The Evolution of Validity and Modern Psychometrics: Do We Need to Revisit Item Validity?

Charles Secolsky – Mississippi Department of Education

William Buchanan – Minnesota Department of Education

Walt Drane – Mississippi Department of Education

Early validity studies were concerned with the discriminative capacity of items. Were examinees with high total test scores answering items correctly and were examinees with low total test scores answering items incorrectly (see, for example, Swineford, 1936)? The statistics used for determining whether an item was valid were the biserial and point-biserial correlations. The school of thought headed up by Lindquist questioned item validity studied this way since a subjective, human ingredient was needed to offset the possibility that ambiguous or structurally deficient items could be missed more by higher ability examinees. It is interesting to note that this view of validity lasted until the works of Cronbach (1971) and Messick (1989). Implied in this earlier view of validity was that it was the responses to items that are validated. When items are validated in this way, unless one is interested in ratings for establishing content validity, of course responses to items are central to item validity claims. Nevertheless, such a view of validity lasted a number of decades through an array of different conceptions regarding criterion-related validity and construct validity, until Cronbach (1971) raised the point from a more psychological perspective that it was the interpretation of test scores that is validated rather than the test per se. This approach to understanding the concept of validity has evolved into Kane's (2013) interpretive argument. Item invalidity and ambiguity were no longer the emphasis in validity studies since the focus of validity was now on the interpretation.

Consonant with developments in the conception of validity, radical shifts took place in the statistical theories on which test scores were analyzed. Lord and

Novick (1968) including the contribution by Birnbaum (1968) and Lord (1980) introduced item response theory (IRT) as an alternative to classical test theory with its idea of a true score (T). Yet, even in discussions of validity in more recent works, the classical-IRT distinction is left out of the argument. In this paper, we compare item validities in classical test theory and IRT. How are they the same and how do they differ? Ultimately, is there a need to revisit item validity for modern psychometrics? Linn's (1989) paper on how IRT has increased the validity of achievement test scores makes the point that IRT in some respects has and in other respects has not increased the validity of achievement testing. Allen et al. (1987) and Way et al (1989) examine effects of altering content characteristics of items on IRT parameter estimates. However, a direct comparison of classical and IRT item statistics is what was proposed in this study. Hambleton and Jones (1993) compared classical test theory and IRT for purposes test development and state "The test characteristic function connects ability scores in item response theory to true scores in classical test theory because an examinee's expected test score at a given ability level is by definition the examinee's true score on that set of test items" (p. 256). However, they do not delve deeply into the topic of validity. Magno (2009) compared the robustness of classical test theory proportion correct scores with the difficulty parameter "b". This paper compares discrimination and information for these two theories.

Method:

In order to check on the comparability of classical test theory and IRT with respect to old and new notions of validity, two sets of comparative analyses were performed. First, using the discriminative capacity of items as the older definition of validity, the "a" parameters for the mathematics section of the third grade assessment of the Mississippi Competency Test were correlated with the point-biserial correlations for the same 55 items of this test. In addition, items that would have been flagged from classical analysis by virtue of having low or negative point-biserials were compared to items that would have been flagged from an IRT perspective. Typically, at many testing firms such as Educational Testing Service, items that are flagged go through analysis by test development staff to determine if low item-test correlations are justified. Since the same type of perception-based analysis from test development staff is likely for the identification of poorly functioning items, or items that do not fit the IRT model well, a comparison of

items that could have been flagged for each of these two theories would be informative.

For a more modern approach to validation, namely whether interpretations of test results are valid, the authors put forth arguments for whether claims about test scores are warranted (see Toulmin, 1958). Hypothetically, if test data are the same, interpretations should be the same and validity arguments can be expressed in the same manner. However, if classical test theory and IRT produce even somewhat different score distributions, interpretations of score results and subsequent arguments could be very different. In fact, there may even be some sort of practical dependency between item and test results for these two different conceptions of validity. From the more modern psychometric conception of validity, Cronbach (1971) stated that even the presence of a few ambiguous items would not necessarily affect a test's validity, yet with high stakes testing becoming more prominent in society, can we even afford to keep any ambiguous items in our scoring of tests?

-

Preliminary Analysis:

Upon producing zero-order correlations among point-biserials, a parameters, and b parameters, it was found that the point-biserials and a parameters were correlated only 0.168 (p=0.271). This suggests that the process for identifying aberrant or malfunctioning items may be considerably different for classical test theory and IRT. And as a result potentially different items may be flagged using the two respective models for scoring.

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Pbis	45	0.41908	0.08254	18.85873	0.15410	0.54872
Aparm	45	0.45413	0.27875	20.43600	0.00200	0.94400
Bparm	45	0.66944	0.60749	30.12500	0.00700	2.44100

Pearson Correlation Coefficients, N = 45
Prob > |r| under H0: Rho=0

	pbis	aparm	Bparm
pbis	1.00l000	0.16759	-0.39968
		0.2712	0.0065
aparm	0.16759	1.00000	-0.24972
	0.2712		0.0980
bparm	-0.39968	-0.24972	1.00000
	0.0065	0.0980	

Of the 55 items on the MCT2 third grade mathematics assessment, 10 items were not calibrated because there was no variance for these items and a limitation of the study was that the "c" parameter was held fixed in the preliminary analysis for all 45 calibrated items at c= 0.159. Given that some of the items may have non-negligibly surpassed this value of c, and that a portion of what is attributable to this "c" parameter in the present context totaling n=9,320 examinees using three different test forms of the exam, the difference between pbis and the "a" parameter may be explainable (see Secolsky, Alqarni, & Rose, 2014).

Another way of comparing item analysis results for classical test theory and IRT is to determine which items were flagged using these two different models. With classical test theory, the approach is to examine those items with low point-biserials and with IRT the approach is to examine those items that do not add item information to the test information. For the MCT2 third grade math test, the ten items with no item variance have point biserials that were not computable and undefined item information functions that contributed in an undefined way to the test information function.

