ASSESSING LEARNING OUTCOMES: AN ANALYSIS OF THE GOALS-2 INSTRUMENT

ANELISE GUIMARAES SABBAG
University of Minnesota
sabb0013@umn.edu

ANDREW ZIEFFLER
University of Minnesota
zief0002@umn.edu

ABSTRACT

The test instrument GOALS-2 was designed primarily to evaluate the effectiveness of the CATALST curriculum. The purpose of this study was to perform a psychometric analysis of this instrument. Undergraduate students from six universities in the United States (n=289) were administered the instrument. Three measurement models were fit and compared: the two-parameter logistic model, the mixed model (comprised of both the two-parameter logistic and the graded-response model), and the bi-factor model. The mixed model was found to most appropriately model students' responses. The results suggested the revision of some items and the addition of more discriminating items to improve total test information.

Keywords: *Statistics education research; Assessment; Item response theory; Simulation/randomization approaches; Statistical inference*

1. INTRODUCTION

Statistics education has experienced many changes in the last decade, one of which is a shift in the focus of learning outcomes in introductory statistics courses. Currently, learning outcomes in many introductory courses are focused more on statistical literacy, thinking, and reasoning than on calculation and procedures (Garfield & Ben-Zvi, 2008). These changes in the field of statistics education are in part due to an increased awareness of curriculum recommendations, such as the *Guidelines for Assessment and Instruction in Statistics Education* (GAISE; ASA, 2005), and Cobb's (2007) suggestions for implementing randomization methods.

One example of a curriculum that was created based on these recommendations is the *Change Agents for Teaching and Learning Statistics* (CATALST) materials. The CATALST materials use a modeling and simulation/randomization approach to teaching statistical inference. Initial evidence regarding students' learning for these types of curricula has been positive (Garfield, delMas, & Zieffler, 2012; Tintle, VanderStoep, Holmes, Quisenberry, & Swanson, 2011; Tintle, Topliff, Vanderstoep, Holmes, & Swanson, 2012). However, there is still more evidence needed to suggest that courses based on curricular recommendations are effective in helping students learn and reason about statistics—especially when it comes to statistical inference.

The Goals and Outcomes Associated with Learning Statistics (GOALS-2) instrument was initially designed to evaluate the effectiveness of the CATALST curriculum in

developing students' conceptual ideas and statistical reasoning. The content of this instrument includes—but it is not limited to—how the use of simulations and randomization tests are contributing to students' understanding of statistical inference.

This study assesses psychometric properties of the investigated instrument. Within the context of this study, several measurement models will be examined to determine which one appropriately models the data from the GOALS-2 instrument. The adopted model will be used to obtain item-, test-, and person-level characteristics as part of the process to evaluate the current instrument and inform decisions to improve it. First, however, a brief review of the literature related to role of assessment in statistics education is presented.

2. THE ROLE OF ASSESSMENT IN STATISTICS EDUCATION

The role of assessment in the field of statistics education has also shifted as a function of the focus on statistical literacy, reasoning, and thinking as new learning outcomes (Garfield & Ben-Zvi, 2008). In their paper about assessing these new outcomes for students, Garfield, delMas, and Zieffler (2010) suggested methods of designing and developing assessments to use in introductory statistics courses. The authors addressed the role of assessment in designing and evaluating curriculum and stressed the importance of the alignment between the assessment and the learning outcomes of the course. The authors also suggested using a blueprint to ensure that this alignment was adequate. Several instruments have been developed to assess the learning outcomes of statistical literacy, reasoning, and thinking. Examples of these instruments include the *Comprehensive Assessment of Outcomes in a First Statistics Course* (CAOS; delMas Garfield, Ooms & Chance, 2007), and the *Assessment of Inferential Reasoning in Statistics* (AIRS; Park, 2012).

Research on how to assess the new learning outcomes and the development of new instruments were not the only changes in the role of assessment in the field of statistics education. Garfield and Ben-Zvi (2008) encouraged instructors to use assessments as part of the learning process rather than only as a method of evaluating students' performance. Garfield and Franklin (2011) also advised educators to broaden their use of assessment. They point out that although traditionally, assessment was classified as formative or summative, current research (Earl & Katz, 2006) categorized assessment as (1) assessment *for* learning, (2) assessment *as* learning, and (3) assessment *of* learning. According to Earl and Katz, assessment *for* learning seeks to determine what students know and what they can do; assessment *as* learning is related to how students think and reflect on what they are learning; and assessment *of* learning is used to know how well students are performing and if they met the desired goals of their programs.

The role of assessment in the field of statistics education is intrinsic not only in how students learn statistics but also in the teaching of statistics. Assessments can provide very important information related to students' learning but it is important to use quality instruments to capture this information.

Assessments are used in research for many different purposes, such as to facilitate student learning, to provide feedback for students, to inform instructors regarding students' achievement, and to evaluate courses. National organizations such as the American Statistical Association (ASA, 2007), American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education (AERA, APA, & NCME, 1999) have outlined several suggestions for developing and improving instruments. In addition, the literature also includes arguments against the use of final exams scores or course grades as indicators of

statistical reasoning (e.g., Chance & Garfield, 2002; Konold, 1995). Despite all of this literature, Zieffler et al. (2008) point out that there are still many studies using these inappropriate measures. In addition, other measurement issues have been reported such as the "lack of enough diversity in available validated tests to allow good alignment between existing tests and intended outcomes in a particular research study" (Pearl et al., 2012). This suggests that there is a need in the field of statistics education for psychometrically sound assessments that measure different learning outcomes for introductory statistics courses.

2.1. EXISTING ASSESSMENTS FOR MEASURING COGNITIVE OUTCOMES

Prior to the development of GOALS-2, there were six published assessments available to measure cognitive outcomes of student learning in introductory statistics courses. These include the *Statistical Reasoning Assessment* (SRA), the *Quantitative Reasoning Quotient* (QRQ), the *Comprehensive Assessment of Outcomes in a First Statistics Course* (CAOS), the *Reasoning about P-values and Statistical Significance* (RPASS), the *Assessment of Inferential Reasoning in Statistics* (AIRS), and the *Statistics Concept Inventory* (SCI). These instruments are summarized in Table 1.

For each of these six instruments, Sabbag (2013) summarized the published validity evidence. In general, content validity evidence and evidence regarding score precision was almost always collected and reported. Other forms of validity evidence, such as evidence of internal structure, the assessment scores relationship with other variables, and consequences of testing, were typically not reported for these assessments. Additionally, the development and validation of these assessments were primarily carried out without the use of appropriate psychometric models, such as item response theory (Park, 2012).

2.2. DEVELOPMENT OF THE GOALS INSTRUMENT

The GOALS instrument was developed from the CAOS assessment (delMas et al., 2007). Similar to CAOS, the new instrument assesses student statistical reasoning in a first course of statistics. One of the primary reasons the instrument was produced was to evaluate the effectiveness of the CATALST curriculum in developing students' conceptual ideas and statistical reasoning. In its first iteration, the authors of GOALS analyzed students' responses to the former instrument using both distractor analysis and item response theory. The results of these analyses were used to identify items that were not performing well and to delete them. Additionally, other items were modified based on research done by Ziegler (2012), who explored how the stem length of forced-choice items affected students' responses.

The new set of items was then examined for alignment with current learning goals for introductory statistics courses. The results of the examination suggested a need to include items that address topics taught in courses using a simulation and randomization approach to inference. It also suggested the need for additional items addressing topics such as the purpose of random assignment, and the interpretation of statistically significant results and *p*-values. As a result, 14 items were added to the test instrument. Of these 14 items, seven were adapted from the NSF-funded *Concepts of Statistical Inference* project, three were adapted from the NSF-funded *Creating a Teaching and Learning Infrastructure for Introductory Statistics Re-design* project, and four items were written by University of Minnesota statistics education faculty members. This resulted in an instrument named GOALS-1, comprising 28 items—three constructed-response items and 25 forced-choice items.

