Cognitive Diagnosis Models

For many years, researchers have been employing traditional testing and assessment techniques in order to obtain examinees' total, average, or individual scores to assess their knowledge (Ranjbaran & Alavi, 2017; Sen & Arican, 2015). Although these scores provide useful insights into the examinees' overall performance in terms of subject areas, they offer no diagnostic information about their strengths and weaknesses in these subject areas. Therefore, in recent years, researchers have been using cognitive diagnosis models (CDMs) to obtain diagnostic information about students' and teachers' test scores (Templin & Bradshaw, 2013).

CDMs, also known as diagnostic classification models (DCMs), are a family of psychometric models that categorize examinees as either a master or nonmaster of an attribute, which is a categorical latent variable, according to their test item responses. "CDMs predict the probability of an observable categorical response from unobservable (i.e. latent) categorical variables" (Ravand & Robitzsch, 2015, p. 2). The term attribute is used to define cognitive skills required in solving a specific item (Common Core State Standards Initiative, 2010). The intent of CDMs is to provide diagnostic feedback with regard to these carefully defined attributes (Bradshaw et al., 2014).

CDMs can be grouped into three categories: compensatory models, noncompensatory models, and general models (Table 1). In the compensatory models (e.g. DINO, C-RUM), mastery of one or some of the attributes required to achieve a correct answer can compensate for nonmastery of the remaining attributes. Hence, mastery of at least one attribute is necessary to achieve a correct answer. On the other hand, in the noncompensatory models (e.g. DINA, NC-RUM), the lack of mastery of one attribute cannot be completely compensated by the mastery of the remaining attributes in terms of item performance. Therefore, possession of all attributes is required to achieve a correct answer. However, the general models (e.g. GDM, LCDM, G-DINA) allow for both compensatory and noncompensatory relationships within the same test (Ravand & Robitzsch, 2015).

In the current study, the statistical reasoning test (SRT) was developed from the perspective of a general CDM, using the LCDM (Henson et al., 2009). Using a generalized linear model, the LCDM maps item responses onto latent attributes (Bradshaw et al., 2014) and therefore helps researchers in detecting patterns of attribute mastery. The LCDM provides more flexibility to the researchers because it can model attribute effects on each item response in a compensatory or noncompensatory manner, which depends on the size and direction of the LCDM item parameters (Bradshaw et al., 2014). Hence, we decided to develop the SRT from the LCDM perspective. The LCDM estimates the probability of an examinee's (e) correct response for an item (i), which measures two attributes ($\alpha_{e1}$and α_{e2}) by applying the following equation:

$$\ln\left(\frac{P(X_{ei}=1|\alpha_e)}{P(X_{ei}=0|\alpha_e)}\right)=\lambda_{i,0}+\lambda_{i,1(1)}(\alpha_{e1})+\lambda_{i,1(2)}(\alpha_{e2})+\lambda_{i,2(1*2)}(\alpha_{e1}\alpha_{e2}) \quad (1)$$

Table 1 Types of cognitive diagnosis models

CDM type	Examples	Description	Author(s)
Compensatory	(a) Deterministic input, noisy-or-gate model (DINO)	Mastery of one or some attributes required to achieve a correct answer can compensate for nonmastery of other attributes (possession of at least one attribute is enough for mastery of an item).	Templin and Henson (2006)
	(b) Compensatory reparameterized unified model (C-RUM)		Hartz (2002)
Noncompensatory	(a) Deterministic input, noisy-and-gate model (DINA)	In noncompensatory models, lack of mastery of one attribute cannot be completely compensated by mastery of other attributes in terms of item performance (possession of all attributes is required for mastery of an item).	Junker and Sijtsma (2001)
	(b) Noncompensatory reparameterized unified model (NC-RUM)		DiBello, Stout, and Roussos (1995); Hartz (2002)
General	(a) General diagnostic model (GDM)	Allows for both compensatory and noncompensatory relationships.	von Davier (2005)
	(b) Log-linear cognitive diagnostic model (LCDM)		Henson et al. (2009)
	(c) Generalized DINA (G-DINA)		de la Torre (2011)

In Eq. 1, the parameter $\lambda_{i,\,0}$ is the intercept of the LCDM and represents the predicted log-odds of a correct response for examinees who have not mastered Attribute 1 or Attribute 2 (Bradshaw et al., 2014). Parameter $\lambda_{i,\ 1(1)}$ is the simple main effect that represents the predicted log-odds of a correct response for examinees who have mastered Attribute 1, but not Attribute 2. Similarly, parameter $\lambda_{i,\,1(2)}$ is the simple main effect that represents the predicted log-odds of a correct response for examinees who have mastered Attribute 2, but not Attribute 1. Finally, parameter $\lambda_{i,\ 2(1*2)}$ is the interaction effect that represents the predicted log-odds of a correct response for examinees who have mastered both Attribute 1 and Attribute 2.

In recent years, some studies (e.g. Choi, Lee, & Park, 2015; Dogan & Tatsuoka, 2008; Im & Park, 2010; Lee, Park, & Taylan, 2011; Sen & Arican, 2015; Toker & Green, 2012) have compared students' performance within international large-scale assessments using CDMs. Two of these studies, Dogan and Tatsuoka (2008) and Sen and Arican (2015), focused on understanding Turkish students' performance in mathematics. Dogan and Tatsuoka (2008) compared Turkish and American 8th grade students' responses to mathematics questions from the Third International Mathematics and Science Study-Repeat (TIMSS-R) 1999. Similarly, Sen and Arican (2015) compared Turkish and South Korean 8th grade students' performance on the TIMSS 2011. Both studies noted that Turkish students were weak in statistics and probability in comparison with their American and Korean peers. In addition, they reported Turkish students' weaknesses in solving open-ended problems. Therefore, the authors of the current study decided to investigate Turkish PSTs' understanding of statistics and probability.

