applied sciences

MDPI

Article

Cognitive Diagnostic Assessment in University Statistics Education: Valid and Reliable Skill Measurement for Actionable Feedback Using Learning Dashboards

Lientje Maas [1,*], Matthieu J. S. Brinkhuis [2], Liesbeth Kester [3] and Leoniek Wijngaards-de Meij [1]

[1] Department of Methodology & Statistics, Utrecht University, 3584 CH Utrecht, The Netherlands; l.wijngaards@uu.nl

[2] Department of Information and Computing Sciences, Utrecht University, 3584 CC Utrecht, The Netherlands; m.j.s.brinkhuis@uu.nl

[3] Department of Education, Utrecht University, 3584 CS Utrecht, The Netherlands; l.kester@uu.nl

* Correspondence: j.a.m.maas@uu.nl

Abstract: E-learning is increasingly used to support student learning in higher education, facilitating administration of online formative assessments. Although providing diagnostic, actionable feedback is generally more effective, in current practice, feedback is often given in the form of a simple proportion of correctly solved items. This study shows the validation process of constructing detailed diagnostic information on a set of skills, abilities, and cognitive processes (so-called attributes) from students' item response data with diagnostic classification models. Attribute measurement in the domain of statistics education is validated based on both expert judgment and empirical student data from a think-aloud study and large-scale assessment administration. The constructed assessments provide a valid and reliable measurement of the attributes. Inferences that can be drawn from the results of these formative assessments are discussed and it is demonstrated how this information can be communicated to students via learning dashboards to allow them to make more effective learning choices.

Keywords: online formative assessment; cognitive modeling; skill measurement; validation; reliability; learning dashboards; university statistics education

Citation: Maas, L.; Brinkhuis, M.J.S.; Kester, L.; Wijngaards-de Meij, L. Cognitive Diagnostic Assessment in University Statistics Education: Valid and Reliable Skill Measurement for Actionable Feedback Using Learning Dashboards. *Appl. Sci.* **2022**, *12*, 4809. https://doi.org/10.3390/app12104809

Academic Editor: Donato Cascio

Received: 14 April 2022
Accepted: 6 May 2022
Published: 10 May 2022

1. Introduction

In higher education, increased autonomy is required from students because external regulation is generally limited [1]. Students need to rely on their own evaluation of performance to regulate their learning processes. Unfortunately, not all students can accurately evaluate their performance, which can result in poor learning choices [2]. The use of e-learning environments in higher education has emerged over the last decades [3], facilitating administration of online formative assessments that can help students interpret the meaning of their past performance and determine what they should be studying and practicing to increase their performance [4,5].

In online formative assessments, students answer assessment items and receive feedback based on their responses via learning dashboards. In current practice, this feedback often consists of information regarding correctness of responses and percentage correct scores. However, if students know the correct answer to an item and can compare it with their given answer, this does not imply that they can accurately infer which knowledge or skills are lacking. For example, Ref. [6] found that students find it difficult to determine which features of statistics items are relevant for learning. Furthermore, the interpretation of percentage correct scores can be problematic, since these scores cannot be directly compared across different item sets and differences in percentage scores cannot be used to express change [7]. It would be beneficial to provide students more detailed, *diagnostic* feedback regarding their knowledge and skills [8]. Diagnostic feedback is information about how students' current levels of performance relate to desired levels of performance, indicating whether or not they have mastered the skills that are required to solve certain tasks [9]. This allows students to determine where they should focus their attention and effort; hence, this information is actionable [10].

In order to provide diagnostic feedback, valid and reliable measurement of students' skills is required, which is the focus of the current study. Diagnostic information can be extracted from students' item responses with cognitive diagnostic assessment, which brings together cognitive science and psychometrics to measure knowledge structures and processing skills in order to provide information about cognitive strengths and weaknesses [8].

1.1. Cognitive Diagnostic Assessment

In cognitive diagnostic assessment, a so-called cognitive model is specified that links understanding to performance [11]. A set of skills, abilities, and cognitive processes required to solve certain items is defined, which are referred to as attributes. The objective of cognitive diagnostic assessment is to classify students as master or nonmaster of each attribute. This classification depends on both expert judgment and empirical evidence. Domain experts encode which attributes are required to solve each item in the Q-matrix, which is a binary matrix that describes the relations between the items and the attributes by indicating for each item whether it measures each attribute or not. The Q-matrix in combination with item response data from diagnostic assessments enables the application of statistical models to classify students. Well-suited models to this end are diagnostic classification models (DCMs; [12]). Previous simulations have shown opportunities to apply these models to item response data from online formative assessments in higher education to obtain diagnostic information if sufficient numbers of students participate (i.e., 100–300 students to assess 3–6 attributes; [13]). Thus, in courses with large groups of students, cognitive diagnostic assessment with DCMs can be a valuable tool to obtain diagnostic feedback at low cost. It enables educational practitioners to design formative assessments that align with the learning objectives of a course, resulting in actionable feedback that aligns with these objectives.

1.2. Cognitive Model Specification

To design educational assessments, one first needs to define the desired results (known as 'backward design'; [14]). One identifies what will be measured, resulting in a cognitive model that consists of a set of attributes. As pointed out by Pellegrino et al. ([15], p. 45), the attributes should "reflect the most scientifically credible understanding of typical ways in which learners represent knowledge and develop expertise in a domain". Attribute specification is ideally based on both expert judgment and research in cognitive and educational science.

1.2.1. Attribute Granularity

Diagnostic assessment can support the analysis of student behavior at various levels of detail by defining attributes at different granularity levels. This granularity level should be valuable from an educational perspective, in the sense that feedback is meaningful and actionable (see e.g., [16]). In addition, the attributes must be psychometrically measurable. If the number of attributes increases, the complexity of the measurement model increases as well, putting higher demands on the data structure. Further, one must be able to construct items that produce observable student behaviors at an appropriate level of specificity.

1.2.2. Item Construction and Q-Matrix Specification

Once the attributes are defined, items are constructed to measure them. It is crucial that the items truly measure the attributes that they are designed to measure, i.e., that students indeed apply (only) those attributes to solve the items. Further, it is essential that the Q-matrix is specified correctly in order to obtain accurate student classifications and, thus, high-quality diagnostic feedback [17,18]. However, establishing the Q-matrix based solely on expert judgment may be subjective and susceptible to specification errors. To address this issue, several qualitative and quantitative methods are available to validate the Q-matrix. Think-aloud studies can provide insight in the processing steps students use to solve the item [19]. Students are asked to verbalize their thoughts while solving items, allowing to verify the relations between items and attributes (see e.g., [20]) and to modify items for better alignment between items and attributes. In addition, assessments can be administered at large scale to gather empirical evidence regarding the extent to which the items elicit the attributes using empirical Q-matrix validation methods to identify misspecified entries (e.g., [21]).

1.3. Diagnostic Assessment in Statistics Education

The current study is focused on the domain statistics education to show how valid and reliable measurements of students' skills can be obtained with cognitive diagnostic assessment. Introductory statistics courses are part of many educational programs at universities, yet students often struggle to understand the abstract concepts in this domain [22]. Diagnostic assessment can support students in developing such conceptual understanding [23].

There has been some research related to attribute specification for statistics in higher education. These studies demonstrate that attribute specification in statistics education should not solely be based on expert judgment, since this may result in invalid attributes [24,25]. To our knowledge, the constructs of interest in assessment in university statistics education have not been extensively studied from a cognitive diagnostic perspective.

1.4. Current Study

In the current study, we construct and validate cognitive diagnostic assessments that can serve as a foundation for designing effective tools for diagnostic assessment of students' conceptual understanding of statistics that can be used to obtain actionable feedback at low cost. Further, it is demonstrated how this information can be communicated to students via learning dashboards to allow them to make more effective learning choices. The following research question is addressed:

> *How can valid and reliable diagnostic information be obtained from online formative assessment to provide actionable feedback via learning dashboards in university statistics education?*

We first identify what attributes are necessary to master for students enrolled in introductory statistics courses in higher education (Section 2). Next, we examine how these attributes can be measured with formative assessment to draw valid and reliable inferences about students' attribute mastery status. Items are constructed and a provisional Q-matrix is specified based on expert judgment (Section 3). Assessments are empirically validated using both qualitative and quantitative methods (Section 4). We discuss what inferences can be drawn from the results and how this can be communicated to students (Section 5). We end with an overall discussion (Section 6). An overview of the current study is presented in Figure 1.

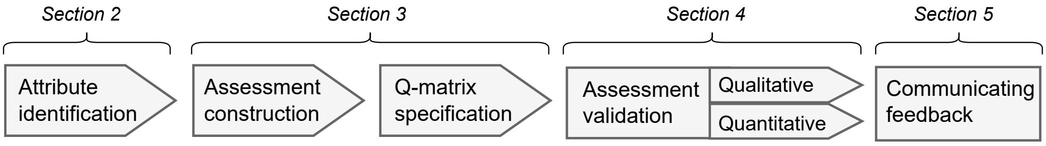


Figure 1. Outline of the current study.

2. Attribute Identification

2.1. Methods

In the attribute identification process, we first defined our domain of interest. We then specified learning objectives for this domain and subsequently defined attributes. The results were refined based on domain expert evaluation.