Table 1. Reliability (internal consistency unless noted otherwise), number of items, content, and topics addressed by existing instruments in the field of statistics education

Instrument	Reliability	No. of items	Content assessed	Topics
AIRS Park (2012)	0.81	34	Inferential Reasoning	Inferential reasoning, informal inference, sampling distribution, design of study, statistical testing, confidence intervals, evaluation of study.
CAOS delMas et al. (2007)	0.82	40	Statistical Reasoning	Data collection and design, descriptive statistics, graphical representations, boxplots, normal distribution, bivariate data, probability, sampling variability, confidence intervals, tests of significance
QRQ Sundre (2003)	0.55	40	Quantitative Reasoning	Correct and incorrect quantitative reasoning involving concepts such as probability, central tendency, sampling variability, design, law of large numbers, equiprobability bias.
RPASS Lane-Getaz (2007ab, 2013)	0.76	34	Inferential Reasoning	Basic terminology and concepts, relationships between inferential concepts, logic of statistical inference and hypotheses, *p*-values, decisions and error
SCI Allen (2006), Stone (2006)	0.67 – 0.77	28	Statistical conceptual knowledge	Descriptive statistics, inferential statistics, probability, graphical methods
SRA Garfield (1998, 2003)	0.70† and 0.75††	20	Statistical Reasoning	Correct and incorrect types of statistical reasoning related to topics such as data, representations of data, uncertainty, samples, association, law of small numbers, outcome orientation.

† Test-retest reliability for correct reasoning †† Test-retest reliability for incorrect reasoning

Content validity evidence of GOALS In order to provide validity evidence about the content assessed by the previous instrument called GOALS-1, statistics instructors doing research in the field of statistics education were identified and invited to provide feedback about the items. With those seven persons who consented to review the items and participate in a follow-up interview, a time during the 2011 United States Conference on Teaching Statistics was arranged for an interview. Prior to the conference, each reviewer was sent an evaluation sheet and a copy of the instrument.

Reviewers were requested to rate the extent to which they agreed that each item measured an important learning outcome for any student who had completed a college-level, non-calculus based, introductory statistics course. Reviewers were also asked to list any learning outcome that was not assessed by the analyzed instrument but at the same time was an important learning outcome for them. Finally, the reviewers were asked to identify and offer modifications for any items that they felt needed improvement. Each set of interviewers took part in a two-hour training session prior to the conference, which consisted of an explanation of the interview protocol and practice interviews. In the interviews, the reviewers were asked about their overall impression of the items and their opinion about the alignment between the items and the learning goals for an introductory statistics course. Reviewers' evaluation feedback sheets, copies of the instrument, and any other material related to the instrument were collected at this time.

Based on the feedback received from the reviewers and discussion with another expert in the field of statistics education, the instrument was modified. Some of the modifications were (1) changing from a constructed-response to a forced-choice format, (2) simplifying the stem of the items by re-wording and/or removing extraneous information, (3) adding phrases to the stem of the item to clarify the meaning of the questions, and (4) re-wording the response options to be more explicit about what they stated. In addition, one item about polling and representative samples was added and four items were deleted. Three items did not change. This led to the second version of the instrument, called GOALS-2. See Sabbag (2013) for more information about the modifications made to the items and about the interviews with statistics instructors.

After reviewing the changes to the new instrument, it was decided that two forms of the instrument were needed. This decision, motivated by feedback from content experts, was made because the assessment of students' reasoning about inference seemed tied to whether students learned inference using a classical or a modeling and simulation/ randomization-based approach. Subsequently, one form was developed for students enrolled in statistics courses with classical content. Another form was developed for students enrolled in simulation/randomization-based courses (e.g., CATALST).

Each of the two forms of the new instrument is comprised of 27 forced-choice items. Apart from the four items with content related to the use of simulation to carry out statistical inference, the remaining 23 items on the two forms are identical. The items on the instrument in the present form address the topics of study design, reasoning about variability, sampling and sampling variability, interpreting confidence intervals and *p*-values, statistical inference, and modeling and simulation. For more information about the measured learning goal for each item, see Sabbag (2013).

3. METHODS

The GOALS-2 instrument was developed to assess how the use of simulations and randomization approaches are contributing to students' reasoning about statistics. This study investigates psychometric properties of this instrument. Specifically, (1) which measurement model is appropriate to model the data from the instrument? And (2) based on the measurement model, how can the instrument be improved further? The psychometric properties of the instrument found in this study can be used as evidence of the validity of the internal structure of the test to support the intended inferences and uses of test scores. To address the research questions posed, the responses of 289 students to the items were analyzed using both a classical test theory (CTT) and an item-response theory (IRT) framework. In this paper, only the results from the IRT analyses are reported. For information related to the CTT analyses, see Sabbag (2013).

3.1. DATA COLLECTION AND STATISTICAL ANALYSIS

The simulation/randomization form of the GOALS-2 instrument was administered to 289 undergraduate students from six universities in the United States. These students were enrolled in a statistics course that was using the CATALST curriculum, a curriculum designed as part of an NSF-funded project, which developed materials, lesson plans, and assessments based on a modeling and simulation/randomization-based approach to statistical inference. For a description of the research foundations of the course and the curriculum, see Garfield, delMas, and Zieffler (2012).

In order to address the research questions, three IRT models were fitted to the data and compared: (1) the two-parameter logistic model, (2) the mixed model composed of

the two-parameter logistic model and the graded-response model, and (3) the bi-factor model. The first model assumes independence between item responses, while the other two allow for modeling item response inter-dependence. Note that because of the small sample size (289 students), the three-parameter logistic model—a model initially considered for the analysis—was not fitted to the data.

Classical test theory (CTT) and item response theory (IRT) are two measurement frameworks that have been widely used in test-development. CTT is a test-based framework that is focused on the sum score of all items, while IRT is an item-based framework that is focused on students' responses to individual items. More specifically, IRT models describe the probability that a student responds correctly to an assessment item given her/his ability level on the latent trait being measured by the assessment. (These models are analogous to the logistic regression model.) The objective of an IRT analysis is to model each item by estimating the properties describing its performance, namely the difficulty and discrimination.

Using an IRT rather than a CTT framework for analysis has many benefits (see Hambleton & Jones, 1993). These authors report that some of the advantages of using an IRT framework are (1) students' ability scores are not dependent on test difficulty, (2) item's characteristics are sample independent, (3) item locations are reported in the same scale as examinee's abilities, and (4) strict parallel tests are not required for reliability estimation. Because of the advantages of using IRT over CTT, in this study, IRT will be used to estimate the students' statistical reasoning ability. Under this framework, a construct of interest can be estimated by using examinee's responses to each item in a test. Therefore, IRT relates examinee's ability to examinee's performance on the test as a whole and on individual items. For each level of the construct of interest – in this case, students' statistical reasoning ability – the probability of correctly answering an item can be modeled by an item characteristic curve (ICC) shaped like the letter "S". ICC are determined based on item characteristics such as item difficulty and item discrimination. For more information about IRT, see de Ayala (2009).

As is typical for IRT models, the origin (mean of ability values) was fixed to zero and the unit (variance of ability values) was fixed to one. The item difficulty and item discrimination were then estimated for each model and evaluated using guidelines suggested by de Ayala (2009).

Evaluating model fit and model comparison The fit of the three IRT models to the GOALS-2 data was evaluated at both the item and the model level. At an item level, the S-X^2 item-fit statistic (Orlando & Thissen, 2000, 2003) was used to assess if each item fits the model. At a model level, likelihood-ratio tests (de Ayala, 2009) were used to compare nested models and fit statistics such as the M_2 statistic, its associated p-value and the root mean square error of approximation (RMSEA; Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Maydeu-Olivares & Joe, 2005, 2006). The Akaike Information Criterion (AIC; Akaike, 1974), and Bayesian Information Criterion (BIC; Schwarz, 1978) were also used for comparison. In addition, the correlation between the person-location parameters and the standard errors of the person-location parameters was examined for each of the three models. Based on these criteria, an optimal model was chosen and the item-response function, item-information function, test-information function, and standard error of measurement (SEM) were computed for the adopted model.

4. RESULTS

Three IRT models were fitted and the assumption of item independence and its plausibility for the GOALS-2 items is discussed below, after which, the results from each IRT model is more fully described. Item-parameter estimates, item-level diagnostics statistics, and fit statistics are reported for each of the three IRT models. The IRT models were estimated using the Bock-Atkin estimation method and the EAP scoring method as carried out in the software IRTPRO 2.1 (Cai, Thissen, & du Toit, 2011).

4.1. ITEM RESPONSE THEORY (IRT) MODELS

Many of the items in the new GOALS-2 instrument share a common stimulus (e.g., Item 13 and Item 14 share a small reading passage). Because of this, students' responses to items sharing a stimulus may be inter-dependent. Other items on the instrument that share a common stimulus are reported in Table 2.