Methods

Test Development and Participants

The first phase of the quantitative part of this current study included the development of a preliminary SRT. In order to develop this test, first, we determined the critical cognitive skills required in solving middle school statistics and probability problems. In doing so, we examined the relevant published literature, the Turkish middle school mathematics curriculum, and we referred to the TIMSS 2011 framework, the Common Core State Standards (CCSS), and the National Council of Teachers of Mathematics (NCTM) standards. Hence, we were able to determine a list of skills (Table 2), some of which are presented as our sub-attributes. Next, we obtained two mathematics teachers' opinions on this list of skills, who provided feedback on the suitability of these skills based on the teaching and learning standards of the Turkish middle school curriculum. Furthermore, we sought constant feedback from a psychometrician with regard to the test development procedure.

After receiving expert feedback, we determined the four core skills (Table 2) that served as our attributes. Finally, we designed a preliminary test, which included 20 multiple-choice and five open-ended items, around the four attributes by following the problem examples provided in the Turkish middle school grades' mathematics curricula, Turkish national tests, and the TIMSS 2011 study. Thus, we were able to obtain construct validity evidence for the preliminary SRT. One of the strengths of CDMs is that they provide highly reliable examinee estimates with a small number of items.

In the second phase, conducted during the Spring 2016 semester, the preliminary SRT was administered to 79 PSTs enrolled at a university located in central Turkey. The PSTs participated on a voluntary basis, and small incentives were provided in recognition of their participation. The purpose of testing the PSTs with the preliminary SRT was to detect any problems with the items. In our analysis, we detected issues with five items, which suggested that removing these items would improve the overall quality of the SRT. Two items did not discriminate between the higher and lower performing groups in which the item discrimination indices were calculated as .11 and −.02, respectively. Using a discrimination cutoff score of .20, we discarded these two items. Moreover, two items were considered to be too difficult (i.e. difficulty indices were .20 or below) in which item difficulty indices were calculated as .16 and .08, respectively. Finally, one item was too easy in which the item difficulty index was calculated as .87 (revising items with item difficulty index .80 or above suggested). Thus, we also removed these three items in the final form of the SRT.

Using the four attributes in Table 2, the final test items were independently coded for their measured attribute by three mathematics educators and two middle school mathematics teachers with 5 and 8 years of teaching experience. This step of the study was crucial because according to Rupp and Templin (2008), correct alignment of items with attributes helps researchers obtain CDM classification accuracy. Three mathematics educators and two mathematics teachers were trained about the coding process. We obtained our *Q-matrix* (Tatsuoka, 1985), which expressed the item-attribute alignment, using the independent codes received from the three mathematics educators and the two mathematics teachers. In the Q-matrix, code "1" represents an item measuring corresponding attribute(s), whereas code "0" represents an item not measuring

Table 2 Critical skills (attributes) required in solving Turkish middle school statistics and probability problems

Attributes			
A1. Representing and interpreting data.	**A2.** Drawing inferences about populations based on samples.	**A3.** Selecting and using appropriate statistical methods to analyze data.	**A4.** Understanding and applying basic concepts of probability.
Sub-attributes			
A1.1. Reading, organizing, and displaying data. **A1.2.** Representing and interpreting data using appropriate representation methods. **A1.3.** Interpreting categorical and quantitative data. **A1.4.** Collecting, organizing, and displaying relevant data to answer questions.	**A2.1.** Making inferences based on data. **A2.2.** Using random sampling to draw inferences about a population. **A2.3.** Drawing informal comparative inferences about two populations. **A2.4.** Making inferences and justifying conclusions. **A2.5.** Developing and evaluating inferences and predictions that are based on data.	**A3.1.** Explaining and calculating measures of central tendency and dispersion. **A3.2.** Summarizing and describing distributions. **A3.3.** Investigating patterns of association in bivariate data. **A3.4.** Developing an understanding of statistical thinking and variability.	**A4.1.** Explaining an event and the probability of its occurrence. **A4.2.** Investigating chance processes and developing, using, and evaluating probability models. **A4.3.** Explaining dependent and independent events and calculating probabilities of these events. **A4.4.** Explaining permutation and combination concepts and calculating permutation and combination in a given event. **A4.5.** Understanding conditional probability and the rules of probability. **A4.6.** Using the rules of probability to compute probabilities of discrete and compound events. **A4.7.** Using probability to make decisions.

corresponding attribute(s). In the Q-matrix, code "1" was applied where at least three coders agreed that an item measured the corresponding attribute, else a code "0" was applied for that attribute. Therefore, in the final SRT, five items measured three attributes, seven items measured two attributes, and eight items measured a single attribute.

In the last phase of the study, conducted during the Fall 2016 semester, we administered the final form of the SRT, which included 15 multiple-choice and five open-ended items, to 456 PSTs (315 females, 108 males, and 33 gender unspecified) from four Turkish universities. Among these 456 PSTs, 106 were freshmen (1st year), 150 were sophomores (2nd year), 158 were juniors (3rd year), and 42 were seniors (4th year). The four universities were located in three different regions of Turkey. Moreover, the universities had different academic rankings (one was high-ranking, one was low-ranking, and two were of moderate-ranking). When deciding the academic ranking of these four universities, we considered the last 5 years' average university entrance scores of their students. In 2016, there were a total of 67 state universities with a middle school mathematics teacher education program in Turkey. According to the students' average university entrance scores, we randomly chose one university from the top 17 universities, two universities that were ranked from 18 to 50, and one university from the lowest ranked 17 universities. The PSTs' correct and incorrect responses to the test items were coded 1 and 0, respectively, and missing responses were coded as 9.

Sample Items

Two items, Item 1 and Item 18, were selected as examples. In Item 1, the PSTs calculated the probability of a mouse who wanted to find some cheese located at the end of a maze. In this item, we expected the PSTs to understand that the mouse chose paths in the maze with equal likelihood. According to our Q-matrix, this item measured Attributes 1, 2, and 4 (see Table 3). In order to solve this item, first, the PSTs needed to interpret the information provided in the figure and draw inferences from it. Next, they were required to understand and apply the basic concepts of probability. In this item, an expert PST, who had mastered all three attributes, could see that the mouse had to choose one of the three paths first, and then chose one of the two paths. Therefore, applying the basic concept of probability, a person could recognize that multiplying one-third by one-half would eventually yield the correct answer.

