AC 2007-1783: GETTING MORE FROM YOUR DATA: APPLICATION OF ITEM RESPONSE THEORY TO THE STATISTICS CONCEPT INVENTORY

Kirk Allen, Purdue University

Kirk Allen is a post-doctoral researcher in Purdue University's Department of Engineering Education. His dissertation research at The University of Oklahoma was the development and analysis of the Statistics Concept Inventory (NSF DUE 0206977), which combines his interest in statistics and assessment methodologies.

Getting More from your Data: Application of Item Response Theory to the Statistics Concept Inventory

Abstract

This paper applies the techniques of Item Response Theory (IRT) to data from the Statistics Concept Inventory (SCI). Based on results from the Fall 2005 post-test (n = 422 students), the analyses of IRT are compared with those of Classical Test Theory (CTT). The concepts are extended to discussions of other applications, such as computerized adaptive testing and Likert-scale items which may be of interest to the engineering education community.

While techniques based on CTT generally yield valuable information, methods of IRT 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". Application of IRT can identify these items as strongly discriminating among students of extreme ability (high or low). The three simplest IRT models (one-, two-, and three-parameter) are compared to illustrate cases where they differ. The theoretical foundations of IRT are provided, extending to validating the assumptions for the SCI dataset and discussing other potential uses of IRT that are applicable to survey design in engineering education.

Introduction

The Steering Committee of the National Engineering Education Research Colloquies[1] identified assessment ("Research on, and the development of, assessment methods, instruments, and metrics to inform engineering education practice and learning") as one of five areas that form the foundation of engineering education research. Further, there are calls for engineering education to become both more inter-disciplinary[2] and rigorous[3].

Item Response Theory[4,5,6,7] (IRT) is commonly used by psychologists in survey design and analysis; effectively "learning the language" opens a conduit for collaboration and dissemination. IRT is therefore useful to place in the engineering education researcher's toolbelt. The mathematical advantages of IRT enhance rigor with procedures to track characteristics of both the test and examinees across such variables as time, gender, major, etc.

This article first describes item response theory from a theoretical perspective, describing the common models for dichotomous (those scored 'correct' or 'incorrect') items and their assumptions. Secondly, interpretation of IRT is presented by application to the Statistics Concept Inventory (SCI). Finally, some extensions of IRT are described which may be of interest to the engineering educator.

Background [4,5,6,7]

Classical test theory (CTT) is based on the model that an individual's observed score on a test depends on a true ability and some amount of measurement error. This is expressed symbolically below.

$$X = T + E$$

where: X is the individual's observed score
T is the individual's true score
E is the error associated with this observation

Similarly, item response theory assumes an under-lying ability influences an individual's response to items. This ability is called a latent trait and symbolized θ. CTT is interested in performance on the test as a whole, with the individual test items being summed to yield a total score (X, above). Item Response Theory (IRT), as the name implies, begins at the item level and is concerned with an examinee's pattern of responses. This bottom-up vantage drives the methods and associated conclusions of IRT.

In its simplest form, item response theory fits a logistic regression model to binary item responses with each examinee's true score (θ) as the independent variable. The 1-parameter logistic (1-PL) model accounts for varying item difficulties. Developer George Rasch, a Danish mathematician, is the eponym of what is commonly called the "Rasch Model." The difficulty parameter, β_j, shifts the item response curve along the abscissa; β_j is the value of true score for which the probability of answering correctly is 0.5. An easy item shifts to the left, signifying a lower ability is able to obtain a high probability of answering correctly. Using the observed total exam scores as an estimate of true score, this model can be fit with any statistical software package. The 1-PL model is shown below:

$$P(X_{ij} = 1 \mid \theta_i) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)}$$

where: X_{ij} is a person i's binary response to item j
θ_i is person i's latent ability level (true score)
β_j is item j's difficulty parameter

The graph of the logistic function is referred to as the item characteristic curve (ICC). Figure 1 shows a family of curves for differing values of β_j.

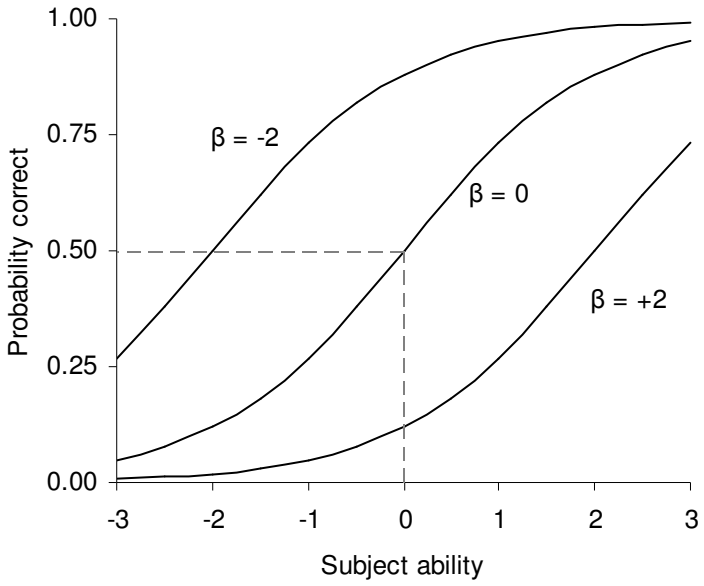


Figure 1: Family of ICC's for 1-PL model
(the calculation of β is highlited for an average ability level of zero)

An additional term can be included to account for the interaction between item and person parameters; the model is now called 2-PL (two-parameter logistic). This additional term is referred to as the item discrimination parameter, α_j, and is equal to the slope of the curve at β_j. A high value will give a steep slope, which means that the item is very good at discriminating between individuals around the inflexion point of the item response curve. Negative slopes indicate low-ability individuals have a higher probability of responding correctly; this is undesirable in achievement testing, though may be encountered in personality scales. The equation and sample graph follow.

$$P(X_{ij} = 1 \mid \theta_i) = \frac{\exp[\alpha_j(\theta_i - \beta_j)]}{1 + \exp[\alpha_j(\theta_i - \beta_j)]}$$

where: α_j is the item discrimination parameter
(note that this equation reduces to the 1-PL when $\alpha_j = 1$)

