Graduate
Management
Admission
Council®

Computerized Adaptive Testing by Mutual Information and Multiple Imputations

Anne Thissen-Roe
Kronos

Presented at the CAT for Classification Paper Session, June 2, 2009

Abstract

Over the years, most computerized adaptive testing (CAT) systems have used score estimation procedures from item response theory (IRT). IRT models have salutary properties for score estimation, error reporting, and next-item selection. However, some testing purposes favor scoring approaches outside IRT. Where a criterion metric is readily available and more relevant than the assessed construct, for example in the selection of job applicants, a predictive model might be appropriate (Scarborough & Somers, 2006). In these cases, neither IRT scoring nor a unidimensional assessment structure can be assumed. Yet, the primary benefit of CAT remains desirable: shorter assessments with minimal loss of accuracy due to unasked items. In such a case, it remains possible to create a CAT system that produces an estimated score from a subset of available items, recognizes differential item information given the emerging item response pattern, and optimizes the accuracy of the score estimated at every successive item. The method of multiple imputations (Rubin, 1987) can be used to simulate plausible scores given plausible response patterns to unasked items (Thissen-Roe, 2005). Mutual information can then be calculated in order to select an optimally informative next item (or set of items). Previously observed response patterns to two complete neural network-scored assessments were resampled according to MIMI CAT item selection. The reproduced CAT scores were compared to full-length assessment scores. Approximately 95% accurate assignment of examinees to one of three score categories was achieved with a 70%-80% reduction in median test length. Several algorithmic factors influencing accuracy and computational performance were examined.

Acknowledgment

Presentation of this paper at the 2009 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Copyright © 2009 by the Authors

Citation

Thissen-Roe, A. (2009). Computerized adaptive testing by mutual information and multiple imputations. In D. J. Weiss (Ed.), ***Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.*** **Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/**

Author Contact

Anne Thissen-Roe, Kronos, 9525 SW Gemini Dr., Beaverton, OR 97008, U.S.A.
Email: anne.thissenroe@kronos.com

Computerized Adaptive Testing by Mutual Information and Multiple Imputations

The purpose of this paper is to describe a general system of computerized adaptive testing (CAT) for classification decisions, capable of accommodating assessment scoring algorithms of arbitrary dimensionality and mathematical form. Assessments developed and used by Kronos, a third-party provider of field hourly hiring solutions, are used to illustrate both the need for and the function of such a system. The system is not specific to selection testing, but is suited to it.

In the systematic selection of job applicants, a primary concern is the predictive validity of the selection method used. That is, if some applicants are favored for selection above others, those applicants should be, as much as possible, those who will perform better on the job in meaningful ways. A pattern of higher performance can lead to quantifiable cost savings or productivity increases. The economic value of the selection method is influenced jointly by the strength of the relationship between the selection method and a specified type of performance, the value of individual differences in that type of performance, and the fraction of the available population selected by the method (see, e.g., Hunt, 1995, p. 69-79).

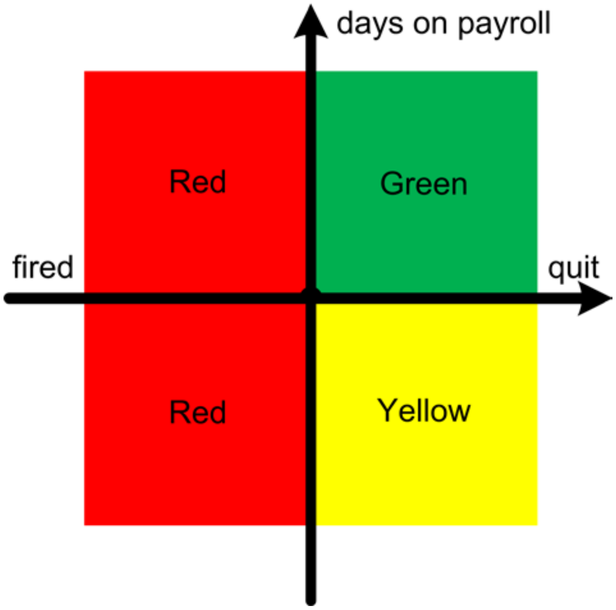
For hourly jobs, individual performance is often straightforward and well-measured. For example, the most salient individual performance dimension for a sales associate is typically *sales*, be it measured in units sold per month or revenue per hour. If the sales associate receives a commission, or if other incentives depend on meeting individual sales goals, an appropriate measure of sales must be tracked consistently for administrative and monitoring purposes. For other positions, such as entry-level service employees, high productivity is less important than low counterproductivity; counterproductivity is reflected in payroll records of involuntary terminations. An organization might study the validity of a selection method by opportunistic use of such objective metrics, as well as ratings of "soft" skills provided by an employee's manager or peers.

