Examining Changes in Item Difficulty Estimates Across Years for a High Stakes Licensure CAT

Jerry Gorham, Pearson VUE
Michelle Reynolds, National Council of the State Boards of Nursing

Introduction

The NCLEX-RN® and NCLEX-PN® exams are high stakes exams used to determine competence for nursing practice, either registered or practical nursing, on the basis of a national standards of nursing practice. The exams are independent of one another but share common features such as the core computer adaptive routine used for administration, the methods for data collection, calibration, scaling, scoring, diagnostic feedback, and passing standard determination. Item development issues such as item writing, reviews, validations, and most administrative procedures are also very similar, if not altogether equivalent.

In the spring of 1994 the NCLEX-RN® and NCLEX-PN® exams were converted from a traditional 300-item paper and pencil exam that had been administered on specific dates each year to a variable length computer adaptive exam (75 to 265 items for RN, and 85 to 205 items for PN, including pretests) that is administered continuously throughout the year at many testing sites across the U.S. and its territories.

Samples

This study will examine item difficulties based on operational CAT data that spans the period from the spring of 1994 until the fall of 2003. In practice, items are embedded in the adaptive tests of examinees (15 items for RN and 25 items for PN) and are delivered randomly to examinees rather than being targeted to examinee ability estimates. Pretest items must meet minimal sample size requirements per item (approximately 400 to 500 reference group examinees) and are calibrated only on a subgroup of examinees – those who are first-time RN test takers who have been educated in the United States. This group has been defined as the “reference group” and is used as the basis for all calibrations.

Generally, the summer testing period provides the largest numbers of RN examinees and the most consistent demographic subgroup for sampling, so this period was chosen to provide year-to-year samples for comparisons.

Method

The Rasch model (Rasch, 1980; Lord, 1980; Wright & Stone, 1979) has been used for calibration and scoring of examinees since the beginning of the testing program. Items generally are not recalibrated unless changes to the item text or item formats justify obtaining new parameter estimates. As a result, some items that were calibrated ten years ago are still being used based on the original Rasch item difficulty estimates. What has not been done is to examine the item difficulties based on operational data to see whether there have been significant changes to many items' difficulty estimates since the items were initially calibrated.

Operational data was collected and reformatted into a sparse data matrix with examinees as rows and items as columns. This data matrix was produced for each operational item pool for a testing quarter (three month testing period). Items were calibrated by pool using the examinees' final CAT ability estimates to fix the scale of the item parameters. Calibration was conducted using Winsteps (Linacre, 2003; Linacre, 2004). A table showing the samples used and the numbers of examinees and items calibrated is shown below (each pool contained three "sample" items that were not scored, so the actual numbers of scored items is N_Op_Items minus 3).

Table 1: NCLEX-RN Samples Used in Study

Sample	N_Ref_Grp_Examinees	N_Operational_Items
July94	44,676	1,798
July95	38,169	1,243
July96	39,329	1,543
July97	40,079	1,529
July98	36,361	1,803
July99	36,012	1,653
July00	23,114	1,703
Apr01	23,566	1,653
July01	45,245	1,653
Oct01	16,647	1,653
Apr02	23,341	1,653
July03	52,549	1,653

During 2002 the program changed vendors and a beta test was conducted during the spring and part of the summer. As a result, testing patterns for the reference group were atypical, so additional samples were chosen around that period to supplement the year 2002 data.

Table 2 below shows the frequency distribution of the number of calibrations generated from the data by item. Items with only one operational calibration were excluded, so the numbers of calibrations ranged from two to eleven per

item. Notice that one-third of the total number of calibrations consisted of only two calibrations per item. The number of calibrations per item can be interpreted *approximately* as the number of years' worth of estimates available for each item since the samples focus on consecutive summers' worth of data that was used.

Table 2: Frequency Distribution of Number of Calibrations by Item

Num_Calibrations	Num_Items	Percent	Cumulative Percent
2	2,220	33.18	33.18
3	1,304	19.49	52.67
4	1,173	17.53	70.20
5	771	11.52	81.72
6	627	9.37	91.09
7	313	4.68	95.77
8	186	2.78	98.55
9	72	1.08	99.63
10	18	0.27	99.90
11	7	0.10	100.00
Total	6,691	100.00	

Table 3 shows mean and standard deviations for items grouped by the total number of difficulty estimates available for an item. The mean differences for each grouping ranges from -0.026864 to +0.010365. The overall mean of the differences between consecutive calibrations for all 6,691 items is +0.000254. There appears to be no evidence of systematic differences between calibration sets from year to year based on the means of the items, however, these averages may not tell the whole story.

Table 3: Mean Differences for Consecutive Operational Item Difficulty Estimates

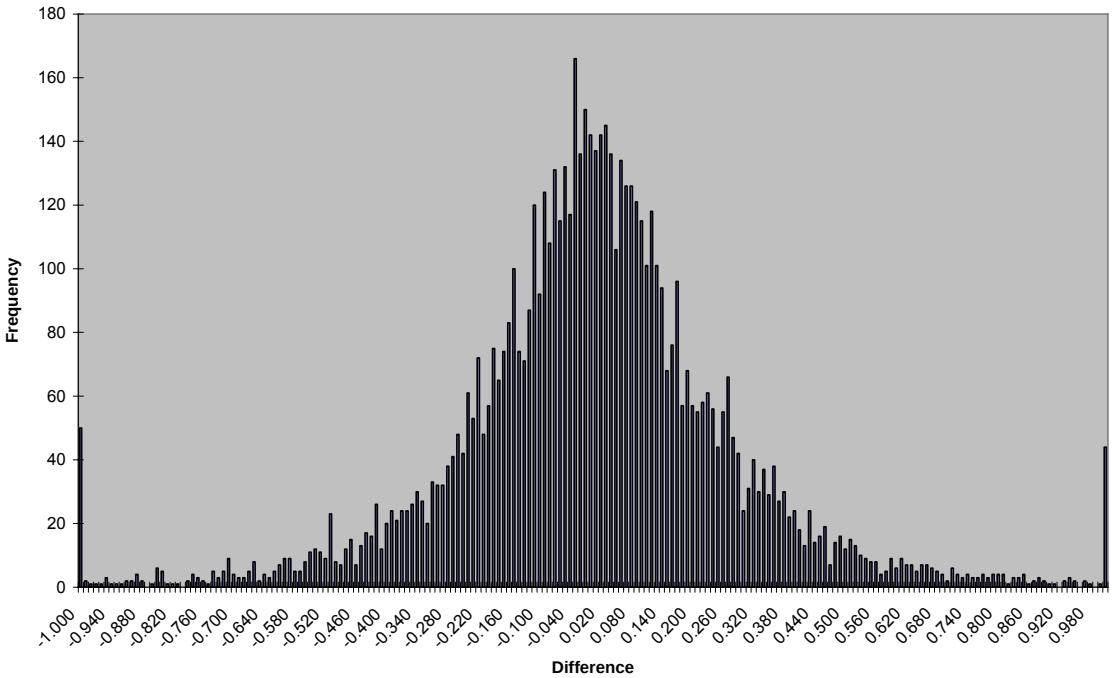
Num_Calibrations	Num_Items	Mean	Std_Dev
2	2,220	-0.013279	0.345241
3	1,304	0.000528	0.324925
4	1,173	0.000374	0.269856
5	771	0.000844	0.222603
6	627	0.005774	0.223937
7	313	0.004246	0.197210
8	186	0.004648	0.162901
9	72	0.000271	0.144088
10	18	0.010365	0.160467
11	7	-0.026864	0.084499

Overall	**6,691**	**0.000254**	**0.262957**

For items that have been exposed across time, we might expect some items to remain essentially the same in difficulty, a large number of items to appear easier because of high item exposures, and possibly a few items to become more difficult because of changes in curriculum. For instance, examinees testing in recent years might not be familiar with older items that emphasize concepts that are now taught less frequently, or with less emphasis because of changes in practice or instruction, making the items appear more difficult.