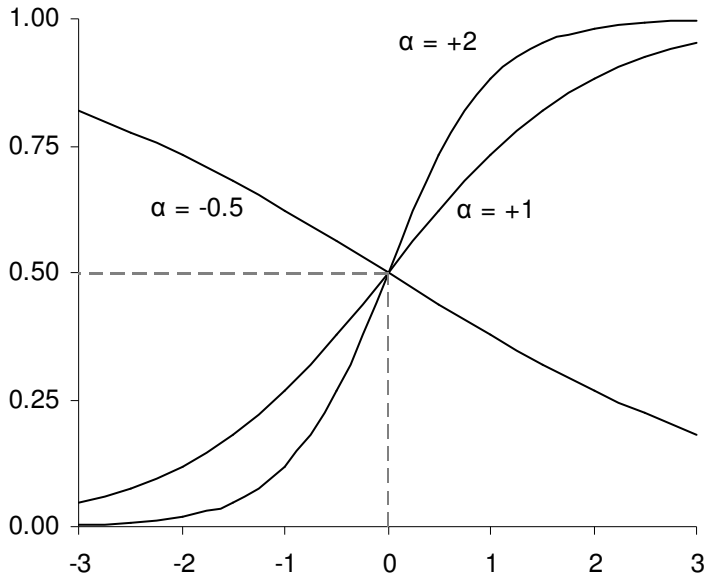


Figure 2: Family of ICC's for 2-PL model
(β constant at zero, again highlited as in Figure 1)

Still more complex, the 3-PL includes a lower asymptote (c) to account for guessing. This means a person of very low (i.e., negative) ability has a probability greater than zero of getting an item correct by pure chance. The graph takes the same form as the 2-PL, but it asymptotes above zero for low ability levels. Consequently, the difficulty parameter (β) is midway between c and 1. The equation is shown below:

$$P(X_{ij} = 1 \mid \theta_i) = c_j + (1 - c_j)\frac{\exp[\alpha_j(\theta_i - \beta_j)]}{1 + \exp[\alpha_j(\theta_i - \beta_j)]}$$

where: c_j is the guessing parameter
(note that this equation reduces to the 2-PL when $c_j = 0$)

Information is the IRT analogue of reliability in classical test theory. Information is defined as being inversely proportional to the standard error of measurement. Each item's information is expressed by the equation below. The information of the test is the sum of the information of the individual items.

$$I_j(\theta) = \frac{[P_j^*(\theta)]^2}{P_j(\theta)(1 - P_j(\theta))}$$

where: I_j is item j's information
P_j is the probability of answering correctly, as a function of θ
P_j^* is the derivative of the ICC at θ

It need be stressed that information is a function of ability. This is an important distinction from classical test theory, wherein reliability takes a single value. The IRT formulation is theoretically and pragmatically more appropriate: for instance, an extremely difficult test where examinees are

forced to guess at best informs the practitioner that the material is not grasped, but the test is unable to differentiate between levels of mastery.

Assumptions

Item Response Theory is based on two inter-related assumptions: unidimensionality and local independence. The former states that a single under-lying trait influences item responses. The latter states that there is no response tendency carry-over from item-to-item aside from examinee ability. A dataset with, for example, two dimensions (i.e., latent traits) violates the dimensionality assumption. It follows that local independence is violated because a different trait determines examinee responses along each dimension. Importantly, these assumptions are based on the data (*de facto*) and are not merely inferred from the instrument design (*de jure*). For practical purposes, local independence can be "taken as a consequence" of a unidimensional item pool[5].

A sort-of "meta-assumption" is that the model fits. There is no consensus on model-fit assessment. Methods typically partition examinees into ability bins and compare observed percent correct to the item response curves. These comparisons can be conducted graphically and/or formally with χ^2 statistics. One of these statistical methods (Q_1) is defined below[8], based on portioning examinees into 10 ability bins, respecting the notation of the original article.

$$Q_{1i} = \sum_{j=1}^{10} \frac{N_j \left(O_{ij} - E_{ij}\right)^2}{E_{ij}\left(1 - E_{ij}\right)}$$

where: subscript i refers to items
subscript j refers to ability-bins
N_j are the number of subjects in bin j (approximately one-tenth of total N)
O_{ij} are the observed proportion correct for item i of subjects in bin j
E_{ij} are the model-predicted proportion correct for item i of subjects in bin j

Q_1 is distributed as χ^2 with degrees of freedom $10 - m$, where m is the number of parameters estimated (e.g., $m = 1$ for 1-PL). Being a goodness-of-fit-test, the null hypothesis is *the model fits*. This contrasts with hypothesis tests typically encountered in introductory statistics, where the null hypothesis is what one is attempting to disprove.

To conclude that the model does not fit, a critical value of 0.05 is assumed. The respective χ^2 critical values are 16.9, 15.5, and 14.1 for the 1-, 2-, and 3-PL models. For example, the value $Q_1 = 11.9$ for a 3-PL is a good fit (it fails to fall in the rejection region of the null distribution).

Computational Details

IRT utilizes maximum likelihood estimation. With the item response curves calibrated, the likelihood function is used to calculate examinee abilities. The likelihood function is defined as follows.

$$L(\theta) = \prod_i P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i}$$

where L is the likelihood function
θ is the ability level
subscript i refers to items
exponent u is 0 or 1 (i.e., incorrect or correct)
P_i is the model-predicted correct probability
Q_i is the model-predicted incorrect probability ($P_i + Q_i = 1$)
[note that P_i (and thus Q_i) is a function of θ]

For an exam with five items, there are thirty-two (2^5) possible response patterns (e.g., 00000, 00001, 00010, 00100, …, 11111). Each response pattern yields has its own likelihood function. The ability estimate for a subject is then θ which yields the *maximum L(θ)*.

Consider a case where the response pattern is 11100, i.e., the first three correct and the final two incorrect. In this case, $L(\theta) = P_1 * P_2 * P_3 * Q_4 * Q_5$, where each P_i is a function of θ. An optimization technique can then be applied to find the maximum L(θ), yielding the θ ability estimate for this response pattern.

Figure 3 is based on this response pattern (11100), using the item parameters from the first five questions on the Statistics Concept Inventory. For comparison, a response pattern of 11000 is included. This crude illustration shows that the former yields an ability estimate of 0.4, while the latter yields -0.5; these are the values along the θ axis corresponding to the maximum likelihood.