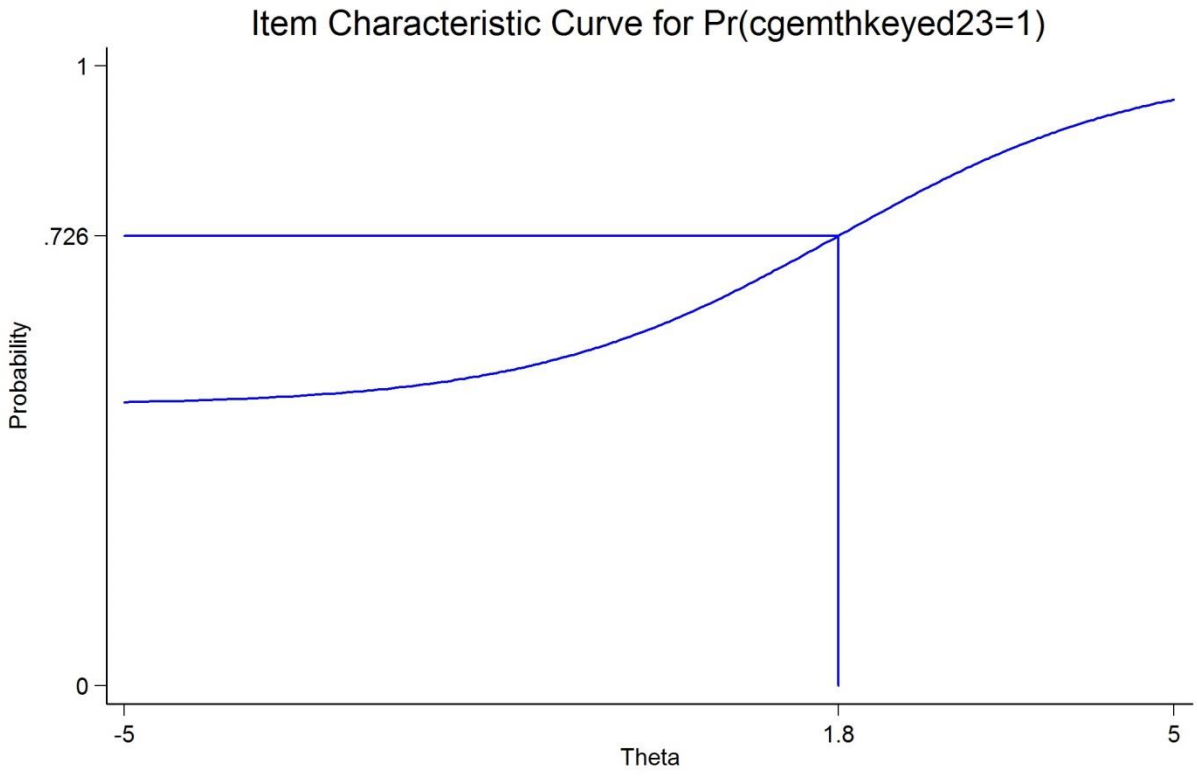
Results:

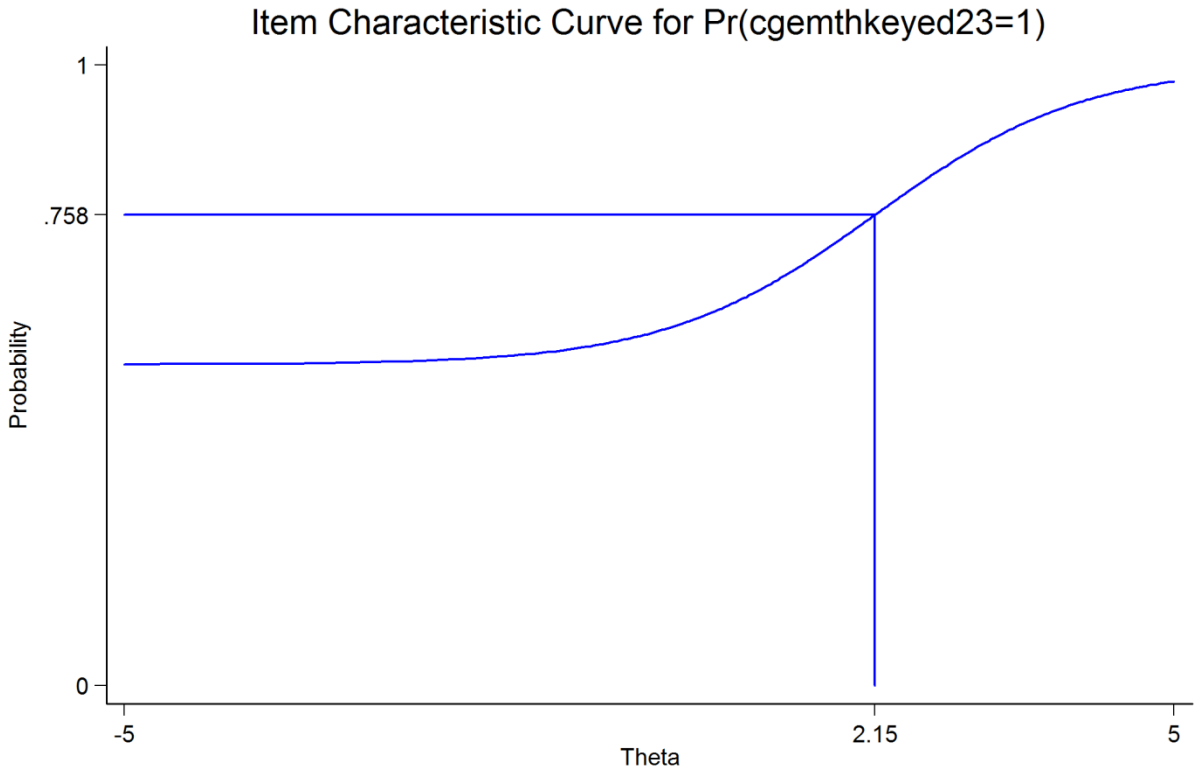
The items with the lowest four point biserials were Item 2 (0.2817), Item 3 (0.2897), item 23 (0.1789), and item 24 (0.2560). It is comforting to see that item 23 has the lowest IRT "a" parameter of any of the 45 variance-laden items (a=0.2612), while a different picture emerges for the "a" parameters for items 2

