Application of Score Information for CAT Pool Development and Its Connection with "Likelihood Test Information" [1]

I. A. Krass
Personnel Testing Division, Defense Manpower Data Center
Seaside, California

In the CAT-ASVAB testing program, estimated information functions are used to determine the test precision of measurement, given a particular set of calibrated items. A target score information function for each test in the battery is set based on previous and current well-established versions of CAT-ASVAB. This target score information function provides a goal used to assess the performance of newly created sets of calibrated items which are destined to become future operational CAT-ASVAB versions. As new candidate items are generated, tested on line, and then calibrated, they are combined into tentative item sets for evaluation of parallelism and information function.

Measuring test precision for examinee ability estimations and his/her score estimation is a rather difficult problem with conventional paper-and-pencil tests. In the case of computerized adaptive tests, this problem is even more complicated. In addition, development of new parallel forms best in precision is also hard, because it becomes a non-linear mathematical optimization problem.

Measurement of Precision in CAT-ASVAB tests

In 1968, Birnbaum (Birnbaum, 1968) introduced his measurement of precision of a test as the inverse to the square of the asymptotic confidence interval of test score y estimating true ability θ. If $\mu_{y|\theta}$ is the mean of score y of the test for the given true θ, and $Var(y|\theta)$ its variation, then the Birnbaum information, or "Score Information" can be computed:

$$I_B(y|\theta) = \frac{(\frac{d}{d\theta}\mu_{y|\theta})^2}{Var(y|\theta)}. \quad (1)$$

Here under the score y we understand any unbiased estimation of true ability θ. If y is the proportion correct score for the test, $p_i(\theta)$ is the item characteristic curve (ICC) of item i, and all the items in the test are locally independent, i. e. independent conditionally on θ, and unidimensional, then it is easy to show (Lord, 1980) that

$$I_B(y/\theta) = \frac{[\sum_{i=1}^{n} p_i'(\theta)]^2}{\sum_{i=1}^{n} p_i(\theta)\cdot(1-p_i(\theta))}, \quad (2)$$

where n is the number of items in the test. Here, true ability θ of an examinee is the unidimensional latent variable. The assumption of local independence and unidimensionality for a considered set of items will be applied in this paper.

[1] All statements expressed in this paper are those of author and not necessarily reflect the official opinions or policies of the U. S. Department of Defense.

About 50 years before Birnbaum's theory, Fisher (Fisher, 1925) introduced the precision measurement of latent variable θ by the maximum likelihood method:

$$I_F(\theta) = E(\frac{d \ln L}{d\theta})^2, \tag{3}$$

where L is the likelihood determined by latent variable θ. Under some condition of regularity (Serfling, 1980), it can be shown that

$$I_F(\theta) = \frac{1}{Var(\tilde{\theta} \mid \theta)},$$

where $\tilde{\theta}$ is a solution of the maximum likelihood problem – MLE of θ.

As it is shown by Lord (1980), if the test is conventional then:

$$I_F(\theta) = \sum_{i=1}^{n} \frac{(p_i'(\theta))^2}{p_i(\theta) \cdot (1 - p_i(\theta))}, \tag{4}$$

where n is the length of the test, and $p_i(\theta)$ is the ICC of the item i. Following Lord we will call the Fisher information function (3) applied for finding examinee ability "Likelihood Test Information." From (4) it follows that the Likelihood Test Information function is additive with respect to items comprising the test, which allows us to introduce the Fisher information of item i as:

$$I_i(\theta) = \frac{(p_i'(\theta))^2}{p_i(\theta) \cdot (1 - p_i(\theta))} \tag{5}$$

and rewrite (4) in the form: $I_F(\theta) = \sum_{i=1}^{n} I_i(\theta)$. From (2) and (4) follows the inequality

$$I_B(y \mid \theta) \le I_F(\theta), \tag{6}$$

where y is the proportion correct score for the test. This inequality also can be obtained as a consequence of the general Cramer-Rao inequality (Kendall & Stuart, 1973).