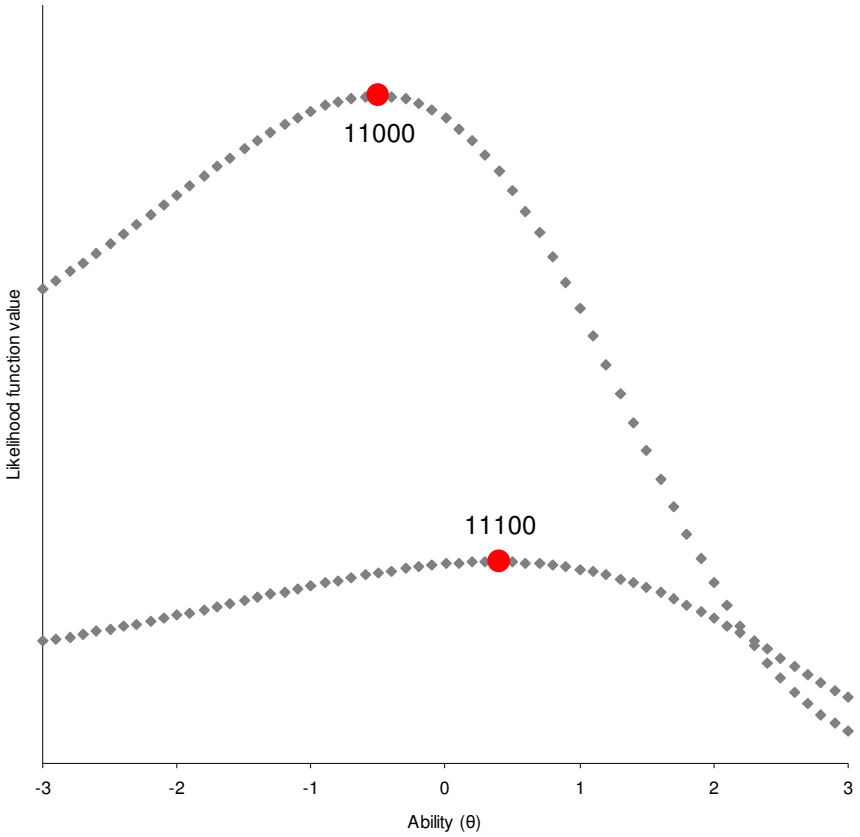


Figure 3: Illustration of maximum likelihood for ability estimation

Given a set of item parameters, estimating examinee abilities thus seems straight-forward. Similar comments can be made about estimating item parameters. Any statistical software package could fit a logistic regression curve to data such as that found in Figure 4, with data points being ordered pairs of the form ($x = \theta$, y = 0 or 1). Clearly, one cannot directly estimate item parameters *and* abilities when both are unknown. Thus, procedures have been devised where, for example, abilities are prescribed starting values, with item parameters estimated in turn. These parameters are then taken as fixed and used to re-estimate the abilities. Such a cyclical procedure continues until a convergent solution is found along both dimensions.

Tests are rarely static entities, continually under-going revision as the uses evolve. Rather than wiping the data slate clean at each iteration, maximum likelihood affords the continual inclusion of all data. For instance, the five-question case described above (e.g., 11100) can derive from an early version. Perhaps the first two items were deleted, due to poor discrimination or content concerns, with two additional items added in place. Now, consider a response pattern with the two additional items both correct, while the original first two were not answered. This can be represented as ××10011, where the × means the item was not presented. Similarly, the original response pattern 11100 can be appended 11100××. The likelihood function remains $L(\theta) = P_1 * P_2 * P_3 * Q_4 * Q_5$ in the first-five case, while $L(\theta) = P_3 * Q_4 * Q_5 * P_6 * P_7$ in the last-five case. The item parameters and subject abilities therefore remain on equal footing as part of a seven-item pool, rather than being analyzed separately as part of two five-item pools.

Methods

The Statistics Concept Inventory (SCI) is a multiple choice assessment instrument designed for use in introductory statistics courses. Prior publications document the validation process[9,10] and the identification of student misconceptions[11]. Though generally classified as one of the engineering concept inventories[12], data presented herein (n = 422 students from Fall 2005) include courses from Mathematics, Meteorology, and Psychology departments. Prior and ongoing collaborations include other disciplines such as Communications and Economics.

Calculations were performed using IRT Command Language (ICL)[13]. This software was selected because it is freely available, thus making it more accessible to the engineering community who may not have access to commercial IRT software which is traditionally utilized by psychologists (e.g., BILOG). Especially for those forging first forays into IRT, the freeness of ICL outweighs the more widespread application of BILOG. Another freeware alternative is PARAM-3PL[14], which performs the full estimation of item and person parameters for the 3-PL but has limited options for lower-order models.

Results

This section describes the item characteristic curves for three questions from the Statistics Concept Inventory (SCI); they are available for review in an Appendix. Figure 4 shows the item response curves using the one-parameter (1-PL; light), two-parameter (2-PL; middle), and three-parameter (3-PL; dark) IRT models. The red dots represent examinees grouped into ten nearly-equally sized bins by ability, to demonstrate fit via Q_1. Little difference is apparent between the three models for this item, due to the steep slope (i.e., highly discriminating) and moderate difficulty.