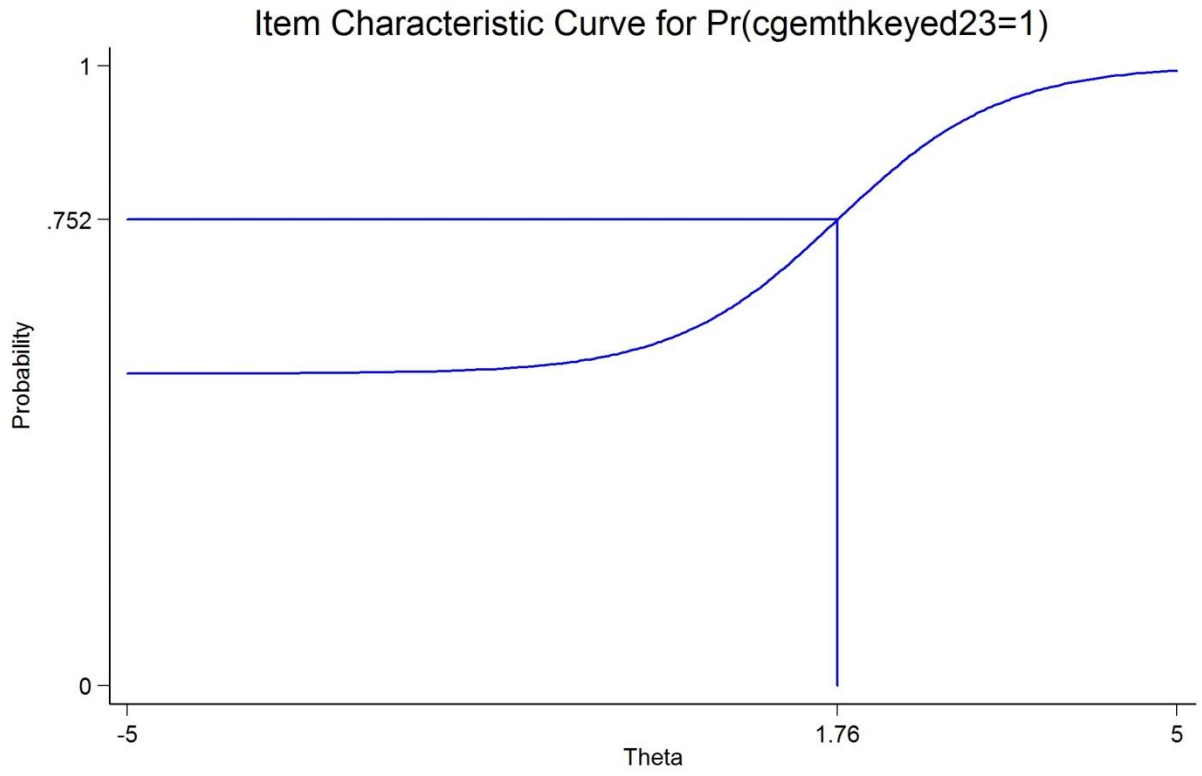
and 3 (1.498 and 1.456, respectively). Also, for Item 24 the ‘a” parameter was low (0.7313). Item 27 had a low “a” parameter (0.7927) and a low point biserial (0.3454). The next two lowest “a” parameters were for items 5 an 6 (0.8442 and 0.8206).And accordingly, the point biserials for these two items were 0.3509 and 0.3374.

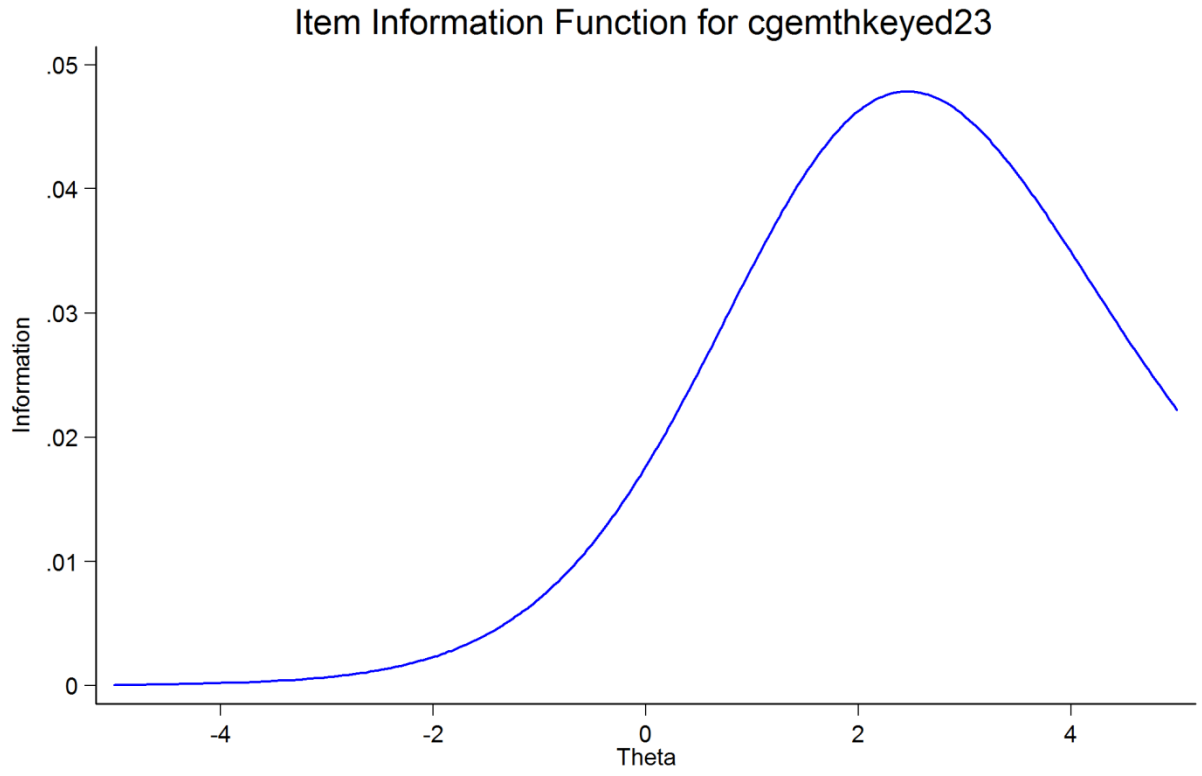
In summary then, when items are not very easy, there seems to be a fairly close correspondence between point biserials and “a” parameters. In contrast when the item difficulty values (“b” parameters) are highly negative on the theta axis (horizontal axis), IRT discriminative capacity is higher but classical test theory discriminative capacity is lower.

