Use of Item Response Theory to Facilitate Concept Inventory Development

Andrea Stone
Consultant, York, PA, USA
astone@cougars.ccis.edu

Teri Reed-Rhoads
Purdue University, West Lafayette, IN, USA
trhoads@purdue.edu

Teri Jo Murphy
Northern Kentucky University, Highland Heights, KY, USA
murphytj1@nku.edu

P.K. Imbrie
Purdue University, West Lafayette, IN, USA
imbrie@purdue.edu

***Abstract**: This paper explores ways that item response theory techniques can be used to gain additional insight into question behavior and student misconceptions, in addition to being valuable tools for development and evaluation practices.*

Context and Research Questions

Numerous studies in engineering and the sciences have concluded that many students lack correct conceptual understanding of fundamental concepts in technical courses, even after they have ostensibly demonstrated mastery of said material through traditional examinations. Therefore, to help instructors identify student misconceptions, a growing number of concept inventories (CI's) have emerged for engineering related subjects as one method for assessing students' understanding of basic concepts.

To guide the CI development process and/or to evaluate the measurement quality of an existing inventory, individual test items from an inventory are traditionally analyzed using tools from classical test theory (CTT). That is, each question on an inventory (or on an inventory subscale) is evaluated in terms of a set of indices such as an item's (a) difficulty level, (b) discriminability, (c) correlation with the total scale score, and (d) scale alpha if deleted. While techniques based on CTT generally yield valuable information, use of item response theory (IRT) methods can reveal unanticipated subtleties in a dataset. For example, items of extreme difficulty (hard or easy) typically attain low discrimination indices (CTT), thus labeling them as "poor". However, application of IRT can identify these items as strongly discriminating among students of extreme ability (high or low).

Therefore, this paper presents how IRT analysis has been used in the development of the Statistics Concept Inventory to make decisions during the revision process, enabling sample independent comparisons of question versions to be made as well as giving insight as to how a question behaves over a range of abilities. Reliability of the instrument will be assessed from a classical test theory perspective and an IRT perspective. Coefficient alpha (a sample dependent measure) has varied from semester to semester, but has generally been near 0.75. The item response theory reliability estimate obtained from all the data (based on the concept of test information) was found to be 0.78. This initial work is now being extended through the development of "clicker" or personal response system questions. These questions are again driven by student misconceptions where each distractor has a meaning to the instructor and is able to alert him/her to specific and/or common misconceptions within their class.

Can concept inventories be improved by applying Item Response Theory methods?

Theoretical Framework

Item Response Theory (IRT) methods model the probabilities of a correct response using nonlinear models. The basic problem remains the same. There exists a latent trait, Θ, which the test is trying to measure. The trait is, as usual, unobservable and the items on the test are used to estimate Θ. By using nonlinear equations to model the item response functions, we can obtain functions that asymptotically approach 1 for high values of Θ and asymptotically approach 0 for low values of theta (Figure 1). Though there is no prescribed function that must be used, there are three models that are typically used.

For each model, the relationship between the latent trait and the observed examinee responses to test items is modeled by a logistic function. The focus of an IRT analysis is on the pattern of responses to the individual test items for each examinee, as opposed to the total test score. The item response patterns are used to determine a set of parameters for each item. These parameters then determine the shape of the item's item characteristic curve, which models the probability that an examinee with a given ability level will answer the item correctly, $P(X_i = 1 | \Theta)$, see Figure 1. The three models that are commonly in use are the one-, two-, and three parameter logistic models, referred to as 1PL, 2PL, and 3PL models respectively. This research will make use of the 2PL model which is discussed below;

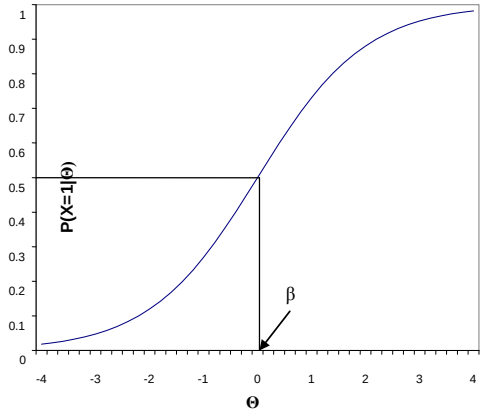


Figure 1: Example of an Item Characteristic Curve (ICC). The threshold parameter β is the value of Θ for which the probability of a correct response is 0.5.