2.1.1. Domain Description

We specified a cognitive model of conceptual understanding of inferential statistics (in a frequentist framework) relevant in introductory nonmathematical statistics courses in higher education. The attributes are focused on principles behind null hypothesis significance testing (NHST), which is a widely used tool in statistical inference that requires students to understand and connect several complex concepts (e.g., variability, sampling distributions, *p*-values). We chose this domain because it is a central topic in many introductory statistics courses, and students often experience difficulty with it due to the large number of involved abstract concepts and the (counter-intuitive) logic of experimental design [26].

2.1.2. Defining Learning Objectives

Learning objectives were defined based on literature and course materials. The literature study was based on four sources: a review of students' misconceptions in statistical inference [22]; a book about the development of students' statistical reasoning [27]; a paper about the development of instruments measuring conceptual understanding of statistics [28]; and the Guidelines for Assessment and Instruction in Statistics Education College Report [29], which lists goals that summarize what students should know and understand after a first course in statistics based on collective beliefs reflected in statistics education literature. After the literature review, we inspected the course materials of (nonmathematical) introductory statistics courses to complement the specified learning objectives.

2.1.3. Defining Attributes

Based on the learning objectives, we aimed to define attributes that are both psychometrically *measurable* and pedagogically *meaningful* and *actionable*. In order to be measurable (i.e., to enable estimation of student attribute profiles), it is recommended to include at least five measurements of each attribute [30]. Evaluating mastery of each learning objective would require administration of huge amounts of items and, moreover, feedback at such fine-grained level may be less meaningful in a domain where students need to develop understanding of interrelations among concepts [31]; therefore, we grouped learning objectives into attributes. We aimed to group learning objectives that address closely related concepts, presuming that students' skills regarding these learning objectives are strongly related. Consequently, concepts that are generally addressed within the same lectures, e-learning modules, or book chapters are encompassed by the same attribute, allowing for actionable feedback (e.g., by recommending relevant learning materials for nonmastered attributes).

2.1.4. Expert Evaluation

Four domain experts were consulted to evaluate the exhaustiveness, redundancy, and grouping of the learning objectives. The experts had 13 to 29 years of experience in statistics education. Based on their feedback, the results were refined.

2.2. Results

The attribute identification process resulted in 9 attributes comprising 33 learning objectives. The attributes are summarized in Table 1 and more extensive results including the learning objectives are presented in Table A1 in Appendix A. In the Supplementary Material, a more detailed description of each attribute is provided including references to empirical studies showing the relevance of addressed concepts and the occurrence of misconceptions.

Table 1. The nine identified attributes.

A.	Understanding center & spread
B.	Interpreting univariate graphical representations
C.	Graphically comparing groups
D.	Understanding sampling variability
E.	Understanding sampling distributions
F.	Understanding the standard error
G.	Understanding principles of hypothesis testing
H.	Evaluating NHST results
I.	Understanding and using confidence intervals

3. Assessment Construction and Q-Matrix Specification

3.1. Item Collection

Assessing conceptual understanding of statistics requires assessment techniques that reflect the nature of students' thinking. Although this is best achieved through one-to-one communication or by examining in-depth student work [32], items with a selected-response format can also be useful to gather limited indicators of statistical reasoning if properly designed [33]. This is demonstrated by instruments designed to measure conceptual understanding of statistics, such as the Statistical Reasoning Assessment (SRA; [34]), the Statistics Concept Inventory (SCI; [35]), the Comprehensive Assessment of Outcomes in Statistics test (CAOS; [28]), and the Assessment Resource Tools for Improving Statistical Thinking (ARTIST; [28]). These instruments consist of multiple-choice items that require thinking and reasoning rather than recalling definitions, computing, or using formulas. In order to measure our attributes, we exploited these sources of existing items. Relying on these sources ensures that the items have been evaluated by experts in earlier research and/or education. Items were selected (and if needed modified) based on validated guidelines for multiple choice item writing [36] and design principles for statistical assessments ([37], p. 139). The set of prototype items consisted of 59 items.

3.2. Q-Matrix Specification

The list of attributes and learning objectives was presented to three independent experts who had 4 to 6 years of experience in statistics education. The experts were not involved in the attribute identification phase of the study. For each item, they were asked to select which learning objective(s) are required to solve it. The expert ratings were combined by means of a majority vote and aggregated to the attribute level, since this is the level at which we want to make inferences about students. Pairwise inter-rater agreement, as indicated by fuzzy kappa [38], ranged from 0.41 to 0.73 at the learning objective level and from 0.79 to 0.91 at the attribute level. Expectedly, agreement is substantially lower at the learning objective level, which stems from the more fine-grained nature of the learning objectives. The lower agreement illustrates that Q-matrix specification based on expert judgment can be susceptible to errors and provides a rationale for the validation.

According to the combined ratings, all learning objectives are measured by the item set, indicating that all aspects of the attributes are measured (i.e., no construct underrepresentation). The aggregated ratings resulted in a provisional Q-matrix. In total, the item set consisted of 41 unidimensional items, 16 two-dimensional items, and 2 three-dimensional items.

4. Assessment Validation

Attribute measurement was validated based on a qualitative evaluation with a student think-aloud study and a quantitative evaluation of students' item response data.

4.1. Qualitative Evaluation

4.1.1. Methods

We let students verbalize their thoughts when solving the items during interviews to verify whether the attributes are a good representation of the skills students rely on when solving the items. The procedures are briefly described below and more extensive descriptions are included in the Supplementary Material.

Participants and Procedures

Participants were 8 university students who recently participated in an introductory statistics course. Students were asked to rate their own performance in these courses on a continuum from basic to excellent. Two students rated their performance as basic, three students as good, and three students as very good.

During the interviews, items were presented one by one and students were asked to read them aloud and, subsequently, explain clearly how they would solve the item. We avoided follow-up questions, because these may evoke thinking patterns that differ from how students think about problems on their own [39]. If necessary, clarifying questions were asked at the end of the interview to successfully ascertain the cognitive processes that the students demonstrated. Only a subset of the 59 items was presented to each student. On average, students solved 27.5 items ($SD = 7.9$, min = 16, max = 41). We ensured each item was answered by at least 3 students by presenting different subsets in different orders.

Analysis

The interviews were recorded and coded by two independent experts who subsequently participated in conversations to reach consensus about students' attribute use (intercoder agreement as indicated by fuzzy kappa was 0.67; [38]). For items with (partly) justified answers, it was coded for each learning objective whether or not students showed (complete or incomplete) evidence of using it. Results were aggregated to the attribute level to evaluate students' attribute use for each item. We verified to what extent this corresponded to the provisional Q-matrix.

In addition, we evaluated clarity of the items from a student perspective (unrelated to their knowledge of statistics), allowing alteration of unclear items. Minor adjustments were made to 17 items, such as reformulation or deletion/addition of context information. The improved items were used in the quantitative evaluation.

4.1.2. Results

The percentages of students who used each attribute when solving each item are presented in Table 2. Colored cells indicate which attributes are measured by the item according to the provisional Q-matrix. It can be seen that the attribute usage to a great extent coincides with the provisional Q-matrix (96.2% agreement based on a majority rule for attribute use). These results were used in the quantitative evaluation of data-driven suggestions for alternative Q-matrix entries.

For one item, students showed indication of engaging in cognitive processes that pointed to the use of domain-relevant attributes that are not included in the set of nine attributes. This was resolved by deleting one distractor from the response alternatives. Apart from this item, students did not show notable indication of using unspecified attributes. Therefore, we believe the nine attributes encompass the knowledge and understanding of statistics required to solve the items.

Table 2. Percentages of students using each attribute (Att.) in their reasoning per item ("-" indicates 0%). Colored cells indicate the attributes that are measured by each item according to the provisional Q-matrix. The last column indicates the number of students who answered each item (*n*).