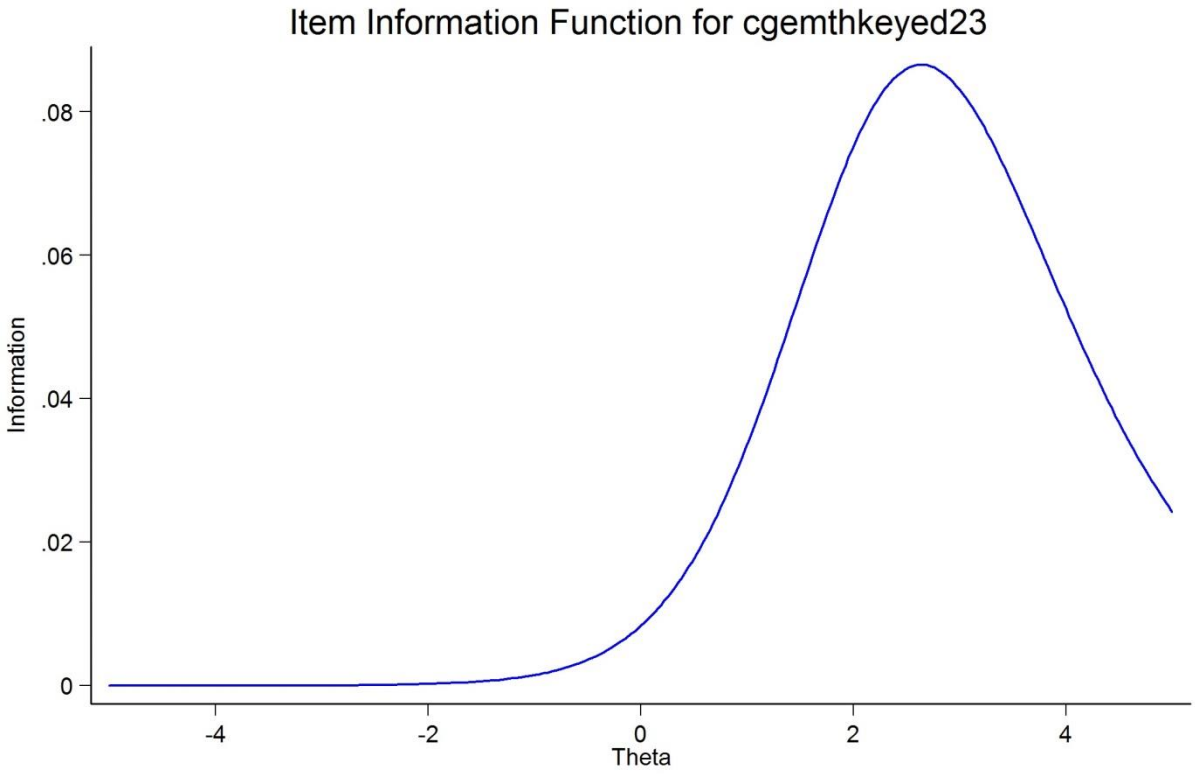
To illustrate the above points, ICCs and item information functions (IIFs) graphs are presented for item 23 and Item 2 for three forms of the MCT2 3rd grade math. Item 23 had the lowest point-biserial and the lowest “a’ parameter. Item 2 had a low point-biserial and a higher than average “a” parameter. Item 23 ICCs are similar on three forms. The three ICCs have a very high value of the c-parameter. The item information shows nearly 0 information in the lower tail which is related to the high values of the c-parameters. The ICCs and IIFs for Item 2, which also has a low point-biserial shows quite a different picture. The c-parameter is close to 0 and there is discrimination at the lower end of the ability scale. In addition, there is considerable information at the low end of the ability scale.