The 2PL model adds an additional parameter, a, which is a discrimination parameter. The model takes the form

$$P(X_i = 1 | \Theta) = \frac{\exp[a_i(\Theta - \beta_i)]}{1 + \exp[a_i(\Theta - \beta_i)]} \qquad [1]$$

where a_i is the value of the slope of the curve at the point Θ = β. The two parameters allow the items to differ in difficulty and discrimination, the ability of the item to differentiate between ability levels. Items which have high a_i values have steep slopes, so that once the threshold ability level is past, the probability of a correct response increases sharply. For lower a_i values, the curves and likewise the probabilities increase gradually, as in Figure Error**! No text of specified style in document.**-1. Steeply increasing curves are more desirable because if a respondent answers a question correctly, then we can be more confident that their ability level is greater than Θ = β. Questions with lower slopes result in more error in the ability estimations.

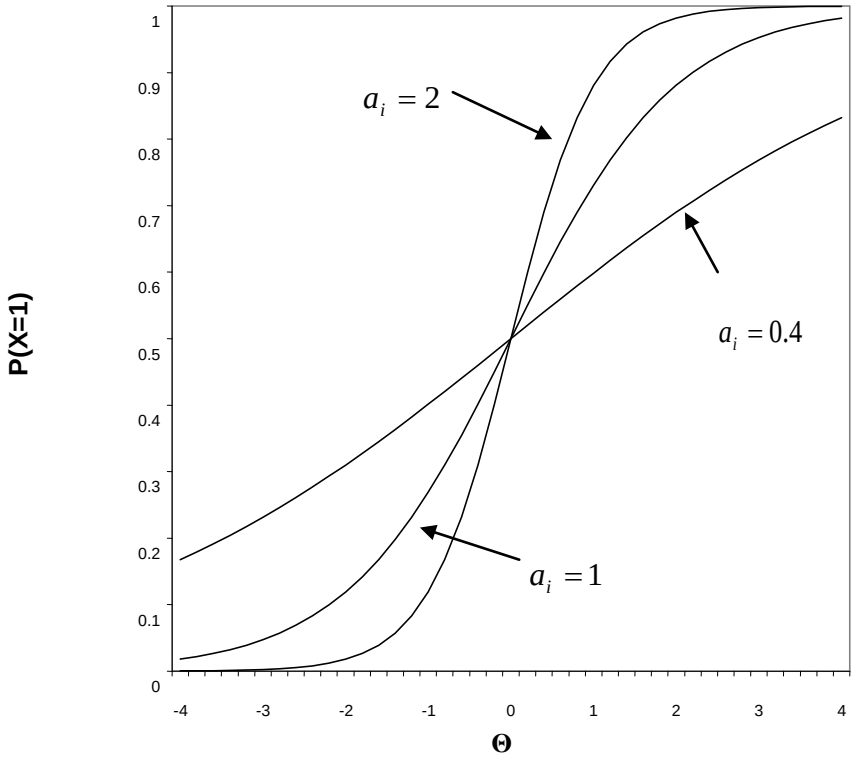


Figure Error! No text of specified style in document.**-1: 2PL item characteristic curves for different values of a, β=0 for all curves.**

The parameter estimates are made using marginal maximum likelihood estimation procedures (Hambleton, Swaminathan, & Rogers, 1991). Under the IRT model, the probability of a correct response depends on the ability and the item parameters, all of which are unknown. What is known is the response pattern for each person. These response patterns are used to select values of the item parameters that maximize the likelihood of obtaining those response patterns. Once the item parameters are known, ability estimates can be obtained for each individual.