Item	Att. A	Att. B	Att. C	Att. D	Att. E	Att. F	Att. G	Att. H	Att. I	*n*
ARTIST_sc_MS_05	67%	-	-	-	-	-	-	-	-	3
ARTIST_sc_MS_01	100%	-	-	-	-	-	-	-	-	4
ARTIST_sc_MC_05	100%	-	-	-	-	-	-	-	-	4
ARTIST_sc_MC_06	100%	-	-	-	-	-	-	-	-	4
ARTIST_db_MS_Q0490	100%	-	-	-	-	-	-	-	-	4
CAOS_14	100%	100%	-	-	-	-	-	-	-	4
CAOS_15	100%	100%	-	-	-	-	-	-	-	4
CAOS_08	67%	100%	-	-	-	-	-	-	-	3
CAOS_09	67%	100%	-	-	-	-	-	-	-	3
CAOS_10	100%	100%	-	-	-	-	-	-	-	3
CAOS_11	-	67%	100%	-	-	-	-	-	-	3
CAOS_12	-	100%	100%	-	-	-	-	-	-	3
CAOS_13	-	-	-	-	-	-	-	-	-	3
SRA_015	-	75%	100%	-	-	-	-	25%	-	4
ARTIST_db_CG_Q0840	50%	100%	100%	-	-	-	-	-	-	4
ARTIST_sc_SV_01	-	67%	-	33%	33%	-	-	-	-	3
CAOS_17	-	-	-	100%	-	-	-	-	-	4
ARTIST_sc_SV_03	-	-	-	100%	-	100%	-	-	-	4
ARTIST_sc_SV_14	67%	-	-	100%	-	33%	-	-	-	3
ARTIST_sc_SV_04	-	-	-	60%	100%	-	-	-	-	5
ARTIST_sc_SV_10	-	-	-	-	100%	-	-	-	-	4
ARTIST_sc_SV_11	-	-	-	25%	100%	50%	-	-	-	4
ARTIST_sc_SV_09	-	50%	-	50%	100%	50%	-	-	-	4
ARTIST_db_SS_Q0061A	-	-	-	50%	100%	-	-	-	-	4
ARTIST_db_SS_Q0061B	-	-	-	-	100%	25%	-	-	-	4
ARTIST_db_SS_Q0061C	-	-	-	-	100%	25%	-	-	-	4
ARTIST_db_SS_Q0061D	-	-	-	-	100%	25%	-	-	-	4
ARTIST_db_SS_Q0061E	50%	-	-	-	75%	100%	-	-	-	4
ARTIST_sc_SV_05	-	-	-	50%	100%	-	-	-	-	4
CAOS_16	-	-	-	100%	-	100%	-	-	-	4
CAOS_32	20%	-	-	40%	-	100%	-	-	-	5
SCI_2004_20	100%	-	-	75%	-	100%	-	-	-	4
GRASPLE_DP_SE_40985	-	-	-	-	-	100%	-	-	-	3
ARTIST_db_SS_Q1437	-	-	-	-	-	67%	-	-	-	3
ARTIST_db_SS_Q0614	-	-	-	-	-	100%	-	-	-	3
CAOS_40	-	-	-	-	-	-	100%	60%	-	5
ARTIST_sc_TS_01	-	-	-	-	-	-	100%	-	-	4
ARTIST_db_TSG_Q1182	-	-	-	-	-	-	100%	25%	-	4
CAOS_23	-	-	-	-	-	25%	100%	75%	-	4
CAOS_24	-	-	-	-	-	-	100%	25%	-	4
ARTIST_db_TSG_Q1392	-	-	-	-	-	-	-	60%	-	5
CAOS_25	-	-	-	-	-	-	-	75%	-	4
CAOS_26	-	-	-	-	-	-	-	100%	-	4
CAOS_27	-	-	-	-	-	-	-	100%	-	4
ARTIST_sc_TS_04	-	-	-	-	-	-	100%	50%	-	4
ARTIST_sc_TS_10	-	-	-	-	-	-	100%	100%	-	4
ARTIST_sc_TS_07	33%	-	-	33%	-	-	33%	100%	-	3
ARTIST_sc_TS_09	-	-	-	-	-	-	75%	100%	-	4
SCI_2004_22	-	-	-	-	-	-	-	100%	-	3
ARTIST_db_TSG_Q1007	-	-	-	-	-	25%	25%	50%	-	4
ARTIST_sc_CI_05	-	-	-	25%	-	-	-	-	75%	4
ARTIST_sc_CI_03	-	-	-	-	-	-	-	-	67%	3
ARTIST_sc_CI_02	-	-	-	-	-	-	-	-	75%	4
ARTIST_sc_CI_01	-	-	-	-	-	-	-	-	100%	3
ARTIST_sc_CI_07	-	-	-	-	-	-	-	-	100%	3
ARTIST_sc_CI_06	-	-	-	-	-	-	-	-	67%	3
ARTIST_sc_CI_10	-	-	-	-	-	-	-	-	100%	3
ARTIST_db_CIOSM_Q1394	-	-	-	-	-	-	25%	-	50%	4
ARTIST_db_CIOSM_Q1387	-	-	-	-	-	-	-	-	100%	3

4.2. Quantitative Evaluation

4.2.1. Methods

In addition to the qualitative evaluation, the items were administered on a large scale to evaluate students' item responses quantitatively.

Participants and Procedures

The items were translated into Dutch and implemented in introductory statistics courses taught during the fall of 2021 in five different bachelor programs at Utrecht University—namely, psychology, educational sciences, pedagogical sciences, sociology, and cultural anthropology. The item set was split into two assessments to reduce the burden on students. The first assessment consisted of 35 items and concerned all topics related to *Samples and spread*, encompassing attributes A–F. The second assessment consisted of 24 items and concerned all topics related to *NHST and confidence intervals (CIs)*, encompassing attributes G–I. The two item sets exclusively measure the attributes that are assessed within each assessment according to the provisional Q-matrix and both Q-matrices meet the conditions for generic identifiability from [40].

The assessments were made available after the relevant topics were addressed in class, but well before the final exam. They were presented as optional, formative assessments. Students were given only one attempt for each assessment to motivate them to make a serious effort. In total, 849 students completed the first assessment and 790 students completed the second assessment.

Analysis

The data were analyzed using the statistical software R [41]. We focused on diagnostic classification models (DCMs) under the log-linear cognitive diagnosis modeling (LCDM) framework because of its modeling flexibility and straightforward interpretation [12]. To validate attribute measurement with DCMs, we first evaluated the assumption of local independence using the local dependence statistic χ^2_{LD} for each item pair [42]. To control for multiple comparisons, p-value adjustments according to the Holm–Bonferroni method are conducted [43].

Next, we empirically evaluated the Q-matrix with the stepwise Wald method [21]. This method provides suggestions for modifications to Q-matrix entries based on the amount of explained variance in the success probabilities for different attribute profiles. This procedure is based on the data at hand, and whether the suggested modifications should be incorporated should be subject to the judgment of domain experts. Therefore, we carefully evaluated the suggested modifications in relation to theoretical considerations and the results from the think-aloud study to make decisions about the final Q-matrix.

We then examined attribute behavior at the item level. This refers to the compensatory nature of the attributes (i.e., whether nonmastery of an attribute can be compensated by mastery of another). The Wald test is used to evaluate whether, for items that require at least two attributes, the LCDM can be replaced by a reduced model without significant loss of fit, reflecting different attribute behaviors at the item level [44]. This can result in higher classification accuracy [45].

We end with an evaluation of the final model. Several fit statistics are evaluated, namely, the M_2 statistic [46], the $RMSEA_2$, and $SRMSR$ [47], and the maximum transformed correlation and log-odds ratio [48]. Finally, the attribute-level classification accuracy index from [49] and the reliability measurement from [50] are reported.

4.2.2. Results

Assumption of Local Independence

The results showed several item pairs with significant χ^2_{LD} values, indicating local dependence. For these item pairs, removal of one item was considered based on the amount of dependence with other items, extreme proportion correct scores, or the presence of items with similar q-vectors (i.e, rows in the Q-matrix). In total, three items were removed from each assessment. The final assessments consist of 32 and 21 items, respectively, which are used in the subsequent analyses and can be found in the Supplementary Material.

Q-Matrix Validation

The stepwise Wald method suggested modifications for the q-vectors of 14 items of the first assessment (21 suggested entry changes) and for the q-vectors of two items of the second assessment (3 suggested entry changes). These suggestions were evaluated based on three aspects. First, the items and attributes were re-evaluated to assess whether the suggested modifications were theoretically defensible. Second, we inspected the results from the think-aloud study to determine the percentage of students who used the attributes under consideration when answering the items, thereby including empirical evidence for whether or not the items measure the attributes. Third, we compared the proportion of variance accounted for (PVAF) of relevant q-vectors to evaluate to what extent the attributes under consideration contribute to explaining variance in success probabilities. Only if suggestions were sensible from a theoretical perspective were modifications considered.

In total, five Q-matrix entries were adjusted for five different items. These adjustments along with extensive argumentation can be found in Table A2 in Appendix A. Table 3 presents the final Q-matrices (along with reduced DCMs, which are discussed next). Both Q-matrices satisfy the conditions for generic identifiability from [40].

Table 3. Final Q-matrices and selected reduced DCMs per item.

Assessment 1			**Attribute**				**Reduced**	**Assessment 2**		**Attribute**		**Reduced**
Item	**A**	**B**	**C**	**D**	**E**	**F**	**Model**	**Item**	**G**	**H**	**I**	**Model**
ARTIST_sc_MS_05	1	0	0	0	0	0	-	CAOS_40	1	0	0	-
ARTIST_sc_MS_01	1	0	0	0	0	0	-	ARTIST_sc_TS_01	1	0	0	-
ARTIST_sc_MC_05	1	0	0	0	0	0	-	ARTIST_db_TSG_Q1182	1	0	0	-
ARTIST_sc_MC_06	1	0	0	0	0	0	-	CAOS_24	1	0	0	-
ARTIST_db_MS_Q0490	1	0	0	0	0	0	-	ARTIST_db_TSG_Q1392	0	1	0	-
CAOS_14	1	1	0	0	0	0	DINA	CAOS_25	0	1	0	-
CAOS_15	1	1	0	0	0	0	DINA	CAOS_26	0	1	0	-
CAOS_08	1	1	0	0	0	0	DINA	ARTIST_sc_TS_04	1	1	0	DINA
CAOS_09	0	1	0	0	0	0	-	ARTIST_sc_TS_10	0	1	0	-
CAOS_10	1	1	0	0	0	0	LLM	ARTIST_sc_TS_07	0	1	0	-
CAOS_12	0	1	1	0	0	0	R-RUM	ARTIST_sc_TS_09	0	1	0	-
CAOS_13	0	0	1	0	0	0	-	SCI_2004_22	0	1	0	-
SRA_015	0	1	1	0	0	0	R-RUM	ARTIST_db_TSG_Q1007	0	1	0	-
ARTIST_db_CG_Q0840	0	1	1	0	0	0	R-RUM	ARTIST_sc_CI_05	0	0	1	-
ARTIST_sc_SV_01	0	1	0	0	1	0	DINA	ARTIST_sc_CI_02	0	0	1	-
CAOS_17	0	0	0	1	0	0	-	ARTIST_sc_CI_01	0	0	1	-
ARTIST_sc_SV_03	0	0	0	1	0	1	LLM	ARTIST_sc_CI_07	0	0	1	-
ARTIST_sc_SV_14	0	0	0	1	0	0	-	ARTIST_sc_CI_06	0	0	1	-
ARTIST_sc_SV_04	0	0	0	1	0	0	-	ARTIST_sc_CI_10	0	0	1	-
ARTIST_sc_SV_10	0	0	0	0	1	0	-	ARTIST_db_CIOSM_Q1394	0	1	1	DINA
ARTIST_sc_SV_11	0	0	0	0	1	1	*A*-CDM	ARTIST_db_CIOSM_Q1387	0	0	1	-
ARTIST_sc_SV_09	0	0	0	0	1	1	R-RUM					
ARTIST_db_SS_Q0061A	0	0	0	0	1	0	-					
ARTIST_db_SS_Q0061B	0	0	0	0	1	0	-					
ARTIST_db_SS_Q0061C	0	0	0	0	1	0	-					
ARTIST_db_SS_Q0061D	0	0	0	0	1	0	-					
ARTIST_sc_SV_05	0	0	0	0	1	0	-					
CAOS_16	0	0	0	1	0	1	LLM					
CAOS_32	0	0	0	0	0	1	-					
SCI_2004_20	1	0	0	0	0	1	LLM					
ARTIST_db_SS_Q1437	0	0	0	0	0	1	-					
ARTIST_db_SS_Q0614	0	0	0	0	0	1	-					