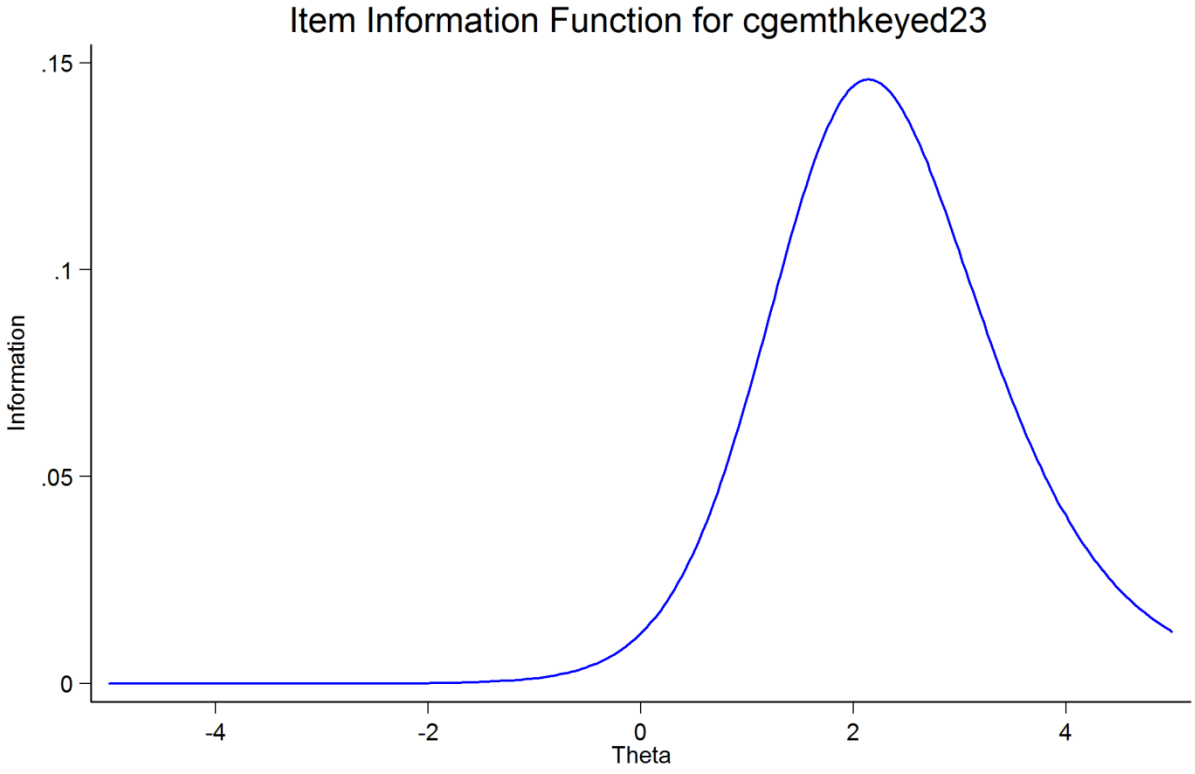

Item Characteristic Curve for Pr(cgemthkeyed23=1)
1
.726
0
Probability
-5
1.8
5
Theta

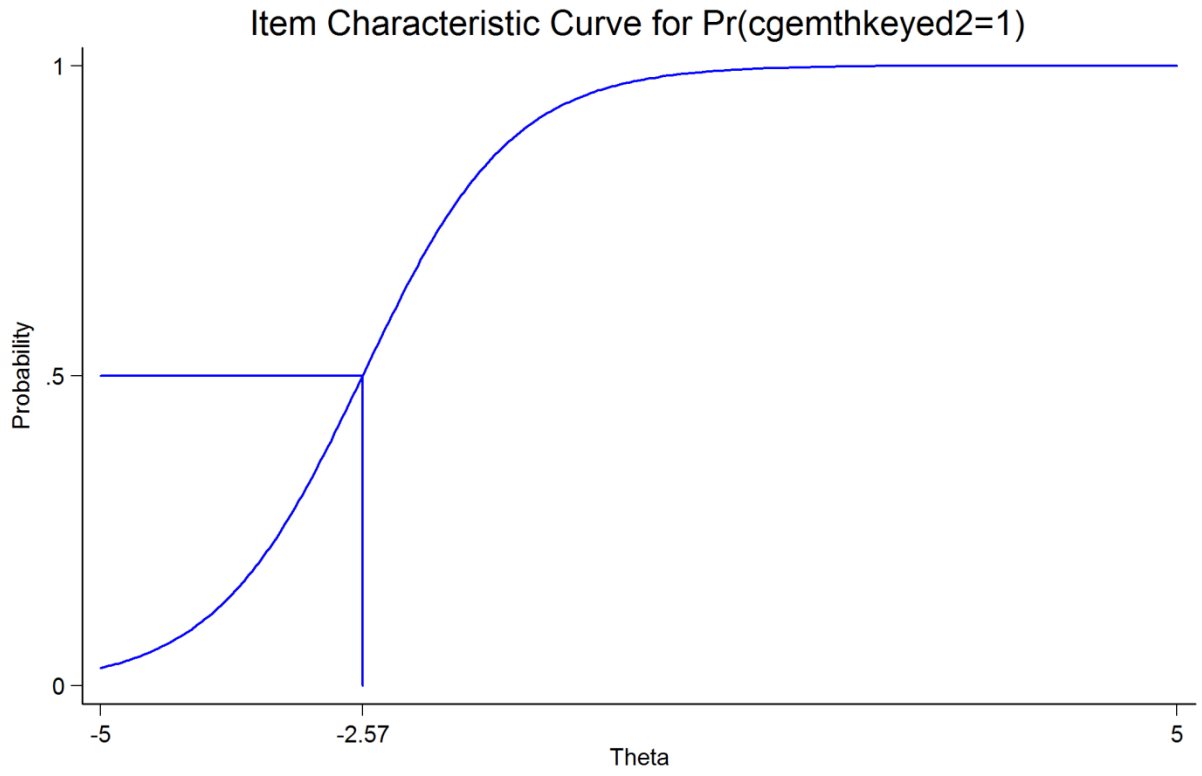

Item Characteristic Curve for Pr(cgemthkeyed23=1)
1
.758
0
Probability
-5
2.15
5
Theta