The main difference between the Fisher (Likelihood Test Information) and Birnbaum (Score Information) functions is that the Fisher information estimates precision of MLE of true ability θ, but the Birnbaum information does not constrain the method of estimation of θ as long as one can numerically estimate all terms included in the definition (1) of $I_B(y \mid \theta)$ and y is an unbiased estimator.

With the CAT-ASVAB, test estimation of true θ of an examinee is done with help of a Baysian estimation of his/her ability, denoted as $\hat{\theta}$ (theta–hat). As it is shown, (Owen, 1975) $\hat{\theta}$ is an unbiased estimation of true ability θ, and due to that the Birnbaum information function $I_B(\hat{\theta} / \theta)$ is applied for precision of the test estimation (Segall, Moreno, & Hetter, 1997). The estimation of the Birnbaum information function is done with simulation and the numerical estimation of the derivative and variance included in (1).

In Figure 1, we present two forms for the CAT-ASVAB Arithmetic Reasoning (AR) test, each containing 45 items chosen from 90 items in the item bank. In this figure, ARnew1 and ARnew2 are estimations of Birnbaum's information functions for two chosen forms. The estimation is done numerically, using a smoothing formula (11-3) from Segall, Moreno, and Hetter (1997). The entire region [-3.0,3.0] of available θ was split into 31 different, equally spaced levels, and for each level,

10,000 simulees for this level of θ were used to estimate numerically the mean and variance of theta-hat, required in the numerator and denominator in (1).

Thus, the curves for ARnew1 and ARnew2 in the figure provide the Birnbaum information: $I_{B\,i}(\hat{\theta}\,|\,\theta),\ i=1,2$ estimation for two chosen forms. Curves denoted as "Scrr Inf1" and "Scrr Inf2" are also provided by Birnbaum's information functions $I_{B\,i}(y\,|\,\theta),\ i=1,2$, but as it is presented in the formula (2) where y is the proportion correct score for chosen test forms (following Lord [1980], we call this type of information "Score Information.") Curves "Test Inf1" and "Test inf2" are Fisher or Likelihood Test Information functions as defined in formula (4) .

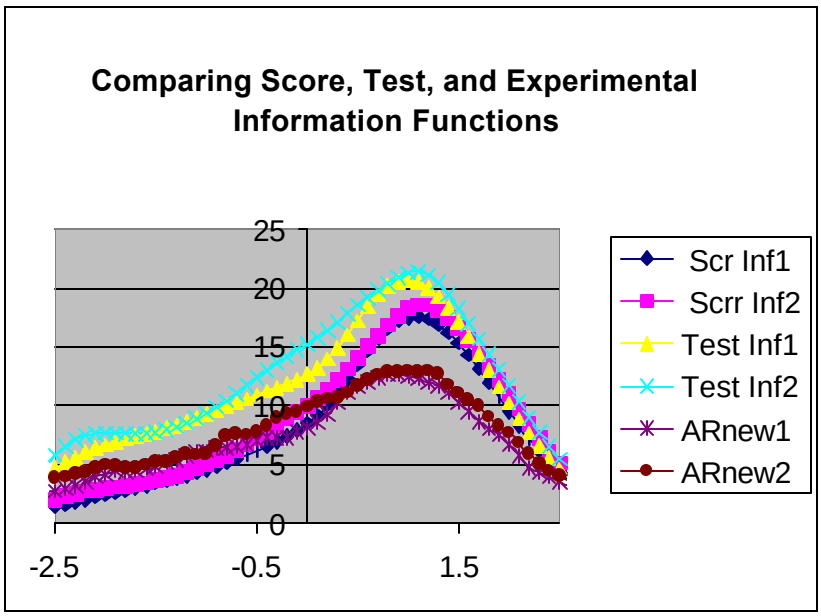

FIGURE 1. Different methods of estimation of information for the given test forms.