A simple measure for observing difficulty changes across time is the difference between the initial calibrated and the final calibrated value, both based on the adaptive data. The shape of the distribution of item difficulties across time may indicate whether there is some systematic bias among items. Figure 1 shows the distribution of these differences in consecutive item difficulty estimates (the tails of the distribution contain large numbers of items simply because the graph was drawn to display the majority of items in the range of -1.0 to +1.0). The mean of the distribution is +0.001180, the standard deviation is 0.315815, and the distribution is slightly negatively skewed (-0.189287). The standard error of the mean is 0.003861, and the mean of the distribution does not differ significantly from zero.

Figure 1: Distribution of Differences in Consecutive Item Difficulty Estimates



That the distribution does not differ from zero might be explained by the fact that the distribution is overwhelmed by the number of N=2 estimate items, which may not display many, if any, changes in item difficulties. Based on the same measure of the difference between the first and final estimates, Table 4 below

shows the number, mean and standard deviations of these differences by the number of per-item estimate categories, which are exclusive of one another (items with only two adaptive estimates, items with only three adaptive estimates, etc.). One might expect that as items continue to be exposed across years, they would become more known and, therefore, less difficult across time.

Note that the mean differences tend to increase from the item categories Est = 2 to Est = 11. Positive differences indicate that the item has become easier while negative differences indicate that the item has become more difficult. This may be an indication of the tendency of items to become less difficult across multiple pool exposures.

Table 4: Means and Std Deviations by Number of Item Estimates

Num_Estimates	N	Mean	Std Dev
Est = 2	2220	-0.013279	0.345241
Est = 3	1304	0.001056	0.322843
Est = 4	1173	0.001121	0.299030
Est = 5	771	0.003377	0.272861
Est = 6	627	0.028868	0.298076
Est = 7	313	0.025478	0.306434
Est = 8	186	0.278900	0.256346
Est = 9	72	0.029800	0.269575
Est = 10	18	0.093289	0.318134
Est = 11	7	-0.268643	0.213880

The exception to this tendency is the last column (Est = 11) in which there are only seven items, each with eleven estimates per item. Figure 2 illustrates the plots of item difficulty by administration quarter for these seven items. Five of the seven items have become more difficult across time while two of the seven items have remained relatively consistent in item difficulty. The item texts cannot be discussed in any detail in a public context, but after review, these items appear to some concepts in nursing that are generally considered more difficult to understand. Some emphasize prioritization of nursing actions, attention to critical signs and symptoms, and seem to contain difficult medical terminology. These characteristics may have contributed to the increasing difficulty of the items across time.

Figure 2: Items with 11 Difficulty Estimates by Date

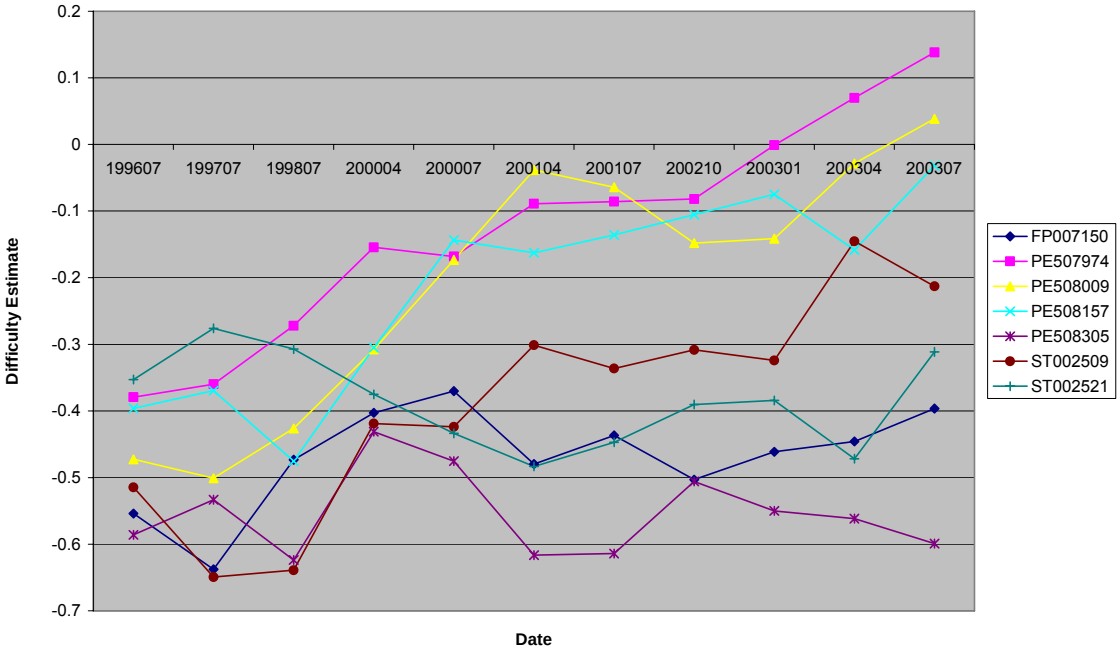
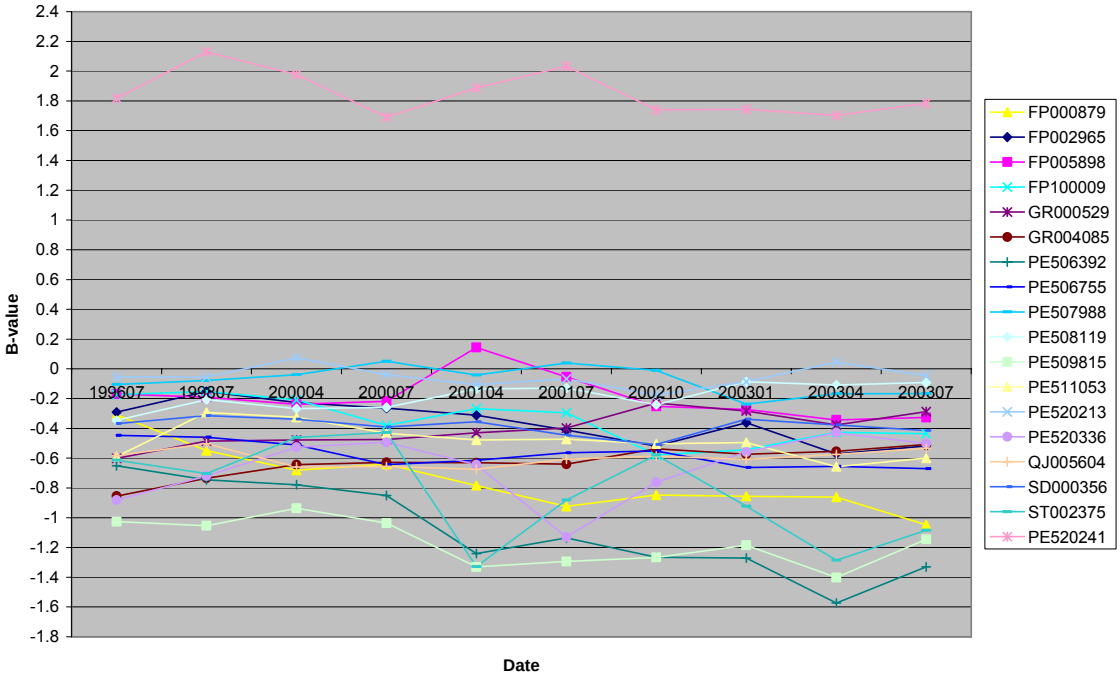


Figure 3 below shows the same type of information for items with ten item difficulty estimates. Note that most items appear relatively stable across time, while a few items have become easier and perhaps one item has become more difficult (GR004085).