In Item 18, which was an open-ended item, the PSTs calculated the probability of having at least one adult and one child in a group of four people randomly selected from a pool of six adults and four children. According to the Q-matrix, this item only measured Attribute 4, understanding and applying the basic concepts of probability. Hence, in order to solve this item, the PSTs had to calculate the possible outcomes of selecting at least one adult and one child from the pool of six adults and four children. Next, they had to divide this number by the total number of possible outcomes in order to calculate the probability.

Results

In the current study, estimates from the data were calculated using Mplus 6.12 (Muthen & Muthen, 1998–2011) statistical software. We conducted the LCDM analysis in Mplus using the Q-matrix and coded responses. The LCDM code for the Mplus was generated using the "Mplus Input Generator," which was written by Dr. Olga Kunina-Habenicht in order to be used with the R 3.3.3 program (R Core Team, 2017). The first

Table 3 The Q-matrix

Attributes	Items																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
A1	1	0	0	1	0	0	1	0	0	1	1	1	1	0	0	1	1	0	0	0	9
A2	1	0	0	0	0	1	1	0	0	1	1	1	1	1	0	0	1	0	0	1	10
A3	0	0	0	0	0	1	0	0	0	1	1	1	0	1	0	1	0	0	1	1	8
A4	1	1	1	1	1	0	1	1	1	0	0	0	0	0	1	0	0	1	0	0	10

Code "1" represents items measuring the corresponding attribute. Code "0" represents items not measuring the corresponding attribute step of the data analysis was to determine the best-fit LCDM model. To determine this best-fit model, we compared several log-linear structural model parameterizations. In Table 4, we provide model comparison indices for three LCDMs. Model A represents the one-way structural model, Model B represents the two-way structural model, and Model C represents the three-way structural model. Model A only included intercepts (i.e. $\lambda_{i,\ 0}$ in Eq. 1) and the main effects (i.e. $\lambda_{i,\ 1(1)}$ and $\lambda_{i,\ 1(2)}$), whereas Model B included intercepts, main effects, and two-way interaction effects (i.e. $\lambda_{i,\ 2(1*2)}$), and Model C, which was the full model, included intercepts, main effects, and both two- and three-way interaction effects.

Conducting a chi-square difference test using the log-likelihood values with the MLR (Maximum Likelihood Robust) estimator (Satorra & Bentler, 2010), these three models were compared with each other. According to Werner and Schermelleh-Engel (2010), a significant chi-square difference ($p < .05$) suggests that the larger model with more freely estimated parameters (i.e. main, two-way, and three-way interaction effects) would better fit the data than the smaller model with less freely estimated parameters. Although an insignificant p value indicates that both models statistically fit the data relatively well, Rupp, Templin, and Henson (2010) suggest that information criteria (i.e. AIC and BIC) should be used to select the most parsimonious model.

Based on the explanations above, Table 4 suggests that Model B fits the data better than Model A ($p = .000 < .05$). Next, comparing Model B with Model C, we determined that Model B was a better fit for the data than Model C ($p = .127$). Therefore, we decided that Model B was a better fit for the data than the one-way and three-way structural models. Model B included all main and two-way interaction effects. The next step was to remove nonsignificant two-way interaction effects, which did not contribute to the estimation of attribute mastery profiles, in order to determine the best-fit model within Model B. In Model B, there were a total of 22 two-way interaction effects, and the LCDM analysis showed that 14 of them were nonsignificant. Hence, we removed these two-way interaction effects one by one, starting from the interaction effect with the highest p value. After removing the interaction effect with the highest p value, we ran the model again and checked the model fit indices (i.e. AIC, Akaike's information criteria; BIC, Bayesian information criteria; and SSA BIC, sample size adjusted Bayesian information criteria) in order to determine if removing this interaction effect improved the model fit. If this new model yielded smaller AIC, BIC, and SSA BIC indices, then we proceeded with this new model. We observed that removing 14

Table 4 Model modifications and fit indices for one-way, two-way, and three-way models and the final model

Model	Description	AIC	BIC	SSA BIC	LL	NPR	Chd	*df*	*p*
A	One-way structural model	10,059.03	10,355.85	10,127.35	−4957.51	72	–	–	–
B	Two-way structural model	10,042.48	10,429.99	10,131.67	−4927.23	94	37.98	1	.000
C	Three-way structural model	10,050.63	10,458.75	10,144.56	−4926.31	99	2.33	1	.127
X	Best-fit model	10,004.24	10,288.69	10,069.71	−4933.12	69	–	–	–

AIC, Akaike's information criteria; *BIC*, Bayesian information criteria; *SSA BIC*, sample size adjusted Bayesian information criteria; *LL*, log-likelihood; *NPR*, number of estimated parameters; *Chd*, chi-square difference; *df*, degrees of freedom nonsignificant two-way interaction effects improved the model fit; thus, only eight two-way interaction effects were found to be significant.

In the last step, nonsignificant main effects were removed. In Model B, there were 37 main effects, and 11 of these main effects were found to be nonsignificant (Item 4: Attribute 4; Item 6: Attribute 3; Item 7: Attribute 2; Item 10: Attributes 1 and 3; Item 11: Attributes 1 and 2; Item 12: Attributes 2 and 3; Item 16: Attribute 3; and Item 20: Attribute 2). The analysis suggested that removing these nonsignificant main effects would improve the model. Thus, by removing these nonsignificant main effects, we obtained our best-fit model, which we called Model X. Model X estimated 69 free parameters, and model fit indices are presented in Table 4. In Model X, we had a total of $190 = \frac{19\times20}{2}$ pairs of items to examine. Item pairs with chi-square values exceeding 3.84 indicated a misfit, because at an alpha level of .05, the chi-square value should be equal to 3.84 ($N = 456$ and $df = 1$). Bivariate model fit information showed that 14 item pairs (7.36%) had significant misfit values. When examined these significant misfit values, there was no indication that a single item was responsible for several of the misfit pairs. Thus, the bivariate analysis did not suggest misfit issues with any particular item.