Precision curves based on the Likelihood Test Information functions are located above the precision curves based on the Birnbaum Score Information functions, which theoretically follows from (6). Further, the precision estimation curves based on the Birnbaum Score Information functions are mostly higher than the precision curves based on the theta-hat estimations done by simulation of the CAT-ASVAB AR test and subsequent estimation information by (1). There are two main reasons for this phenomenon. First, in the Score Information computation (2), we are using all 45 items comprised of the corresponding form of the test, but every CAT-ASVAB AR test is only 15 items long, which is enough for an estimation of theta-hat of a particular simulee. Second, in the CAT-ASVAB simulation, all items in the form are subject to an exposure control which is even stricter (as it will be explained below) than the usual exposure control developed by the Sympson-Hetter (Hetter & Sympson, 1985) algorithm.

Goal Approach to Form Assembly With a Weighted Information Functions

One of the major problems in test theory application is the creation of some number of parallel test forms with given test precision out of the given set (item bank) of n items. As in the case of conventional paper-and-pencil tests, we assume that there are the goal information curves $G(\theta)$ which should be reached, or exceeded, by any information curves for a particular form to provide needed

precision for the test (Van der Linden, 1998; Krass & Thomason, 1999). Further, the different forms should be parallel because different tests should estimate ability of an examinee coherently and measures of parallelism between different test forms are estimated by the closeness of their information curves. As an information goal criterion, we chose the Birnbaum theta-hat information function $I_B^j(\hat{\theta}\,|\,\theta), \quad j=1,\ldots 4$ of the existing CAT-ASVAB forms. To be more specific, as an information goal $G(\theta)$, we chose the information curve of CAT1 $I_B^1(\hat{\theta}\,|\,\theta)$ of the correspondent CAT-ASVAB test, the precision of which has been tested and approved in the many years of use.

Because the Birnbaum information function is not linear, we cannot apply a Linear Programming technique for the assembly of individual test forms (Van der Linden, 1998), but rather, we develop an heuristic "Bottle-Neck" or goal approach (which was also partially applied for assembly of previous paper-and-pencil forms [Krass & Thomason, 1999]).

The Experimental Score Information function (1) is not additive, as in the case of a Test Information function (4). Nevertheless, if an item with an ICC $p_i(\theta)$ has a large positive derivative $p_i(\theta)$ in the neighborhood of some point $\theta_0 \in [-3,+3]$ and its variation $p_i(\theta)\cdot(1-p_i(\theta))$ of theta-hat estimation is small, then adding this item to the existing items included in the form will increase the value of the Score Information of the form in θ_0. Really, if the CAT test that is derived from the given item bank is not divergent (Krass, 2000), then without randomization caused by exposure there is a unique sequence of items $i_j(\theta);\ j=1,\ldots,k$ which gives trajectory of the CAT test for an examinee with true ability θ. Then for this sequence $\{i_j(\theta);\ j=1,\ldots,k\}$, we can apply formula (4) for estimation information for an examinee with ability θ. If the new item has considerably more information at θ it can "kick out" an old item from the sequence $\{i_j(\theta);\ j=1,\ldots,k\}$ in the CAT selection process, substitute the new one, and increase information for $\hat{\theta}$ estimation conditional on θ. In the opposite case, the information for this estimation will remain the same as old value. The randomization slightly smooths this effect. This small observation leads to the formulation of our iterative heuristic algorithm.

Let $J_{i,t} \in \{1,\ldots,n\}$ be the set of items included in the form i on the iteration number t. (In this paper we consider the case of just two forms ($i=1,2$), though all discussion can be easily generalized to any number of needed forms.) Let $E_{i,t}(\theta)$ be the estimation of the Birnbaum information with respect to the theta-hat estimation of θ for the form $J_{i,t}$, which we call the Experimental Information for form i on iteration t. As we have already described, the computation of Experimental Information requires the estimation of exposure rates for any item in the form $J_{i,t}$ and subsequent extensive simulation. This is a rather computer-intensive computation, which takes about 4 minutes on a 600-MHz PC with 128 Mgb memory.

