A Comparison of Traditional and IRT based Item Quality Criteria

Brian D. Bontempo, Ph.D.

Mountain Measurement, Inc.

Jerry Gorham, Ph.D.

Pearson VUE

April 7, 2006

A paper presented at the Annual Meeting of the National Council on Measurement in Education: San Francisco, CA

Executive Summary

This study compared the survival rates of new items based on traditional item statistics (point-biserial correlation coefficient) to the survival rates of those based on IRT item statistics (fit statistics). The assessment data used in this study came from two (2) different large-scale, paper-and-pencil licensure and certification programs. Each exam's item pool contained at least 1,500 new items which were administered to at least 400 first-time test takers. Comparable traditional and IRT based item selection criteria (e.g., pt bis >= 0 or infit mean square fit < 2.0) were developed based on the survival rates and the percentage of items that are classified identically using each set of type of item selection criteria. Lastly, conditions, such as the difficulty of the item, that encourage or discourage the use of one type of item selection criteria over the other were explored.

Introduction

Many large scale assessment programs have switched from a classical test theory paradigm to an IRT based paradigm in recent years. However most of these programs still use classical test theory statistics such as the point-biserial correlation coefficient to evaluate the statistical quality of newly developed items. This research aims to provide IRT evaluation statistics in the form of infit and outfit Mean Square fit statistics that mimic the more traditional stats. The goal of this study is to illustrate how item survival rates can be used to develop a conversion table that will make it easy for the practitioner to convert their present CTT item evaluation criteria to IRT based criteria.

Since it is well known that the point-biserial correlation coefficient and other similar classical test theory item statistics are less than ideal, it is NOT the goal of this study to develop IRT evaluation criteria that produce the same outcomes as the classical test theory criteria. Rather, it is the goal of this study to develop criteria that yield a similar item survival rate. This goal is important since many large-scale assessment programs base their item selection criteria on real world considerations such as the number or percentage of items that need to survive in order to construct the next version of the assessment.

Data

The data for this study came from two different large-scale, paper-an-pencil certification exams. The first exam (Exam A) contained 1,983 items that were assembled into 24 different overlapping test forms each containing 150 items. The other exam (Exam B) contained 1,632 items that were assembled into 13 different overlapping test forms each containing 180 items. Each form was administered to a sample of at least 400 representative first-time test-takers. Each of the exams was in the field during 2005.

The characteristics of the set of items from each exam are displayed in Tables 1 and 2. The exams were similar. Both exams had a large range in their item difficulty. Mins were below -6 and maxs were above 4. Exam A was slightly more difficulty challenging and Exam B had a larger variance in item difficulty. One difference between the exams is that the set of items of Exam B was more diverse with respect to item quality. The range of point-to-measure correlation coefficients, infit, and outfit were all greater than Exam A and as was the variance of outfit.

Table 1. Characteristics of Exam A items

Exam A	Min	Max	Mean	Std. Dev
Item Difficulty	-6.43	6.14	0.00	1.38
PtMe Correlation	-0.16	0.45	0.19	0.09
Infit Mean Square	0.89	1.16	0.99	0.04
Outfit Mean Square	0.52	1.58	0.97	0.10

Table 2. Characteristics of Exam B Items

Exam B	Min	Max	Mean	Std. Dev
Item Difficulty	-6.46	4.81	-0.04	1.64
PtMe Correlation	-0.24	0.52	0.17	0.10
Infit Mean Square	0.87	1.19	1.00	0.04
Outfit Mean Square	0.25	2.45	0.99	0.12

Methodology

For each of these exams, the item-level result (1/0 correct/incorrect result) was concurrently calibrated using the 1 PL dichotomous Rasch model using Winsteps. The point-to-measure correlation coefficient without inclusion was calculated for each item as well as the infit mean square and outfit mean square statistics. The calculation of each of these statistics was easily conducted by using the options available in Winsteps.