The assumptions of the IRT models are that the test is unidimensional, there is only one trait that accounts for the test performance. In practice this assumption is considered to be met if there is a single dominant trait that influences the item responses, this is the trait that is measured by the test. The second assumption is that of local independence. This requires that an examinee's response to one item is independent of their response to another item, once ability has been taken into consideration. Essentially, this means that questions should not give clues to other questions, build on previous questions, etc.

There are several major advantages that IRT provides over CTT and factor analytic models. Assuming that the model fits the data, the parameter estimates are not sample dependent. Furthermore, estimates of examinee ability are also independent of the specific items chosen. The model also allows the measurement error to vary across the ability distribution. These advantages allow for the construction of shorter, more reliable tests, the possibility of adaptive testing, and tests that can be more tailored to specific needs (for example to distinguish between examinees at a narrow part of the ability distribution). It also provides better methods for test equating and detecting test bias.

Despite all the advantages of IRT, there are still important disadvantages. The model assumptions are more restrictive than for the other test models reviewed here. The estimation procedures are much more difficult to employ: they require many computer intensive calculations and special software that is expensive, not widely available, and not particularly easy to use. In addition, large data sets are required in order to estimate the item parameters.

Methodology

The Statistics Concept Inventory (SCI) is one of several concept inventories currently being developed in a variety of engineering disciplines (Evans *et al*., 2003; Foundation Coalition, 2001). Statistics is an increasingly important topic in many disciplines and is receiving increased attention in the K-12 curriculum (National Council of Teachers of Mathematics, 2000). Within engineering, statistics is recognized as an important component of the engineering curriculum and is explicitly included in the ABET accreditation criteria (Engineering Accreditation Commission, 2003).

Enrollment in undergraduate statistics has been rapidly increasing over the last ten years (Loftsgaarden & Watkins, 1998; Schaeffer & Stasny, 2004). During this same time, the reform movement in statistics education has been gaining momentum. The direction of the current reform movement is toward an emphasis on conceptual learning instead of focusing on procedural and computational skills (Ben-Zvi & Garfield, 2004; Cobb, 1993; Gal & Garfield, 1997a; Garfield, Hogg, Schau, & Whittinghill, 2002; Moore, 1997)

Topic selection for the Statistics Concept Inventory (SCI) began with an instructor survey to identify critical topics (Stone et al., 2003). In addition, the Advanced Placement Exam syllabus for statistics, widely used textbooks, and research from statistics education literature identifying student misconceptions were utilized. This information was used to draft question and response sets, incorporating known student misconceptions when possible. The target audience for the SCI was the engineering student population, however due to the homogeneity of content within introductory statistics courses and intentionally limiting the engineering contexts and jargon, the SCI should be able to be widely used.

The SCI consists of 38 items categorized into 4 sub-areas based on the content: probability, descriptive statistics, inferential statistics, and graphical representations. In addition to the item classification, a taxonomy of errors and misconceptions with their associated responses has been compiled.

More than 1200 students have taken the SCI in a variety of statistics courses from engineering, mathematics, psychology, and communication departments. The majority of students have been engineering majors taking an engineering statistics course. Revisions were made to the instrument after each administration based on item analysis including item discrimination, item difficulty, the item response distribution, comments from student focus groups, evaluation for test-wiseness cues, and reliability analysis. Eighteen of the items have a discrimination index higher than 0.4 (considered good), 14 items have moderate values between 0.2 and 0.4, while six items have poorer values of less than .2.

Post-test scores have been consistently low ranging from 45-50% each semester since the pilot version in fall 2002. Scores vary more by course, with courses serving non-engineering student populations tending to score lower. These courses generally have younger students with less mathematics and science backgrounds. Where pre-test scores are available, gains have been minimal (normalized gains range from 1-25%), consistent with the range found with other concept inventories.