We begin the process from an empty set ($J_{i,0}=\varnothing, i=1,2$). On iteration t we found the most "troubled" or "Bottle Neck" $\bar{\theta}$ that which maximizes the function $[G(\theta)-E_{i,t}(\theta)]^+$, where $[x]^+ = x$, if $x>0$, and $[x]^+ = 0$, otherwise. Let this maximum be reached at $\bar{\theta}$ for form $i=1$. Then we found from the set $\{1,\ldots,n\}-J_{1,t-1}-J_{2,t-1}$ of "not used" items, the item which best fitted to the Bottle Neck theta in such a way that value

$$\hat{I}_k = \alpha \cdot I_k(\bar{\theta}) + (1-\alpha) \cdot I_k(\theta_{\max}) \tag{7}$$

be maximum among all available items. Here weight $\alpha \in [0,1]$ is the tuning parameter, depending on the test (for example $\alpha = 0.5$ for AR test), and $\theta_{\max}$ is value of θ which maximizes the item k Fisher information (5). By choosing this method of best fitting to the ability $\bar{\theta}$, we try to emphasize the importance of increasing the Score Information, not only in the "Bottle Neck" point of $\bar{\theta}$, but also in the overall height of Score Information for the form. Let item k_0 provide the maximum in (7) for the form $i = 1$; then in the form $i = 2$ we can find item k_1 whose ICC is closest to the ICC of item k_0 among all the items in the set of "not used" items. (For the closeness of two curves $f(\theta), g(\theta), \quad \theta \in [-3,+3]$ we are estimating, as usual, by their distance in L^2, or $d(f,g) = \sqrt{\int_{-3}^{+3} (f(\theta) - g(\theta))^2 \, d\theta}$).

Generally speaking, if the computer is fast enough, this process should soon finish iteration t and we would proceed to the next iteration. But as we explained before, the process of estimating the Experimental Information, $E_{i,t}(\theta)$, is very "expensive," so instead of computing the exact value $E_{i,t}(\theta)$, we, for the next τ steps, compute "approximate" Experimental Information:

$$\tilde{E}_{i,t+k}(\theta) = \sum_{l=0}^{t} w_l \cdot \hat{I}_l + E_{i,t}(\theta), \tag{8}$$

where $w_l = \sqrt{\frac{m}{M}}$, m is the number of items currently in the form (number of items in the set $J_{i,t}$), and M is the maximum number of items for the form. The value $w_l \cdot I_i(\theta)$ we call the "Weighted Information" of the item i. Multiplier $w_l < 1$ scales down the Likelihood Information of the item, which allows us to "compare," or put together, both the Fisher item information, and Birnbaum test information for the estimation. Usually we choose the size of approximation step $\tau = 10$, but it depends of the speed of computing and the type of the test.

After τ approximate steps, we compute the value of the Birnbaum information and go on to the next iteration. This process continues until it exhausts the given item bank or the next iteration does not essentially increase values of the Experimental Information curve.

As we have explained, before computing the Experimental Information $E_{i,t}(\theta)$ for form i on iteration t, we estimate the value of exposure parameters for the items included in the form $J_{i,t}$. The estimation is done with the Sympson-Hetter algorithm (Hetter & Sympson, 1985) in such a way that after the application of exposure control, the usage coefficient of the items will be not more than 0.33 for the CAT-ASVAB tests AR, WK (Word Knowledge), PC (Paragraph Comprehension), MK (Mathematical Knowledge), (the basic ASVAB tests for selection into military service), and 0.66 for all other tests in the battery (for classification into military job training). Originally, the upper bound of the exposure control coefficient was accepted as 1.0, which meant that if the CAT selection algorithm chose the item with $h_i = 1$, the item would be presented to the examinee as his/her next item for the test. However, if an item had an exposure coefficient of $h_i < 1$, and the item was chosen by the selection algorithm, it would have been presented to the examinee with the probability h_i.