Table 2. Testlets and items correspondence

Testlet	1	2	3	4	5	6	7
Items	5-6	10-11	13-14	15-18	19-22	23-25	26-27

According to Sireci, Thissen, and Wainer (1991), context dependency among items might lead to the violation of the assumption of local independence—one of the assumptions of an item response theory model. Disregarding the assumption of local independence can lead to overestimation of the reliability coefficient, inappropriate standard-error (SE) estimates of the general factor of the test, and overestimation of test information functions (Sireci et al., 1991), errors in equating and scaling and improper estimation of discrimination parameters and item misfit indices (see DeMars, 2012). To accommodate potential violations of this assumption, items that share a stimulus can be grouped into "testlets". Responses to each testlet, which are assumed to relate to a common trait, are modeled as nuisance factors since these factors are usually of no interest to the authors of the test (DeMars, 2006).

Two-parameter logistic model This model – henceforth abbreviated as 2PL model – specifies the probability of a correct response to an item as a logistic distribution in which items are allowed to vary in terms of their difficulty and discrimination. In the 2PL model, the probability of a correct response is given by

$$p\left(x_j = 1 \middle| \theta, \alpha_j, \delta_j\right) = \frac{\exp\left(\alpha_j\left(\theta - \delta_j\right)\right)}{1 + \exp\left(\alpha_j\left(\theta - \delta_j\right)\right)} = \frac{\exp\left(\alpha_j \theta + \gamma_j\right)}{1 + \exp\left(\alpha_j \theta + \gamma_j\right)},$$

where θ is the latent trait (or person location parameter), α_j is the discrimination parameter for item *j*, and δ_j is the difficulty for item *j*. The first version of the formula has parameters easier to interpret, and the second version with the intercept γ_j will be useful for the generalization to the bi-factor model.

For the 2PL model, the test-information function (also known as total test information) indicates how much information an instrument provides for calculating person-level parameter estimates. Each item in an instrument contributes independently to decrease uncertainty in the estimation of the person-level location parameter. de Ayala (2009) provides detail regarding calculation of the item information function for the 2PL model.

Table 3 provides the item parameters estimates (intercept, item discrimination, and item difficulty) and standard error of the estimates obtained from fitting the 2PL model to the data. The estimated item discrimination parameters ranged from –0.45 to 2.51. Item difficulty values ranged from –2.36 to 2.11, except for eight items which had unusually low or high difficulty values (these items are flagged in Table 3).

Table 3. Item parameters for the 2PL model

Item	Intercept		Discrimination		Difficulty	
	Estimate	SE	Estimate	SE	Estimate	SE
1	0.84	0.15	1.01	0.20	–0.84	0.18
2	1.87	0.23	1.32	0.27	–1.42	0.22
3	2.08	0.26	1.38	0.29	–1.50	0.22
4	–0.08	0.13	0.90	0.18	0.09	0.15
5	0.34	0.12	–0.45[a]	0.15	0.77	0.37
6	0.43	0.12	–0.08[a]	0.14	5.16[b]	8.87
7	5.94	1.87	2.51	1.22	–2.36	0.47
8	–1.29	0.15	0.61[a]	0.18	2.11	0.59
9	–0.26	0.12	0.31[a]	0.14	0.84	0.55
10	1.43	0.23	1.49	0.33	–0.96	0.15
11	1.24	0.21	1.43	0.31	–0.87	0.15
12	2.52	0.34	1.82	0.37	–1.38	0.17
13	–1.56	0.16	0.00[a]	0.19	661.18[b]	52078.48
14	–1.50	0.15	–0.09[a]	0.19	–17.38[b]	37.55
15	–0.98	0.13	0.29[a]	0.16	3.39[b]	1.86
16	1.24	0.14	0.13[a]	0.17	–9.65[b]	12.38
17	0.14	0.12	0.39[a]	0.15	–0.36	0.33
18	2.20	0.20	0.15[a]	0.23	–14.63[b]	21.97
19	1.15	0.15	0.68[a]	0.18	–1.71	0.43
20	0.70	0.14	0.88	0.19	–0.80	0.20
21	0.35	0.14	0.99	0.19	–0.36	0.14
22	1.85	0.30	2.03	0.38	–0.91	0.11
23	1.39	0.15	0.19[a]	0.17	–7.33[b]	6.68
24	0.37	0.15	1.30	0.23	–0.28	0.11
25	1.04	0.16	1.04	0.21	–1.00	0.19
26	–0.22	0.12	0.02[a]	0.14	13.36[b]	110.72
27	0.17	0.12	0.55[a]	0.16	–0.31	0.23

[a] indicates item with discrimination value lower than 0.8;
[b] indicates item with unusual item parameter value (> 3 or < –3).

Mixed model A common approach in modeling responses to multiple items that make up a testlet is to consider the testlet as a single polytomous item, which is scored by summing the number of correct items in the testlet that the student responded to. The second model used to model the data was the unidimensional model considering testlets as a polytomous item (Sireci et al., 1991; Wainer, 1995; Wainer & Wang, 2000). The mixed model is a mixture of two models. The unidimensional 2PL model was used to model the dichotomous items not included in a testlet and the graded response model (henceforth GR model; Samejima, 1969) was used to model the items that make up testlets. Items that had context dependency were grouped as a testlet, thus the decision regarding which items would constitute a testlet was done prior to data analysis. For the

mixed model, we used eight dichotomous items (items 1, 2, 3, 4, 7, 8, 9, and 12) and seven testlets. Table 2 shows the attribution of items to testlets.

Responses from polytomous items are not scored as 0 (incorrect) or 1 (correct), but are instead divided into category scores with higher-category scores indicating a higher level of overall performance. If m_j represents the number of steps needed to correctly respond to item j, then the response to item j can be scored as x_j with values $k = \{0, 1, 2, \ldots, m_j\}$. The GR model considers the probability of a person responding in category k or higher versus responding in categories lower than k. According to the GR model, the probability of obtaining k or higher is,

$$p\left(x_j \geq k \,\middle|\, \theta, \alpha_j, \delta_{jk}\right) = \frac{\exp\left(\alpha_j\left(\theta - \delta_{jk}\right)\right)}{1 + \exp\left(\alpha_j\left(\theta - \delta_{jk}\right)\right)},$$

where θ is the latent trait (or person location parameter), α_j is the discrimination parameter for item j, and δ_{jk} is the category boundary location for category k on item j, which indicates the difficulty (or level of ability) at which a person is more likely to respond with k (rather than $k - 1$).

To estimate the item information function for polytomous models, the information of each response category must be considered. De Ayala (2009) provides detail regarding calculation of the item-information function for the GR model. One of the limitations of considering a testlet as a single polytomous item is the loss of information since each testlet is scored as the number correct of the items contained in the testlet. According to Yen (1993), one way to overcome this loss of information is to use few items inside each testlet so that more information can be retained (as cited in Park, 2012).

Table 4 provides item parameters estimates (intercept, item discrimination, and item difficulty) and standard errors of the estimates for the dichotomous items that were estimated under the 2PL model. Item discrimination values ranged from 0.28 to 2.81. Item difficulty values ranged from −2.29 to 2.25.

Table 4. Item parameters for dichotomous items for the mixed model

Item	Intercept		Discrimination		Item difficulty	
	Estimate	SE	Estimate	SE	Estimate	SE
1	0.89	0.16	1.19	0.23	−0.75	0.15
2	1.93	0.25	1.42	0.30	−1.36	0.21
3	2.19	0.29	1.55	0.35	−1.41	0.21
4	−0.09	0.13	0.92	0.19	0.09	0.14
7	6.43	1.50	2.81	1.01	−2.29	0.35
8	−1.28	0.15	0.57[a]	0.18	2.25	0.68
9	−0.26	0.12	0.28[a]	0.14	0.92	0.64
12	2.31	0.30	1.54	0.34	−1.50	0.22

Note. [a] indicates items with discrimination values lower than 0.8.

The intercept and item discrimination values, along with their standard errors, were also estimated for each of the seven testlets using the GR model. These estimates are shown in Table 5. Item discrimination parameters ranged from −0.23 to 1.61.

Estimates for the category boundary locations, item difficulties, and their standard errors for the seven testlets are presented in Table 6. Item difficulty values for the GR model are computed by averaging across the category boundary locations for each testlet. The difficulty estimates ranged from −2.31 to 1.84. Testlet 3 (items 13 and 14), presented

unusual values for the category boundaries and item difficulties. These values are probably related to the very low discrimination value for this testlet.