The SCI moved to online administration during the fall of 2005. No systematic differences have been found between the paper and online administrations of the instrument. As part of the online version, participants were asked to rank their confidence in their answers on a scale from 1 (not confident at all) to 4 (very confident) (Allen, Rhoads, and Terry, 2006; Allen, 2006). This confidence ranking was then compared to the percent correct for each question using rank order. There was a significant positive correlation between the two rankings ($r = 0.334$, $p = 0.020$). Items were categorized as having over-confidence or under-confidence when the rank order of their confidence rating differed from the rank order of the fraction correct by greater than 10 (Table 6). Probability is an area where many known misconceptions have been identified. Of the 11 questions covering probability topics, 5 fell into the over-confident category with none in the under-confident category. This type of analysis may help identify not only legitimate misconceptions, but also areas of guessing and mastery.

In order to achieve a large enough sample size to carry out an IRT analysis, the questions on the fall 2004 version were divided into groups by topic area and assigned a master number so they could be tracked backward through the previous versions of the instrument. The questions on each previous

version of the SCI were compared to the fall 2004 version. Then a new data set was created for each semester that included the item responses for those questions that were the same as the fall 2004 version. The questions that were different or that were no longer on the fall 2004 version were marked as not presented. Finally these data sets were combined into a single master data set. The same method was followed for subsequent semesters so that a master data set has been created with all data from fall 2002 to summer 2005 with each question having a unique identifier.

A few questions had undergone minor revisions for fall 2004 and had been unchanged for several semesters prior to fall 2004. These questions were included in the data set but were divided into their two versions, for example P2 (earlier version) and P2a (newer version). The data included in the master set are shown in Table 4-1. By including both versions of these questions, we can evaluate the changes that were made and decide whether the changes were an improvement or not. This method is essentially a horizontal equating scheme with common items and non-equivalent groups (Kolan and Brennan 1995). The common items serve as "anchor items" and item parameters are estimated simultaneously. The two forms of the questions can then be compared.

Once the data set had been created, the IRT analysis was carried out using the analysis software BILOG-MG (Zimowsky, Muraki, Mislevy and Bock 2003). The data were modeled with a 2-parameter logistic model. In this model, two parameters for each item are estimated that define the item characteristic curve (ICC) for that item; a slope or discrimination parameter, a, and a threshold parameter, β. The threshold parameter is the value of theta (the ability level) for which the probability of answering the question correctly is 0.5. The discrimination parameter is the slope of the ICC at the point $\Theta=\beta$. For the estimation routine, Bayesian priors were used for both the slope and the threshold parameters. The following analysis is in the logit metric, for which the model is depicted in equation (1).

Table Error**! No text of specified style in document.**-1 contains the item statistics and item parameter estimates. Recall that higher values of the discrimination parameter a are desirable, the normal range of values is from 0 to 2. The threshold parameter β is a measure of the item difficulty and it the point along the ability distribution where the probability of answering correctly is 0.5. For example, consistent with previous findings, the parameter estimates for question P4 indicate a very difficult question (β=5.752) with low discrimination (α=0.329). Similar results are found for question G2 (β=5.251, α=0.307).

Table Error! No text of specified style in document.**-1: Item Statistics and Parameter Estimates**