Beginning with the CAT-ASVAB Form 3, the upper bound of exposure control coefficients was set at 0.7 to take into account that the Sympson-Hetter algorithm estimate of exposure control was set for the population of examinees with a normal $N(0,1)$ distribution of abilities, whereas in real life the distribution very often differs far from the normal distribution. In this study we apply a "mixed" approach, using the upper bound 1.0 if the item difficulty is $|b_i| < 0.5$, and using the upper bound 0.7 for all other items. We found that this approach saved the target usage for a majority of available items for the experimental distribution of abilities and was not so strict as with the upper bound 0.7 applied universally. Of course, the application of exposure control decreases the height of the Experimental Information curves.

In Figure 2 we present three views of consequent development of two forms with increasing number of iterations out of an existing item bank for the CAT-ASVAB WK test, where CAT Form 1 WK is the goal information curve. Simultaneously with the Experimental Information curve, we put Weighted Information curves based on the Fisher information (see above). As we can see, the Weighted Information curves, which are used for the estimation of information in intermediate steps, are rather good approximation of the final curves.

Because the described mechanism is able only to add new items to an existing form without taking out unused items, after the final iteration we take out "not used" items from the forms. An item is defined as "not used" if its usage is less than the given value (lesser than 0.001 in the case of ASVAB tests) for any of the 10,000-examinee population whose ability is subsequently fixed on one of the 31 levels of theta range. For example, in the above case, after iteration 202 we took away unused items, which left only 156 and 155 items per form, not 202 as it should be after 202 iterations. The process of removing unused items does not change the Experimental Information curves because the usage of unused items is negligible.