Table 5. Item discrimination and intercept for the testlets for the mixed model

Testlet		Discrimination		Intercept 1		Intercept 2		Intercept 3		Intercept 4	
No	Items	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
1	5-6	–0.23[a]	0.13	1.11	0.14	–0.26	0.12				
2	10-11	0.81	0.19	1.17	0.16	1.00	0.15				
3	13-14	–0.06[a]	0.18	–1.45	0.15	–1.62	0.15				
4	15-18	0.53	0.14	5.10	0.71	2.11	0.19	0.03	0.12	–2.31	0.20
5	19-22	1.61	0.25	3.88	0.37	2.27	0.24	0.69	0.16	–1.32	0.19
6	23-25	1.29	0.20	4.13	0.39	1.11	0.16	–0.58	0.15		
7	26-27	0.71[a]	0.17	1.90	0.18	–2.07	0.19				

Note. [a] indicates items with discrimination values lower than 0.8.

Table 6. Category boundaries and item difficulty for the testlets for the mixed model

Testlet		Category boundary 1		Category boundary 2		Category boundary 3		Category boundary 4		Item difficulty
No	Items	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate
1	5-6	4.81	2.74	–1.14	0.80					1.84
2	10-11	–1.45	0.32	–1.24	0.28					–1.35
3	13-14 *	–25.08	76.58	–27.95	85.33					–26.52
4	15-18	–9.55	2.73	–3.96	1.01	–0.06	0.23	4.32	1.11	–2.31
5	19-22	–2.41	0.29	–1.41	0.17	–0.43	0.10	0.82	0.13	–0.86
6	23-25	–3.19	0.45	–0.85	0.14	0.45	0.13			–1.20
7	26-27	–2.68	0.59	2.91	0.65					0.12

Note. * indicates items with unusual item parameter values.

Bi-factor model The third model fitted to the data was the multidimensional bi-factor model (Gibbons & Hedeker, 1992) with testlets. In this model, there is a general dimension θ_g that is measured by all items, and F other dimensions that are related to the F testlets. The bi-factor model is a multidimensional extension of the 2PL model, where the probability of a correct response on item j by person i is given by

$$p\left(x_{ij}=1\middle|\underline{\theta}_i,\underline{\alpha}_j,\gamma_j\right)=\frac{\exp(\underline{\alpha}_j'\,\underline{\theta}_i+\gamma_j)}{1+\exp(\underline{\alpha}_j'\,\underline{\theta}_i+\gamma_j)}.$$

The intercept parameter γ_j is not directly interpretable as item difficulty. The vector $\underline{\theta}_i$ corresponds to the location parameters of person i on each of the F dimensions. The vector $\underline{\alpha}_j = (\alpha_{j1}, \alpha_{j2}, \ldots, \alpha_{jF})$ contains the discrimination parameters of item j on each of the F dimensions. For this study, eight factors were considered: a general factor, which all items were allowed to load on, and seven nuisance factors (related to the seven testlets). De Ayala (2009) provides detail regarding calculation of the item information function for the bi-factor model.

Table 7 provides the item parameters estimates (intercepts and item discriminations, as well as the standard errors for these estimates) after fitting the bi-factor model to the data. In this model, there was a single general dimension, which all items were loaded on to, and seven specific dimensions related to the seven testlets. Item discrimination values

for the general dimension ranged from –0.74 to 2.55. Items 10, 11, and 24 had unusually high values for both the discrimination and intercept.

Table 7. Item parameters for the general dimension for the bi-factor model

Item	Intercept		Discrimination		Item	Intercept		Discrimination	
	Estimate	SE	Estimate	SE		Estimate	SE	Estimate	SE
1	0.89	0.15	1.15	0.17	15	–1.70	0.22	0.63	0.18
2	1.96	0.22	1.46	0.23	16	1.37	0.16	0.13	0.15
3	2.18	0.24	1.52	0.24	17	0.15	0.12	0.44	0.13
4	–0.08	0.13	0.93	0.15	18	2.22	0.20	0.23	0.20
5	0.53	0.15	–0.74	0.16	19	1.84	0.23	1.11	0.21
6	0.67	0.15	–0.09	0.15	20	0.72	0.14	0.89	0.16
7	6.00	1.12	2.55	0.71	21	0.37	0.13	1.07	0.16
8	–1.29	0.15	0.60	0.15	22	2.04	0.26	2.32	0.30
9	–0.26	0.12	0.29	0.12	23	1.66	0.18	0.39	0.16
10*	16.67	0.57	11.38	0.58	24*	41.66	2.51	134.51	4.23
11*	14.66	0.54	11.19	0.58	25	1.18	0.17	1.14	0.18
12	2.48	0.28	1.76	0.27	26	–0.22	0.12	0.01	0.12
13	–23.81	0.61	1.52	0.58	27	0.17	0.12	0.53	0.13
14	–23.36	0.68	0.57	0.62					

Note. * indicates items with unusual item parameter values.

4.2. EVALUATING MODEL FIT AND MODEL COMPARISON

Based on the examination of the results (see Table 8) the 2PL model's $S\text{-}X^2$ statistic indicates misfit for two items (items 13 and 14). The $S\text{-}X^2$ statistic from the mixed model suggests item misfit for Testlet 7, which was comprised of items 26 and 27. It also indicates borderline fit for Testlet 3 (items 13 and 14). The $S\text{-}X^2$ statistic from the bi-factor model, on the other hand, does not present with any item misfit.

The fit of the 2PL model and the bi-factor model was compared using a likelihood ratio test. (Note: Since only nested models can be used in this part of the analysis, the mixed model was not included here.) The results of this test, $\Delta G^2 = 529.80$ ($p < 0.001$; $df = 15$) indicates that the bi-factor model likely fits the data better than the 2PL model. The results for the M_2 fit statistic are significant for all three models. Based on these results (Table 9), all models have significant misfit to the data. However, as mentioned before, the M_2 fit statistic is overly sensitive to small deviations in the model–data fit. Therefore, the RMSEA is also reported and compared for the three models. Based on guidelines suggested by Browne and Cudeck (1993), the RMSEA values for both the mixed model and the bi-factor model indicate close fit to the data, while the RMSEA value for the 2PL model indicates only mediocre fit to the data.

Rank ordering of the AIC and BIC measures (Table 9) indicates that after accounting for model complexity and sample size, the bi-factor model fits the data better than the 2PL model. Note that the AIC and BIC statistics for the mixed model should not be compared with the other two models since the data set used for the mixed model is not the same as the dataset used for 2PL model and bi-factor model.

Lastly, the correlation matrix based on the three models' person-location parameters was also examined. The correlations were all very high (0.96–0.98), indicating that each of the models leads to very similar trends in the estimates of students' abilities. The correlation matrix based on the models' standard errors of the person location parameters

also suggests that the three models produce similar trends in the size of the standard errors for the ability estimates as well. Only the correlation between the 2PL model and the mixed model (0.88) was below 0.90.

Table 8. Item-level diagnostics for the investigated models

Item	2PL model			Mixed model			Bi-factor model		
	$S\text{-}X^2$	*df*	*p*-value	$S\text{-}X^2$	*df*	*p*-value	$S\text{-}X^2$	*df*	*p*-value
1	12.56	13	0.484	9.99	11	0.532	11.45	12	0.492
2	8.74	11	0.646	9.07	11	0.617	9.29	11	0.597
3	3.71	11	0.978	3.73	11	0.977	3.79	11	0.976
4	16.56	13	0.219	16.02	13	0.248	16.24	13	0.236
5	14.32	14	0.428	17.29	25	0.872	11.71	13	0.552
6	12.98	13	0.451				9.52	13	0.733
7	2.92	3	0.404	3.09	3	0.379	2.86	3	0.415
8	16.13	12	0.184	19.13	12	0.085	16.08	12	0.187
9	14.42	14	0.420	13.10	15	0.596	14.10	15	0.519
10	14.86	11	0.188	7.14	18	0.989	14.03	11	0.231
11	13.13	11	0.283				12.31	10	0.264
12	6.54	10	0.769	7.18	10	0.709	6.60	10	0.763
13	25.76	13	0.018*	27.23	17	0.055	11.48	12	0.490
14	27.25	13	0.011*				13.53	13	0.410
15	15.43	13	0.280	30.93	35	0.666	14.95	12	0.244
16	21.52	14	0.088				20.35	13	0.087
17	10.04	14	0.759				9.10	13	0.766
18	10.25	11	0.509				10.12	10	0.432
19	8.57	12	0.740	46.37	39	0.194	8.68	11	0.652
20	8.44	13	0.814				8.39	12	0.754
21	15.06	13	0.303				14.57	12	0.265
22	15.14	10	0.126				15.39	9	0.081
23	9.04	13	0.770	23.04	28	0.732	9.22	12	0.685
24	10.80	12	0.547				9.48	11	0.579
25	12.02	13	0.527				11.42	12	0.495
26	14.92	13	0.314	38.18	23	0.024*	14.92	12	0.245
27	16.90	14	0.260				16.10	13	0.243

*p-values < 0.05.