		Item Statistics			Item Parameters		
Item	**N**	**% Correct**	**Pearson Correlation**	**Biserial Correlation**	**Slope Parameter *a***	**Threshold Parameter β**	**Factor Loading**
XD1	**272**	**31.2**	**0.121**	**0.158**	**0.454**	**1.754**	**0.413**
XD2	**483**	**56.1**	**0.039**	**0.049**	**0.303**	**-0.871**	**0.29**
XD3	**1146**	**72.3**	**0.208**	**0.279**	**0.755**	**-1.43**	**0.603**
XD4	**1146**	**69.8**	**0.276**	**0.363**	**1.073**	**-0.969**	**0.731**
XD5	**593**	**63.6**	**0.26**	**0.333**	**0.733**	**-0.884**	**0.591**
XD6	**976**	**71.4**	**0.299**	**0.397**	**1.044**	**-1.059**	**0.722**
XD7	**976**	**61.5**	**0.167**	**0.213**	**0.533**	**-0.921**	**0.47**
XD8	**390**	**46.2**	**0.314**	**0.394**	**0.872**	**0.249**	**0.657**
XD8A	**483**	**41.4**	**0.254**	**0.321**	**0.778**	**0.436**	**0.614**
XD9	**483**	**64.8**	**0.229**	**0.294**	**0.732**	**-0.986**	**0.591**
XD10	**873**	**67.6**	**0.341**	**0.444**	**1.152**	**-0.824**	**0.755**
XG1	**499**	**26.5**	**0.305**	**0.411**	**0.943**	**1.206**	**0.686**
XG2	**499**	**16**	**0.018**	**0.028**	**0.307**	**5.251**	**0.294**
XG3	**499**	**54.5**	**0.062**	**0.078**	**0.307**	**-0.643**	**0.293**
XG4	**873**	**38**	**0.243**	**0.31**	**0.654**	**0.801**	**0.547**
XG5	**499**	**67.3**	**0.064**	**0.083**	**0.353**	**-2.122**	**0.333**
XG6	**873**	**20.6**	**0.216**	**0.307**	**0.711**	**2.072**	**0.58**
XG7	**499**	**43.7**	**0.103**	**0.13**	**0.396**	**0.602**	**0.368**
XI1	**483**	**47.8**	**0.146**	**0.183**	**0.42**	**0.152**	**0.387**

XI2	**593**	**41**	**0.069**	**0.087**	**0.324**	**1.098**	**0.308**
XI3	**873**	**43.5**	**0.255**	**0.321**	**0.642**	**0.427**	**0.54**
XI4	**878**	**26.7**	**0.151**	**0.204**	**0.553**	**1.98**	**0.484**
XI4A	**268**	**32.1**	**-0.004**	**-0.005**	**0.332**	**2.133**	**0.315**
XI05	**499**	**30.3**	**0.029**	**0.039**	**0.306**	**2.671**	**0.292**
XI06	**873**	**36.2**	**0.291**	**0.373**	**0.888**	**0.727**	**0.664**
XI07	**976**	**41.6**	**0.315**	**0.398**	**0.921**	**0.443**	**0.678**
XI08	**483**	**88.8**	**0.21**	**0.348**	**0.884**	**-2.681**	**0.662**
XI09	**483**	**43.3**	**0.11**	**0.139**	**0.346**	**0.731**	**0.327**
XI10	**663**	**42.1**	**0.171**	**0.215**	**0.484**	**0.738**	**0.435**
XI10A	**483**	**43.1**	**0.307**	**0.387**	**0.834**	**0.316**	**0.64**
XI11	**374**	**39.6**	**-0.031**	**-0.039**	**0.256**	**1.54**	**0.248**
XI11A	**109**	**49.5**	**0.167**	**0.209**	**0.554**	**0.04**	**0.485**
XI12	**166**	**57.8**	**0.2**	**0.253**	**0.591**	**-0.599**	**0.509**
XP1	**374**	**58.6**	**0.231**	**0.292**	**0.704**	**-0.622**	**0.576**
XP1A	**109**	**45**	**0.147**	**0.184**	**0.579**	**0.371**	**0.501**
XP2	**773**	**34.3**	**0.346**	**0.447**	**1.018**	**0.821**	**0.714**
XP2A	**203**	**41.4**	**0.297**	**0.375**	**0.806**	**0.357**	**0.628**
XP3	**483**	**14.7**	**0.086**	**0.133**	**0.49**	**3.629**	**0.44**
XP4	**1037**	**13.2**	**0.046**	**0.072**	**0.329**	**5.752**	**0.312**
XP4A	**166**	**7.8**	**0.172**	**0.316**	**0.724**	**3.543**	**0.586**
XP5	**483**	**56.1**	**0.063**	**0.079**	**0.322**	**-0.825**	**0.307**
XP6	**499**	**56.9**	**0.134**	**0.169**	**0.466**	**-0.675**	**0.423**
XP7	**493**	**29.6**	**0.257**	**0.34**	**0.71**	**1.44**	**0.579**
XP7A	**483**	**48.9**	**0.291**	**0.364**	**0.857**	**-0.005**	**0.651**
XP8	**976**	**34.3**	**0.373**	**0.482**	**1.196**	**0.704**	**0.767**
XP9	**976**	**67**	**0.315**	**0.409**	**1.035**	**-0.826**	**0.719**