The viability of each item from a traditional classical test theory perspective was determined by running each item through a point-to-measure correlation coefficient item selection filter. Since licensure programs vary widely in the stringency of the filtration, three levels of filtration were applied:

- Low Item Performance Benchmark (Low Level of Filtration)
    - Point to Measure Correlation Coefficient > 0.0
- Medium Item Performance Benchmark (Medium Level of Filtration)
    - Point to Measure Correlation Coefficient > 0.05
- High Item Performance Benchmark (High Level of Filtration)
    - Point to Measure Correlation Coefficient > .1

Under each one of these levels of filtration, the survival rate of the items was calculated. Using these survival rates, a comparable set of IRT filtration criteria (based solely on the infit and outfit MNSQ values) were determined. Specifically, four (4) IRT filters were applied a.) heavier filtration on infit and less on outfit b.) heavier filtration on outfit and less on infit c.) a balanced approach to both infit and outfit d.) a compensatory approach where the sum of the infit and outfit mean square was used. After the comparable IRT filter values had been determined, the items that were filtered out were queried for further comparison.

Table 3 and 4 display the survival rates of exam for the five item quality criteria. Inspection of these tables reveals that it was not possible to determine the criteria so that the exact survival rate was obtained. This was due to the fact that the item quality statistics were rounded to the hundredths place, logical for these types of statistics, which created a discrete frequency distribution. In lieu of an exact match, the closest survival rate was utilized.

One interesting finding was that the unbalanced approach that favored heavier filtration on outfit was unable to yield a survival rate that approximated the balanced approach. The survival rate was always lower (the failure rate was always higher). Therefore, this technique was abandoned.

Table 3. Item Survival Rates for Exam A

	PtMe Correlation	Percent Flagged	Number Flagged
Low Filtration	>=0	1.7%	33
Med Filtration	>=.05	4.2%	84
High Filtration	>=.1	12.3%	243

	Sumfit		
Low Filtration	<=2.24		34
Med Filtration	<=2.17		80
High Filtration	<=2.11		250

Balanced	Infit	Outfit	
Low Filtration	<1.18	<1.18	32
Med Filtration	<1.12	<1.12	87
High Filtration	<1.07	<1.07	258

Heavy Infit Light Outfit	Infit	Outfit	
Low Filtration	<1.10	<1.23	36
Med Filtration	<1.08	<1.15	86
High Filtration	<1.05	<1.10	245

Light Infit Heavy Outfit	Infit	Outfit	
Low Filtration	<1.19	<1.17	36
Med Filtration	<1.13	<1.11	105
High Filtration	<1.08	<1.06	331

Table 4. Item Survival Rates for Exam B

	PtMe Correlation	Percent Flagged	Number Flagged
Low Filtration	>=0	3.4%	55
Med Filtration	>=.05	10.7%	175
High Filtration	>=.1	23.9%	390

	Sumfit		
Low Filtration	2.24		58
Med Filtration	2.14		185
High Filtration	2.07		395

Balanced	Infit	Outfit	
Low Filtration	<1.18	<1.18	56
Med Filtration	<1.10	<1.10	187
High Filtration	<1.05	<1.05	416

Heavy Infit Light Outfit	Infit	Outfit	
Low Filtration	<1.10	<1.23	52
Med Filtration	<1.07	<1.15	184
High Filtration	<1.03	<1.10	387

Light Infit Heavy Outfit	Infit	Outfit	
Low Filtration	<1.19	<1.17	70
Med Filtration	<1.11	<1.09	224
High Filtration	<1.06	<1.04	464

Analysis

For each set of filtered items, the percentage of items that were flagged by one method and not by the other were identified as well as the percentage of flagged items that were flagged by both methods. These statistics are displayed for Exam A in tables 5-10 and for For Exam B in Table 11-16.

For Exam A, the methods that were the most similar were the sum of infit and outfit mean square and the heavy infit or light outfit. Over 90% of the items flagged by one of these methods were also flagged by the other method as well. That is except when the filtration level was low, in which case only 75% of the items flagged by one method were also flagged by the other.