Table 9. Model-based measures of fit for the investigated models

Fit measures	Model		
	2PL	Mixed	Bi-factor
Akaike Information Criterion (AIC)	8495.02	6182.77	7995.22
Bayesian Information Criterion (BIC)	8693.01	6336.76	8248.20
RMSEA	0.09	0.02	0.03
p-value for M_2 statistic	0.0001	0.0210	0.0008

4.3. ADDITIONAL INFORMATION FOR THE MIXED MODEL

The SEM and total test information function are shown in Figure 1. The SEM, represented by the dotted line, is inversely related to the information function. The SEM

was larger (less precision) for ability values outside the range of –3 and –0.5. This suggests that GOALS-2 provided the most information for students with ability values (theta) between –3 and –0.5. The item-response and item-information functions for the dichotomous items and testlets (based on the mixed model) are displayed in Appendix A.

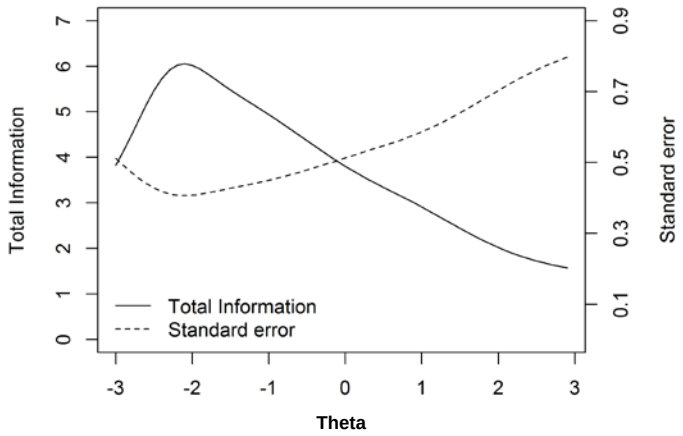


Figure 1. Test information and standard error of measurement (SEM).

Dimensionality Understanding the internal structure of a test is one piece of validity evidence that can be used to support the intended inferences and uses of test scores. In addition, one of the assumptions for many IRT models states that responses to items in a test are exclusively a function of a single continuous latent variable (de Ayala, 2009). According to Reckase (1985) and Ansley and Forsyth (1985), violation of this unidimensionality assumption can lead to biased estimates of both item and person parameters (as cited by Finch and Monahan, 2008). A confirmatory factor analysis (CFA) using mean-adjusted weighted least-squares was fitted to the students' responses and standardized factor loadings and fit indices such as RMSEA, Bentler's Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) were examined to determine the plausibility of the unidimensionality assumption. The CFA specifying a single factor was fitted to the responses using MPLUS (Muthén & Muthén, 2010). Standardized factor loadings are presented in Table 10. Fit indices were also calculated: TLI was 0.875, CFI was 0.892, and RMSEA was 0.058 with a 90% confidence interval of (0.045, 0.070). Therefore, the parameter estimates and fit indices indicate good fit for the unidimensional model.

Table 10. Factor loadings and fit indices for CFA

Item	Testlet	Factor Loadings	SE	*p*-value	Item	Testlet	Factor Loadings	SE	*p*-value
1	–	0.550	0.074	0.000	10-11	2	0.455	0.075	0.000
2	–	0.597	0.066	0.000	12	–	0.517	0.048	0.000
3	–	0.633	0.070	0.000	13-14	3	–0.015*	0.095	0.877
4	–	0.541	0.071	0.000	15-18	4	0.285*	0.068	0.000
5-6	1	–0.161*	0.077	0.036	19-22	5	0.737	0.044	0.000
7	–	0.548	0.077	0.000	23-25	6	0.572	0.061	0.000
8	–	0.325	0.098	0.001	26-27	7	0.327	0.079	0.000
9	–	0.152*	0.087	0.083					

* indicates items with factor loadings < 0.3.

Eleven out of the 15 factor loadings presented values greater than 0.3 (cutoff suggested by McDonald, 1997) and only two factor loadings were not significant at α = 0.05. According to the guidelines from Browne and Cudeck (1993), the RMSEA fit statistic indicated close to fair fit of the unidimensional model. Furthermore, the indices TLI and CFI suggested appropriate fit of the unidimensional model according to the cutoff suggested by Hu and Bentler (1999; TLI and CFI greater than 0.85). In conclusion, the parameter estimates and fit indices indicated good fit for the unidimensional model.

5. DISCUSSION

This study set out to investigate the psychometric properties of the GOALS-2 instrument. Several measurement models were examined to determine which was appropriate to model the data from this instrument. The adopted model, the mixed model, was then used to obtain item-, test-, and person-level characteristics as part of the process to evaluate the developed instrument and to inform decisions about how to improve it.

5.1. ADOPTING A MEASUREMENT MODEL

One of the assumptions of many IRT models is that responses to items in a test are exclusively a function of a single continuous latent variable (de Ayala, 2009). To this point, no validity evidence had been provided about the construct being measured by the GOALS-2 test. This instrument is assumed to measure students' statistical reasoning in a first course of statistics but there is no documented evidence to support this assumption. The standardized factor loadings and the fit indices from the CFA suggest that the internal structure of the investigated instrument is unidimensional in nature. While this finding does help us meet one of the assumptions for using IRT models to analyze the data, there is no information available regarding the relationship between items and the unidimensional construct being measured by the test. Therefore, validity evidence is missing to support the assumption that GOALS-2 is indeed measuring students' statistical reasoning in a first course of statistics. On the other hand, validity evidence was gathered regarding the content assessed by the previous GOALS-1 instrument (see Section 2.2).

Another assumption of IRT analysis is local independence. Context-dependency among items might lead to the violation of this assumption. For this reason, some of the items in the instrument that have a common stimulus were grouped to form a testlet.

Section 4 presented the results for the three IRT models used to fit the data. For both the 2PL and the bi-factor models, around half of the items (44%: 12 out of 27 items for 2PL; 41%: 11 out of 27 items for bi-factor) presented item discrimination values between 0.8 and 2.5, which according to de Ayala (2009) indicates good item discrimination. The mixed model had a slightly higher percentage of items with good item discrimination values (53%: 8 out of 15 items). All three models also produced item discrimination values for some of the items that were unusual, negative, or near zero. The unusual values for item discrimination could be due to the small sample size or lack of fit; however, further research is needed to explore this issue. Most items from the 2PL model and the mixed model had item difficulty values between –2 and 2 (average difficulty). In addition, both models produced a few item difficulty values higher than 3 or lower than –3. (Note: Item difficulty values were not available for the bi-factor model.)

When considering only the item-level diagnostics, the most appropriate fitted model would be the bi-factor model, which presented no item misfit. The likelihood-ratio test, and the AIC and BIC fit statistics also favored the bi-factor model when compared to the 2PL model. While the M_2 statistic indicated significant misfit to the data for all models, it

is noted that M_2 is overly sensitive to small deviations from the model (like all chi-squared statistics). Thus, even though M_2 suggested a statistically significant deviation from the model, the RMSEA values—which indicated close fit for both the mixed and bi-factor models— suggest that this deviation was not of practical significance.

The correlation between the ability parameters for the mixed and the bi-factor models was very high, indicating that both models had similar estimates. The high correlation between the two models' standard errors of the ability parameter also suggested comparable estimates. Therefore, considering the discussion above and the preference for parsimony, the mixed model seems to be the most appropriate for modeling students' responses to GOALS-2.

5.2. IMPROVEMENT OF THE GOALS-2 INSTRUMENT

To explore how to improve the present instrument, the item parameters' estimates and item-level diagnostics from the mixed model, as well as, the test information function, and SEM were computed and analyzed. The item response and item information functions, displayed in Appendix A, were also analyzed. Based on guidelines available in de Ayala (2009), items 8 and 9 have discrimination values lower than 0.8, indicating that these items may not discriminate between low and high ability students. These items address students' understanding of sampling variability. In addition to having a very low discrimination value, Item 8 is the most difficult item on the analyzed instrument.