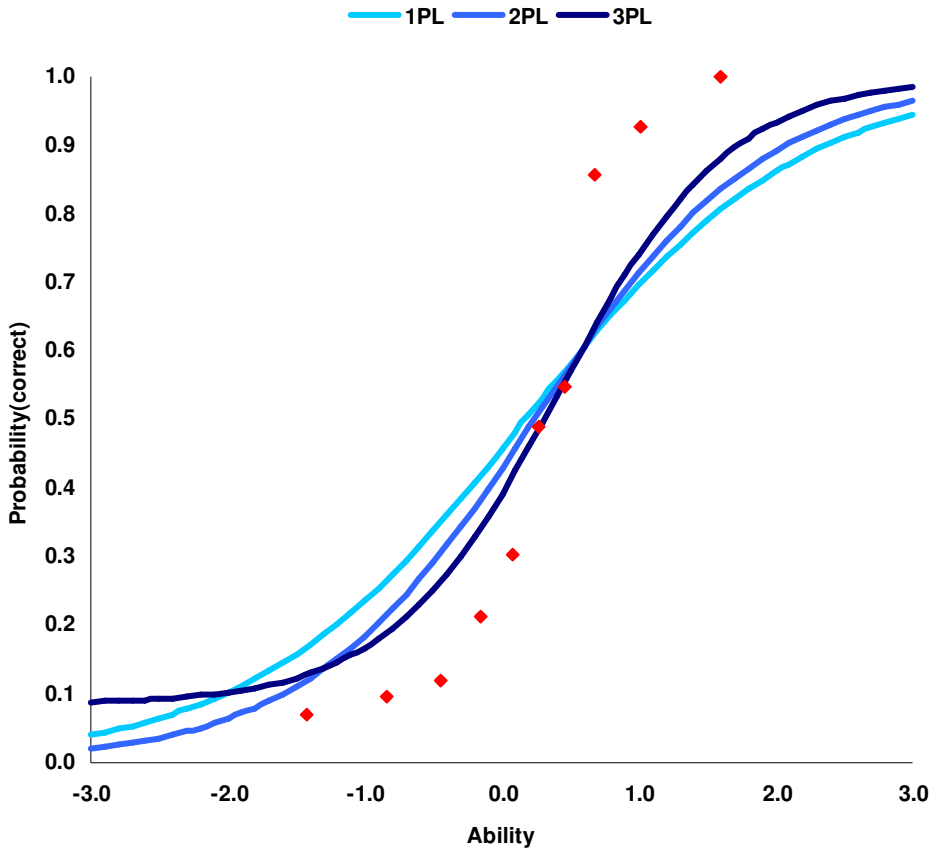


Figure 4: Item Characteristic Curves for 1-, 2-, and 3-PL;
little difference for this item of high discrimination and moderate difficulty

Figure 5 shows the inadequacy of the 1-PL model for some items. The 2-PL and 3-PL capture the poor discrimination across ability levels (e.g., the lowest and highest ability bins differ by only 14%). The discrimination index (CTT) is 0.20, which is at best borderline for retention if the item pool were revised. Because discrimination varies within the item pool, the 1-PL is not appropriate for this dataset.

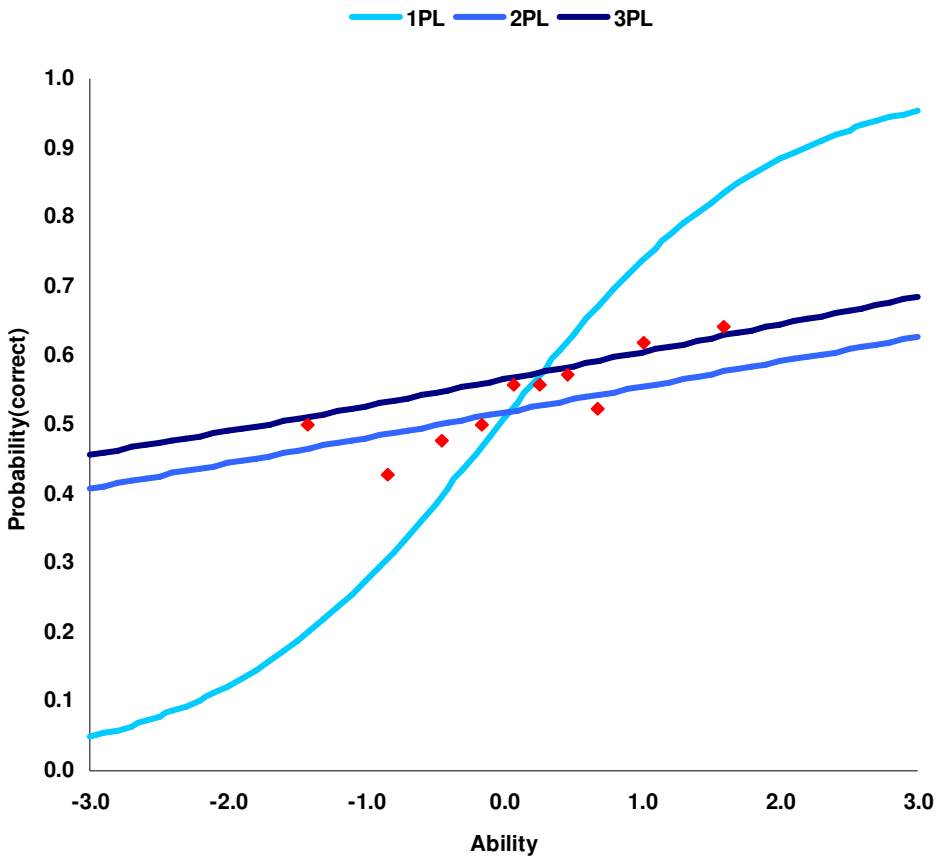


Figure 5: Item Characteristic Curves for 1-, 2-, and 3-PL;
1-PL is a poor fit for this item of low discrimation

Figure 6 shows the advantage of the 3-PL over lower-dimensional IRT models. By allowing non-zero lower asymptote, the tendency of low-ability students to respond correctly is captured. The 3-PL is also considered a good fit statistically ($Q_1 = 11.9 < 14.1$). By CTT, this is considered a relatively poor item, with only 29% correct and a discrimination index of 0.25 (rank 28 of 38). For the 3-PL, this item has a quite strong discrimination (rank 5 of 38), inferring its utility at differentiating between students with standardized ability estimates above 1.0. (The relatively high percent correct of the lowest ability bin appears anomalous when viewed against a more consistent result for other low-ability examinees.)