For Exam A, the most dissimilar methods were the infit or outfit filtration when one method was balanced and the other was heavy on infit and light on outfit. For the lowest level of filtration, only 56% of the items flagged by one method were also flagged by the other.

For Exam B, the most similar method was the sum of infit and outfit and the infit or outfit (balanced). For the highest level of filtration 94% of the items flagged by one method were also flagged by another. As with Exam A, over 90% of the items flagged by sumfit (sum of infit and outfit) were also flagged by the infit or outfit (balanced) rule. That is except when the filtration level was low, in which case 83% of the items flagged by one method were also flagged by another. All pairwise method comparisons yielded at least 60% of the items flagged by one method were also flagged by the other.

Table 5. Exam A Survival Rate Similarity Point-to-Measure and Sumfit

Point-to-Measure Correlation Coefficient	Infit or Outfit Balanced		
Low	Flagged	Surviving	
Flagged	66%	36%	
Surviving	34%		
Med	Flagged	Surviving	
Flagged	72%	25%	
Surviving	28%		
High	Flagged	Surviving	
Flagged	75%	20%	
Surviving	25%		

Table 6. Exam A Survival Rate Similarity between Point-to-Measure and Sumfit

Point-to-Measure Correlation Coefficient	Sum of Infit and Outfit		
Low	Flagged	Surviving	
Flagged	82%	15%	
Surviving	18%		
Med	Flagged	Surviving	
Flagged	82%	15%	
Surviving	18%		
High	Flagged	Surviving	
Flagged	83%	28%	
Surviving	17%		

Table 7. Exam A Survival Rate Similarity between Point-to-Measure and Heavy Infit or Light Outfit

Point-to-Measure Correlation Coefficient	Heavy Infit or Light Outift		
Low	Flagged	Surviving	
Flagged	69%	24%	
Surviving	31%		
Med	Flagged	Surviving	
Flagged	84%	14%	
Surviving	16%		
High	Flagged	Surviving	
Flagged	70%	30%	
Surviving	30%		

Table 8. Exam A Survival Rate Similarity between Sumfit and Infit or Outfit Balanced

Sum of Infit and Outfit Mean Square	Infit or Outfit Balanced		
Low	Flagged	Surviving	
Flagged	81%	24%	
Surviving	19%		
Med	Flagged	Surviving	
Flagged	80%	20%	
Surviving	20%		
High	Flagged	Surviving	
Flagged	82%	1%	
Surviving	18%		

Table 9. Exam A Survival Rate Similarity between Sumfit and Heavy Infit or Light Outfit

Sum of Infit and Outfit Mean Square	Heavy Infit or Light Outift		
Low	Flagged	Surviving	
Flagged	75%	21%	
Surviving	25%		
Med	Flagged	Surviving	
Flagged	91%	10%	
Surviving	9%		
High	Flagged	Surviving	
Flagged	94%	4%	
Surviving	6%		

Table 10. Exam A Survival Rate Similarity between Infit or Outfit Balanced and Heavy Infit or Light Outfit

Infit or Outfit Balanced	Heavy Infit or Light Outift		
Low	Flagged	Surviving	
Flagged	56%	38%	
Surviving	44%		
Med	Flagged	Surviving	
Flagged	73%	28%	
Surviving	27%		
High	Flagged	Surviving	
Flagged	81%	23%	
Surviving	19%		

Table 11. Exam B Survival Rate Similarity Point-to-Measure and Sumfit

Point-to-Measure Correlation Coefficient	Infit or Outfit Balanced	
Low	Flagged	Surviving
Flagged	66%	33%
Surviving	34%	
Med	Flagged	Surviving
Flagged	71%	25%
Surviving	29%	
High	Flagged	Surviving
Flagged	67%	28%
Surviving	33%	

Table 12. Exam B Survival Rate Similarity between Point-to-Measure and Sumfit

Point-to-Measure Correlation Coefficient	Sum of Infit and Outfit	
Low	Flagged	Surviving
Flagged	82%	16%
Surviving	18%	
Med	Flagged	Surviving
Flagged	68%	28%
Surviving	32%	
High	Flagged	Surviving
Flagged	68%	31%
Surviving	32%	