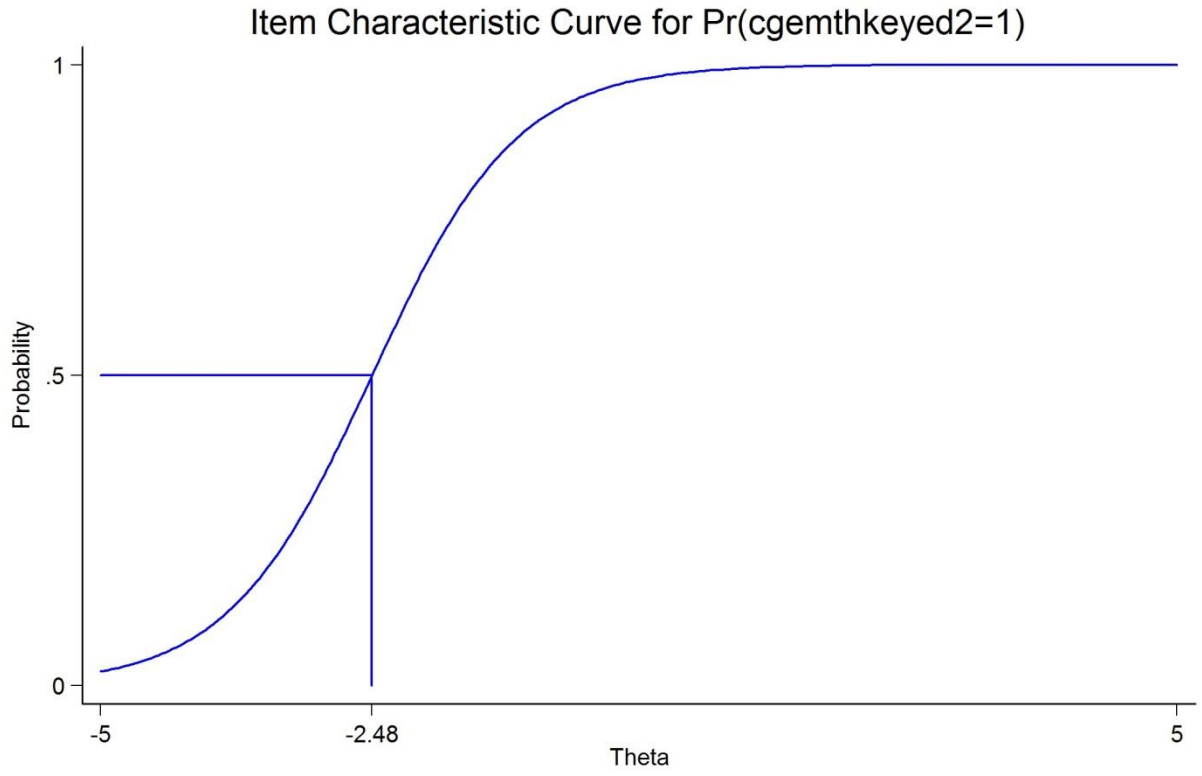

Item Characteristic Curve for Pr(cgemthkeyed23=1)
1
.752
0
Probability
-5
1.76
5
Theta

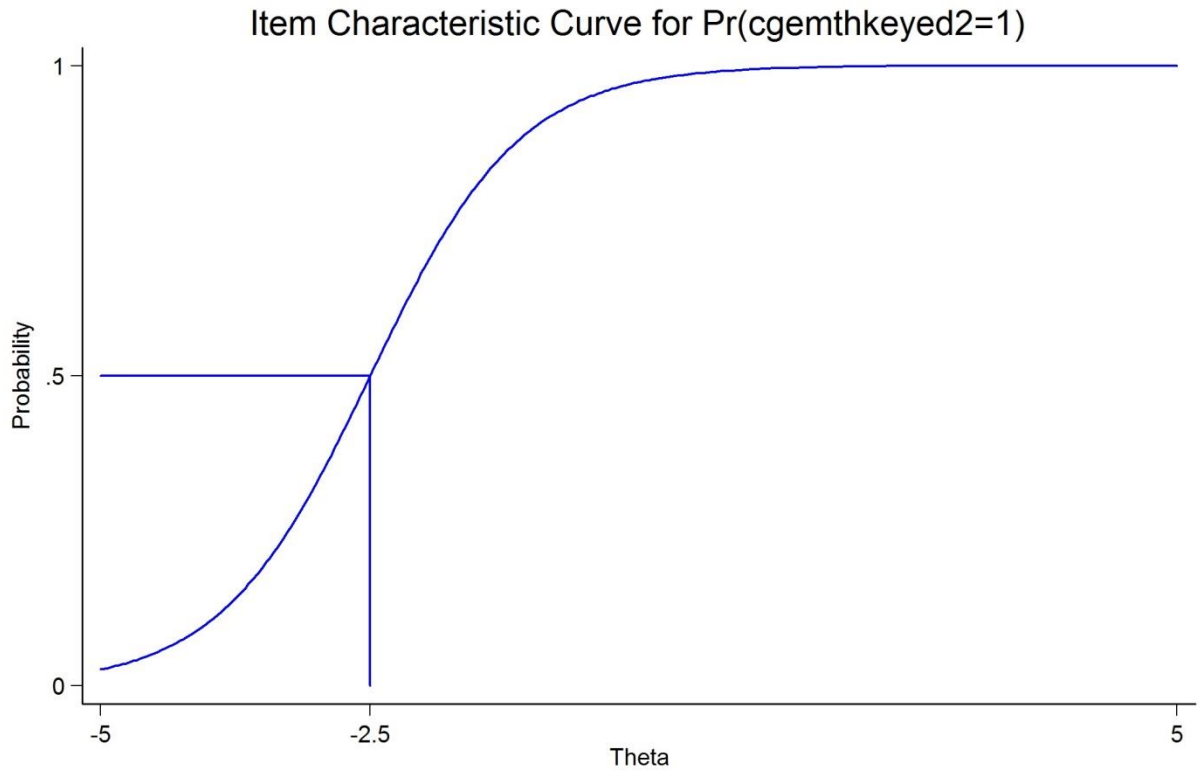

Item Information Function for cgemthkeyed23
.05
.04
.03
.02
.01
0
Information
-4
-2
0
2
4
Theta