The LCDM places test-takers into some latent classes based on their mastery of attributes. In the current study, there were $2^4 = 16$ distinct latent classes because four attributes were estimated. Class counts (i.e. numbers of PSTs belonging to each class) and proportions for the latent classes based on the estimated model are presented in Table 5. The LCDM provides class counts in decimals. Hence, we rounded up these

Table 5 Class counts and proportions for latent classes based on the estimated model

Class	Attribute profile	Counts	Proportions
1	0000	5	.011
2	0001	2	.004
3	0010	93	.204
4	0011	0	.000
5	0100	0	.000
6	0101	0	.000
7	0110	61	.134
8	0111	0	.000
9	1000	36	.079
10	1001	142	.311
11	1010	19	.042
12	1011	29	.063
13	1100	46	.101
14	1101	8	.018
15	1110	1	.002
16	1111	14	.031

Codes with four numbers in each attribute profile show the PSTs' mastery of Attributes 1 through 4. For example, attribute profile 0101 indicates the PSTs' mastery of Attributes 2 and 4, and nonmastery of Attributes 1 and 3 decimals to the closest whole number. As can be ascertained from Table 5, approximately 142 PSTs (31.1%) belonged to the Latent Class 10, which indicated mastery of Attributes 1 and 4; whereas, roughly 93 PSTs (20.4%) belonged to the Latent Class 3 that showed their mastery of Attribute 3. Moreover, about 61 PSTs (13.4%) belonged to the Latent Class 7 that showed their mastery of Attributes 2 and 3. On the other hand, only 14 PSTs (3.1%) mastered all four attributes, and five PSTs (1.1%) did not master any of the four attributes. In addition, none of the PSTs belonged to the Latent Classes 4 (*0011*), 5 (*0100*), 6 (*0101*), and 8 (*0111*). This result suggested that none of the PSTs mastered Attribute 2 alone, Attributes 3 and 4 together, Attributes 2 and 4 together, or Attributes 2, 3, and 4 together. Using the proportions in Table 5, we calculated that 98.97% of the PSTs who had mastered Attribute 4 also had mastered Attribute 1. On the other hand, only 11.28% and 22.05% of the PSTs who had mastered Attribute 4 also had mastered Attributes 2 and 3, respectively. Furthermore, 53.07% and 58.4% of the PSTs who had mastered Attribute 2 also had mastered Attribute 1 and Attribute 3, respectively. In addition, 29.03% of the PSTs who had mastered Attribute 3 also had mastered Attribute 1.

We calculated attribute mastery using proportions in Table 5. Accordingly, Table 6 shows that 64.7% of the PSTs mastered Attribute 1, 28.6% of the PSTs mastered Attribute 2, 47.6% of the PSTs mastered Attribute 3, and 42.7% of the PSTs mastered Attribute 4. Table 6 suggests that although many PSTs were able to represent and interpret data, they had difficulty especially in drawing inferences about populations based on samples. Moreover, more than half of the PSTs had difficulty selecting and using appropriate statistical methods when analyzing data and understanding and applying the basic concepts of probability. Furthermore, using the DCM measure of reliability from Templin and Bradshaw (2013), we calculated reliability of these attribute mastery classifications. This measure relies upon "the correlation of mastery statuses between two hypothetical independent administrations of the same test" (Templin & Bradshaw, 2013, p. 259). Reliability indices in Table 6 show that the SRT was a highly reliable source in estimating the PSTs' mastery of the attributes.

When developing a test from the perspective of classical test theories, item analysis needs to be performed in order to detect problematic items by examining item difficulty and discrimination, and there is a need to sustain overall reliability of the test. Because DCMs are designed for diagnostic purposes (not specifically designed for measuring individuals' success rates in a test), they use different measures for defining an item as good or bad. According to DCMs, a reliable test is one that correctly estimates examinees' profiles (Templin, 2008). DCMs examine test quality by determining item-attribute discrimination indices (e.g. Henson & Douglas, 2005; Henson, Roussos, Douglas, & He, 2008) that highlight how well an item estimates the required attribute or attributes. Using an executable DCM file specifically designed to determine

Table 6 Proportions of the attribute mastery and attribute classification reliabilities

	Attribute 1	Attribute 2	Attribute 3	Attribute 4
Mastery	.647	.286	.476	.427
Reliability	.89	.82	.83	.90

item-attribute discrimination indices, we calculated item difficulty and item-attribute discrimination indices of the SRT items.

The item difficulty index, which ranges between 0 and 1, expresses the proportion of students that answered an item correctly. Where the index is close to 0, this depicts that an item is difficult, whereas an index value close to 1 indicates that an item is easy. In the current study, the item difficulty index ranged between .13 and .86, with a mean of .49 (with nine medium difficulty items [index values between .40 and .60], six difficult items [index values less than .40], and five easy items [index values more than .60]). Hence, there was a good balance among the items in terms of their level of difficulty. On average, while 52.07% of the PSTs were able to solve multiple-choice items, 38.6% of them solved open-ended items. In the SRT, Item 18 was perceived as the most difficult item, which was solved by only 13% of the PSTs. On the other hand, Items 6 and 13 were regarded as relatively easy items, having been answered correctly by 86% and 82% of the PSTs, respectively. As the purpose of the SRT was to determine the PSTs' strengths and weaknesses and did not aim to measure their academic achievement, no high or low level outliers were deleted at this level (Table 7).