The stimulus for Item 8 describes a company that claims that 50% of the candies they produce are brown and that candy pieces are randomly placed into bags. Students are then asked whether a person who bought a small or large bag of candy would be more likely to have more than 70% brown candies. Only 23% of the respondents chose the correct option; 72% chose an option which stated that both bags have the same probability of containing 70% brown candy because the bags are both random samples of candy pieces. One possible explanation as to why the majority of students chose this option might be related to how students were learning statistics in the CATALST course, the reference course used here, which emphasizes the study of random processes and the patterns that emerge in repeated trials of randomly generated data. It is possible that this response option is being selected because it includes the word "random". Another explanation is that students might be selecting this option because they are inferring that the *random* process of filling the bags produces an equal chance of 70% brown candies (anything is possible with a random process). While the CATALST course emphasizes the differences in variability of the estimates for different sample sizes, more emphasis on the effect of varying sample size in the curriculum may be needed. Garfield et al. (2012) reported that CATALST students had difficulty reasoning about the effect of sample size on drawing inferences about group differences. Yet another possible explanation might be related to the wording of the item. In the stem, the words "more likely" may lead students to different interpretations of what the question is asking. Further research is needed to fully understand why students are responding with this incorrect response option.

The second item that has a low discrimination value, Item 9, also addresses students' reasoning and understanding of sampling variability. This item asks students to consider 10 random samples of 20 candies drawn from a population in which 35% of the candies are yellow. Each of the four response options provides a potential range for the percentage of yellow candies obtained in the 10 samples: (A) About 0% to 100%, (B) About 15% to 55%, (C) About 30% to 40%, and (D) About 35% to 65%. For this item, 44% of the students assessed selected the correct response option (B). None of the 289 students in the sample selected option C, so this response option could be removed from

the instrument as it is not a plausible distractor. The remaining incorrect response options were selected by roughly the same number of students. Again, why students are selecting these options is unknown, but in part may be due to the fact that the different response options are not mutually exclusive. In addition, the item may not be clear to students since it uses a broad statement such as, "you can reasonably expect".

Two of the seven testlets (Testlets 1 and 3) had very low values of discrimination. Testlet 1 (composed of questions 5 and 6) has a discrimination value that is negative, which indicates that students with high overall ability levels are not performing well on this testlet. Therefore, the items composing this testlet need to be rewritten or replaced. The items included in this testlet measure students' reasoning about statistical significance. One possible reason why the discrimination is very low might be related to how students answered Item 5. This item has two alternatives (A and B) and the proportion of students choosing alternatives A and B is roughly 0.5; therefore, students might be guessing and this would contribute to the low discrimination value. More research would be needed to understand why students would be guessing. Contextually, this item requires students to understand that a statistically non-significant result does not guarantee that there is no effect. A possible explanation might be related to the fact that the reference curriculum does not focus on how the sample size might affect the significance of a test. The CATALST curriculum gives more emphasis on how sample size affects the precision of an estimate.

Testlet 3 (composed of items 13 and 14) also has a low item discrimination value (near zero), suggesting that the testlet is not discriminating well among students with high and low ability levels. The items in this testlet assess students' reasoning about how sampling error is used to make an informal inference about a sample mean. Of note is that the majority of students (81%) answered both of the items in this testlet incorrectly. It is, perhaps, not surprising that students did not perform well on these items given that the topics of sampling distributions and standard error are both topics that have been cited in the research literature as challenging for students (e.g., Kahneman & Tversky, 1972; Tversky & Kahneman, 1971; delMas, Garfield, & Chance, 1999; Chance, delMas, & Garfield, 2005). Consequently, the low discrimination values observed for this testlet may be the result of students' misconceptions rather than poor item writing. In addition, the CATALST curriculum gives great emphasis to using the *p*-value to determine if a result is surprising or not. Therefore, students might not be used to the idea of using the standard error directly to make inferences.

Testlet 4 and Testlet 7 also had low discrimination values, although not as low as those for Testlets 1 and 3. However, based on the information functions, these testlets contribute very little to the reliability of the test scores. Therefore, the items can be dropped to shorten the assessment or revised/replaced with better items to increase information and reliability.

Almost all testlets can be considered of average difficulty. Testlet 4, on the other hand, would be considered easy, since its discrimination value was below –2 (de Ayala, 2009). Testlet 4 assesses students' ability to interpret a confidence interval. Testlet 3 had unusual values for the category boundary locations and item difficulty. Since item difficulty is a function of item discrimination, a possible explanation for these unusual values might be related to item discrimination. As mentioned above, the discrimination value for Testlet 3 was close to zero explaining why the item difficulty for this testlet was so unusual.

Examining the item-level diagnostics statistics suggest that the mixed model only has item misfit for Testlet 7. The items in this testlet (item 6 and 27) assess the ability of students to recognize a misinterpretation of a statistically significant result. Seventy-two

percent of the students answered only one of the items in the testlet correctly, which might be the reason for the misfit. Nevertheless, it is important to note that the discrimination for the testlet was 0.71, and although the item did not fit the model, it still discriminated moderately among students.

From Figure 1, it is clear that GOALS-2 is most informative (scores are most precise) for students with ability values between –3 and –0.5. This plot also suggests that this instrument provides limited information about students' reasoning when their ability level is greater than –0.5. This limitation is because items located higher on the ability scale (e.g., Item 8, Item 9, and Testlet 1) are not discriminating well among students with high and low ability. Therefore, these items provide very little information for estimating students' ability. Consequently, items located in the higher end of the ability scale items with higher difficulty) need to be added to the instrument. These items also need to have good discrimination so that there is an increase in total information for higher ability students.

5.3. LIMITATIONS

Although the research presented yielded some interesting findings, there were limitations in this study. One of these limitations is related to the sample used. The sample consisted of only 289 students, and a larger sample size would lead to more precise estimates of item difficulty, discrimination and person abilities. In addition, the administration of the GOALS-2 instrument was not uniform among all six universities in the United States. Some instructors used this instrument as the final exam in their course, while others used it as an extra-credit opportunity. The mode of administration also differed. Some instructors administered the online version of the instrument, and others administered the paper-and-pencil version. These differences in test administration may contribute to an increase in the error variance in student responses. While it would have been beneficial to account for the nested structure of classes, instructors, and test formats, these hierarchies were not incorporated in the model because of the small sample size.

5.4. IMPLICATIONS FOR FUTURE RESEARCH

The results and suggestions provided in this study can be used to develop a more appropriate version of the GOALS-2 instrument. For instance, items that were flagged due to low discrimination can be reviewed, re-written, or deleted to improve the quality of the instrument. Also more research can be done to better understand why students poorly responded to some of the items.

The instrument was designed to assess students' statistical reasoning after taking an introductory statistics course. However, further research is needed to provide validity evidence regarding the relationship between items and the unidimensional construct being measured by this instrument. In other words, the instrument is *assumed* to measure students' statistical reasoning in a first course of statistics, but there is no evidence to support this assumption. According to the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999), this type of evidence could "include logical or empirical analyses of the adequacy with which the test content represents the content domain and of the relevance of the content domain to the proposed interpretation of test scores. Evidence based on content can also come from expert judgment of the relationship between parts of the test and the construct."

One important characteristic of GOALS-2 is that it is the only published instrument that measures learning outcomes based on how the use of simulations and randomization

tests are contributing to students' understanding of statistical inference. However, only four items (or one testlet) are included on this instrument to assess this learning goal. Additional items addressing this content could be added to the next version. Including new items will broaden the content assessed by this instrument and help to improve the precision of scores. Additional items would also, hopefully, increase the range of ability levels that the instrument could provide information for.

One important consideration is that GOALS-2 seems limited in challenging students with higher ability levels. Many of the difficult questions have a very low discrimination value and likely need to be re-written. Any new items added to the instrument should be written so they are located in the higher part of the ability scale. These items would also need to have good discrimination so that more information will be obtained for estimating students' ability. Therefore, further research can focus on the development of quality items to improve properties of the developed instrument. Additionally, research is needed to understand why students are struggling with some items. Perhaps these items need to be re-written. However, the issue may be with lack of understanding of the content being assessed by these items, the match between content coverage and emphasis and how it was assessed, guessing, or the ability of the student. Understanding why students are guessing on items could help re-write those items to improve the instrument in order to provide more information regarding students' misconceptions related to introductory statistical concepts.