Figure 3: Items with 10 Difficulty Estimates by Date



Comparison with Initial Calibrated Pretest Estimates

Regardless of the overall consistency of estimates based on the adaptive data, items are nevertheless selected by the CAT routine and scored with maximum likelihood scoring based on their initial pretest estimates. Some of these estimates may be many years old and in fact, for most items the pretest estimates have not been updated because of concerns over adverse impacts on the overall scale and other unknowns in online recalibration.

In light of these stationary estimates, quality control measures have been put in place to ensure that items are behaving appropriate to their initial non-adaptive estimates. One important measure is a model-data fit statistic that is calculated for each operational item. Items that are outside a confidence interval of fit are permanently eliminated from the live CAT pools. The statistic for calculating model-data fit with the NCLEX CAT is described below (NCLEX Technical Reports, Appendix A,1994-2004).

The statistic Z is a standardized residual for item i and a restricted ability group j as follows:

$$Z_{ij} = \frac{N_j^{1/2}\left[P_{+ij} - E(P_{+ij})\right]}{\left[E(P_{+ij})(1-E(P_{+ij}))\right]^{1/2}}$$

where,

$P_{+ij} = \frac{1}{N_j} \sum_{g \in j}^{N_j} u_{ig}$ = observed proportion correct for the g candidates in group j, and

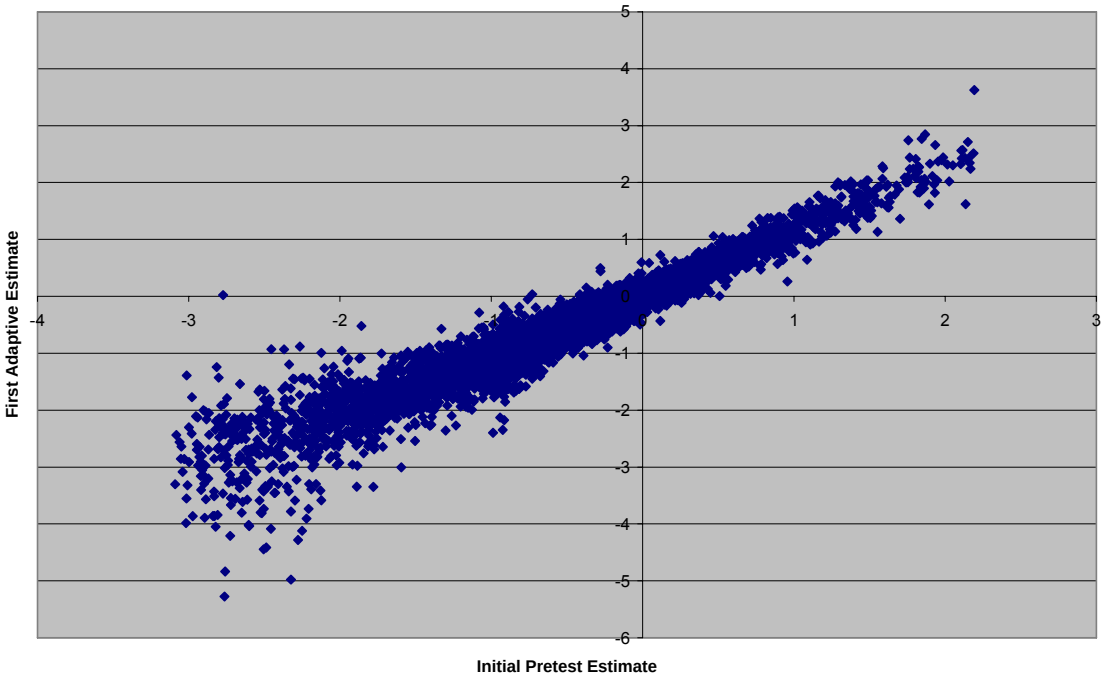
$E(P_{+ij}) = \frac{1}{N_j} \sum_{g \in j}^{N_j} P_i(\hat{\Theta}_g)$ = the expected proportion of g candidates in group j correctly answering item i as predicted by the Rasch model, and Θ_g is in group j and $b_i - \zeta \leq \hat{\Theta}_g \leq b_i + \zeta$, where b_i is the estimated difficulty of item i and ζ (zeta) is a specified distance on the ability metric, where ζ is set at 0.5. To compensate for wide variations in sample sizes that exist in CAT data, the Z statistic is adjusted for items with N > 400 observations by the following:

$$Zadj_{ij} = Z_{ij} \left(\frac{N_{REF}}{N_j} \right)^{1/2}$$

This adjustment provides a sort of referential statistic for comparisons among items with wide variations in sample sizes. The general procedure for using this statistic is to eliminate items whose Z statistics across a six-month operational pool are greater than or equal to an absolute value of 4.0. This ensures that items no longer fitting their Rasch difficulty parameters will be weeded out of the active item pools. Most items remain well within these Z parameters and are not removed from the active pools. Typically, about two to three hundred items are removed annually from the pools on the basis of a misfit of data to model. These items are permanently deleted from the pools and are generally not re-written or re-pretested.

Figure 4 shows a scatterplot of the initial pretest estimates by the first adaptive (based on adaptive responses) difficulty estimate for 5,234 items. Although the correlation is high (r = +0.9567), the variability occurs at the ends of the distributions, particularly at the lower end of the difficulty scale. This is to be expected and reflects the larger standard errors that typically occur with examinees at the highest and lowest ends of the scale.

Figure 4: Initial Pretest Estimates by First Adaptive Estimates



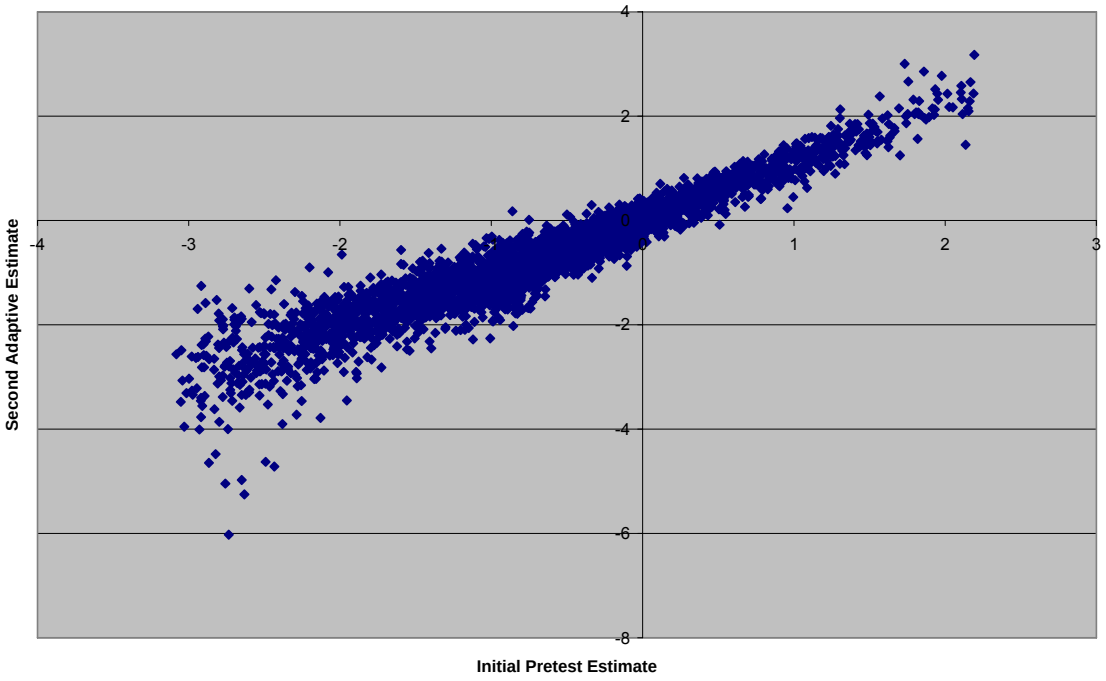
Mean (pretest estimates) = -0.61622
Std Dev (pretest estimates) = 0.91452