Rather than providing classical item discrimination indices, CDMs provide item-attribute discrimination indices that express how well an item discriminates between masters and nonmasters of an attribute. If the item-attribute discrimination index is 0, masters and nonmasters of attribute(s) have the same probability of answering the item correctly. Furthermore, an item-attribute discrimination index with a value of 1 indicates that the correct answer rate is higher for masters of the attribute(s). On the other hand, a negative index indicates that the correct answer rate is higher for nonmasters of the attribute(s). Although there is no clear cutoff score mentioned in the literature for determining poor discrimination indices, de la Torre (2008) stated a discrimination index of .31 as being low. The item-attribute discrimination indices in Table 8 show that except for Items 6, 15, and 18, the remaining items discriminated well between masters and nonmasters of an attribute. Therefore, as required, while the masters of attribute(s) tended to answer the items correctly, nonmasters tended to answer them incorrectly.

The item parameter estimates, standard errors, and estimated probabilities are provided in Table 9. Parameter $\lambda_{i,\,0}$ is the intercept and represents the predicted log-odds of a correct response for examinees who did not master any of the required

Table 7 Item difficulty indices

Item	Index	Item	Index
Item 1	.50	Item 11	.75
Item 2	.43	Item 12	.68
Item 3	.46	Item 13	.82
Item 4	.57	Item 14	.52
Item 5	.44	Item 15	.23
Item 6	.86	Item 16	.56
Item 7	.33	Item 17	.53
Item 8	.67	Item 18	.13
Item 9	.31	Item 19	.29
Item 10	.24	Item 20	.42

Items 1 to 15 are multiple-choice, and Items 16 to 20 are open-ended attributes for an item. We calculated estimated probabilities using these parameter values. Table 9 shows that 89.2% of the nonmasters of any attributes were able to solve Item 6, and 83.5% and 81.9% of them solved Items 14 and 11, respectively. According to Table 7, Items 6 and 11 were easy items, and Item 14 had a medium difficulty. Moreover, estimated probabilities in Table 9 also show the nonmasters' difficulties in solving Items 18 and 9 that were solved by only 5.5% and 14.4% of the nonmasters, respectively. In Table 7, both items were determined as difficult items. Thus, considering the nonmasters' facility with solving easy items and difficulty with solving challenging items, their success rates in solving the SRT items appeared to be related to item difficulties.

One of the advantages of CDMs is that they can be used in providing diagnostic feedback on individual performance. In CDMs, classifications of respondents into latent classes are "the direct result of the application of a psychometric model" (Rupp et al., 2010, p. 86). Mplus calculates posterior probabilities of the attribute profiles for each respondent using an expected a posteriori (EAP) estimation and places respondents into one of these latent classes based on these probabilities. However, DCMs do not use a specific cutoff score when placing respondents into these latent classes (Bradshaw, 2015). Hence, a model-internal latent class classification criterion is used in DCMs (Rupp et al., 2010). Each respondent's proportion of attribute mastery is also estimated using these posterior probabilities. For instance, the proportion of mastery for Attribute 1 can be estimated by the sum of posterior probabilities beginning

Table 8 Item-attribute discrimination indices

Item	Attribute 1	Attribute 2	Attribute 3	Attribute 4
Item 1	.55	.63		.39
Item 2				.69
Item 3				.78
Item 4	.61			.56
Item 5				.86
Item 6		.27	.23	
Item 7	.58	.45		.73
Item 8				.65
Item 9				.73
Item 10	.52	.45	.41	
Item 11	.45	.43	.35	
Item 12	.53	.55	.38	
Item 13	.41	.38		
Item 14		.75	.59	
Item 15				.21
Item 16	.51		.54	
Item 17	.44	.42		
Item 18				.22
Item 19			.68	
Item 20		.50	.63	

Table 9 Item parameter estimates and estimated probabilities

Item	$\lambda_{i,0}$	Prop	$\lambda_{i,1(1)}$	Prop	$\lambda_{i,1(2)}$	Prop	$\lambda_{i,1(3)}$	Prop	$\lambda_{i,1(4)}$	Prop	$\lambda_{i,2(1,2)}$	Prop	$\lambda_{i,2(2,3)}$	Prop	$\lambda_{i,2(2,4)}$	Prop
1	.83 (.35)	.698	−1.99 (.65)	.239	−1.90 (.62)	.255			1.37 (.61)	.901	3.06 (1.00)	.499				
2	−.97 (.17)	.274							1.57 (.31)	.645						
3	−.70 (.17)	.331							1.68 (.31)	.727						
4	−.07 (.20)	.482	.75 (.29)	.662												
5	−.63 (.15)	.347							1.21 (.29)	.642						
6	2.10 (.21)	.892			−.81 (.40)	.785										
7	−1.14 (.26)	.242	−25.88 (.73)	.00					27.65 (.73)	1.00	26.66 (.89)	.412			−29.17 (2.4)	.065
8	.39 (.16)	.596							1.30 (.33)	.844						
9	−1.78 (.39)	.144							2.35 (.48)	.637						
10	−.73 (.14)	.323			−87.4 (.87)	.00					47.54 (1.00)	.00	64.71 (1.00)	.00		
11	1.51 (.22)	.819					−.69 (.33)	.692								
12	−1.00 (.82)	.269	2.61 (.81)	.833									178.1 (.85)	1.00		
13	.62 (.34)	.649	1.31 (.42)	.873	1.28 (.75)	.869					24.15 (1.27)	1.00				
14	1.62 (.37)	.835			−4.02 (2.59)	.083	−2.12 (.53)	.376					5.87 (2.68)	.794		
15	−1.09 (.19)	.250							.54 (.31)	.365						
16	−.90 (.27)	.289	1.89 (.33)	.729												
17	−.75 (.34)	.322	2.53 (.58)	.855	−2.33 (.77)	.044										
18	−2.84 (.39)	.055							1.87 (.46)	.276						
19	.30 (.24)	.575					−2.98 (.58)	.065								
20	1.03 (.23)	.736					−2.57 (.45)	.175								

Numbers between parentheses are standard errors from Latent Class 9 and ending up with Latent Class 16 (see Table 5). Three PSTs' item responses and estimated proportions for attribute mastery are presented in Table 10. Table 10 shows that PST 41 was able to correctly answer seven items, and among the three PSTs, she obtained the lowest estimated attribute mastery proportions for Attributes 1, 2, and 4. Nevertheless, she had an 80.2% chance of mastering Attribute 3, which was higher than PST 256's expected mastery of this attribute. Similarly, both PST 193 and PST 256 obtained the same number of correct responses; however, PST 193 had higher estimated proportions for mastering all four attributes than PST 256. Thus, proportions of attribute mastery may not depend on respondents' number of correct responses.