Findings

Question I10 (#35), which is about confidence intervals, was reworded in both the stem and the response set for fall 04. The response distribution appears to be similar for both forms of the question:

> I10a. When calculating a confidence interval on a given population with a fixed significance level, using a larger sample size will make the confidence interval:
>
> a) Smaller (Correct)
>
> b) Larger
>
> c) No Change
>
> d) It depends on the significance level
>
> I10. Two confidence intervals are calculated for two samples from a given population. Assume the two samples have the same standard deviation and that the confidence level is fixed. Compared to the smaller sample, the confidence interval for the larger sample will be:
>
> a) Narrower (Correct)
>
> b) Wider
>
> c) The same width
>
> d) It depends on the confidence level

The discrimination index for this question had generally been around 0.3 and it has usually had a midrange alpha-if-item-deleted ranking. The new version had a considerably higher discrimination index, over 0.5 and one of the highest alpha-if-item deleted rankings. The item characteristic curves are shown in Figure **Error! No text of specified style in document.**-2. The item difficulty for the two questions was about the same, but the newer version of the question was more discriminating. The

questions are virtually the same at face value, but their behaviour is different. The newer version I10a, which is more precise and seems to work better, was retained.

Future Research

Item analysis of this type can help to guide further refinements of the SCI. Being able to make comparisons based on information that is derived from all the data at once instead of from a single semester can lend increased confidence to the subsequent decisions. The item characteristic curves also provide a sense of the question behaviour over the entire ability distribution.

The SCI is a unique instrument for evaluating statistics understanding. There is no other instrument currently available which focuses on conceptual understanding and which covers the scope of a typical introductory statistics course. It has been demonstrated to be a reasonably reliable instrument for research use. The SCI should be used in classroom settings as a posttest and optionally as a pretest for the purposes of evaluating instructional methods. Baseline data is available that can be used as a benchmark for comparison.

It is hoped that instructors find that the content on the SCI corresponds to what they expect their students to have mastered upon leaving the introductory statistics course. As such, the SCI can fulfill the role that other concept inventories have in initiating widespread interest in instructional research and innovations for statistics within the classroom setting.