Table 13. Exam B Survival Rate Similarity between Point-to-Measure and Heavy Infit or Light Outfit

Point-to-Measure Correlation Coefficient	Heavy Infit or Light Outift	
Low	Flagged	Surviving
Flagged	71%	33%
Surviving	29%	
Med	Flagged	Surviving
Flagged	68%	29%
Surviving	32%	
High	Flagged	Surviving
Flagged	64%	36%
Surviving	36%	

Table 14. Exam B Survival Rate Similarity between Sumfit and Infit or Outfit Balanced

Sum of Infit and Outfit Mean Square	Infit or Outfit Balanced	
Low	Flagged	Surviving
Flagged	73%	27%
Surviving	27%	
Med	Flagged	Surviving
Flagged	84%	15%
Surviving	16%	
High	Flagged	Surviving
Flagged	93%	2%
Surviving	7%	

Table 15. Exam B Survival Rate Similarity between Sumfit and Heavy Infit or Light Outfit

Sum of Infit and Outfit Mean Square	Heavy Infit or Light Outift	
Low	Flagged	Surviving
Flagged	83%	23%
Surviving	17%	
Med	Flagged	Surviving
Flagged	91%	10%
Surviving	9%	
High	Flagged	Surviving
Flagged	92%	10%
Surviving	8%	

Table 16. Exam B Survival Rate Similarity between Infit or Outfit Balanced and Heavy Infit or Light Outfit

Infit Balanced	Heavy Infit or Light Outift	
Low	Flagged	Surviving
Flagged	63%	41%
Surviving	37%	
Med	Flagged	Surviving
Flagged	76%	25%
Surviving	24%	
High	Flagged	Surviving
Flagged	90%	16%
Surviving	10%	

Each of the items was plotted on a graph (see Figure 1 and 2) to illustrate the relationship between the difficulty of the item (in logits), the point-to-measure correlation coefficient, and the functionality of the low filtration flagging criteria. The main finding from these graphs is that the sumfit rule was better than the infit or outfit rule at detecting those above average difficulty items that had negative point-to-measure correlation coefficients. The same pattern was evident for the medium and high level of filtration.

Figure 1. Exam A Point-to-Measure Correlation plotted against Item Difficulty by Flagging Rule