Discussion and Conclusions

The purpose of this study was to investigate Turkish preservice middle school mathematics teachers' understanding of statistics and probability concepts, and to provide diagnostic feedback on their strengths and weaknesses. By carefully examining the current published statistics education literature, national and international standards, and large-scale tests, we were able to determine a list of teacher skills (i.e. *attributes*) that were deemed necessary for solving middle school statistics and probability problems in the Turkish middle school mathematics curricula. Among with these skills, we determined four core attributes and developed a multidimensional test, the SRT, around these four attributes. We were then able to provide diagnostically reliable interpretations of the PSTs' understanding of statistics and probability using the LCDM in analyzing the PSTs' responses to the SRT items. In this quantitative study, we described the development of the SRT and explained the results that were subsequently obtained from the application of the test.

Based on the PSTs' mastery of four attributes, we provided cognitive feedback for their particular strengths and weaknesses. The LCDM results suggested that although many PSTs were able to master Attribute 1 (representing and interpreting data), most of them especially experienced difficulty in mastering Attribute 2 (drawing inferences about populations based on samples). Furthermore, more than half of the PSTs did not master Attribute 3 (selecting and using appropriate statistical methods when analyzing data) and Attribute 4 (understanding and applying basic concepts of probability). In order to master Attribute 2, the PSTs were required to draw inferences for more general situations. Similarly, the mastery of Attributes 3 and 4 required their recognition of certain statistical formulas or rules and application of these formulas. Therefore, Attributes 2, 3, and 4 entailed cognitively more challenging skills than Attribute 1.

Table 10 Three preservice teachers' item response patterns and estimated proportions for attribute mastery

ID	Response pattern	P(Att1)	P(Att2)	P(Att3)	P(Att4)
41	0000010100111 1*01000	.454	.413	.802	.011
193	101101*1111111110000	.988	.882	.931	.975
256	10110110111111010100	.983	.632	.768	.966

Thus, this interpretation of attributes could explain the PSTs' facilities with mastering Attribute 1 and difficulties with mastering the remaining attributes.

Among the four attributes, the lowest proportion of mastery was for Attribute 2. This finding suggested the PSTs' weaknesses in drawing inferences about populations based on samples. According to Table 9, except for Items 6 and 13, the mastery of Attribute 2 did not contribute enough to the PSTs' success rates in solving the SRT items. Overall, the PSTs' weaknesses in mastering Attribute 2 appeared to be related to Attribute 2 involving utilization of a sophisticated reasoning, inferential reasoning. The sophistication of inferential reasoning stems from requiring PSTs to look at the data first in order to identify underlying patterns and then to look beyond the data to draw inferences (Leavy, 2010, p. 47). In agreement with this sentence, Leavy (2010) reported the PSTs' difficulties developing pedagogical contexts for advancing their future students' informal inferential reasoning. For her, the PSTs' difficulties suggested issues with their pedagogical content knowledge.

In terms of Attribute 3, although the calculation of the mean, mode, and median values was expected to be easy for the PSTs, we recognized that many PSTs confused the mode and median concepts with each other. For example, in Item 16, we presented test scores of 10 students in a table and asked the PSTs to determine the mean, mode, and median of these test scores. As indicated in Table 7, 44% of the PSTs could not determine all three values. Similar to the findings of Mathews and Clark (2003), we found that while many PSTs were able to calculate the mean, most of them had difficulty determining the mode and median. As stated by Groth and Bergner (2006), although calculating mean, mode, and median seems to be a simple activity for teachers, they may lack the common knowledge (e.g. Hill et al., 2004) necessary to compute these concepts. Therefore, this finding confirms difficulties that PSTs have in understanding and teaching core ideas of probability and statistics (Garfield & Ben-Zvi, 2007) and suggests deficiencies in their content knowledge.

Regarding the PSTs' mastery of Attribute 4, we observed that 98.97% of the PSTs who had mastered Attribute 4 had also mastered Attribute 1. Hence, there was a high positive association between the two attributes that shows the PSTs' mastery of the two attributes develop together. This result might have occurred due to the fact that the mathematics curriculum provides experience with Attribute 1 before providing experience with Attribute 4. The LCDM analysis showed that 57.3% of the PSTs could not master Attribute 4, and two of the most difficult items (i.e. Items 18 and 15) measured this attribute. In Item 18, we recognized that although many PSTs were able to calculate possible outcomes of selecting at least one adult and one child from a pool of six adults and four children, they did not divide this number by the total number of possible outcomes. Therefore, as stated by Stohl (2005), the conceptual complexity of probability was a major issue for the development of the PSTs' mathematical knowledge.

We found that on average, the PSTs had more difficulty answering open-ended items than multiple-choice items. Overall, Item 18 was the most difficult item, which was answered by 13% of the PSTs (see Table 7). Similarly, Dogan and Tatsuoka (2008) and Sen and Arican (2015) also stated that Turkish 8th grade students experienced difficulties in solving open-ended items. Hence, Turkish students' and PSTs' weaknesses indicated that both groups were similar in terms of their ability to answer open-ended items. The open-ended items were included in the SRT because by including these items, we expected to decrease the guessing factor (Panackal & Heft, 1978) and obtain detailed information about their understanding. However, as stated by Haladyna, Downing, and Rodriguez (2002), well-constructed multiple-choice items can also provide as much information about PSTs' understanding as open-ended items. Therefore, both types of items were used in the SRT.

The item parameter estimates in Table 9 showed that Items 6, 11, and 14 had high intercept values. A high intercept value suggests high correct response rates from examinees who have not mastered any of the attributes. As previously explained, the high intercept values of Items 6 and 11 appeared to be related to their item difficulty levels. Table 7 presents that Items 6 and 11 were easy items. On the other hand, Item 14 had a medium difficulty. However, as in Item 6, it estimated the PSTs' abilities to make inferences about a given sample, and both items could be answered without making any calculations. Thus, nonsmasters appeared to solve these two items without requiring the knowledge of Attributes 2 and 3.