Mean (adaptive estimates) = -0.65696
Std Dev (adaptive estimates) = 1.02288

The mean of the adaptive estimates is slightly lower than the pretest estimates and the standard deviation of the adaptive estimates is larger than that of the pretest estimates.

A similar pattern can be seen for 4,513 items from the initial pretest and second adaptive estimates (Figure 5, below). The correlation is high (+0.9516), the mean of the adaptive estimates is slightly lower and the standard deviation of the adaptive estimates is larger than the pretest estimates.

Figure 5: Initial Pretest Estimates by Second Adaptive Estimates

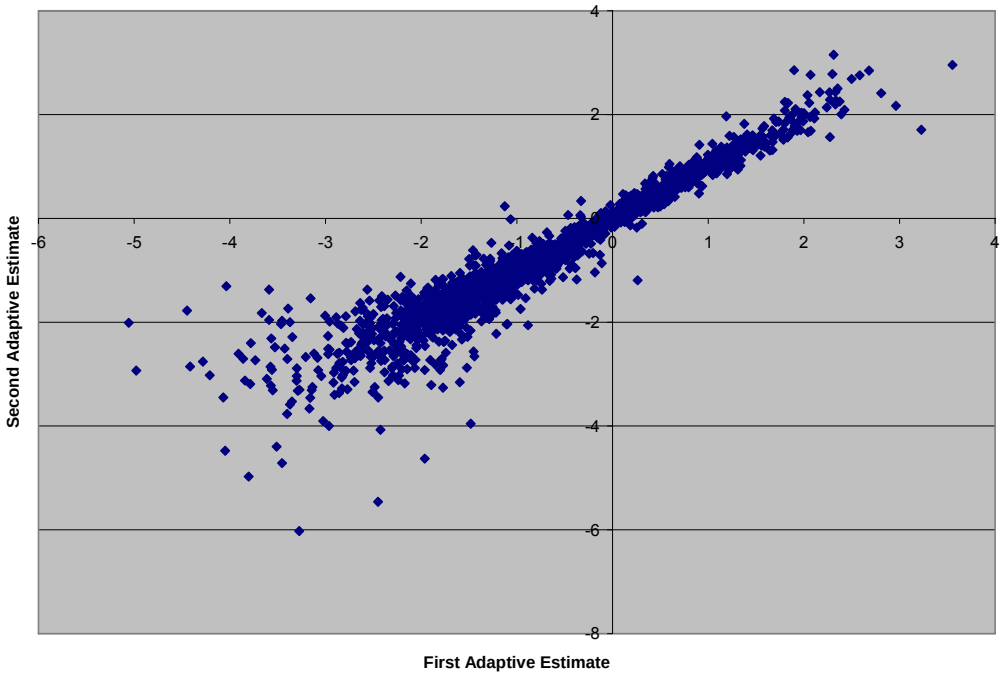


Mean (pretest estimates) = -0.62879
Std Dev (pretest estimates) = 0.893181

Mean (adaptive estimates) = -0.67197
Std Dev (adaptive estimates) = 0.98954

Figure 6 shows the relationship between the first and second adaptive estimates for 2,220 items. The correlation is slightly higher (r = +0.9606) and the means and standard deviations between the adaptive estimates are much closer than the mean and standard deviation of the pretest estimates compared to the means and standard deviations of either adaptive estimate.

Figure 6: First Adaptive Estimates by Second Adaptive Estimates

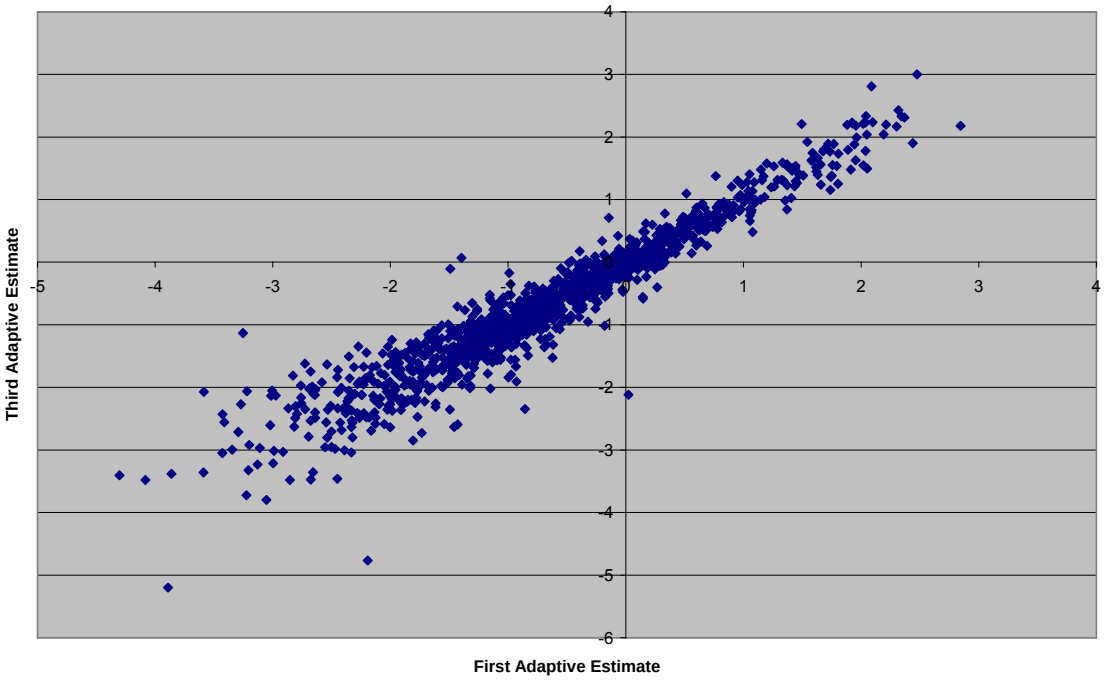


Mean (adaptive estimate #1) = -0.66714
Std Dev (adaptive estimate #1) = 1.232898

Mean (adaptive estimate #2) = -0.65386
Std Dev (adaptive estimate #2) = 1.227723

Figure 7 plots the first and third adaptive estimates for 1,304 items.