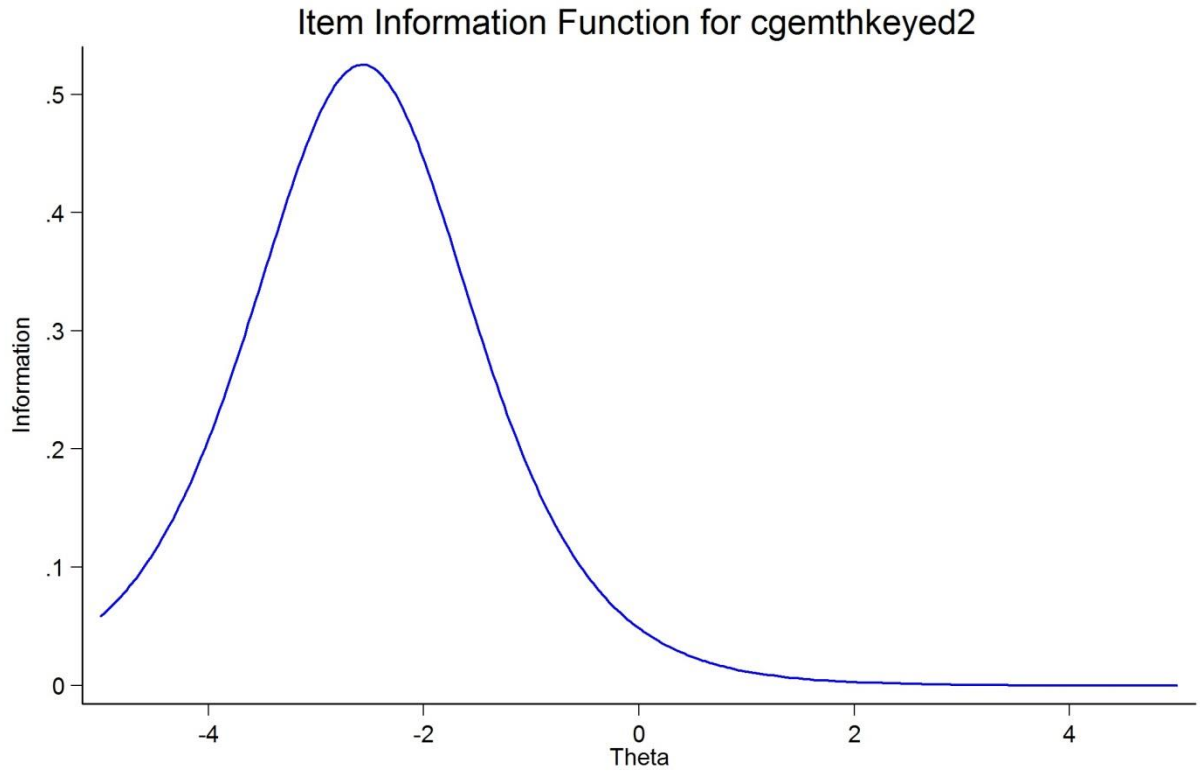

Item Information Function for cgemthkeyed23
Information
0
.02
.04
.06
.08
-4
-2
0
2
4
Theta

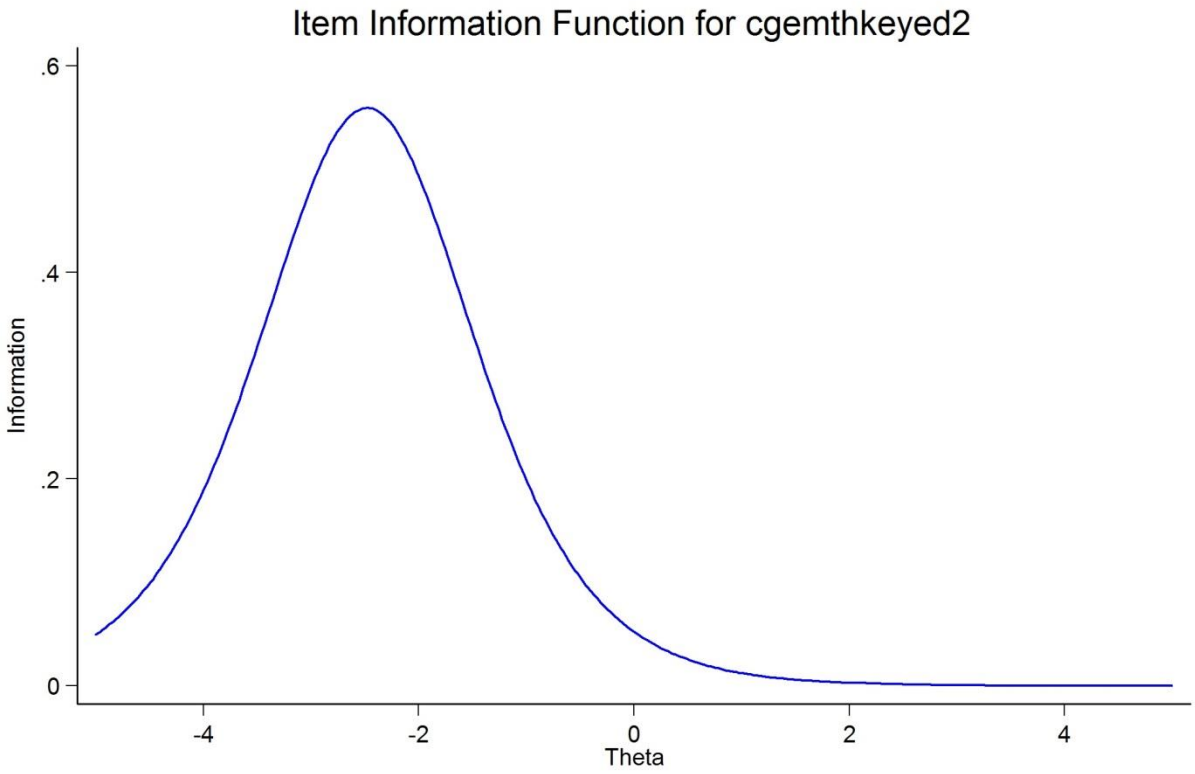

Item Information Function for cgemthkeyed23
Information
0
.05
.1
.15
-4
-2
0
2
4
Theta