Implications for Teaching and Suggestions

Due to classical test theories providing only a single overall score for each student, they offer limited information about students' strengths and weaknesses (Sen & Arican, 2015). Hence, in recent years, researchers have been paying more attention to CDMs in order to provide diagnostic feedback on students' performance. CDMs can also be used in providing diagnostic feedback on individual performance as in Table 10. Therefore, university educators and school teachers can use CDMs in detecting strengths and weaknesses of individuals and can then provide diagnostic feedback if necessary. University educators should work to expand PSTs' knowledge of statistics and probability in order that they can develop meaningful understanding of these subjects. We strongly believe that results obtained from CDM analysis may provide new insights into PSTs' understanding of statistics and probability concepts.

In Turkey, national tests that are conducted at the end of 8th and 12th grades only included multiple-choice items for many years. Hence, teachers usually focus on teaching rules and rote computations rather than developing their students' meaningful understanding of mathematics. Therefore, Turkish PSTs' weaknesses on the open-ended items might be a reflection of the Turkish testing system. We should note that in the 2017 university entrance exam, the national testing association decided to include open-ended items. Although educators found the inclusion of open-ended items to be a promising step, they questioned the quality and ratio of the number of open-ended items used. Thus, findings obtained from research such as the current study can influence policymakers' decision-making on curricular choices and can therefore help policymakers in developing effective educational systems.

In this study, we were unable to provide partial credits for the open-ended items. Providing partial credits to the open-ended items could improve the overall reliability of the attribute estimations. Moreover, this study was conducted with 456 PSTs from four universities in Turkey. Therefore, further studies should examine PSTs' understanding of statistics and probability using a larger sample.

Acknowledgements Parts of this study were presented at the 2017 European Educational Research Conference held in Copenhagen/Denmark.

Authors' Contributions This study was conducted in equal collaboration between the two authors.

Funding The authors disclose receipt of the following financial support for the research, authorship, and publication of this study: This study was supported by the Ahi Evran University Scientific Research Projects Coordination Unit. Project Number *EGT.A3.16.014*.

Compliance with Ethical Standards

Conflict of Interest The authors declare that there are no conflicts of interest.

References

Ball, D. L., Lubienski, S. T., & Mewborn, D. S. (2001). Research on teaching mathematics: The unsolved problem of teachers' mathematical knowledge. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 433–456). Washington, DC: American Educational Research Association.

Batanero, C., & Díaz, C. (2010). Training teachers to teach statistics: What can we learn from research? *Statistique et Enseignement, 1*(1), 5–20.

Batanero, C., & Díaz, C. (2012). Training school teachers to teach probability: Reflections and challenges. *Chilean Journal of Statistics, 3*(1), 3–13.

Batanero, C., Godino, J. D., & Roa, R. (2004). Training teachers to teach probability. *Journal of Statistics Education, 1*(1), 5–20. Retrieved February 3, 2018, from http://www.amstat.org/publications/jse/v12n1/batanero.html.

Baumert, J., Kunter, M., Blum, W., Brunner, M., Voss, T., Jordan, A., ... Tsai, Y. (2010). Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal, 47*(1), 133–180.

Bradshaw, L. (2015). *PARCC diagnostic assessments for mathematics comprehension: A diagnostic classification model approach.* Paper presented at the Council of Chief State School Officers (CCSSO) 2015 National Conference on Student Assessment (NCSA) in San Diego, California.

Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice, 33*(1), 2–14.

Brown, A., & Silver, E. (1989). Data organization and interpretation. In M. M. Lindquist (Ed.), *Results from the fourth mathematics assessment of the national assessment of educational progress* (pp. 28–34). Reston, VA: National Council of Teachers of Mathematics.

Burgess, T. A. (2007). *Investigating the nature of teacher knowledge needed and used in teaching statistics* (Unpublished doctoral dissertation). Massey University, Auckland, New Zealand.

Choi, K. M., Lee, Y. S., & Park, Y. S. (2015). What CDM can tell about what students have learned: An analysis of TIMSS eighth grade mathematics. *Eurasia Journal of Mathematics, Science & Technology Education, 11*(6), 1563–1577.

Common Core State Standards Initiative. (2010). *The common core state standards for mathematics*. Washington, DC: Author.

Curcio, F. R. (1987). Comprehension of mathematical relationships expressed in graphs. *Journal for Research in Mathematics Education, 18*(5), 382–393.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement, 45*(4), 343–362.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*(2), 179–199.

delMas, R., Garfield, J., & Ooms, A. (2005, July). *Using assessment items to study students' difficulty with reading and interpreting graphical representations of distributions.* Paper presented at the Fourth Forum on Statistical Reasoning, Thinking, and Literacy (SRTL–4), Auckland, New Zealand.

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal, 6*(2), 28–58.

Dereli, A. (2009). *The mistakes and misconceptions in probability of eighth grade students* (Unpublished Master's Thesis). Eskisehir Osmangazi University, Turkey.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood–based classification techniques. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–390). Hillsdale, NJ: Lawrence Erlbaum.

Dogan, E., & Tatsuoka, K. (2008). An international comparison using a diagnostic testing model: Turkish students' profile of mathematical skills on TIMSS–R. *Educational Studies in Mathematics, 68*(3), 263–272.

Fennema, E., & Franke, M. L. (1992). Teachers' knowledge and its impact. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 147–164). New York, NY: Macmillan.

Franklin, C., Kader, G., Mewborn, D. S., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre–K–12 curriculum framework*. Alexandria, VA: American Statistical Association. Retrieved February 8, 2018, from http://www.amstat.org/asa/education/Guidelines-for-Assessment-and-Instruction-in-Statistics-Education-Reports.aspx.