Figure 7: First Adaptive Estimates by Third Adaptive Estimates

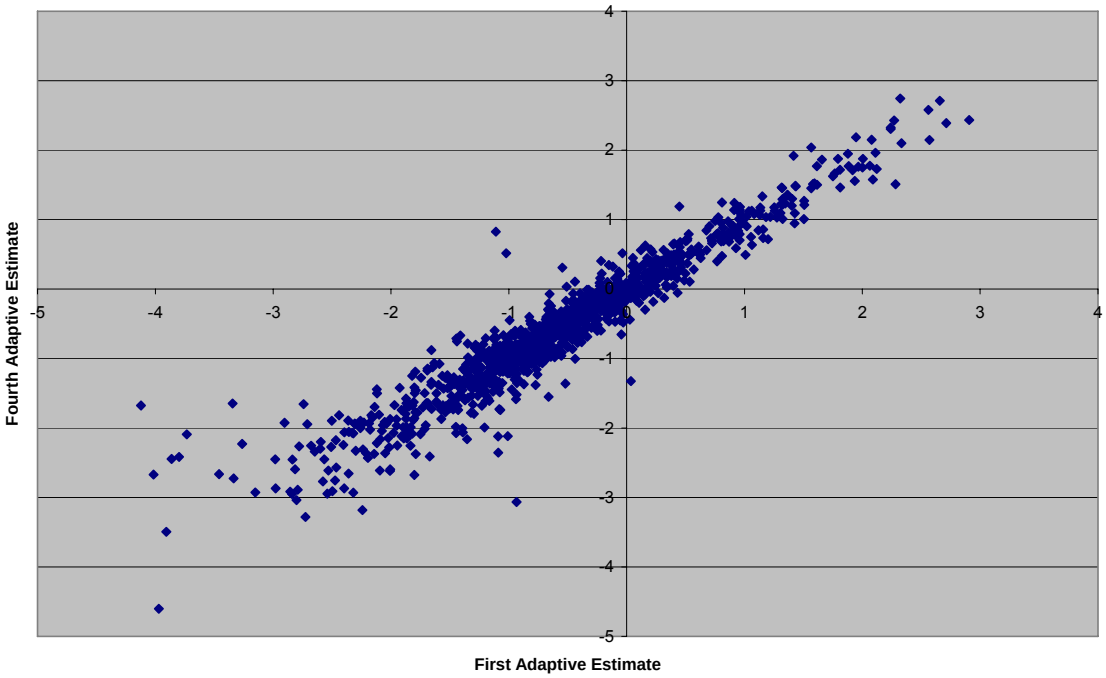


Mean (adaptive estimate #1) = -0.67107
Std Dev (adaptive estimate #1) = 1.102962

Mean (adaptive estimate #3) = -0.67213
Std Dev (adaptive estimate #3) = 1.100462

Figure 8 shows the first and fourth adaptive estimates for 1,173 items.

Figure 8: First Adaptive Estimates by Fourth Adaptive Estimates

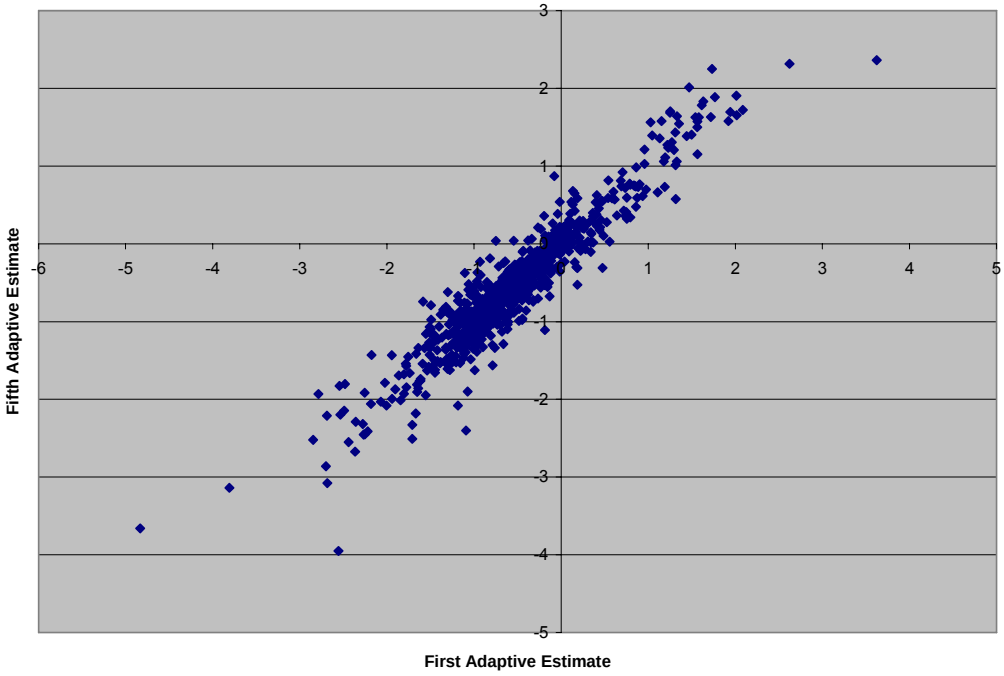


Mean (adaptive estimate #1) = -0.55839
Std Dev (adaptive estimate #1) = 1.001088

Mean (adaptive estimate #4) = -0.55951
Std Dev (adaptive estimate #4) = 0.978122

Figure 9 plots the first and fifth adaptive estimates for 771 items.

Figure 9: First Adaptive Estimates by Fifth Adaptive Estimates

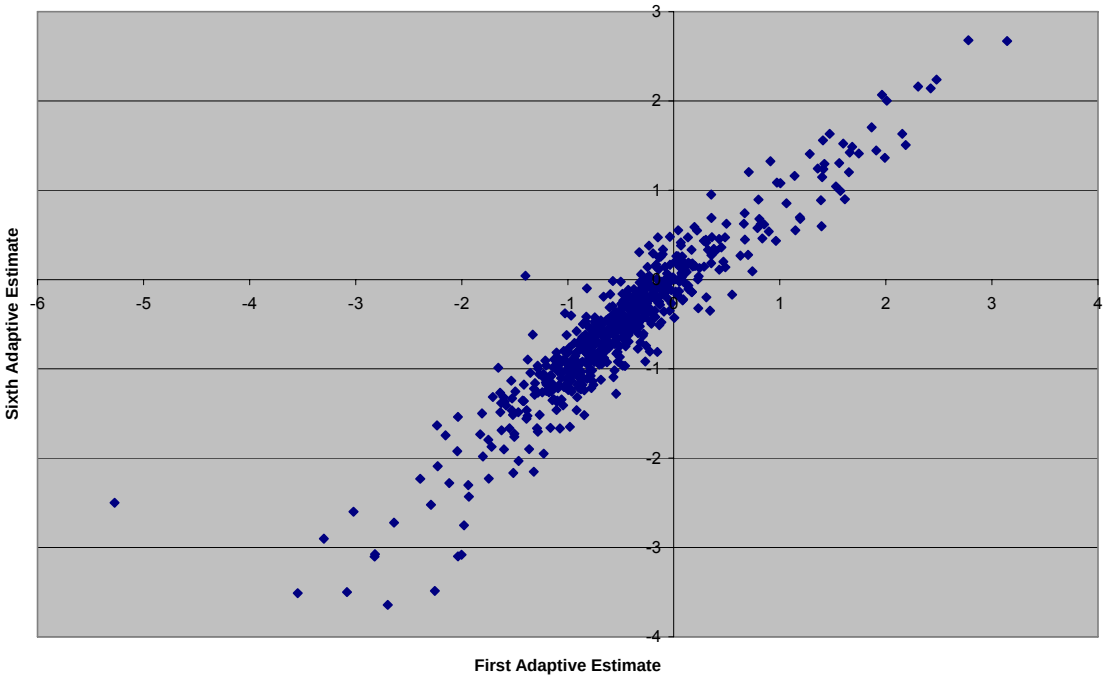


Mean (adaptive estimate #1) = -0.52007
Std Dev (adaptive estimate #1) = 0.827429

Mean (adaptive estimate #5) = -0.52345
Std Dev (adaptive estimate #5) = 0.820738

Figure 10 shows the relationship between the first and sixth adaptive estimates for 627 items.