Attribute Behavior at the Item Level

The Wald test was used to conduct item-level comparisons of the LCDM and reduced DCMs that it subsumes [44], namely, the deterministic inputs, noisy "and" gate (DINA) model [51]; the deterministic inputs, noisy "or" gate (DINO) model [52]; the additive cognitive diagnosis model (*A*-CDM; [53]); the linear logistic model (LLM; [54]); and the reduced reparameterized unified model (R-RUM; [55]). The Wald test suggested reduced models for all two-dimensional items: 13 and 2 items in assessments 1 and 2, respectively. The suggested models are presented in Table 3. Fitting these models did not significantly reduce model fit compared with fitting the saturated LCDM to all items, $\chi^2(17) = 9.8$, $p = 0.91$ (assessment 1) and $\chi^2(4) = 5.1$, $p = 0.28$ (assessment 2). The reduced models are used in the subsequent analyses.

Final Model

The fit statistics to evaluate the absolute fit of the final model are shown in Table 4. The statistics indicate that the model fits adequately (for the $RMSEA_2$, the cutoff values 0.030 and 0.045 indicate excellent and good fit for the LCDM [56]; for the $SRMSR$, values below 0.05 indicate acceptable model fit in DCMs, e.g., [57]). Further, classification accuracy and reliability are mostly high, as shown in Table 5. Detailed model results are included in the Supplementary Material.

Table 4. Absolute fit statistics of the final models.

	M_2	$RMSEA_2$	$SRMSR$	Max Transformed Correlation *	Max Log-Odds Ratio *
Assessment 1	591.4 ($p < 0.001$)	0.025	0.042	0.128 ($p = 0.10$)	0.982 ($p = 0.19$)
Assessment 2	212.1 ($p = 0.06$)	0.014	0.038	0.104 ($p = 0.76$)	0.850 ($p = 0.49$)

* To control for multiple comparisons, p-value adjustments according to the Holm–Bonferroni method are conducted [43].

Table 5. Attribute-level classification accuracy and reliability.

			Assessment 1					Assessment 2	
Attribute	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Accuracy	0.89	0.83	0.90	0.89	0.82	0.82	0.82	0.84	0.88
Reliability	0.87	0.69	0.89	0.87	0.68	0.67	0.67	0.76	0.85

5. Communicating Feedback

With the final model results, students' mastery status can be determined based on their item responses resulting in diagnostic information. Diagnostic assessments should provide meaningful information to students that can be readily understood and used to improve learning. In this section, we pay attention to the psychometric interpretation of the results, i.e., to the inferences that can be made based on results from cognitive diagnostic assessment with DCMs. In addition, we discuss score reporting based on a mock-up dashboard to show how this information can be communicated to students.

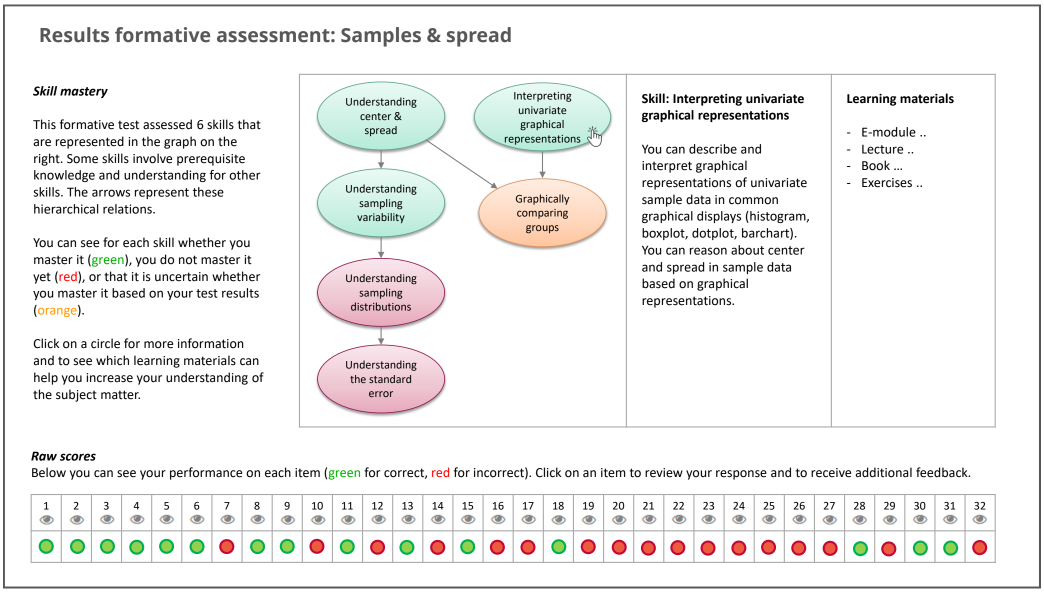
5.1. Psychometric Interpretation

Before students can be supported to interpret assessment results, assessment developers must accurately interpret the results from a psychometric perspective. DCMs assume that the latent variables predicting the item responses (i.e., the attributes) are discrete variables. For each attribute, two groups are distinguished: those who master the attribute and those who do not. With DCMs, one can estimate the mastery probability for each attribute (expected a posteriori estimates). Classification as master or nonmaster is based on these marginal mastery probabilities. For example, probabilities above 0.5 result in classification as master and below 0.5 as nonmaster (although some contexts may call for a different cut-off value).

The mastery probability is not an estimate of the amount of mastery, nor of a student's progress from nonmastery to mastery. Rather, it indicates the (un)certainty of the classification, which is determined by students' consistency in providing either correct or incorrect answers to different items measuring an attribute [58]. The closer the mastery probability is to 0 or 1, the more certainty there is about the classification, whereas a probability of 0.5 reflects complete uncertainty. Although these models do not allow to break the probability scale into more than two groups to indicate more than two mastery levels, it is possible to break the scale to indicate uncertainty. One could for example not classify students on attributes with mastery probabilities between 0.4 and 0.6, but instead add an "undetermined" category for these uncertain situations (as proposed by [58], and as will be demonstrated below), or one could flag uncertain classifications when reporting results.

5.2. Reporting Results

If the assessments and model results are implemented in online learning environments, feedback can be reported directly after completion of the assessments. This is a great advantage, since timeliness is important to encourage stakeholders to use diagnostic information [9]. Further, web-based reporting allows to manage and organize large amounts of information via interactive dashboards [59,60]. Ref. [61] present a framework for developing score reports for cognitive diagnostic assessments. Following their guidelines, we created a mock-up student dashboard that is personalized, contains a description of the presented information, provides a visual summary of student performance that can be readily interpreted, and outlines how this information can guide study behavior. The mock-up dashboard is presented in Figure 2.



Results formative assessment: Samples & spread

Skill mastery

This formative test assessed 6 skills that are represented in the graph on the right. Some skills involve prerequisite knowledge and understanding for other skills. The arrows represent these hierarchical relations.

You can see for each skill whether you master it (green), you do not master it yet (red), or that it is uncertain whether you master it based on your test results (orange).

Click on a circle for more information and to see which learning materials can help you increase your understanding of the subject matter.

Understanding center & spread
Interpreting univariate graphical representations
Understanding sampling variability
Graphically comparing groups
Understanding sampling distributions
Understanding the standard error

Skill: Interpreting univariate graphical representations

You can describe and interpret graphical representations of univariate sample data in common graphical displays (histogram, boxplot, dotplot, barchart). You can reason about center and spread in sample data based on graphical representations.

Learning materials

- E-module ..
- Lecture ..
- Book ...
- Exercises ..

Raw scores

Below you can see your performance on each item (green for correct, red for incorrect). Click on an item to review your response and to receive additional feedback.