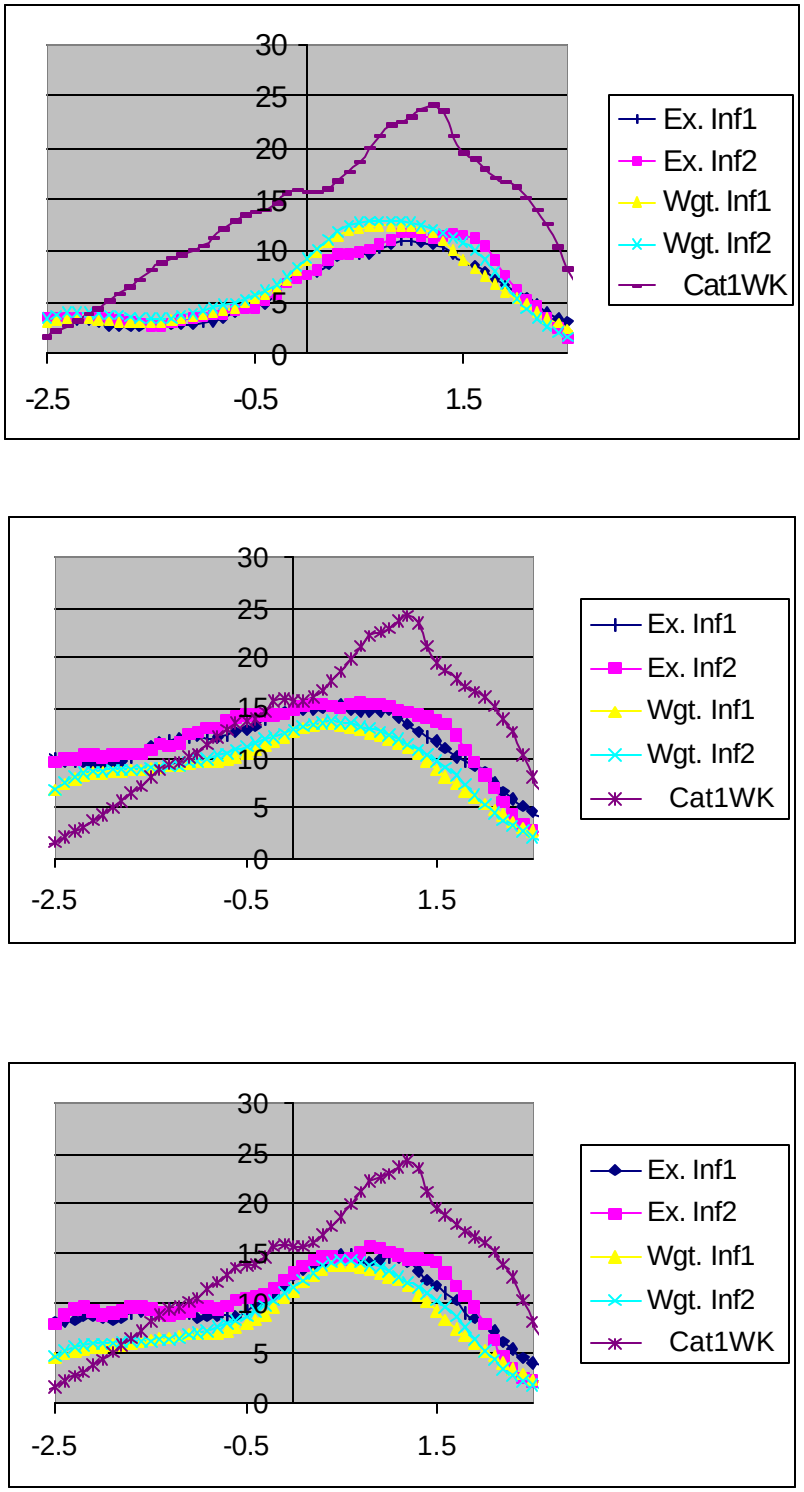

FIGURE 2. Development of the form for CAT-ASVAB WK test: 41, 71, and 201 iterations.

Another “unpleasant” phenomenon in the above process is the deterioration of parallelism for the created forms. Even in the body of the algorithm, we try to add to different forms “parallel” items; this decision is done locally so it is not surprising that the final curves can be rather unparallel. To make the forms more parallel we apply a "swapping" mechanism. We identify the $\bar{q}$ point where the forms are most different and exchange items with different values of criteria (7) to make the final forms more close from the point of view of their precision curves. In Figure 3 we show the result of ten applications of the swapping mechanism to the final forms of the CAT-ASVAB WK test shown in Figure 2. As we can see, the forms after swapping are much more parallel.

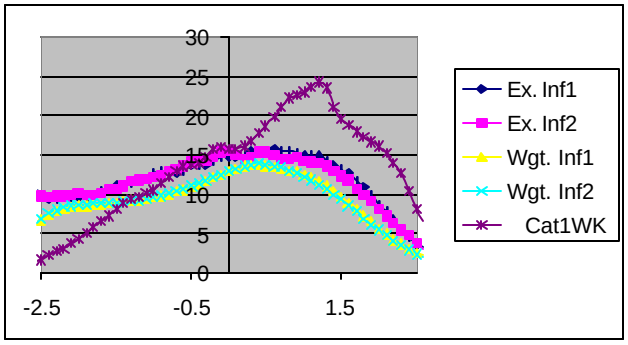

FIGURE 3. CAT-ASVAB WK forms after swapping items.

In Figure 4 we present the result of parallel-intended swapping for the CAT-ASVAB PC test. On the left side of Figure 4 we show graphs for two forms of the PC test after all the assembly algorithm iterations. Forms contain 104 and 105 items correspondingly, but they do not look very parallel. On the right side are the same forms after four applications of swapping. The new forms look more parallel than those on the left side. Note, however, the number of items in the final forms, after swapping, equals 107. This number of item increase happens because the swapping algorithm can use not only items currently chosen for the forms, but also items not used in the forms. Thus, in this case, the swapping algorithm brought in two and three new items from the “unused” set. Also, in the case of this PC test, we can see in Figure 4 that the Weighted Information curves are considerably lower than the Experimental Information curves. This means, that for this case, the “Scale” constant $w_l = \sqrt{\frac{m}{M}}$ is too small and requires adjustment, possibly by decreasing the constant M for this test in the formula.