Figure 10: First Adaptive Estimates by Sixth Adaptive Estimates

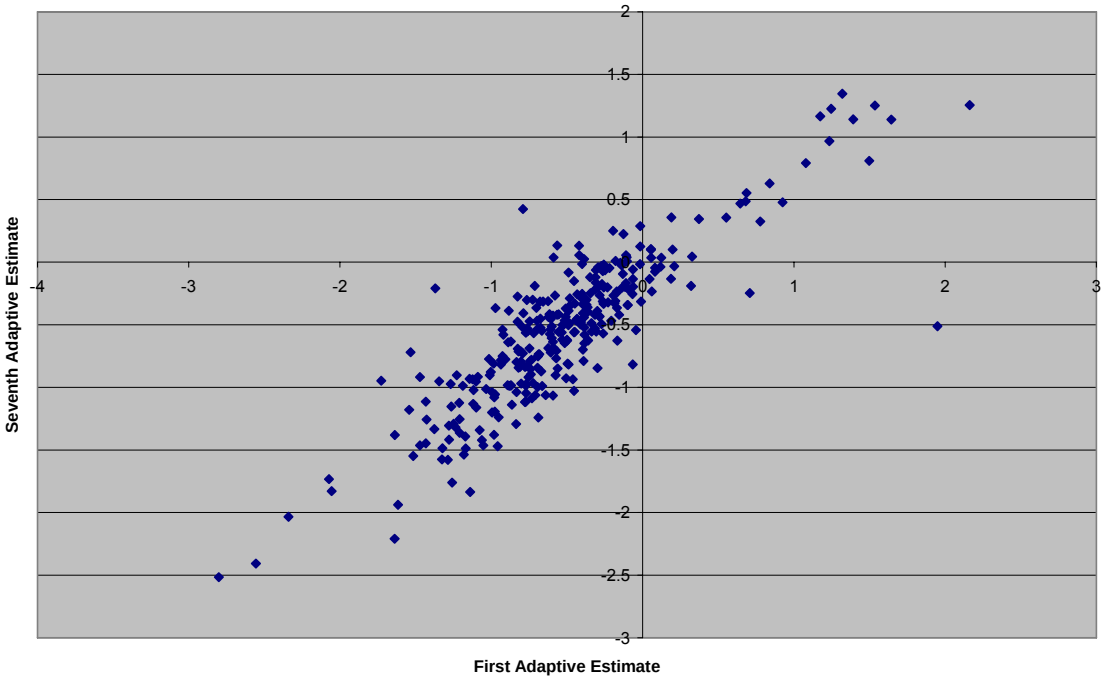


Mean (adaptive estimate #1) = -0.47985
Std Dev (adaptive estimate #1) = 0.850014

Mean (adaptive estimate #6) = -0.50872
Std Dev (adaptive estimate #6) = 0.839082

Figure 11 shows the relationship between the first and seventh adaptive estimates for 313 items.

Figure 11: First Adaptive Estimates by Seventh Adaptive Estimates

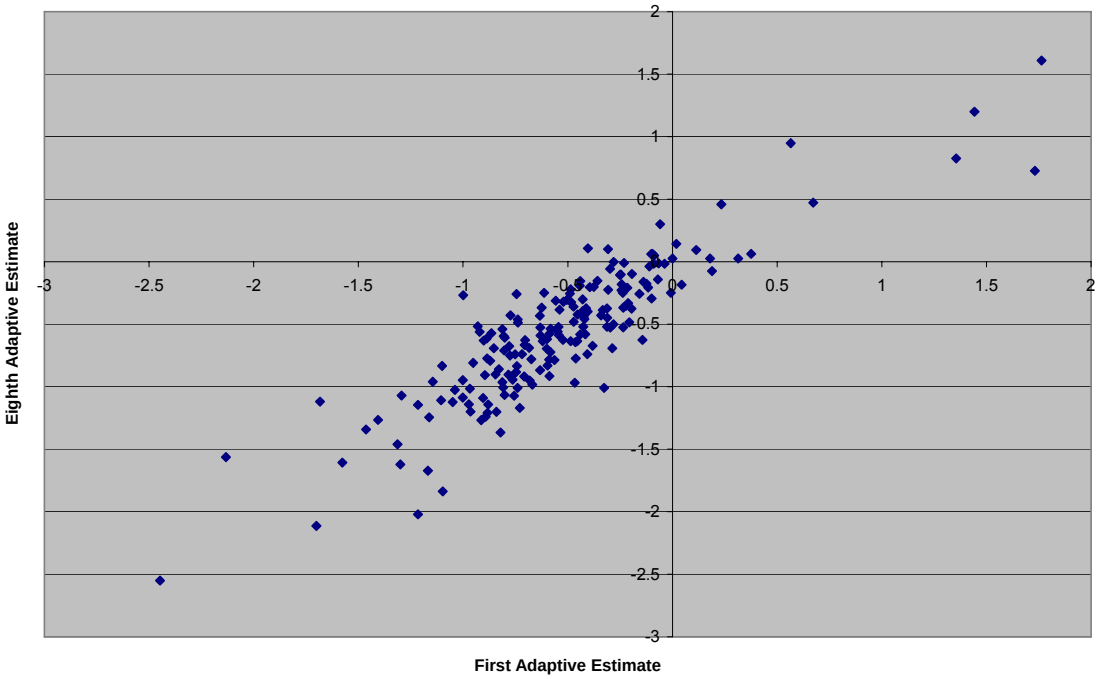


Mean (adaptive estimate #1) = -0.52769
Std Dev (adaptive estimate #1) = 0.636703

Mean (adaptive estimate #7) = -0.54923
Std Dev (adaptive estimate #7) = 0.59028

Figure 12 shows the first and eighth adaptive estimates for 186 items.