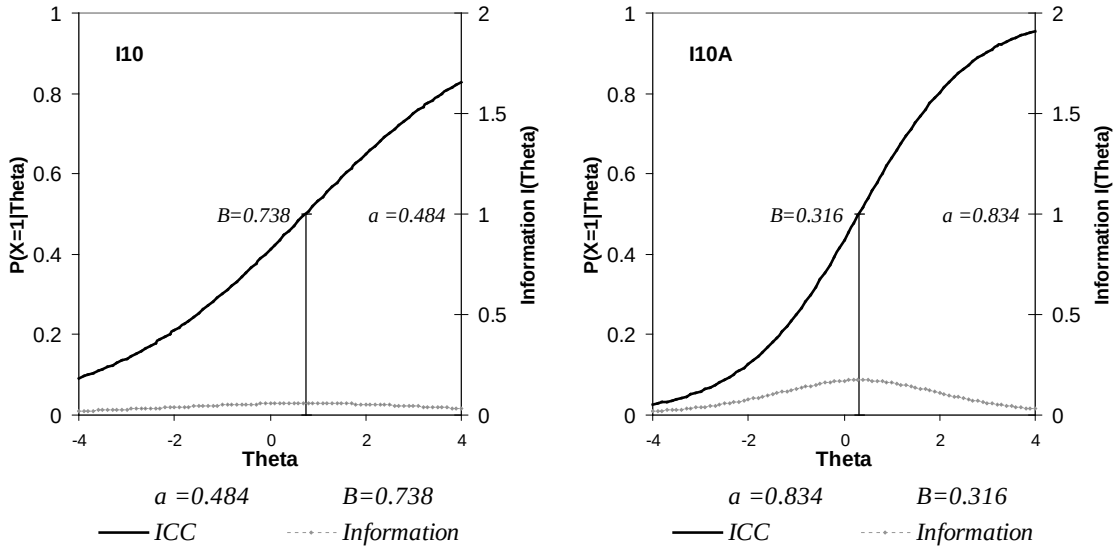


Figure Error! No text of specified style in document.**-2: Item characteristic and information curves for items I10 and I10a .**

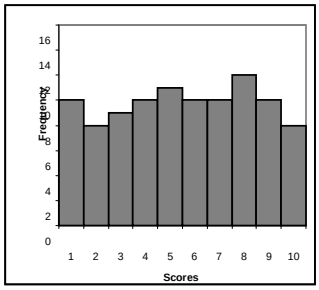
Another item response theory model which has potential to be helpful in the future development of the SCI and other concept inventories is Bock's nominal response model (Bock 1972). When multiple choice items are dichotomously scored, information contained in the incorrect responses is essentially lost because all the incorrect answers are collapsed into one category. One of the main ideas underlying the concept inventory movement is that important information about student understanding is contained in the incorrect responses as well as the correct responses.

For each item, the nominal response model provides a response curve for every response alternative, not simply the correct one. In this way all the information in the response pattern is used and this can help increase the accuracy of the theta estimates. In addition, it provides a more accurate picture of the item behavior across the theta distribution, including which distractors are more likely to be chosen at each point along the distribution.

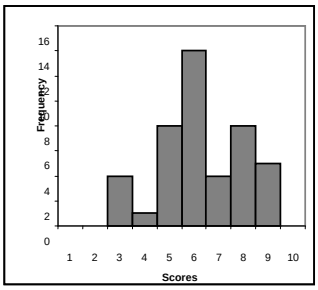
The same data set was used for this analysis as for the 2PL model discussed before. The parameter estimation was conducted using MULTILOG (Thissen 2003). An example that can result from this additional analysis is the following;

In question G6 (#30), students were asked to identify which distribution would have the greatest variance. Routinely response (b) was chosen by almost 60% of examinees. Focus group interviews indicated students focus on the bumpiness or raggedness of the shape of the distribution with no thought given to any notion of spread or wideness or relation to center. This type of reasoning would indicate a fundamental lack of understanding about variance or at the very least a lack of visual representation for the concept.

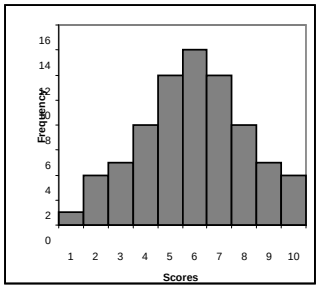
G6. The following are histograms of quiz scores for four different classes. Which distribution shows the most variability?



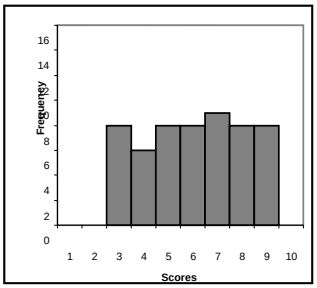
I



II



III



IV

a) I (Correct)

b) II

c) III

d) IV

The response curves, shown in **Figure** Error! No text of specified style in document.**-3** indicate that this belief is widespread throughout the theta distribution. Since variation is one of the key ideas of statistics, this is an important misconception that could be addressed in a statistics course.