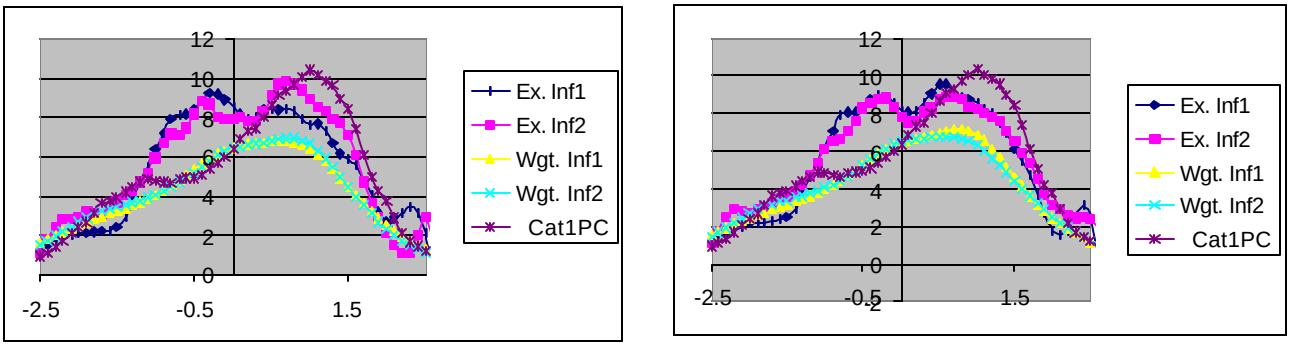

FIGURE 4. Improving form parallelism for the CAT-ASVAB PC test.

Reaching Precision Goal in "Generalized" Iterations

The main resource of new items for the CAT-ASVAB is a seeded-item mechanism that is imbedded in the CAT algorithm (Segall, Moreno, & Hetter, 1997). Every four or five months, our group of editors create batches of 100 items per test. After calibration with on-line tryout data, and rejection of deficient items (Krass & Thomasson, 1999) we use all available (to the given moment) items to assemble two preliminary forms to see if we have already reached our Precision Goal.

If the Precision Goal curve is not reached, we tell editors where on the scale it is especially critical to get needed precision, and the editors focus on those needs as they create new batches of items. In Figure 5 we present two stages of CAT-ASVAB MC test development. On the left side is an early stage, and we can see that the new form information curves are lower than the information curves of the CAT-ASVAB Form 1 MC test in the area of positive abilities (area of higher-scoring examinees). The editors therefore focused on writing more difficult MC items, and as is shown on the right side of Figure 5, the information curves for the next stage of form development nearly reached Goal Precision. (Although the goal curve is not yet reached in the area of highest ability $\theta > 2.0$, this is not a big concern for test developers because there are only few examinees in this high-ability area.)