Figure 12: First Adaptive Estimates by Eighth Adaptive Estimates

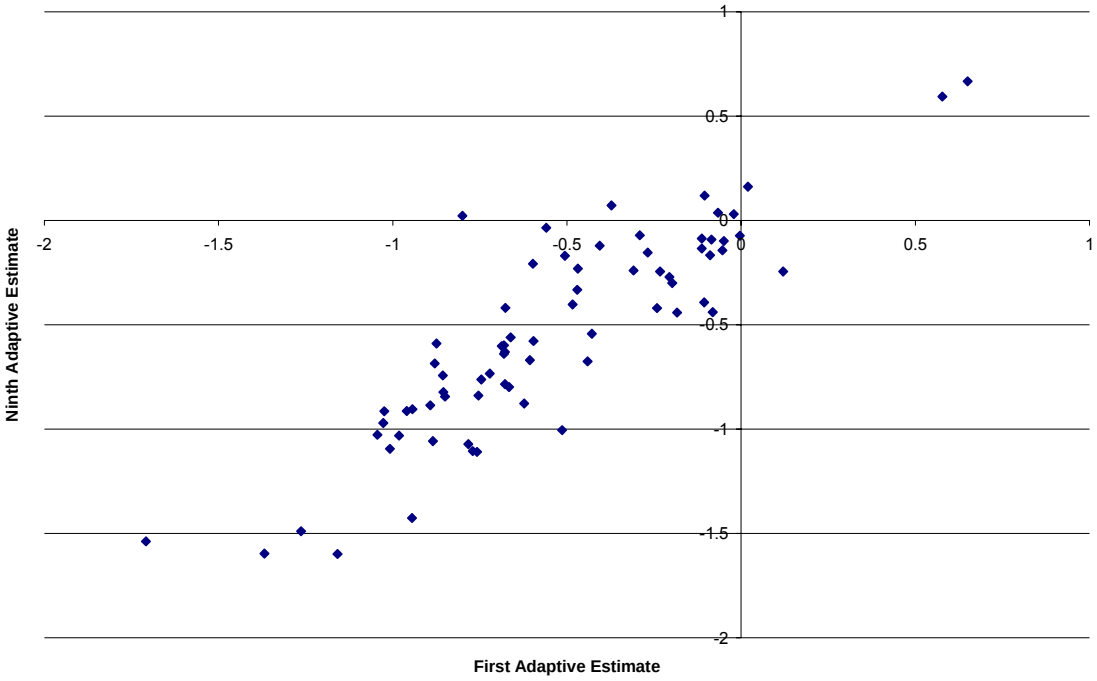


Mean (adaptive estimate #1) = -0.53247
Std Dev (adaptive estimate #1) = 0.533191

Mean (adaptive estimate #8) = -0.56501
Std Dev (adaptive estimate #8) = 0.545334

Figure 13 shows the first and ninth adaptive estimates for 72 items.

Figure 13: First Adaptive Estimates by Ninth Adaptive Estimates

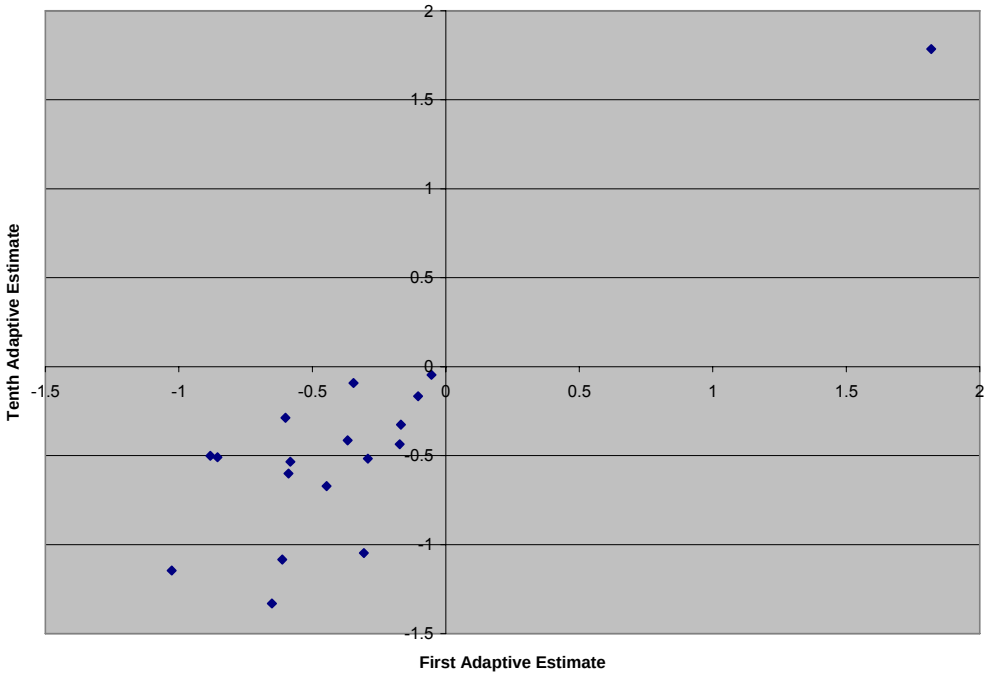


Mean (adaptive estimate #1) = -0.53868
Std Dev (adaptive estimate #1) = 0.423413

Mean (adaptive estimate #9) = -0.54085
Std Dev (adaptive estimate #9) = 0.482284

Figure 14 plots the first and tenth adaptive estimates for 18 items.

Figure 14: First Adaptive Estimates by Tenth Adaptive Estimates

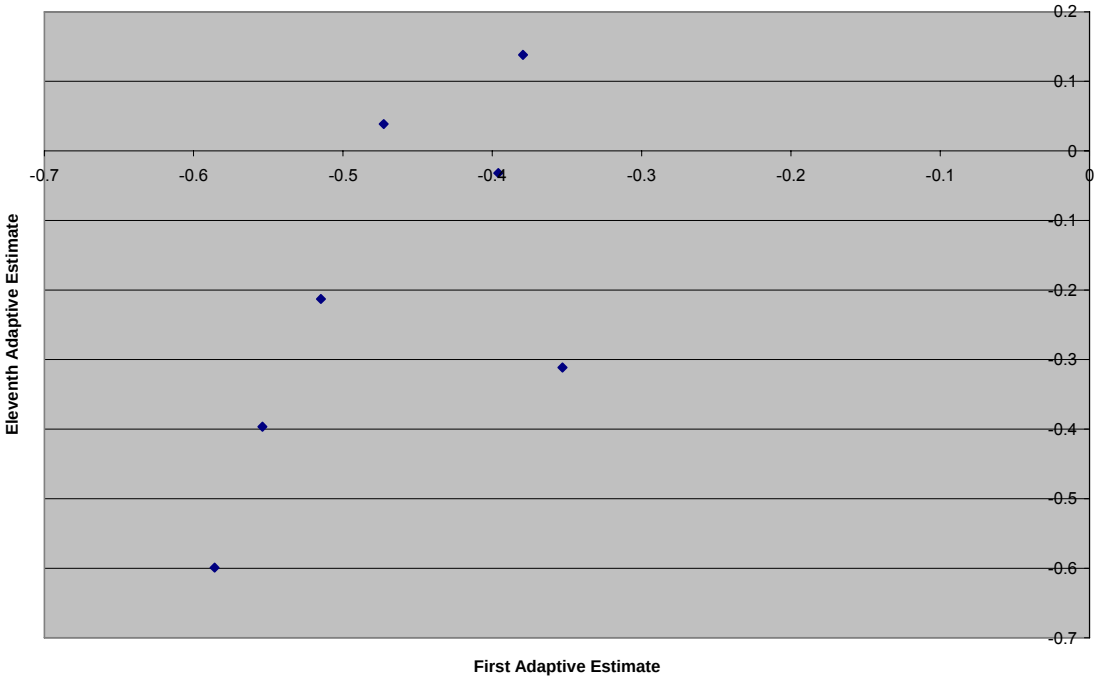


Mean (adaptive estimate #1) = -0.34671
Std Dev (adaptive estimate #1) = 0.606662

Mean (adaptive estimate #10) = -0.43999
Std Dev (adaptive estimate #10) = 0.664884

Figure 15 is provided for purposes of completeness even though there are only seven observations for items that have eleven adaptive estimates.

Figure 15: First Adaptive Estimates by Eleventh Adaptive Estimates



Mean (adaptive estimate #1) = -0.46511
Std Dev (adaptive estimate #1) = 0.091021

Mean (adaptive estimate #11) = -0.196471
Std Dev (adaptive estimate #11) = 0.2614012

From this somewhat dry, repetitive series of charts there is some suggestion that as items continue to be administered across multiple pool administrations, there is a tendency for those items to become slightly easier. However, there are exceptions to the rule, such as in the case of items with 11 estimates. As Figures 2 and 3 (earlier) suggested, the actual trend plots of items with many estimates across time still show some items becoming easier over time and some items remaining relatively stable in their difficulty estimates across time.