1 👁	2 👁	3 👁	4 👁	5 👁	6 👁	7 👁	8 👁	9 👁	10 👁	11 👁	12 👁	13 👁	14 👁	15 👁	16 👁	17 👁	18 👁	19 👁	20 👁	21 👁	22 👁	23 👁	24 👁	25 👁	26 👁	27 👁	28 👁	29 👁	30 👁	31 👁	32 👁
🟢	🟢	🟢	🟢	🟢	🟢	🔴	🟢	🟢	🔴	🟢	🔴	🟢	🔴	🟢	🔴	🔴	🟢	🔴	🔴	🔴	🔴	🔴	🔴	🔴	🔴	🔴	🟢	🔴	🟢	🟢	🔴

Figure 2. Mock-up student dashboard to report cognitive diagnostic assessment results.

The dashboard consists of two sections to report both attribute-level scores and raw scores. We focus on the first section, titled "Skill mastery", which presents attribute-level performance. On the left, it is explained what information is presented in this section. On the right, a graph is provided that visualizes the attributes (the presented hierarchy reflects attribute dependencies, indicating whether mastery of an attribute is a prerequisite to mastery of another attribute; this hierarchy was determined in consultation with domain experts). If a student clicks on an attribute, brief descriptions of the measured skills are shown, as well as recommended learning materials to improve these skills. Indicating skill mastery is substantively meaningful, because the inferences about student performance are made with reference to the cognitive skills measured by the assessment, allowing for criterion-referenced interpretations (cut-scores in DCMs are set to maximize the reliable separation of respondents, i.e., the criterion is set statistically; [12], Chapter 5). We have used the familiar terms "mastery" versus "nonmastery" to label the classifications throughout this paper. Alternatively, one can choose labels that express in-progress learning, such as "on track" versus "needs attention". The labels can influence how the information is interpreted and used by students and it is important to select labels that are appropriate for a given context [58].

6. Discussion

The current study showed the validation process of constructing detailed diagnostic information on a set of skills, abilities, and cognitive processes from students' item response data. This integration of learning analytics and educational assessment promotes valid and reliable formative assessment. We constructed cognitive diagnostic assessments to measure attributes relevant for nonmathematical introductory statistics courses in higher education using diagnostic classification models. To construct the assessments, we did not write new items to measure the identified attributes, but we relied on existing items, which may limit the scope of what can be measured [62]. However, we did not define attributes based on the items but searched for items based on the prespecified attributes. We exploited multiple sources for item collection and evaluated whether all aspects of the attributes were measured. Moreover, we validated a new measurement scale for actionable feedback based on the existing items, which can be viewed as a contribution of our work. The final models showed adequate model fit, classification accuracy, and reliability. The constructed assessments provide a valid and reliable measurement of the specified attributes and allow provision of actionable feedback to students via learning dashboards.

6.1. Implications for Educational Practice

If the assessments are implemented in online learning environments of introductory statistics courses, the mastery status of new students can be determined directly after they complete an assessment based on their automatically scored item responses. Based on timely, actionable diagnostic feedback, students can make more effective learning choices [10].

The value of feedback depends on how it is interpreted by the recipient and, in turn, influences learning choices. Although both interpretation and use are critical to the effectiveness of formative assessment [63], only few studies have examined actual uses of score reports (see [64] for a review). It is an interesting avenue for future research to explore how diagnostic feedback impacts students' learning processes and whether there are individual differences in the effects of diagnostic feedback on learning choices (e.g., due to differences in expertise level or self-regulated learning skills; [65,66]).

6.2. Limitations

Although the constructed assessments provide a valid and reliable measurement of the specified attributes, this study presents several limitations. One limitation lies in the small sample size in the think-aloud study due to limited resources, which may limit the generalizability of the qualitative evaluation. Furthermore, verbal reports may not completely reflect students' cognitive processes [67]. Especially if constructs are not well-understood, verbal reports can be subject to bias and error. The participants in the think-aloud mostly rated their performance as (very) good. Although it would be interesting to include students with low performance, it is difficult to collect high-quality verbal report data from those students. Even for students with high performance, verbal reports may provide an incomplete record of their knowledge and cognitive processes as they solved the items. Therefore, the results from the think-aloud were not used as hard evidence, but rather as supportive information in the empirical Q-matrix validation.

Further, in the quantitative evaluation, the assessments were offered as optional study material. This might have influenced students' motivation, for example, resulting in rapid-guessing behavior that can lead to differential item functioning [68]. Moreover, it could cause selection bias, for example, if the assessments were mostly made by students who did not grasp the materials yet, resulting in a low base rate of attribute mastery. However, Ref. [69] showed that DCM item parameters are theoretically invariant with respect to the calibration sample; thus, the estimates are not influenced by the base rate of attribute mastery. Since we found adequate model-data fit, the calibrated assessments can be used to appropriately classify new students. Nevertheless, changes in parameters can be expected [70] and regular monitoring of parameter invariance is recommended.

6.3. Concluding Remarks

To conclude, cognitive diagnostic assessment can be a valuable tool within e-learning environments to obtain timely, diagnostic feedback on cognitive attributes to support student learning. Since poor quality assessment leads to less-effective learning choices, it is important to validate whether assessments are adequate and appropriate for their intended interpretation and use [71]. The current study was focused on validating the interpretation of assessments in university statistics education (i.e., the claims that can be made about students), but it is important to also verify whether appropriate actions follow if these interpretations are presented to students (e.g., following [72]).

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app12104809/s1.

Author Contributions: Conceptualization, L.M., M.J.S.B., L.K. and L.W.-d.M.; data curation, L.M.; formal analysis, L.M.; investigation, L.M.; methodology, L.M., M.J.S.B., L.K. and L.W.-d.M.; supervision, L.W.-d.M.; validation, L.M.; writing—original draft preparation, L.M.; writing—review and editing, M.J.S.B., L.K. and L.W.-d.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was approved by the Ethics Review Board of the Faculty of Social and Behavioural Sciences at Utrecht University (protocol number 21-0115, date of approval 1 April 2021).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy/ethical restrictions.

Acknowledgments: The authors sincerely thank the experts who advised them and supported the study.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DCM	Diagnostic classification model
LCDM	Log-linear cognitive diagnosis model
DINA	Deterministic inputs, noisy "and" gate
DINO	Deterministic inputs, noisy "or" gate
A-CDM	Additive cognitive diagnosis model
LLM	Linear logistic model
R-RUM	Reduced reparameterized unified model

Appendix A

This section presents more extensive results from (1) the attribute identification process and (2) the empirical Q-matrix validation:

- Table A1 presents an overview of the 9 attributes that resulted from the attribute identification process, which are comprised of 33 learning objectives. The table includes references to empirical studies that show the relevance of the addressed concepts and the occurrence of misconceptions for each learning objective. In the Supplementary Material, a more detailed description of each attribute and the references to the empirical studies is provided.
- Table A2 presents all items for which modifications to the Q-matrix were suggested in the empirical Q-matrix validation. Argumentation to retain or modify entries is provided along with the final q-vectors.

Table A1. Identified attributes and learning objectives with references showing their relevance.

A.	**Understanding center & spread**			
	1.	Understanding of the idea of variability in samples and populations.	-	[73]
	2.	Ability to describe and interpret measures of center (mean, median, mode).	-	[27], Chapter 9
	3.	Ability to describe and interpret measures of spread (range, interquartile range, variance, standard deviation).	-	[73]
	4.	Understanding of how center and spread are affected by data transformations.	-	[28]
	5.	Understanding of how center and spread are affected by handling of outliers.	-	[27], Chapters 9 and 10
B.	**Interpreting univariate graphical representations**			
	6.	Ability to describe and interpret graphical representations of sample data in common univariate graphical displays (histogram, boxplot, dotplot, bar chart), including the ability to reason about center and spread in sample data based on graphical representations.	-	[28,74]
C.	**Graphically comparing groups**			
	7.	Ability to compare groups on a continuous outcome variable by focusing on data as an aggregate with characteristics such as center, spread, and shape.	-	[27], Chapters 9 and 10; [28,75]
	8.	Ability to make informal inferences about group differences based on graphical representations by comparing variability between and within groups, i.e., by considering both center and spread.	-	[27], Chapter 11; [76]
D.	**Understanding sampling variability**			
	9.	Understanding that a sample provides incomplete information about the population from which it is drawn and that random sampling forms the basis for statistical inference.	-	[27], Chapter 12; [77]
	10.	Understanding how the set of all possible random samples (with the given size) from the population of interest is considered in inferential statistics, i.e., that the unit of analysis is the entire sample rather than a single observation.	-	[27], Chapters 12 and 13; [78]
	11.	Understanding that not all samples are identical, so not all of them will resemble the population in the same way and to the same extent every time.	-	[79]
	12.	Understanding of expected patterns in sampling variability, in which some values are more or less likely than others to be drawn from a particular population.	-	[27], Chapter 12; [28]
E.	**Understanding sampling distributions**			
	13.	Ability to describe and distinguish between the sample distribution, sampling distribution, and population distribution.	-	[27], Chapter 12; [79,80]
	14.	Knowing that, given a sufficiently large sample size, the sampling distribution of the mean for a variable will approximate a normal distribution.	-	[27], Chapter 12; [81,82]
	15.	Knowing how the (unknown) population mean, the sample mean, the possible sample mean values for different random samples of the given size, and the theoretical mean of these possible values relate to each other.	-	[27], Chapter 12; [83]
F.	**Understanding the standard error**			
	16.	Ability to interpret the standard error as a measure of variability of sample means.	-	[27], Chapter 12
	17.	Understanding that statistics from small samples vary more than statistics from large samples and knowing how the standard error decreases with sample size.	-	[27], Chapter 12; [28,79]
	18.	Understanding of how population variance influences sampling variability and, thus, the standard error.	-	[27], Chapter 12
	19.	Ability to make informal inferences about sample means based on measures of sampling variability.	-	[28]
G.	**Understanding principles of hypothesis testing**			
	20.	Understanding of the goals and the logic of significance tests; understanding of the theoretical idea of finding evidence against a null hypothesis.	-	[84]
	21.	Understanding that statistical inferences are not deterministic and NHST results do not prove the null hypothesis to be true or false.	-	[84,85]
H.	**Evaluating NHST results**			
	22.	Ability to take a correct decision about the null hypothesis based on the significance level and *p*-value.	-	[86]
	23.	Ability to correctly interpret *p*-values as the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is true.	-	[28,87–90]
	24.	Ability to describe the concepts Type I error, Type II error, significance level, and statistical power.	-	[86,91]
	25.	Understanding of how increasing the sample size increases power by reducing the standard error.	-	[27], Chapter 13; [86]
	26.	Understanding of how increasing the significance level increases power by enlarging the rejection region.	-	[86]
	27.	Understanding of how larger effect sizes result in higher power by enlarging the difference between the means of the null and alternative distribution.	-	[86]
	28.	Understanding that statistical significance does not imply practical relevance and vice versa.	-	[28,92,93]
I.	**Understanding and using confidence intervals**			
	29.	Understanding that CIs are used to provide an indication of the precision of an estimate.	-	[28,94,95]
	30.	Ability to correctly interpret CIs in terms of repeated sampling (e.g., if repeated samples were taken and the 95% confidence interval was computed for each sample mean, 95% of the intervals would contain the population mean).	-	[96]
	31.	Understanding of how increasing the sample size decreases CI width by reducing the standard error.	-	[97]
	32.	Understanding of how increasing the confidence level increases CI width by reducing the precision of the estimate.	-	[97]
	33.	Understanding of connections between CIs and hypothesis tests; ability to draw a conclusion about the null hypothesis based on a CI (i.e., if the CI does not contain the null hypothesis value, the results are statistically significant).	-	[98]