Franklin, C., & Mewborn, D. (2006). The statistical education of PreK–12 teachers: A shared responsibility. In G. Burrill (Ed.), *NCTM 2006 Yearbook: Thinking and reasoning with data and chance* (pp. 335–344). Reston, VA: NCTM.

Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International Statistical Review, 75*(3), 372–396.

Garfield, J., & Ben–Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Berlin, Germany: Springer Science & Business Media.

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for research in Mathematics Education, 19*(1), 44–63.

Groth, R. E. (2007). Toward a conceptualization of statistical knowledge for teaching. *Journal for Research in Mathematics Education, 38*(5), 427–437.

Groth, R. E., & Bergner, J. A. (2006). Preservice elementary teachers' conceptual and procedural knowledge of mean, median, and mode. *Mathematical Thinking and Learning, 8*(1), 37–63.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–333.

Hartz, S. (2002). *A Bayesian framework for the Unified Model for assessing cognitive abilities: Blending theory with practice* (Unpublished doctoral dissertation). University of Illinois at Urbana–Champaign.

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnostics. *Applied Psychological Measurement, 29*(4), 262–277.

Henson, R., Roussos, L., Douglas, J., & He, X. (2008). Cognitive diagnostic attribute–level discrimination indices. *Applied Psychological Measurement, 32*(4), 275–288.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log–linear models with latent variables. *Psychometrika, 74*(2), 191–210.

Hill, H., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*(2), 371–406.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *The Elementary School Journal, 105*(1), 11–30.

Im, S., & Park, H. J. (2010). A comparison of US and Korean students' mathematics skills using a cognitive diagnostic testing method: Linkage to instruction. *Educational Research and Evaluation, 16*(3), 287–301.

Jones, G. A. (2005). *Exploring probability in school: Challenges for teaching and learning*. New York, NY: Springer.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258–272.

Leavy, A. M. (2010). The challenge of preparing preservice teachers to teach informal inferential reasoning. *Statistics Education Research Journal, 9*(1), 46–67.

Lee, H. S., & Hollebrands, K. F. (2011). Characterising and developing teachers' knowledge for teaching statistics with technology. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics-challenges for teaching and teacher education* (pp. 359–369). Dordrecht, Netherlands: Springer.

Lee, Y. S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing, 11*(2), 144–177.

Li, K. Y., & Shen, S. M. (1992). Students' weaknesses in statistical projects. *Teaching Statistics, 14*(1), 2–8.

Makar, K., & Rubin, A. (2009). A framework for thinking about informal statistical inference. *Statistics Education Research Journal, 8*(1), 82–105.

Mathews, D., & Clark, J. (2003). *Successful students' conceptions of mean, standard deviation, and the central limit theorem*. Unpublished paper. Retrieved July 4, 2018 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.636.8870&rep=rep1&type=pdf.

Muthen, L. K., & Muthen, B. O. (1998–2011). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthen & Muthen.

O'Connell, A. A. (1999). Understanding the nature of errors in probability problem–solving. *Educational Research and Evaluation, 5*(1), 1–21.

Panackal, A. A., & Heft, C. S. (1978). Cloze technique and multiple choice technique: Reliability and validity. *Educational and Psychological Measurement, 38*(4), 917–932.

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved January 3, 2018, from http://www.R-project.org/.

Ranjbaran, F., & Alavi, S. M. (2017). Developing a reading comprehension test for cognitive diagnostic assessment: A RUM analysis. *Studies in Educational Evaluation, 55*, 167–179.

Ravand, H., & Robitzsch, A. (2015). Cognitive diagnostic modeling using R. *Practical Assessment, Research & Evaluation, 20*(11), 1–12.

Rupp, A., & Templin, J. (2008). Effects of Q–matrix misspecification on parameter estimates and misclassification rates in the DINA model. *Educational and Psychological Measurement, 68*(1), 78–98.

Rupp, A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi–square test statistic. *Psychometrika, 75*(2), 243–248.

Sen, S., & Arican, M. (2015). A diagnostic comparison of Turkish and Korean students' mathematics performances on the TIMSS 2011 assessment. *Journal of Measurement and Evaluation in Education and Psychology, 6*(2), 238–253.

Shaughnessy, J. M. (2007). Research on statistics learning and reasoning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 957–1009). Reston, VA: The National Council of Teachers of Mathematics.

Stohl, H. (2005). Probability in teacher education and development. In G. A. Jones (Ed.), *Exploring probability in school: Challenges for teaching and learning* (pp. 345–366). Boston, MA: Springer.

Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics, 10*(1), 55–73.

Templin, J. (2008). *Test construction item discrimination*. Lecture presented at the Diagnostic Modelling Seminar at the University of Georgia, Athens. Retrieved February 2, 2018, from https://jonathantemplin.com/files/dcm/ersh9800f08/ersh9800f08_lecture11.pdf.

Templin, J., & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30*(2), 251–275.

Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287–305.

Toker, T., & Green, K. (2012). *An application of cognitive diagnostic assessment on TIMMS–2007 8th grade mathematics items*. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, British Columbia, Canada.

von Davier, M. (2005). *A general diagnostic model applied to language testing data. ETS Research Report*. Princeton, NJ: Educational Testing Service.

Watson, J. M. (2006). *Statistical literacy at school: Growth and goals*. Mahwah, NJ: Lawrence Erlbaum.

Werner, C., & Schermelleh-Engel, K. (2010). Deciding between competing models: Chi–square difference tests. In *Introduction to structural equation modeling with LISREL* (pp. 1–3). Frankfurt, Germany: Goethe University.

Zawojewski, J. S., & Heckman, D. S. (1997). What do students know about data analysis, statistics, and probability? In P. A. Kenny & E. A. Silver (Eds.), *Results from the sixth mathematics assessment of the National Assessment of Educational Progress* (pp. 195–223). Reston, VA: National Council of Teachers of Mathematics.

Zawojewski, J. S., & Shaughnessy, J. M. (2000). Mean and median: Are they really so easy? *Mathematics Teaching in the Middle School, 5*(7), 436–440.