There is still much work to be done to improve the GOALS-2 instrument. However, it is clear that in undertaking this work, the present level already has the potential to become a useful tool in studying and improving students' statistical reasoning.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

American Statistical Association (2005). *Guidelines for assessment and instruction in statistics education: College report*. Alexandria, VA: Author. [Online: http://www.amstat.org/education/gaise/GaiseCollege_full.pdf]

American Statistical Association (2007). *Using statistics effectively in mathematics education research*. Alexandria, VA: Author. [Online: http://www.amstat.org/education/pdfs/UsingStatisticsEffectivelyinMathEdResearch.pdf]

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Allen, K. (2006). *The Statistics Concept Inventory: Development and analysis of a cognitive assessment instrument in statistics*. (Doctoral dissertation). [Online: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2130143]

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, *9*(1), 37–48.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen, & J. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited information goodness-of-fit testing of item response theory models for sparse 2^p tables. *British Journal of Mathematical and Statistical Psychology*, *59*(1), 173-194.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO for Windows 2.1. Lincolnwood, IL: Scientific Software International.

Chance, B. L., & Garfield, J. B. (2002). New approaches to gathering data on student learning for research in statistics education. *Statistics Education Research Journal*, *1*(2), 38–41.
[Online: http://iase-web.org/documents/SERJ/SERJ1%282%29.pdf]

Chance, B., delMas, R., & Garfield, J. (2005). Reasoning about sampling distributions. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 295–323). Dordrecht, The Netherlands: Springer.

Cobb, G. W. (2007). The introductory statistics course: A Ptolemaic curriculum? *Technology Innovations in Statistics Education*, *1*(1), 1–15.
[Online: http://www.escholarship.org/uc/item/6hb3k0nz]

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.

delMas, R. C., Garfield, J., & Chance, B. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education, 7*(3).
[Online: http://www.amstat.org/publications/jse/secure/v7n3/delmas.cfm]

delMas, R. C., Garfield, J. B., Ooms, A., & Chance, B. L. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28–58.
[Online: http://iase-web.org/documents/SERJ/SERJ6(2)_delMas.pdf]

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, *43*(2), 145–168.

DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement*, *36*(2), 104–121.

Earl, L. M., & Katz, M. S. (2006). *Rethinking classroom assessment with purpose in mind: Assessment for learning, assessment as learning, assessment of learning*. Winnipeg, Canada: Manitoba Education, Citizenship and Youth.
[Online: http://www.edu.gov.mb.ca/k12/assess/wncp/full_doc.pdf]

Finch, H., & Monahan, P. (2008). A bootstrap generalization of modified parallel analysis for IRT dimensionality assessment. *Applied Measurement in Education*, *21*(2), 119–140.

Garfield, J. B. (1998). The statistical reasoning assessment: Development and validation of a research tool. In L. Pereira-Mendoza (Ed.), *Proceedings of the Fifth International Conference on Teaching Statistics* (pp. 781–786). Voorburg, The Netherlands: International Statistical Institute.
[Online: http://iase-web.org/documents/papers/icots5/Topic6u.pdf]

Garfield, J. (2003). Assessing statistical reasoning. *Statistics Education Research Journal*, *2*(1), 22–38.
[Online: http://iase-web.org/documents/SERJ/SERJ2(1).pdf]

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching practice*. Dordrecht, The Netherlands: Springer.

Garfield, J., delMas, R., & Zieffler, A. (2010). Assessing important learning outcomes in introductory tertiary statistics courses. In P. Bidgood, N. Hunt, & F. Jolliffe (Eds.), *Assessment methods in statistical education: An international perspective* (pp. 75–86). Chichester, UK: John Wiley & Sons.

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM: The International Journal on Mathematics Education*, *44*(7), 883–898.

Garfield, J., & Franklin, C. (2011). Assessment of learning, for learning, and as learning in statistics education. In C. Batanero, G. Burrill, & C. Reading (Eds.), *Teaching statistics in school mathematics: Challenges for teaching and teacher education* (pp. 133–145). Dordrecht, The Netherlands: Springer.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423–436.

Hambleton, R. K. & Jones, R. W. (1993). An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454.

Konold, C. (1995). Issues in assessing conceptual understanding in probability and statistics, *Journal of Statistics Education*, *3*(1).
[Online: http://www.amstat.org/publications/jse/v3n1/konold.html]

Lane-Getaz, S. J. (2007a). *Development and validation of a research-based assessment: reasoning about p-values and statistical significance* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis, MN.
[Online: http://iase-web.org/documents/dissertations/07.Lane-Getaz.Dissertation.pdf]

Lane-Getaz, S. J. (2007b). Toward the development and validation of the Reasoning about *p*-values and Statistical Significance scale. In B. Phillips & L. Weldon (Eds.), *Proceedings of the IASE Satellite Conference on Assessing Student Learning in Statistics*. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://iase-web.org/documents/papers/sat2007/Lane-Getaz.pdf]

Lane-Getaz, S. J. (2011). A comparison of students' inferential reasoning in three college courses. In *Proceedings of the Joint Statistical Meetings, Section on Statistical Education* (pp. 5467–5481). Alexandria, VA: American Statistical Association.
[Online: http://www.meetingproceedings.us/2011/asa-jsm/contents/papers/303371_70177.pdf]

Lane-Getaz, S. J. (2013). Development of a reliable measure of students' inferential reasoning ability. *Statistics Education Research Journal*, *12*(1), 20–47.
[Online: http://iase-web.org/documents/SERJ/SERJ12(1)_LaneGetaz.pdf]

Maydeu-Olivares, A., & Joe, H. (2005). Limited and full-information estimation and testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*(471), 1009–1020.
[Online: http://www.ub.edu/gdne/documents/maydeu-joe05_jasa.pdf]

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732.

McDonald, R. (1997). Goodness of approximation in the linear model. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests*? (pp. 199–219). Hillsdale, NJ: Erlbaum.

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.
[Online: http://www.statmodel.com/download/usersguide/Mplus%20Users%20Guide%20v6.pdf]

Orlando, M., & Thissen, D. (2000). Likelihood-based item fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of $S\text{-}X^2$: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*(4), 289–298. [Online:

http://www.researchgate.net/publication/247742999_Further_Investigation_of_the_Performance_of_S_-_X2_An_Item_Fit_Index_for_Use_With_Dichotomous_Item_Response_Theory_Models]

Park, J. (2012). *Developing and validating an instrument to measure college students' inferential reasoning in statistics: An argument-based approach to validation* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis, MN. [Online: http://conservancy.umn.edu/handle/11299/165057]

Pearl, D. K., Garfield, J. B., delMas, R., Groth, R. E., Kaplan, J. J., McGowan, H., & Lee, H. S. (2012). *Connecting research to practice in a culture of assessment for introductory college-level statistics*. [Online: http://www.causeweb.org/research/guidelines/ResearchReport_Dec_2012.pdf]

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*(4), 401–412. [Online: conservancy.umn.edu/bitstream/handle/11299/102195/v09n4p401.pdf?sequence=1]

Sabbag, A. (2013). *A psychometric analysis of the Goals and Outcomes Associated with Learning Statistics (GOALS) instrument* (Unpublished master's thesis). University of Minnesota, Minneapolis, MN.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. (Psychometric Monograph 17). Richmond, VA: Psychometric Society. [Online: http://www.psychometricsociety.org/sites/default/files/pdf/MN17.pdf]

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. [Online: http://projecteuclid.org/euclid.aos/1176344136]

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237–247. [Online: dx.doi.org/10.1111/j.1745-3984.1991.tb00356.x]

Stone, A. (2006). *A psychometric analysis of the statistics concept inventory* (Doctoral dissertation). University of Oklahoma, Norman, OK. [Online: http://shareok.org/handle/11244/1013]

Sundre, D. L. (2003, April). Assessment of quantitative reasoning to enhance educational quality. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. [Online: http://apps3.cehd.umn.edu/artist/articles/AERA_2003_QRQ.pdf]

Tintle, N., VanderStoep, J., Holmes, V.-L., Quisenberry, B., & Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education*, *19*(1), 1–25. [Online: http://www.amstat.org/publications/jse/v19n1/tintle.pdf]

Tintle, N., Topliff, K., VanderStoep, J., Holmes, V.-L., & Swanson, T. (2012). Retention of statistical concepts in a preliminary randomization-based introductory statistics curriculum. *Statistics Education Research Journal*, *11*(1), 21–40. [Online: http://iase-web.org/documents/SERJ/SERJ11(1)_Tintle.pdf]

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education*, *8*(2), 157–186.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*(3), 203–220.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187–213.