Table A2. Items for which the stepwise Wald method suggested q-vector modifications. The table shows the original and suggested q-vectors that represent rows in the Q-matrix. Further, the table shows (1) the item content re-evaluation indicating whether suggestions to include or exclude attributes could be theoretically defensible (yes/no), (2) the percentage of students in the think-aloud (TA) who used the attributes for which modifications are suggested, (3) the PVAF for relevant q-vectors, (4) the argumentation to retain or modify q-vector entries, and (5) the final q-vectors (modifications in boldface).

Item	q-Vector		Content Re-Evaluation: Theoretical Defensibility	Attribute Use in TA	PVAF	Argumentation	Final q-Vector
CAOS_15	Orig. Sugg.	110000 100000	Exclude *B*: no	*B*: 100%	100000: 0.328 110000: 0.816	Att. *B* should be included from a theoretical perspective and explains a substantial proportion of additional variance.	110000
CAOS_08	Orig. Sugg.	110000 010000	Exclude *A*: no	*A*: 67%	010000: 0.464 110000: 0.835	Att. *A* should be included from a theoretical perspective and explains a substantial proportion of additional variance.	110000
CAOS_12	Orig. Sugg.	111000 001000	Exclude *A*: yes Exclude *B*: no	*A*: 0% *B*: 100%	001000: 0.959 011000: 0.964 111000: 0.985	Although att. *B* explains little additional variance, it cannot be excluded from a theoretical perspective. Att. *A* can be excluded and explains little additional variance; therefore, it is excluded.	**0**11000
CAOS_13	Orig. Sugg.	011000 001000	Exclude *B*: yes	*B*: 0%	001000: 0.999 011000: 0.999	Att. *B* can be excluded from a theoretical perspective and explains no additional variance; therefore, it is excluded.	0**0**1000
SRA_15	Orig. Sugg.	001000 010000	Include *B*: yes Exclude C: no	*B*: 75% *C*: 100%	001000: 0.002 010000: 0.437 011000: 0.469	Although att. *C* explains a very small proportion of variance, it cannot be excluded from a theoretical perspective. Att. *B* can be included from a theoretical perspective and explains a substantial proportion of additional variance; therefore, it is included.	0**1**1000
ARTIST_db_CG_Q0840	Orig. Sugg.	011000 001000	Exclude *B*: no	*B*: 100%	001000: 0.876 011000: 0.969	Att. *B* should be included from a theoretical perspective and explains some additional variance.	011000
ARTIST_sc_SV_04	Orig. Sugg.	000100 001000	Include *C*: no Exclude *D*: no	*C*: 0% *D*: 60%	000100: 0.056 001000: 0.183	Including att. *C* and excluding att. *D* is not theoretically defensible. Despite the higher PVAF for att. *C*, the original q-vector is retained.	000100
ARTIST_sc_SV_10	Orig. Sugg.	000110 000010	Exclude *D*: yes	*D*: 0%	000010: 0.945 000110: 0.986	Att. *D* can be excluded from a theoretical perspective and explains only little additional variance; therefore, it is excluded.	000**0**10
ARTIST_sc_SV_11	Orig. Sugg.	000011 000010	Exclude *F*: no	*F*: 50%	000010: 0.949 000011: 0.951	Although att. *F* explains only a small proportion of additional variance; it cannot be excluded from a theoretical perspective.	000011
ARTIST_db_SS_Q0061A	Orig. Sugg.	000010 100000	Include *A*: no Exclude *E*: no	*A*: 0% *E*: 100%	000010: 0.016 100000: 0.597	Including att. *A* and excluding att. *E* is not theoretically defensible. Despite the higher PVAF for att. *A*, the original q-vector is retained.	000010
ARTIST_db_SS_Q0061C	Orig. Sugg.	000010 001000	Include *C*: no Exclude *E*: no	*C*: 0% *E*: 100%	000010: 0.232 001000: 0.288	Including att. *C* and excluding att. *E* is not theoretically defensible. Despite the higher PVAF for att. *C*, the original q-vector is retained.	000010
ARTIST_db_SS_Q0061D	Orig. Sugg.	000010 001000	Include *C*: no Exclude *E*: no	*C*: 0% *E*: 100%	000010: 0.121 001000: 0.182	Including att. *C* and excluding att. *E* is not theoretically defensible. Despite the higher PVAF for att. *C*, the original q-vector is retained.	000010
CAOS_32	Orig. Sugg.	000001 000010	Include *E*: no Exclude *F*: no	*E*: 0% *F*: 100%	000001: 0.113 000010: 0.379	Including att. *E* and excluding att. *F* is not theoretically defensible. Despite the higher PVAF for att. *E*, the original q-vector is retained.	000001
SCI_2004_20	Orig. Sugg.	100001 100000	Exclude *F*: no	*F*: 100%	100000: 0.908 100001: 0.958	Although att. *F* explains only a small proportion of additional variance; it cannot be excluded from a theoretical perspective.	100001
ARTIST_sc_TS_04	Orig. Sugg.	110 010	Exclude *G*: no	*G*: 100%	010: 0.849 110: 0.999	Att. *G* should be included from a theoretical perspective and explains some additional variance.	110
ARTIST_db_CIOSM_Q1394	Orig. Sugg.	001 010	Include *H*: yes Exclude *I*: no	*H*: 0% *I*: 50%	001: 0.180 010: 0.885 011: 0.897	Att. *H* can be included from a theoretical perspective and explains a substantial proportion of variance; therefore, it is included. Although att. *I* explains only little additional variance, it cannot be excluded from a theoretical perspective.	0**1**1