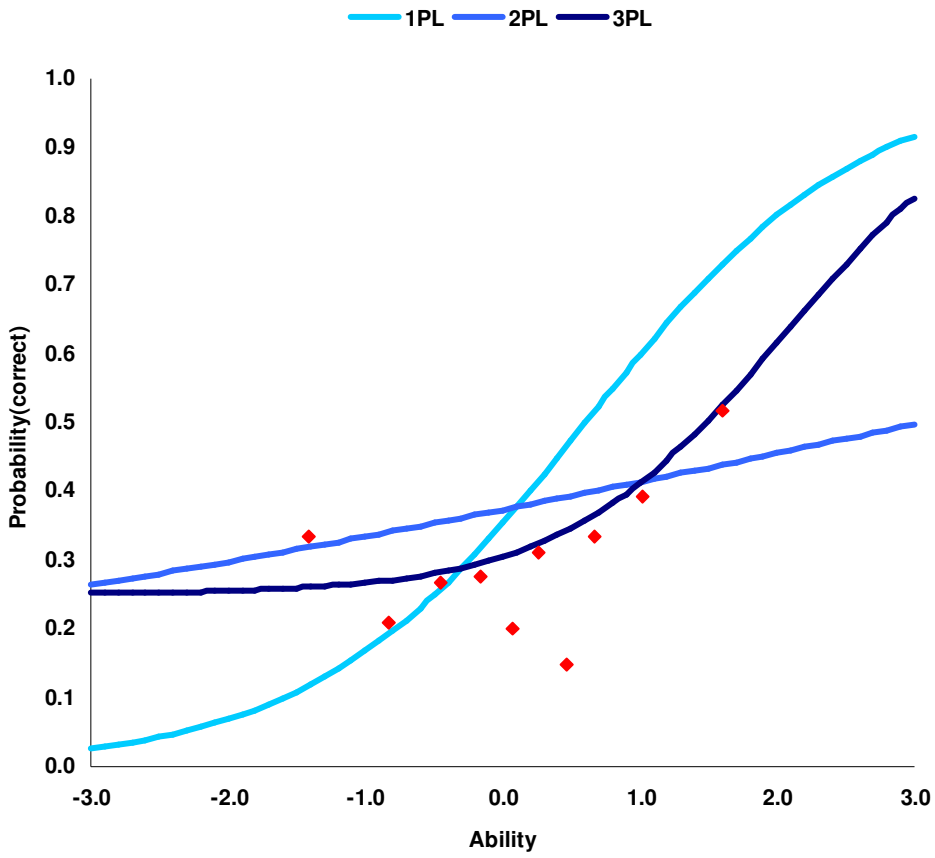


Figure 6: Item Characteristic Curves for 1-, 2-, and 3-PL;
3-PL captures higher discrimination of high-ability subjects

Figure 7 demonstrates item information, the analogue to reliability, for the 3-PL in Figure 6. The item is maximally informative in the area of steepest slope, the region where the item discriminates between examinees of adjacent abilities. At the extremes, however, the item is relatively uninformative, where examinees are essentially guessing (low ability) or nearly assured of a correct response (high ability).

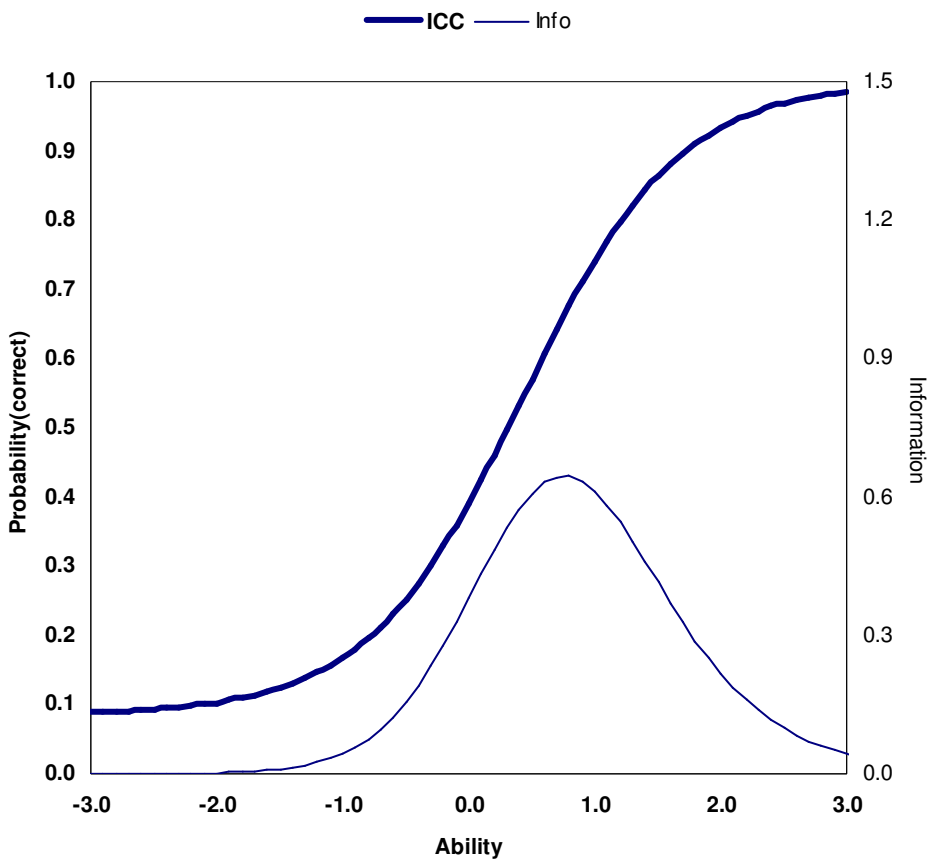


Figure 7: Information function and Item response curve for one item

Figure 8 shows the information function for the full test, calculated by summing the item information functional values at each ability level. The maximum occurs at a value of approximately $\theta = +1.0$, meaning that the SCI is best tailored to examinees with ability levels somewhat above average. Addition of more easy items to the item pool could more reliably assess lower-level individuals. This is an important consideration for the SCI, because statistics is found at many points in curricula due to its ubiquity in many disciplines.