An assessment given to job applicants can be used as a selection method. If an assessment is used, there is a mathematical function that relates any given pattern of assessment item responses to a score; that score then has a statistical relationship to performance, which is its validity. Score validity can be optimized for a particular set of items if the mathematical function—the scoring algorithm—is based on a sufficiently complex and robust predictive model.

The best predictor of a job performance dimension might not be a single individual difference variable measurable at the time of application, no matter how well that variable is measured. Concrete business outcomes such as sales, counterproductivity and retention are influenced by multiple individual behaviors as well as environmental factors. To predict these outcomes, Kronos has adopted an empirical approach. The Kronos Service Reliability Assessment, for example, uses neural networks to model the relationships between assessment response patterns and two business outcomes: days on payroll and voluntary vs. involuntary termination. Some individual assessment items are direct neural net inputs; in other cases, groups of related items are collapsed into latent trait estimates in order to improve measurement quality and avoid collinearity of model features. The assessment's measurement space is arbitrarily multidimensional, and the scoring algorithm is of arbitrary mathematical form. Large samples, including over half a million employees in total, and rigorous cross-validation designs are used to ensure that the scoring algorithm is of the highest possible validity given the set of items administered.

Selection assessments are decision support tools. Scores on an assessment contribute, often alone, to the classification of a job applicant as "passing," "failing," or similar values. The Service Reliability Assessment scores applicants as "green," "yellow," or "red," with green representing the most positive recommendation given to a hiring manager. Classifications need not represent ordered segments of any single dimension; indeed, the Service Reliability Assessment distinguishes "red" from "yellow" and "green" based on its termination type prediction, and "yellow" from "green" based on its retention prediction (see Figure 1).

Figure 1. An Example of Red, Yellow and Green Score Zones Based on Two Underlying Variables



Where scores are reported only as classifications, business utility comes from the differences in outcomes of examinees assigned to each classification. If "green" applicants outperform "red" applicants, for example, the test has practical utility as well as valid underlying measures or scores. Then, the validity of a scoring algorithm implies savings or revenue for an organization only through the value of a higher classification relative to a lower one. Differences between individuals within a classification take part in neither the selection decision nor the value proposition.

As with other classes of assessment, the time spent by job applicants on answering items is valuable. It is desirable to administer a short assessment, if possible, while maintaining high validity. The set of computerized adaptive testing (CAT) methods produce a highly informative set of items tailored to each individual examinee, rendering a short assessment which is at least as accurate as, or more accurate than, any single subset of items can be. However, traditional CAT methods make assumptions about the form of the scoring algorithm which are not compatible with an arbitrary, multidimensional approach to scoring.

A modified, generally applicable method for next-item selection is needed. Such a method must permit multidimensionality in the measurement space, should take advantage of the classification form of reported scores, and must not depend on the particulars of the measurement model, the scoring algorithm, or the partition procedure for reporting classifications. Items might be unidimensional, grouped into non-overlapping scales or subtests, or a multidimensional "cloud," high on unique variance. Classifications might represent strictly ordered ranges of values on a single dimension, or they might not. Traditional CAT and computerized classification testing are not flexible enough to address all of these possibilities.

Several useful results have been discussed previously by other authors. Segall (1996, 2000) considered CAT for a multidimensional latent trait space. He suggested that item selection algorithms for items in such a space should minimize the credible region of true scores. Under certain assumptions about the posterior distribution of true scores given a set of item responses, such as a multivariate normal approximation, Segall proposed a simple calculation permitting next-item selection. Segall's method will be used as a comparison standard upon whose efficiency the present study seeks to improve.

Another class of statistic that holds promise for next-item selection derives from information theory (Shannon, 1948; Kullback & Leibler, 1951). Information theory concerns itself with the relationships of distributions of discrete values, a set which includes classification decisions as well as item responses. Shannon, Kullback and Leibler extended the work of Fisher (1925), whose conception of information has been used in CAT item selection for decades (Samejima, 1977).

A number of recent authors have championed the use of information theory metrics in item selection, particularly for computerized classification testing and cognitive diagnosis, wherein the ultimate scores are dichotomous or polytomous. Chang and Ying (1996) used Kullback-Leibler divergence (Kullback & Leibler, 1951), or mutual information, in their global information method. Xu, Chang and Douglas (2003) studied both Kullback-Leibler information and entropy-based (Shannon, 1948) methods for cognitive diagnosis. The two information theoretic methods are closely related both algebraically and historically (Kullback, 1959; Cheng, 2009). Similarly, Weissman (2003, 2007) selected items for a computerized classification test using mutual information. Although he studied unidimensional assessments, he noted that mutual information item selection methods could be applied more generally, including to multidimensional assessments. Several additional studies have since then extended mutual information methods to multidimensional CAT (e.g. Diao & Reckase, 2009; Wang & Chang, 2009).