Zieffler, A., Garfield, J., Alt, S., Dupuis, D., Holleque, K., & Chang, B. (2008). What does research suggest about the teaching and learning of introductory statistics at the

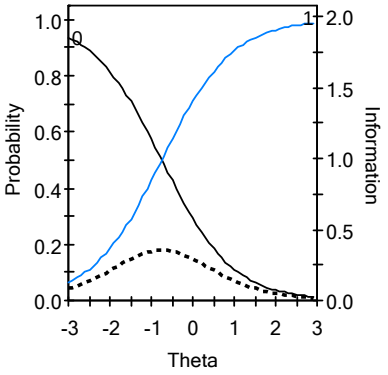
college level? A review of the literature. *Journal of Statistics Education*, *16*(2), 1–25. [Online: http://www.amstat.org/publications/jse/v16n2/zieffler.pdf]

Ziegler, L. (2012). The effect of length of an assessment item on college student responses on an assessment of learning outcomes for introductory statistics (Pre-dissertation paper). University of Minnesota, Minneapolis, MN.

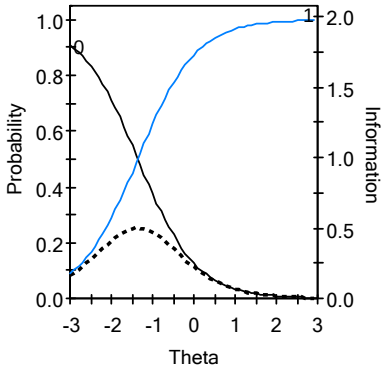
ANELISE G. SABBAG
Room 192, Educational Sciences Building
56 East River Road
Minneapolis, MN 55455

APPENDIX A: ITEM-RESPONSE AND ITEM-INFORMATION FUNCTIONS

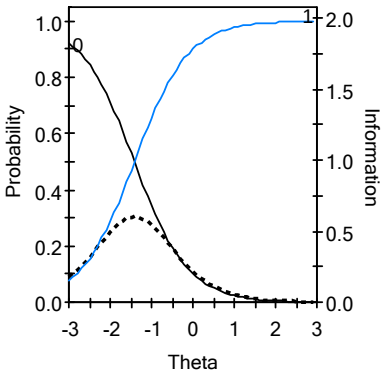
Item 1



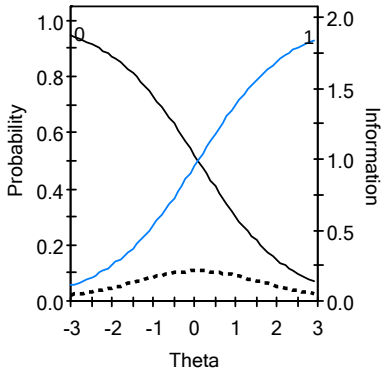
Item 2



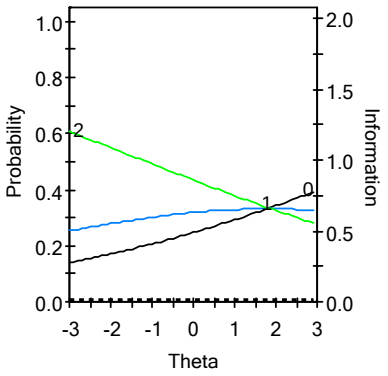
Item 3



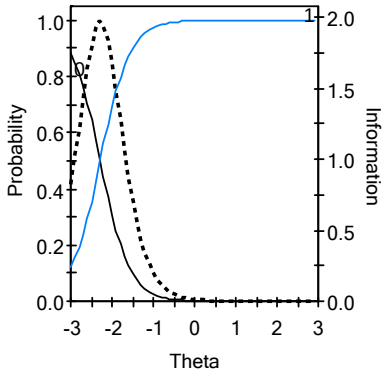
Item 4



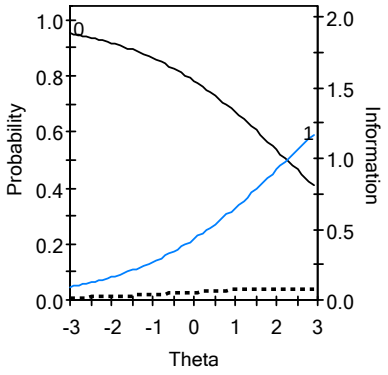
Testlet 1
(It. 5-6)



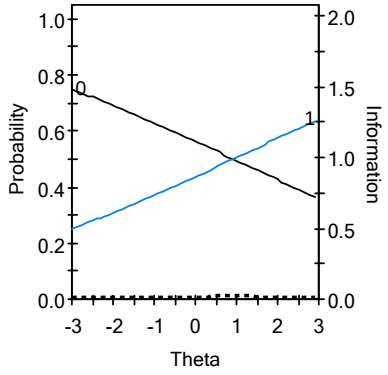
Item 7



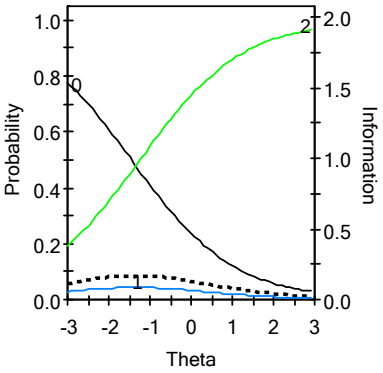
Item 8



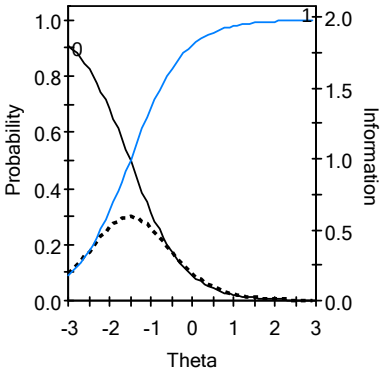
Item 9



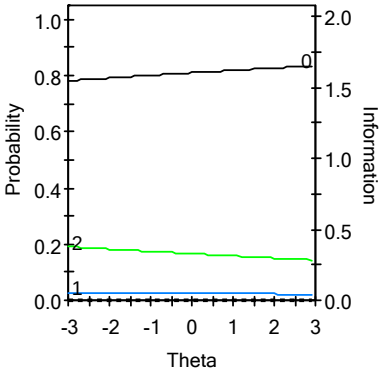
Testlet 2
(It. 10-11)



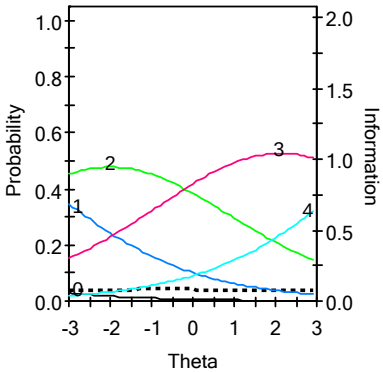
Item 12



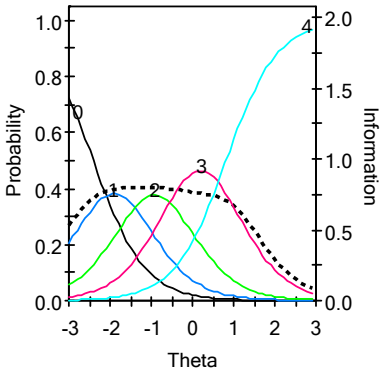
Testlet 3
(It. 13-14)



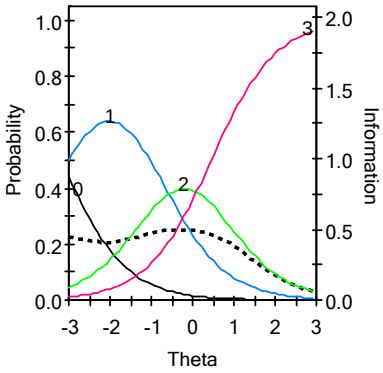
Testlet 4
(It. 15-18)



Testlet 5
(It. 19-22)



Testlet 6
(It. 23-25)



Testlet 7
(It. 26-27)