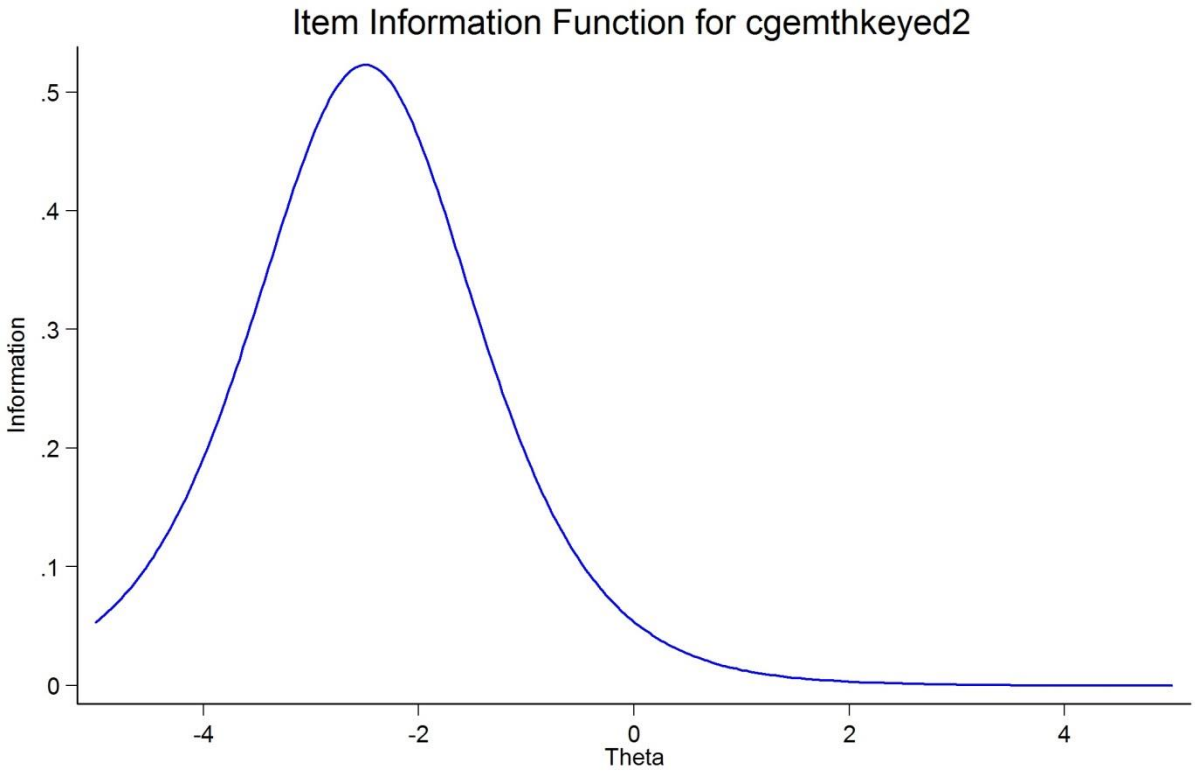

Item Characteristic Curve for Pr(cgemthkeyed2=1)
Probability
1
.5
0
-5
-2.57
5
Theta


Item Characteristic Curve for Pr(cgemthkeyed2=1)
Probability
1
.5
0
-5
-2.48
5
Theta


Item Characteristic Curve for Pr(cgemthkeyed2=1)
Probability
1
.5
0
-5
-2.5
5
Theta


Item Information Function for cgemthkeyed2
Information
.5
.4
.3
.2
.1
0
-4
-2
0
2
4
Theta


Item Information Function for cgemthkeyed2
.6
.4
.2
0
Information
-4
-2
0
2
4
Theta


Item Information Function for cgemthkeyed2
.5
.4
.3
.2
.1
0
Information
-4
-2
0
2
4
Theta

Now to consider the modern psychometric approach to validation, one is interested in interpretations of test results and not in item statistics, IRT or otherwise. The approach to argument is the same, yet the models differ and, as we found out in the results, different interpretations of test data are possible. It depends what the inferences we make from test results.

Discussion:

Different interpretations of test results can occur using classical and IRT models and therefore so would validation of results. In this paper, we found that classical test theory discriminative capacity and IRT can produce profound differences if the b and c parameters are taken into account. And different items would result in non-scoring. Furthermore, consequences for test results are different because different examinees are likely to surpass benchmarks given a particular standard setting method. With respect to validity, Cronbach (1971) was concerned with interpretations of test scores in classical test theory. With the greater precision of IRT, Cronbach may have reconsidered his statement that a few ambiguous items did not necessarily have to affect validity.

Conclusion:

IRT opens the door for new types of item validity conceptions. With IRT, more than with classical test theory, validity hinges on item statistics.

References:

Allen, N. L., Ansley, T. N. & Forsyth, R. A. (1987). The effect of deleting content-related items on IRT ability estimates. *Educational and Psychological Measurement,* 47, 1141 – 1152.

Birnbaum, A. (1968) Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statisticfal theories of mental test scores (*pp. 397 – 470). Reading, MA: Addison Wesley.

Cronbach, L. J. (1971). Test Validation. In R. L. Thorndike (Ed)., *Educational Measurement (2nd Ed.),* Washington, DC: American Council on Education.

Hambleton, R. K. & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. Amherst, MA: National Council on Measurement in Education.

Kane,M.K (2013). Validating the Interpretations and Uses of Test Scores. Journal of Educational Measurement., 50(1), pp 1 – 73.

Lindquist, E.F. (1936). The theory of test construction. In H. E. Hawkes, E.F. Lindquist, & C.R. Mann (Eds.), The construction and use of achievement examinations. Boston, MA: Houghton-Mifflin Company.

Linn, R. L. (1989) Has IRT increased validity of achievement test scores? Center for Research on Standards, Evaluation, and Student Testing : UCLA.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Lord, F. M. (1980). A*pplications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd ed., PP 13 – 103). New York, NY: American Council on Education and Macmillan.

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. International Journal of Educational and Psychological Assessment, 1 (1), 1 – 11.

Secolsky, C., Alqarni, A., & Rose, S. (2014). Approaches for incorporating theta estimation in generating information functions of observed score on true score or the tail of two test theories. Paper presented at the annual meeting of the Northeastern Educational Research Association. , Turnbull, CT.

Swineford, F.M. (1936). Validity of test items. Journal of Educational Psychology. 27(1), 68 – 78.

Toulmin, S. (1958). *The Uses of Argument*. Cambridge, MA: Cambridge University Press.

Way, W. D., Forsyth, R. A., & Ansley, T. N. (1989). IRT ability estimates from customized achievement tests without representative content sampling. *Applied Measurement in Education*, 2m 15 -35.