In parallel, Chambless and Scarborough (2001) used methods from information theory, including a measure of entropy and mutual information to select items singly and jointly for a behavioral assessment scored with a neural net. Although Chambless and Scarborough's work was aimed at constructing a static assessment, not an adaptive test, their research established the suitability of information theory methods for selecting important input variables to an arbitrarily structured nonlinear predictive model in the industrial selection domain. Chambless and Scarborough also demonstrated the feasibility and utility of selecting multiple complementary items simultaneously through information theory.

Finally, the method of multiple imputations (Rubin, 1987) may be used to estimate the probable distribution of full-test scores available to a mid-test examinee, for the purposes of next-item selection and the application of stopping rules (Mislevy et al., 1992; Thissen-Roe, 2005). Unobserved responses to unasked items can be considered a form of nonresponse to a survey which consists of the full-length assessment. Whether an item is asked is under the control of the CAT program, and depends only on observed variables; therefore, nonresponse to unasked items is *ignorable nonresponse*. The assessment may be provisionally completed $m>2$ times by sampling random responses to each unasked item from an appropriately constructed distribution informed by the observed responses. From these, scores may be generated by the usual algorithm, and the sensitivity of the score to variability in responses determined.

CAT by Mutual Information and Multiple Imputations

Employers ask for, and benefit from, selection tests that produce systemic savings or revenue, given as small an investment in items administered, and therefore applicant time, as possible. The present study assumes that an assessment exists with adequate validity and business utility, but that the assessment is considered too long.

A system is needed which can administer shorter tests with minimal loss of classification accuracy due to unasked items, regardless of the dimensionality or mathematical form of the scoring algorithm. Two assumptions will be made:

1. The score is a discrete category, nominal or ordinal.
2. The degree to which the score changes when a particular item response is given varies based on responses to other items.

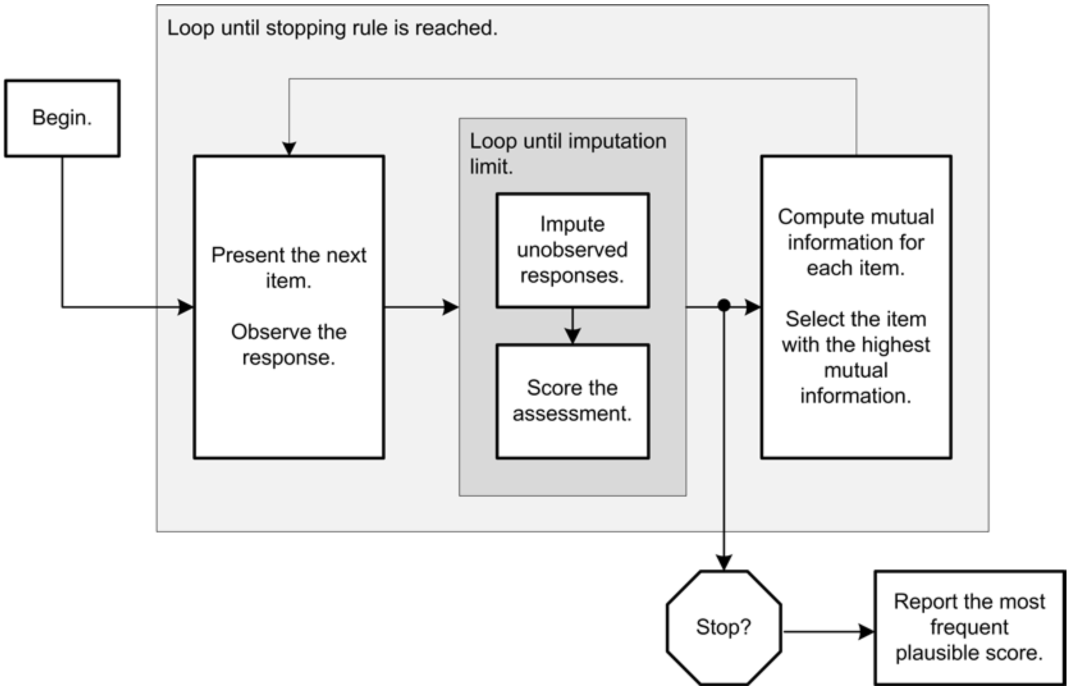
Classification accuracy, in this case, refers to the congruence of the classification made by the short test with the classification made by the full-length test. Minimizing the loss of classification accuracy protects the business utility of the assessment.