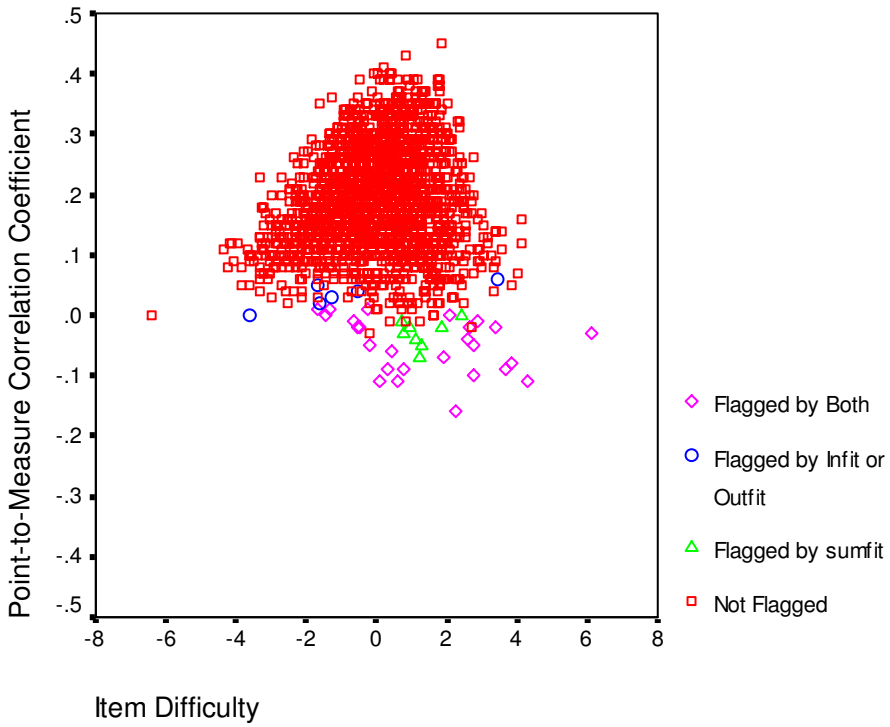
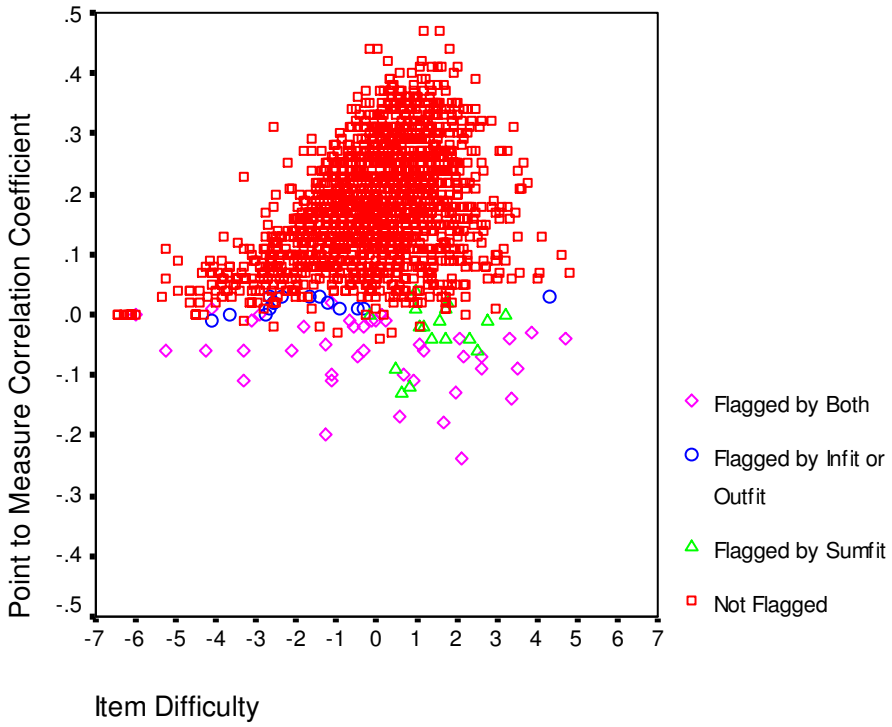


Figure 2. Exam B Point-to-Measure Correlation plotted against Item Difficulty by Flagging Rule



Conclusion

This was a short and simple study that used item survival rates as the benchmark by which to carry over traditional classical test theory item quality statistics to the item response theory world.

Item survival rates are commonly used and discussed by item developers and psychometricians. In the real world of large-scale licensure and certification assessment, item survival rates are fairly stable from one item development effort to the next. For programs interested in transitioning from classical test theory item quality criteria to item response theory criteria, the most important question asked by program stakeholders is "How will this transition effect the item survival rate?" By starting with this question, we were able to adjust the flagging criteria *values* until we were certain that item survival rates would not change.

Of the four different IRT methods tested, the sum of infit and outfit means square yielded the most similar outcomes to the point-to-measure correlation coefficient. Programs interested in mimicking classical test theory may wish to use this method. Programs adopting this method will still benefit from calculating outfit and infit since these stats individually can be used to troubleshoot and revise weak items to bolster their viability.

Since each method has statistical merits and weaknesses, future research should aim to probe the perceived quality of each method by investigating the questionable items. That is, subject matter experts should review the set of item flagged by one method and not by the other and determine which pile of questionable items is more desirable.

In conclusion, licensure and certification programs that are still clinging to classical test theory based statistics such as the point-biserial correlation coefficient, should use item survival rates and the method outlined in this study, when modifying their item quality selection criteria. By doing so, programs will be able to choose a method of filtration that best suits their program while also ensuring item pool stability.