References

1. Sitzmann, T.; Ely, K. A meta-analysis of self-regulated learning in work-related training and educational attainment: What we know and where we need to go. *Psychol. Bull.* **2011**, *137*, 421–442. [CrossRef] [PubMed]
2. Dunlosky, J.; Rawson, K.A. Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learn. Instr.* **2012**, *22*, 271–280. [CrossRef]
3. Lee, K. Rethinking the accessibility of online higher education: A historical review. *Internet High. Educ.* **2017**, *33*, 15–23. [CrossRef]
4. Gikandi, J.W.; Morrow, D.; Davis, N.E. Online formative assessment in higher education: A review of the literature. *Comput. Educ.* **2011**, *57*, 2333–2351. [CrossRef]
5. Brinkhuis, M.J.S.; Cordes, W.; Hofman, A. Governing games: Adaptive game selection in the Math Garden. *ITM Web of Conf.* **2020**, *33*, 03003. [CrossRef]
6. Quilici, J.L.; Mayer, R.E. Teaching students to recognize structural similarities between statistics word problems. *Appl. Cogn. Psychol. Off. J. Soc. Appl. Res. Mem. Cogn.* **2002**, *16*, 325–342. [CrossRef]
7. Guskey, T.R. The case against percentage grades. *Educ. Sch. Couns. Psychol. Fac. Publ.* **2013**, *71*, 68–72.
8. Leighton, J.P.; Gierl, M.J. Why cognitive diagnostic assessment? In *Cognitive Diagnostic Assessment for Education*; Leighton, J.P., Gierl, M.J., Eds.; Cambridge University Press: New York, NY, USA, 2007; pp. 3–18. [CrossRef]
9. Huff, K.; Goodman, D.P. The demand for cognitive diagnostic assessment. In *Cognitive Diagnostic Assessment for Education*; Leighton, J.P., Gierl, M.J., Eds.; Cambridge University Press: New York, NY, USA, 2007; pp. 19–60.
10. Kanar, A.M.; Bell, B.S. Guiding learners through technology-based instruction: The effects of adaptive guidance design and individual differences on learning over time. *J. Educ. Psychol.* **2013**, *105*, 1067–1081. [CrossRef]
11. Norris, S.P.; Macnab, J.S.; Phillips, L.M. Cognitive modeling of performance on diagnostic achievement tests: A Philosophical Analysis and Justification. In *Cognitive Diagnostic Assessment for Education*; Leighton, J.P., Gierl, M.J., Eds.; Cambridge University Press: New York, NY, USA, 2007; pp. 61–84. [CrossRef]
12. Rupp, A.A.; Templin, J.; Henson, R.A. *Diagnostic Measurement: Theory, Methods, and Applications*; The Guilford Press: New York, NY, USA, 2010.
13. Maas, L.; Brinkhuis, M.J.S.; Kester, L.; Wijngaards-de Meij, L. Diagnostic classification models for actionable feedback in education: Effects of sample size and assessment length. *Front. Educ.* **2022**, *7*, 36. [CrossRef]
14. Wiggins, G.; McTighe, J. *Understanding by Design*; Association for Supervision and Curriculum Development: Alexandria, VA, USA, 2005.
15. Pellegrino, J.W.; Chudowsky, N.; Glaser, R. *Knowing What Students Know: The Science and Design of Educational Assessment*; National Academy Press: Washington, DC, USA, 2001. [CrossRef]
16. Thompson, K.; Yonekura, F. Practical guidelines for learning object granularity from one higher education setting. *Interdiscip. J.-Learn. Learn. Objects* **2005**, *1*, 163–179. [CrossRef]
17. Rupp, A.A.; Templin, J. The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educ. Psychol. Meas.* **2008**, *68*, 78–96. [CrossRef]
18. Kunina-Habenicht, O.; Rupp, A.A.; Wilhelm, O. The impact of model misspecification on estimation accuracy in diagnostic classification models. *J. Educ. Meas.* **2012**, *49*, 59–81. [CrossRef]
19. Leighton, J.P.; Gierl, M.J. Verbal reports as data for cognitive diagnostic assessment. In *Cognitive Diagnostic Assessment for Education*; Leighton, J.P., Gierl, M.J., Eds.; Cambridge University Press: New York, NY, USA, 2007; pp. 146–172. [CrossRef]
20. Tjoe, H.; de la Torre, J. The identification and validation process of proportional reasoning attributes: An application of a cognitive diagnosis modeling framework. *Math. Educ. Res. J.* **2014**, *26*, 237–255. [CrossRef]
21. Ma, W.; de la Torre, J. An empirical Q-matrix validation method for the sequential generalized DINA model. *Br. J. Math. Stat. Psychol.* **2020**, *73*, 142–163. [CrossRef]
22. Castro Sotos, A.E.; Vanhoof, S.; Van den Noortgate, W.; Onghena, P. Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educ. Res. Rev.* **2007**, *2*, 98–113. [CrossRef]
23. Garfield, J.B.; Ben-Zvi, D.; Chance, B.; Medina, E.; Roseth, C.; Zieffler, A. Assessment in statistics education. In *Developing Students' Statistical Reasoning*; Springer: Berlin, Germany, 2008; pp. 82–114. [CrossRef]
24. Tacoma, S.; Sosnovsky, S.; Boon, P.; Jeuring, J.; Drijvers, P. The interplay between inspectable student models and didactics of statistics. *Digit. Exp. Math. Educ.* **2018**, *4*, 139–162. [CrossRef]
25. Cui, Y.; Roduta Roberts, M. Validating Student Score Inferences With Person-Fit Statistic and Verbal Reports: A Person-Fit Study for Cognitive Diagnostic Assessment. *Educ. Meas. Issues Pract.* **2013**, *32*, 34–42. [CrossRef]
26. delMas, R. A comparison of mathematical and statistical reasoning. In *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*; Ben-Zvi, D., Garfield, J.B., Eds.; Springer: Dordrecht, The Netherlands, 2004; pp. 79–95. [CrossRef]
27. Garfield, J.B.; Ben-Zvi, D. *Developing Students' Statistical Reasoning: Connecting Research and Teaching Practice*; Springer Science & Business Media: Berlin, Germany, 2008. [CrossRef]
28. delMas, R.; Garfield, J.B.; Ooms, A.; Chance, B. Assessing students' conceptual understanding after a first course in statistics. *Stat. Educ. Res. J.* **2007**, *6*, 28–58. [CrossRef]
29. GAISE. *Guidelines for Assessment and Instruction in Statistics Education College Report 2016*; American Statistical Association: Alexandria, VA, USA, 2016.

30. Madison, M.J.; Bradshaw, L. The effects of Q-matrix design on classification accuracy in the log-linear cognitive diagnosis model. *Educ. Psychol. Meas.* **2015**, *75*, 491–511. [CrossRef]
31. Garfield, J.B.; Ben-Zvi, D. Research on statistical literacy, reasoning, and thinking: Issues, challenges, and implications. In *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*; Ben-Zvi, D., Garfield, J.B., Eds.; Springer: Dordrecht, The Netherlands, 2004; pp. 397–409. [CrossRef]
32. Garfield, J.B.; Chance, B. Assessment in statistics education: Issues and challenges. *Math. Think. Learn.* **2000**, *2*, 99–125. [CrossRef]
33. Gal, I.; Garfield, J.B. Curricular goals and assessment challenges in statistics education. In *The Assessment Challenge in Statistics Education*; Gal, I., Garfield, J.B., Eds.; IOS Press: Amsterdam, The Netherlands, 1997; pp. 1–13.
34. Garfield, J.B. Assessing statistical reasoning. *Stat. Educ. Res. J.* **2003**, *2*, 22–38. [CrossRef]
35. Allen, K. The Statistics Concept Inventory: Development and Analysis of a Cognitive Assessment Instrument in Statistics. Ph.D. Dissertation, University of Oklahoma, Norman, OK, USA, 2006.
36. Haladyna, T.M.; Downing, S.M.; Rodriguez, M.C. A review of multiple-choice item-writing guidelines for classroom assessment. *Appl. Meas. Educ.* **2002**, *15*, 309–333. [CrossRef]
37. Garfield, J.B.; Franklin, C. Assessment of learning, for learning, and as learning in statistics education. In *Teaching Statistics in School Mathematics-Challenges for Teaching and Teacher Education*; Batanero, C., Burrill, G., Reading, C., Eds.; Springer: Dordrecht, The Netherlands, 2011; pp. 133–145. [CrossRef]
38. Kirilenko, A.P.; Stepchenkova, S. Inter-coder agreement in one-to-many classification: Fuzzy kappa. *PLoS ONE* **2016**, *11*, e0149787. [CrossRef]
39. Ericsson, K.A.; Simon, H.A. How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind Cult. Act.* **1998**, *5*, 178–186. [CrossRef]
40. Gu, Y.; Xu, G. Sufficient and Necessary Conditions for the Identifiability of the Q-matrix. *Stat. Sin.* **2021**, *31*, 449–472. [CrossRef]
41. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019.
42. Chen, W.H.; Thissen, D. Local dependence indexes for item pairs using item response theory. *J. Educ. Behav. Stat.* **1997**, *22*, 265–289. [CrossRef]
43. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
44. de la Torre, J.; Lee, Y.S. Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *J. Educ. Meas.* **2013**, *50*, 355–373. [CrossRef]
45. Ma, W.; Iaconangelo, C.; de la Torre, J. Model similarity, model selection, and attribute classification. *Appl. Psychol. Meas.* **2016**, *40*, 200–217. [CrossRef]
46. Maydeu-Olivares, A.; Joe, H. Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika* **2006**, *71*, 713. [CrossRef]
47. Maydeu-Olivares, A.; Joe, H. Assessing approximate fit in categorical data analysis. *Multivar. Behav. Res.* **2014**, *49*, 305–328. [CrossRef]
48. Chen, J.; de la Torre, J.; Zhang, Z. Relative and absolute fit evaluation in cognitive diagnosis modeling. *J. Educ. Meas.* **2013**, *50*, 123–140. [CrossRef]
49. Wang, W.; Song, L.; Chen, P.; Meng, Y.; Ding, S. Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. *J. Educ. Meas.* **2015**, *52*, 457–476. [CrossRef]
50. Templin, J.; Bradshaw, L. Measuring the reliability of diagnostic classification model examinee estimates. *J. Classif.* **2013**, *30*, 251–275. [CrossRef]
51. Haertel, E.H. Using restricted latent class models to map the skill structure of achievement items. *J. Educ. Meas.* **1989**, *26*, 301–321. [CrossRef]
52. Templin, J.; Henson, R.A. Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* **2006**, *11*, 287. [CrossRef]
53. de la Torre, J. The generalized DINA model framework. *Psychometrika* **2011**, *76*, 179–199. [CrossRef]
54. Maris, E. Estimating multiple classification latent class models. *Psychometrika* **1999**, *64*, 187–212. [CrossRef]
55. DiBello, L.V.; Stout, W.F.; Roussos, L.A. Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In *Cognitively Diagnostic Assessment*; Nichols, P.D., Chipman, S.F., Brennan, R.L., Eds.; Erlbaum: Hillsdale, NJ, USA, 1995; pp. 361–390.
56. Liu, Y.; Tian, W.; Xin, T. An application of M2 statistic to evaluate the fit of cognitive diagnostic models. *J. Educ. Behav. Stat.* **2016**, *41*, 3–26. [CrossRef]
57. Liu, R.; Huggins-Manley, A.C.; Bulut, O. Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educ. Psychol. Meas.* **2018**, *78*, 357–383. [CrossRef]
58. Bradshaw, L.; Levy, R. Interpreting probabilistic classifications from diagnostic psychometric models. *Educ. Meas. Issues Pract.* **2019**, *38*, 79–88. [CrossRef]
59. Aljohani, N.R.; Davis, H.C. Learning analytics and formative assessment to provide immediate detailed feedback using a student centered mobile dashboard. In Proceedings of the 2013 Seventh International Conference on Next Generation Mobile Apps, Services and Technologies, Prague, Czech Republic, 25–27 September 2013; pp. 262–267. [CrossRef]