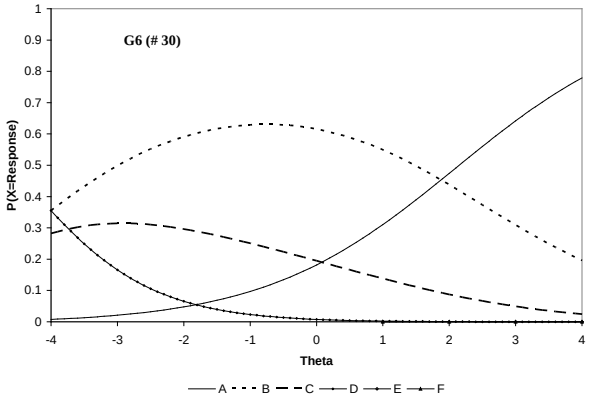


Figure Error! No text of specified style in document.**-3: Response curves for item G6 for the nominal response model.**

This type of analysis will be used when making further revisions to the SCI. In addition, it may help pinpoint specific errors or misconceptions that may be useful in developing instructional strategies. Due to the large sample size required to employ these techniques, they typically could not be used in the beginning phases of the development of a concept inventory. However, once a suitably large data set has been amassed, the techniques are promising.

References

Allen, K. 2006. Statistics Concept Inventory. Doctoral Dissertation, Industrial Engineering, University of Oklahoma.

Allen, K, Reed-Rhoads, T., and Terry, R. 2006. Work in Progress: Assessing Student Confidence of Introductory Statistics Concepts. Paper presented at the 36th ASEE/IEEE Frontiers in Education Conference, October 28-31, at San Diego, CA.

Ben-Zvi, D., & Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben_Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 3-15): Kluwer.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29-51.

Cobb, G. W. (1993). Reconsidering statistics education: A national science foundation conference, *Journal of Statisitcs Education* (Vol. 1).

Engineering Accreditation Commission. (2003). 2004-2005 criteria for accrediting engineering programs. Retrieved April 12, 2005, from http://www.abet.org/criteria_eac.html

Evans, D. L., Gray, G. L., Krause, S., Martin, J., Midkiff, C., Notaros, B. M., et al. (2003, November 5-8). *Progress on concept inventory assessment tools.* Paper presented at the 33rd ASEE/IEEE Frontiers in Education Conference, Boulder, CO.

Foundation Coalition. (2001, 2/22/2005). Foundation coalition key components: Concept inventories. Retrieved February 22, 2005, from http://www.foundationcoalition.org/home/keycomponents/concept/index.html

Gal, I., & Garfield, J. (1997a). Curricular goals and assessment challenges in statistics education. In I. Gal & J. B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 1-16). Amsterdam: IOS Press.

Garfield, J., Hogg, B., Schau, C., & Whittinghill, D. (2002). First courses in statistical science: The status of educational reform efforts, *Journal of Statistics Education* (Vol. 10).

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

Kolan, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.

Loftsgaarden, D. O., & Watkins, A. E. (1998). Statistics teaching in colleges and universities: Courses, instructors, and degrees in fall 1995. *The American Statistician, 52*(4), 308-314.

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review, 65*(2), 123-137.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

Schaeffer, R. L., & Stasny, E. A. (2004). The state of undergraduate education in statistics: A report from the CBMS 2000. *The American Statistician, 58*(4), 265-271.

Stone, A. 2006. A Psychometric Analysis of the Statistics Concept Inventory. Doctoral Dissertation, Mathematics, University of Oklahoma, Norman, OK.

Stone, A., Allen, K., Reed-Rhoads, T., Murphy, T.J., Shehab, R. L.and Saha, C. 2003. The Statistics Concept Inventory: A Pilot Study. Paper presented at the ASEE/IEEE Frontiers in Education Conference, November 5-8, at Boulder, CO.

Thissen, D. (2003). MULTILOG (Version 7.0.2327.3): Scientific Software International, Inc.

Zimowsky, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG (Version 3.0): Scientific Software International.

Acknowledgements

The National Science Foundation has supported this research through many grants, such as DUE-0731232 and DUE-0206977**.**

Copyright statement