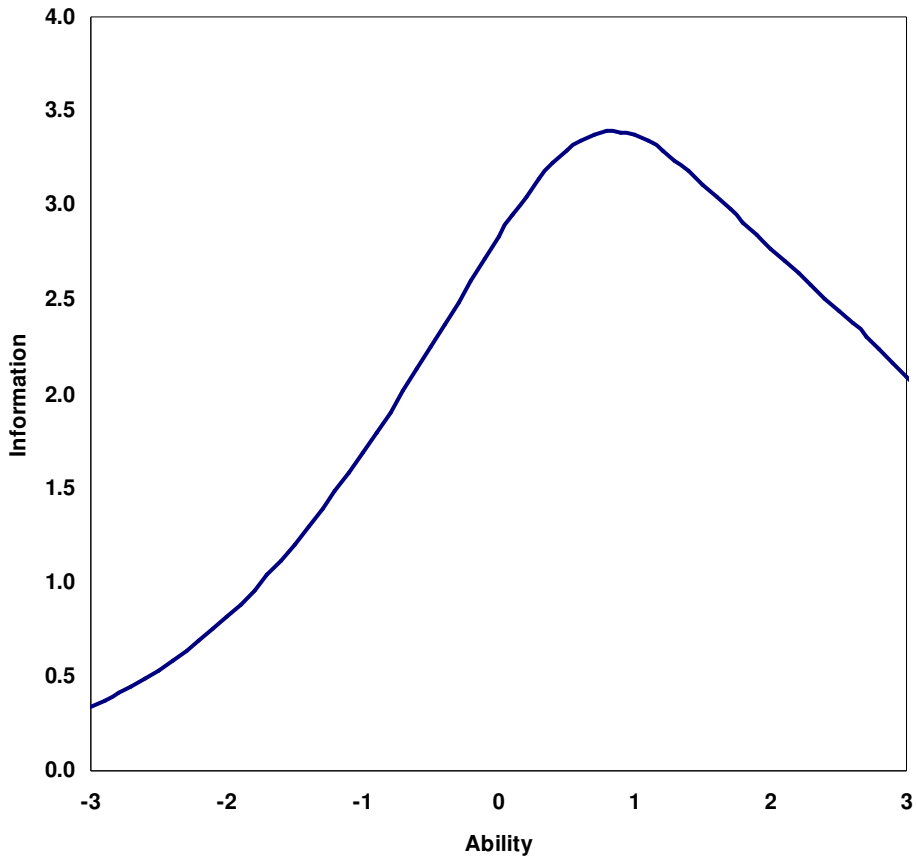


Figure 8: Information function for full test

Assumptions

Unidimensionality is typically assessed via principal components analysis. Figure 9 below shows the scree plot (eigenvalues vs. extracted factor number) for this dataset. Though the first factor explains a relatively small proportion of variance (12.3%), it is nearly three times larger than the second factor (4.3%) and there is no apparent cut-off between adjacent higher-order factors.

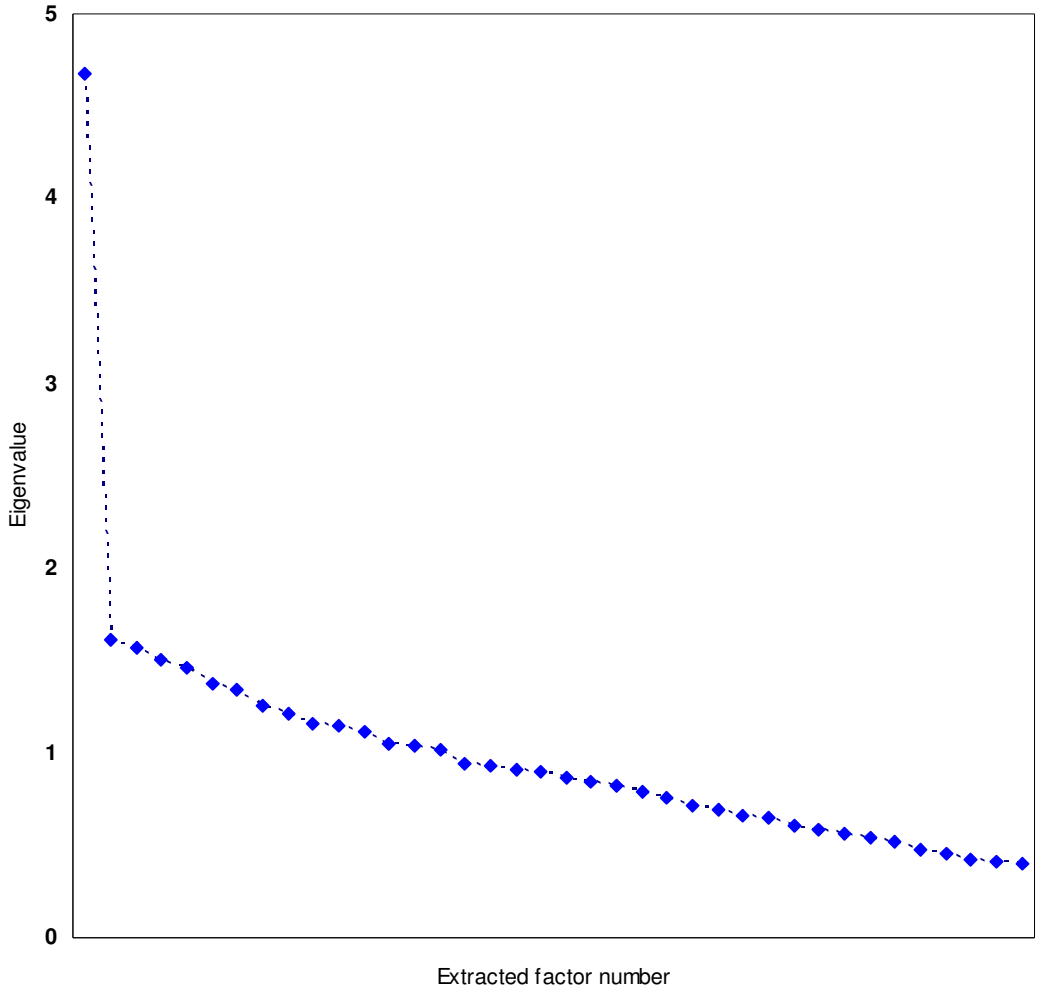


Figure 9: Scree plot used to assess unidimensionality assumption

Local independence is less straight-forward to assess. Evidence for unidimensionality may be considered sufficient to retain the local independence assumption[5]. In a finer scrutiny of the factor analysis (as in Figure 9), coupled with a structural equation model, it was shown that only a handful of highly similar items are correlated at a level strong enough to be considered non-chance[15]. The test is also designed such that information from one item is not required to answer any other item, which favors local independence; items are only similar to the extent that some concepts are assessed by multiple items.

At a gross level, the IRT models are considered poor fits. For example, the 3-PL is judged as non-fitting on 31 of 38 items. At a finer level, the misfit most often arises from a large residual in one or two (out of ten) of the ability bins. Availability of a larger sample or use of fewer bins would likely wash out some of this misfit (recall that each bin contains around 42 examinees).

Extensions of IRT

Item Response Theory can be extended beyond dichotomous item formats. Figure 10 shows the item response curves for a hypothetical three-category Likert-style item. Subjects of low ability are more likely to endorse the "Low" category (dark blue, e.g., 0.84 probability at θ = -2.0). Middle-ability subjects are relatively evenly split between the three options, while high-ability subjects display a pattern similar to that encountered for dichotomous items. At each θ value, the predicted values sum to 1. In fact, a dichotomous response is a simplification, with only two response categories (*correct* symbolized as *P*, and *incorrect* symbolized *Q*, where $P + Q = 1$).