60. Verbert, K.; Govaerts, S.; Duval, E.; Santos, J.L.; Van Assche, F.; Parra, G.; Klerkx, J. Learning dashboards: An overview and future research opportunities. *Pers. Ubiquitous Comput.* **2014**, *18*, 1499–1514. [CrossRef]
61. Roduta Roberts, M.; Gierl, M.J. Developing score reports for cognitive diagnostic assessments. *Educ. Meas. Issues Pract.* **2010**, *29*, 25–38. [CrossRef]
62. de la Torre, J.; Minchen, N. Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicol. Educ.* **2014**, *20*, 89–97. [CrossRef]
63. Bennett, R.E. Formative assessment: A critical review. *Assess. Educ. Princ. Policy Pract.* **2011**, *18*, 5–25. [CrossRef]
64. Gotch, C.M.; Roduta Roberts, M. A review of recent research on individual-level score reports. *Educ. Meas. Issues Pract.* **2018**, *37*, 46–54. [CrossRef]
65. Roelle, J.; Berthold, K.; Fries, S. Effects of feedback on learning strategies in learning journals: Learner-expertise matters. In *Virtual Learning Environments: Concepts, Methodologies, Tools and Applications*; Khosrow-Pour, M., Ed.; IGI Global: Hershey, PA, USA 2012; pp. 710–725. [CrossRef]
66. Clarebout, G.; Horz, H.; Schnotz, W.; Elen, J. The relation between self-regulation and the embedding of support in learning environments. *Educ. Technol. Res. Dev.* **2010**, *58*, 573–587. [CrossRef]
67. Leighton, J.P. Rethinking think-alouds: The often-problematic collection of response process data. *Appl. Meas. Educ.* **2021**, *34*, 61–74. [CrossRef]
68. DeMars, C.E.; Wise, S.L. Can differential rapid-guessing behavior lead to differential item functioning? *Int. J. Test.* **2010**, *10*, 207–229. [CrossRef]
69. Bradshaw, L.; Madison, M.J. Invariance properties for general diagnostic classification models. *Int. J. Test.* **2016**, *16*, 99–118. [CrossRef]
70. Brinkhuis, M.J.S.; Maris, G. Tracking Ability: Defining Trackers for Measuring Educational Progress. In *Theoretical and Practical Advances in Computer-Based Educational Measurement*; Veldkamp, B.P., Sluijter, C., Eds.; Methodology of Educational Measurement and Assessment; Springer International Publishing: Cham, Switzerland, 2019; chapter 8, pp. 161–173.
71. Kane, M.T. Validating the interpretations and uses of test scores. *J. Educ. Meas.* **2013**, *50*, 1–73. [CrossRef]
72. Hopster-den Otter, D.; Wools, S.; Eggen, T.J.; Veldkamp, B.P. A general framework for the validation of embedded formative assessment. *J. Educ. Meas.* **2019**, *56*, 715–732. [CrossRef]
73. delMas, R.; Liu, Y. Exploring students' conceptions of the standard deviation. *Stat. Educ. Res. J.* **2005**, *4*, 55–82. [CrossRef]
74. Bakker, A.; Gravemeijer, K.P.E. Learning to reason about distribution. In *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*; Ben-Zvi, D., Garfield, J.B., Eds.; Springer: Dordrecht, The Netherlands, 2004; pp. 147–168.
75. Konold, C.; Pollatsek, A.; Well, A.; Gagnon, A. Students analyzing data: Research of critical barriers. In *Research on the Role of Technology in Teaching and Learning Statistics*; Springer: Dordrecht, The Netherlands, 1997; pp. 151–167.
76. Garfield, J.B. The challenge of developing statistical reasoning. *J. Stat. Educ.* **2002**, *10*. [CrossRef]
77. Tversky, A.; Kahneman, D. Belief in the law of small numbers. *Psychol. Bull.* **1971**, *76*, 105. [CrossRef]
78. Schuyten, G. Statistical thinking in psychology and education. In *Proceedings of the 3rd International Conference on Teaching Statistics: Vol. 2. Teaching Statistics Beyond School Level*; Vere-Jones, D., Ed.; ISI Publications in Statistical Education: Dunedin, New Zealand, 1991; pp. 486–490.
79. Chance, B.; delMas, R.; Garfield, J.B. Reasoning about sampling distribitions. In *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*; Ben-Zvi, D., Garfield, J.B., Eds.; Springer: Dordrecht, The Netherlands, 2004; pp. 295–323.
80. Lipson, K. The role of computer based technology in developing understanding of the concept of sampling distribution. In Proceedings of the 6th International Conference on Teaching Statistics, Cape Town, South Africa, 7–12 July 2002.
81. Batanero, C.; Tauber, L.M.; Sánchez, V. Students' reasoning about the normal distribution. In *The Challenge of Developing Statistical Literacy, Reasoning and Thinking*; Ben-Zvi, D., Garfield, J.B., Eds.; Springer: Dordrecht, The Netherlands, 2004; pp. 257–276. [CrossRef]
82. Bower, K.M. *Some Misconceptions about the Normal Distribution*; Six Sigma Forum; American Society for Quality: Milwaukee, WI, USA, 2003.
83. Batanero, C.; Godino, J.D.; Vallecillos, A.; Green, D.e.; Holmes, P. Errors and difficulties in understanding elementary statistical concepts. *Int. J. Math. Educ. Sci. Technol.* **1994**, *25*, 527–547. [CrossRef]
84. Vallecillos, A. Understanding of the logic of hypothesis testing amongst university students. *J.-Math.-Didakt.* **2000**, *21*, 101–123. [CrossRef]
85. Falk, R.; Greenbaum, C.W. Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory Psychol.* **1995**, *5*, 75–98. [CrossRef]
86. Perezgonzalez, J.D. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Front. Psychol.* **2015**, *6*, 223. [CrossRef]
87. Haller, H.; Krauss, S. Misinterpretations of significance: A problem students share with their teachers. *Methods Psychol. Res.* **2002**, *7*, 1–20.
88. Falk, R. Misconceptions of statistical significance. *J. Struct. Learn.* **1986**, *9*, 83–96.
89. Vallecillos, A.; Batanero, C. Conceptos activados en el contraste de hipótesis estadísticas y su comprensión por estudiantes universitarios [Activated concepts in statistical hypothesis testing and their understanding by university students]. *Rech. Didact. Math.* **1997**, *17*, 29–48.

90. Williams, A.M. Students' understanding of the significance level concept. In Proceedings of the 5th International Conference on Teaching Statistics, Singapore, 21–26 June 1998; pp. 743–749.
91. Mittag, K.C.; Thompson, B. Research news and Comment: A National Survey of AERA Members' Perceptions of Statistical Significance Tests and Other Statistical Issues. *Educ. Res.* **2000**, *29*, 14–20. [CrossRef]
92. Gliner, J.A.; Leech, N.L.; Morgan, G.A. Problems with null hypothesis significance testing (NHST): What do the textbooks say? *J. Exp. Educ.* **2002**, *71*, 83–92. [CrossRef]
93. Gagnier, J.J.; Morgenstern, H. Misconceptions, misuses, and misinterpretations of *p* values and significance testing. *J. Bone Jt. Surg.* **2017**, *99*, 1598–1603. [CrossRef]
94. Cumming, G.; Williams, J.; Fidler, F. Replication and researchers' understanding of confidence intervals and standard error bars. *Underst. Stat.* **2004**, *3*, 299–311. [CrossRef]
95. Fidler, F. Should psychology abandon p-values and teach CIs instead? Evidence-based reforms in statistics education. In Proceedings of the 7th International Conference on Teaching Statistics, Salvador, Brazil, 2–7 July 2006.
96. Hoekstra, R.; Morey, R.D.; Rouder, J.N.; Wagenmakers, E.J. Robust misinterpretation of confidence intervals. *Psychon. Bull. Rev.* **2014**, *21*, 1157–1164. [CrossRef]
97. Kalinowski, P. Identifying misconceptions about confidence intervals. In Proceedings of the 8th International Conference on Teaching Statistics, Ljubljana, Slovenia, 11–16 July 2010.
98. Belia, S.; Fidler, F.; Williams, J.; Cumming, G. Researchers misunderstand confidence intervals and standard error bars. *Psychol. Methods* **2005**, *10*, 389. [CrossRef]