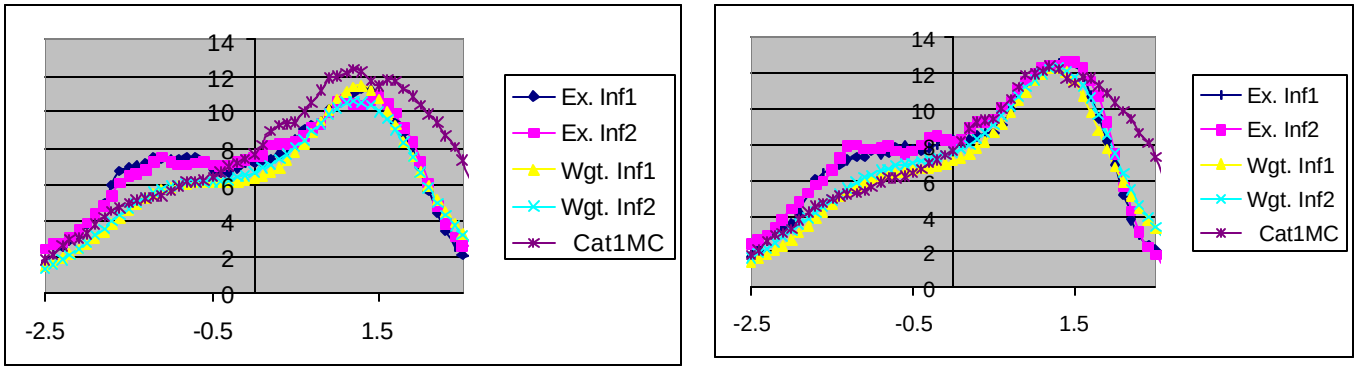

FIGURE 5. Two stages of development for CAT-ASVAB MC test.

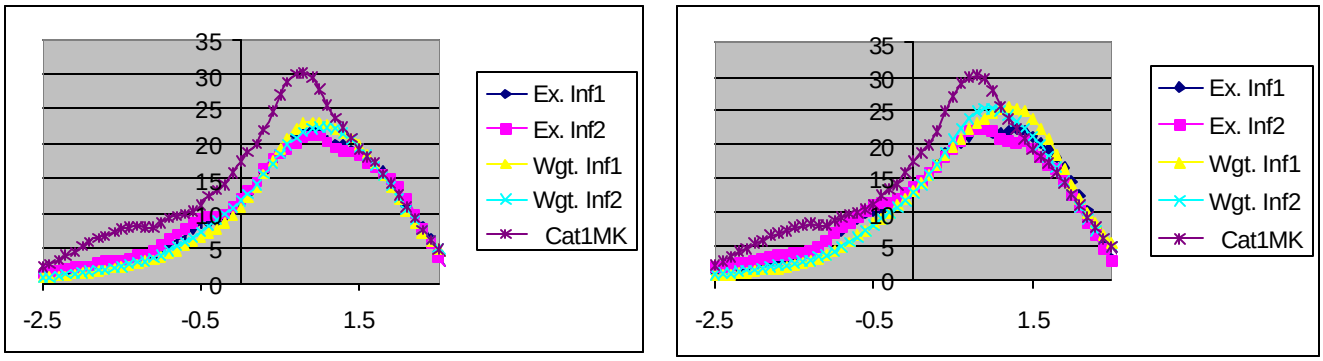

FIGURE 6. Subsequent stages of development for CAT ASVAB MK test.

In Figure 6 we present the analogous case for the CAT-ASVAB MK test. On the left side, we see the information curves are below the Precision Goal curves in the area $\theta < 1.5$ for the preliminary stage of the test form. The editors developed a new batch of items which begins to correct the deficit. Note, however, in the area of $-2.5 < \theta < -1.0$ and $-0.5 < \theta < +1.0$ we still need more informative new items. These shortages may be taken care of with the next set of new items.

References

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.)*, Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Fisher, R. A. (1925). Theory of statistical estimation*. Journal of the Royal Statistical Society, Series B 1,.* 175–185.

Hetter, R. D., & Sympson, J. B. (1985). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing* (pp. 141-144). Washington, DC: American Psychological Association.

Kendall, M. G., & Stuart, A. (1973). *The advanced theory of statistic*. (Vol. 2) New York: Hafner

Krass, I. A., & Thomasson, G. (1999). *Defining deficient items by IRT analysis of calibration data.* Paper presented at the annual meeting of the National Council on Measurement in Education in Montreal, Canada.

Krass, I. A. (2000). *Change in distribution of latent ability with item position in CAT sequence.* Paper presented at the annual meeting of the National Council on Measurement in Education in New Orleans, LA.

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

Owen, R. J. (1975). A baysian sequential procedure for quantal response of adaptive mental testing. *Journal of the American Statistical Association, 70, 351-356.*

Segall, D. O., Moreno, B. M., & Hetter, R. D. (1997). Item pool development and evaluation. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing* (pp. 131-140). Washington, DC: American Psychological Association.

Serfling, R. J. (1980). *Theorems of mathematical statistics.* New York: Wiley.

Van der Linden, W. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement, 22*(3), 195-211.