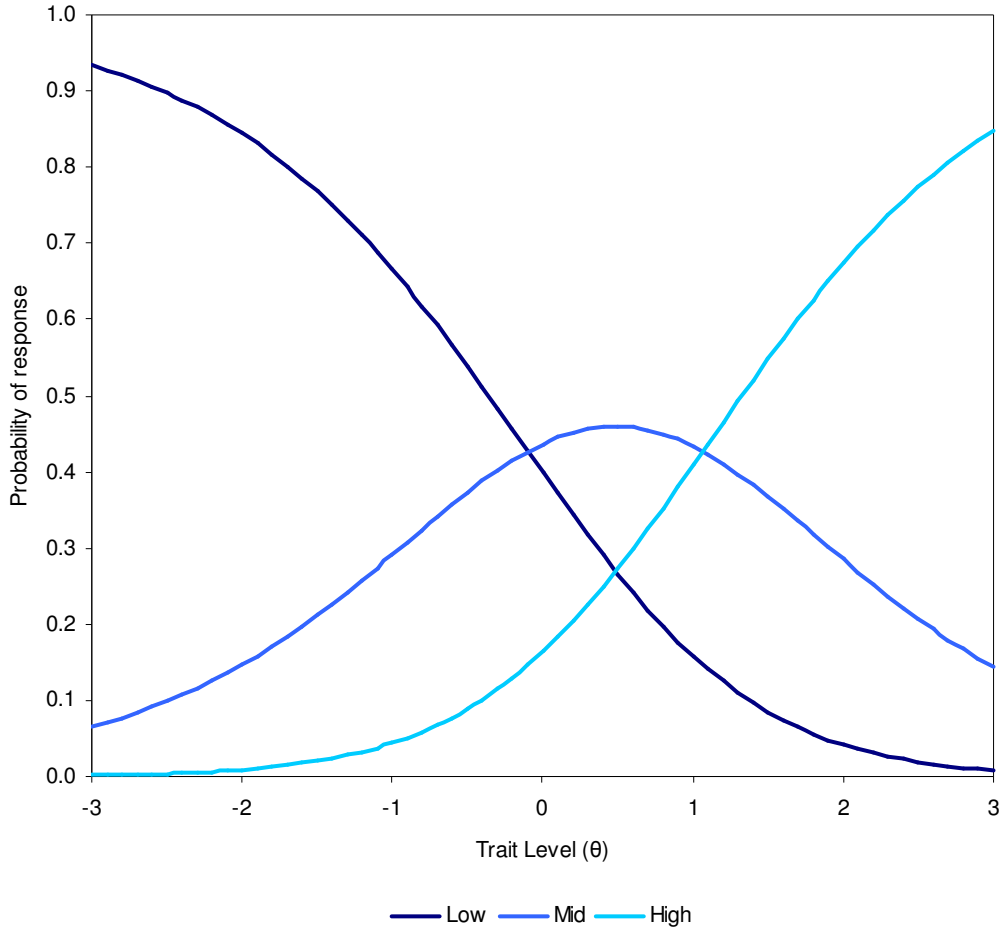


Figure 10: Item Response Curves for Likert-type item; three categories, Low-Mid-High

Along similar lines, the *nominal response model* is used when the categories are the individual response choices for a multiple choice test. In Figure 10, this could be the case for choices A, B, and C rather than options low, middle, and high. In the nominal case, the correct response will be identical to the dichotomous item characteristic curve, while the incorrect options will be parsed rather than collapsed into the category *zero*. This type of analysis was conducted with a different dataset for the SCI[16]. A non-IRT version, where total test score serves as ability, was recently

published for the Force Concept Inventory[17], although this method relies on prohibitively large sample sizes (4500 students in this case) which are impractical for most researchers.

Item response theory also incorporates procedures for detection of item bias, referred to as Differential Item Functioning (DIF)[18], providing a statistical basis for the comparison of item response curves from different groups (e.g., male vs. female). This is advantageous over classical test theory, which could at best allow comparison of percent correct and discrimination separately. Simultaneous comparison of all estimated parameters across the full ability range may reveal additional subtleties. Similar to fit assessment, a non-rigorous DIF can be conducted by inspecting item response curves from different groups.

DIF implicitly emphasizes the principle of *parameter invariance*, which states that item parameters are constant regardless of not only variables such as gender but also across ability levels. Combining maximum likelihood estimation (i.e., missing data is allowed) and parameter invariance, IRT is used in adaptive testing, wherein an examinee answers items selected to efficiently identify his ability level, such as on the GRE. With item parameters previously calibrated, ability can be re-estimated after each item, with the subsequently presented item one that matches the present ability estimate as to provide maximal information.

Conclusion

This paper described the background of Item Response Theory (IRT) with the understanding that it is a method not yet commonly used by engineering educators. The application and interpretation of IRT was highlighted for the Statistics Concept Inventory (SCI), using the simplest models for dichotomously scored items. Assumptions of IRT were evaluated for the SCI dataset and shown to approximately hold; model-fit was imperfect but not egregious. Finally, more advanced IRT methods were described, which have application to Likert-scale instruments and detection of item bias. As the field of engineering education matures, so must its tools. The theoretical advantages of IRT over Classical Test Theory allow a more robust yet subtle analysis, thus heeding the calls for rigorous educational research.

Appendix

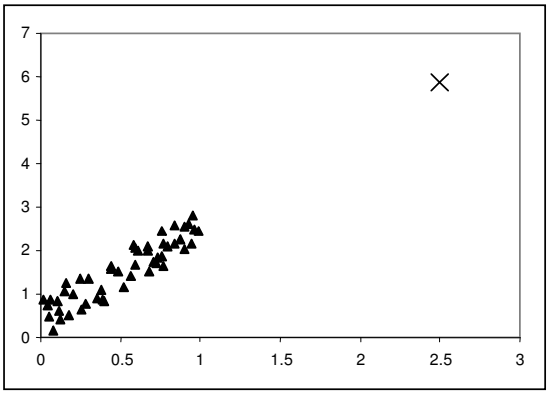
The items analyzed in the Results section are shown below.