Changes in Item Difficulty Estimates Across Time

A categorization was created to identify items that have become “less difficult”, “more difficult”, or “relatively stable” across time. For the 6,692 items discussed earlier, the difference between the first adaptive and final adaptive estimates was compared to the standard error of the initial adaptive estimates to roughly identify items that appear to have become much easier, much more difficult, or items that have had no change, across multiple administrations. Items that changed by two or more standard errors of the initial adaptive estimate were categorized as significantly different in difficulty from their initial estimate.

Table 5 below summarizes the results of categorizing these items. The majority of items (58.9%) do not appear to have any significant changes in item difficulty. Approximately 21.7% of the items have become less difficult and approximately 19.4% of the items have become more difficult. What is somewhat interesting is that items without major changes in difficulty tend to be items that have fewer cumulative exposures and have been used in the active pools for a fewer number of quarters of testing. Increased exposure tends to shift item difficulties in either direction, although this process is likely very complex. Note the initial estimates for each group of items. The items that have become less difficult are items whose initial estimates began slightly above the cutscore (which has ranged from about -0.4700 to more recently, -0.2800). Items that have become more difficult are items whose initial estimates began slightly below the cutscore, and items that have not moved significantly in their estimates as a group are items that are well below (based on the mean) the cutscore.

Table 5: Summary of Items Categorized by Significant Shifts in Difficulty

	No Difference	Less Difficult	More Difficult	Overall
Num_Items	**3,942**	**1,453**	**1,296**	**6,691**
Percent_Items	**58.9%**	**21.7%**	**19.4%**	**100.0%**
Mean_Initial_Estimate	**-0.7708**	**-0.2988**	**-0.4257**	**-0.6015**
Mean_Final_Estimate	**-0.7525**	**-0.6100**	**-0.1394**	**-0.6028**
Mean_Difference_Initial, Final	**-0.0183**	**0.3112**	**-0.2863**	**0.0014**
Mean_Cumulative_Exposures	**11,757**	**24,528**	**26,974**	**17,478**
Mean_Number_Quarters	**7.2**	**9.8**	**10.0**	**8.3**

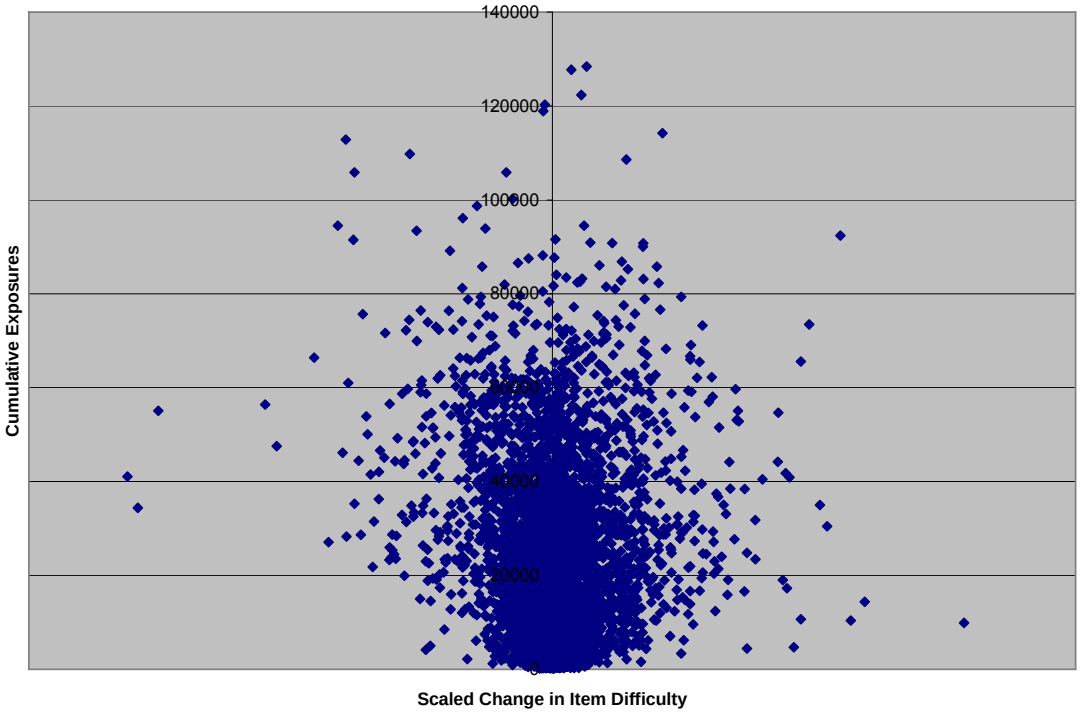
One possible explanation for these data is that items in the less difficult or more difficult categories are simply regressing toward the cutscore and are not as a group changing all that much. There are certainly individual items whose estimates appear to be changing, but as a whole, the pools of items may be behaving fairly well as a group.

Another possible explanation for this item estimate behavior is related to ability estimate bias near the cutscore. For the CAT to stop at a minimum test length (60 scored items), the examinee's ability estimate needs to be well outside the confidence interval. This may create ability estimate bias in either direction for minimum length exams near the cutscore. All items are calibrated using these ability estimates to fix the scale, so items just above and just below the cutscore will carry that bias. This could explain why items just below the cutscore appear to become more difficult and items just above the cutscore appear to become easier when calibrated with the adaptive data. What is interesting is that the mean difference for these two groups of items is very close (0.3112 for the less difficult group, and -0.2863 for the more difficult group).

Conclusions

For the most part, many items appear to be relatively stable across multiple administrations. Figure 16 shows scaled changes in item difficulties by the cumulative number of exposures per item. The graph has been scaled by the standard error of item estimates to allow direct comparison of item difficulty changes. Note that there are many items with 20,000 to 60,000 cumulative exposures whose item difficulties have not changed dramatically. There are also over 110 items that have been administered over 50,000 times per item across a period of over 14 testing quarters without any noticeable change in item difficulty. This does not mean that increased item exposures do not impact item difficulty. Earlier data presented in the paper seems to suggest that increased item exposure does have an effect on item difficulty *in general*. The point is that the relationship between item difficulty changes and item exposure is more complex than we may have been led to believe. What seems more important are the conditional cumulative exposures that occur among various subgroups and among different ability levels.

Figure 16: Changes in Item Difficulties by Cumulative Number of Exposures



As a whole, the items that remain in the active pools appear relatively stable across time. Items that do not perform according to their expected item difficulties are routinely removed from the active pools. There are individual items that have become much easier or much more difficult (note the outliers in Figure 16). These items can be identified, reviewed for content validity and relevance, and re-pretested in a non-adaptive manner to validate their changes

in item difficulties. Although there are currently no limits established for the number of times that an item may be administered, it might be useful to create a set of criteria for limiting the number of exposures and / or quarterly administrations of a particular item. We might also use the old agricultural principle of rotating fields (allowing a field to rest for a year before planting a new crop) to create a more systematic use, rest, and re-use of items in the live pools.

References

Linacre, J.M. (2003). A users guide to WINSTEPS: Rasch measurement computer program, Chicago, IL: MESA Press.

Linacre, J. M. (2004). WINSTEPS Rasch Measurement. Version 3.50 (February, 2004). Chicago.

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

NCLEX® Technical Reports. National Council of State Boards of Nursing (NCSBN). NCLEX-RN® and NCLEX-PN® examinations using computerized adaptive testing. (April 1994 to December 2004). Educational Testing Service, Chauncey Group, and Pearson VUE.

Rasch, G. (1980). Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danmarks Paedogogiske Institute; reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press.

Wright, B. D., and Stone, M. H. (1979). Best test design. Chicago: MESA Press.