First item (Figure 4)

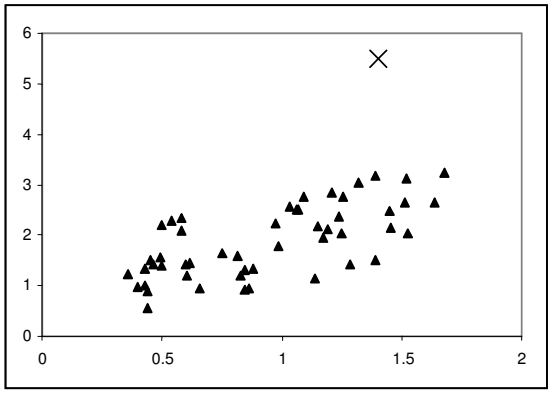
> For the past 100 years, the average high temperature on October 1 is 78° with a standard deviation of 5°. What is the probability that the high temperature on October 1 of next year will be between 73° and 83°?
>
> a) 0.68 (Correct)
> b) 0.95
> c) 0.997
> d) 1.00

Second item (Figure 5)

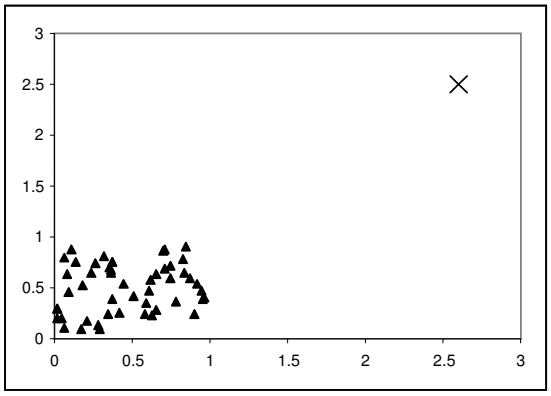
Consider the correlation coefficients of the scatter plots below. If the data point that is marked by an × is *removed*, which of the following statements would be true?



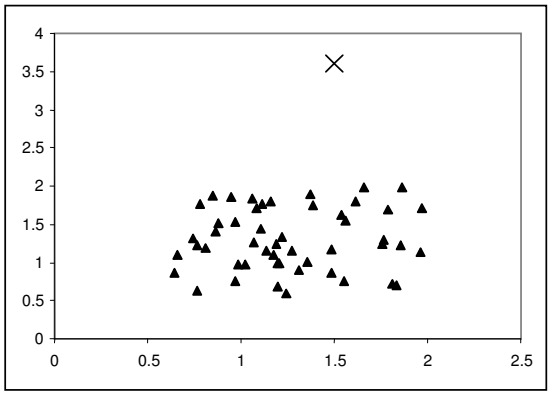
I



II



III



IV

a) correlation of (I) decreases, correlation of (II) stays the same
b) correlation of (III) increases, correlation of (IV) increases
c) correlation of (I) stays the same, correlation of (III) decreases (Correct)
d) correlation of (II) increases, correlation of (III) increases

Third item (Figure 6)

A researcher performs a t-test to test the following hypotheses:

$$H_0 : \mu \leq \mu_0$$
$$H_1 : \mu > \mu_0$$

He rejects the null hypothesis and reports a p-value of 0.10. Which of the following must be correct?

a) The test statistic fell within the rejection region at the $\alpha = 0.05$ significance level
b) The power of the test statistic used was 90%
c) Assuming H_0 is true, there is a 10% possibility that the observed value is due to chance (Correct)
d) The probability that the null hypothesis is not true is 0.10
e) The probability that the null hypothesis is actually true 0.9

References

[1] Steering Committee of the National Engineering Education Research Colloquies. 2006. The Research Agenda for the New Discipline of Engineering Education. *Journal of Engineering Education*. 95 (4 / October): 259-261.

[2] Fortenberry, N.L. 2006. An Extensive Agenda for Engineering Education Research. *Journal of Engineering Education*. 95 (1): 3-5.

[3] Streveler, R.A., and K.A. Smith. 2006. Conducting Rigorous Research in Engineering Education. *Journal of Engineering Education*. 95 (2, April): 103-105.

[4] Embretson, S.E., and S.P. Riese. 2000. *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates: Mahwah, NJ.

[5] Hulin, C.J., F. Drasgow, and C.K. Parsons. 1983. *Item Response Theory: Application to Psychological Measurement*. Dow-Jones Irwin: Homewood, IL.

[6] Partchev, I. 2004. *A visual guide to item response theory*. Available at http://www.metheval.uni-jena.de/irt/VisualIRT.pdf; verified December 11, 2006.

[7] van der Linden, W.J., and R.K. Hambleton, eds. 1997. *Handbook of Modern Item Response Theory*. Springer: New York.

[8] Yen, W.M. 1981. Using Simulation Results to Choose a Latent Trait Model. *Applied Psychological Measurement*. 5 (2 / Spring): 245-262.

[9] Stone, A., K. Allen, T.R. Rhoads, T.J. Murphy, R.L. Shehab, and C. Saha. 2003. The Statistics Concept Inventory: A Pilot Study. *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conferenc*e. Session T3D-6.

[10] Allen, K., A. Stone, T.R. Rhoads, and T.J. Murphy. 2004. The Statistics Concept Inventory: Developing a Valid and Reliable Instrument. *Proceedings of the 2004 American Society for Engineering Education Annual Conference and Exposition*. Session 3230.

[11] Allen, K., T.R. Rhoads, and R. Terry. 2006. Misconception or Misunderstanding? Assessing Student Confidence of Introductory Statistics Concepts. *Proceedings of the 36th ASEE/IEEE Frontiers in Education Conference*. Session S2E.

[12] Evans, D.L., G.L. Gray, S. Krause, J. Martin, C. Midkiff, B.M. Notaros, M. Pavelich, D. Rancour, T.R. Rhoads, P. Steif, R.A. Streveler, and K. Wage. 2003. Progress On Concept Inventory Assessment Tools. *Proceedings of the 33rd ASEE/IEEE Frontiers in Education Conference*. Session T4G-8.

[13] Hanson, B.A. 2002. IRT Command Language. Available from http://www.b-a-h.com/software/irt/icl/index.html; verified December 19, 2006.

[14] Rudner, L.M. 2005. PARAM-3PL Calibration Software for the 3 Parameter Logistic IRT Model (freeware). Available: http://edres.org/irt/param; verified December 18, 2006.

[15] Allen, K. 2006. *The Statistics Concept Inventory: Development and Analysis of a Cognitive Assessment Instrument in Statistics*. Dissertation. University of Oklahoma.

[16] Stone, A. 2006. *A Psychometric Analysis of the Statistics Concept Inventory*. Dissertation. University of Oklahoma.

[17] Norris, G.A., L. Branum-Martin, N. Harshman, S.D. Baker, E. Mazur, S. Dutta, T. Mzoughi, V. McCauley. 2006. Testing the test: Item response curves and test quality. *American Journal of Physics*. 74 (5): 449-453.

[18] Holland, P.W., and H. Wainer, eds.. 1993. *Differential Item Functioning*. Lawrence Erlbaum Associates: Mahwah, NJ.