It is worth noting that if there are one or more underlying scales that contribute to the classification, their validity is not optimized; to maximize scale validity requires minimizing the loss of precision at all points in score space, even those points which are not near category boundaries. That is, the correlation between the short test's estimate and the full-length test's scale score must exist *within* as well as between classifications. Segall's credible region method is a system that minimizes error equally at all points in the scale score space, and therefore minimizes loss of scale validity. Such a solution, while elegant, is not necessary to produce savings or revenue for an employer. Imprecision that does not affect the classification cannot affect the hiring manager's decision, and therefore cannot affect the business utility of the system.

It is possible to create a CAT system that satisfies the requirement of optimizing classification accuracy at every successive item. Combined with an appropriate stopping rule, such a system can shorten the administered assessment considerably while retaining high classification accuracy. The system is capable of producing an estimated score from any subset of available items, recognizing differential item information given the emerging item response pattern, and selecting the most informative next item given that information.

One such system combines the method of multiple imputations with mutual information item selection. Figure 2 shows a simple mutual information/multiple imputations (MIMI) CAT algorithm. It consists of an item administration loop with an imputation loop nested inside it.

Figure 2. Architecture of an Algorithm for Item Selection by Mutual Information and Multiple Imputations


Loop until stopping rule is reached.
Begin.
Present the next item.
Observe the response.
Loop until imputation limit.
Impute unobserved responses.
Score the assessment.
Compute mutual information for each item.
Select the item with the highest mutual information.
Stop?
Report the most frequent plausible score.

The algorithm works as follows:

1. An item is administered and a response is observed.
2. Unobserved responses are imputed at random, according to an appropriate distribution, a specified large number of times, generating *plausible response patterns*. An appropriate distribution for response imputation might be static population response frequencies to each item, a full item model that takes account of observed responses, or anything in between.
3. Each plausible response pattern is scored, generating a set of *plausible scores*. Plausible scores are indexed and classified according to the plausible responses that yielded them.
4. Stopping rule conditions are tested. If the stopping rule applies, the most frequent plausible score is reported and the assessment ends.
5. The mutual information of each unobserved item is calculated based on the plausible response patterns and scores. The item yielding the highest mutual information is selected.
6. Loop back to step 1 with an item selected for administration.

In a MIMI CAT system, the box reading "score the assessment" is singular and unelaborated. No details of the scoring algorithm need be reflected in the remainder of the system. The only true requirement is that it produce a score given any complete response pattern.

Any scoring algorithm can be used with MIMI CAT. However, extremely simple scoring algorithms such as sum scores will derive no benefit from CAT; summed items will be consistently ordered based on population parameters, regardless of those responses already observed. MIMI CAT is most effective when used with a scoring algorithm in which some or all items possess variable degrees of influence on the score depending on the responses to other items.

The degree of an item's influence on the score is calculated in the form of mutual information. Although earlier mutual information item selection algorithms have analyzed the information in continuous distributions (Chang & Ying, 1996; Xu et al., 2003; Weissman, 2003; Weissman, 2007; Cheng, 2009; Diao & Reckase, 2009; Wang & Chang, 2009), it is straightforward to calculate mutual information from discrete data.

The equation for mutual information is

$$I(\vec{x},\vec{y}) = \sum_{x\in\vec{x}}\sum_{y\in\vec{y}} p(x,y)\log\left(\frac{p(x,y)}{p_x(x)p_y(y)}\right) \tag{1}$$

It can be computed from the observed marginal distribution $p_y(y)$ across plausible scores, the observed marginal distribution $p_x(x)$ across plausible responses to one item, and the joint distribution $p(x,y)$ of plausible item responses and plausible scores. The more the joint distribution varies in each cell from the product of the marginal distributions, the greater the information contributed by that item.

MIMI natively handles polytomous scoring; to distinguish three or more categories, ordered or unordered, requires no more assumptions and no more complicated mathematics than to distinguish "pass" from "fail." Weissman (2007) noted that this was an advantage of mutual information over methods historically used in classification testing, which tend to treat each partition separately.

The sequential probability ratio test, or SPRT (Wald, 1947), is the usual choice of stopping rule for non-fixed-length classification testing. MIMI can be used with the SPRT, but the two are not perfectly matched. MIMI does not evaluate the difference between responses at particular points on either side of a cut score, but sampled across the whole category. MIMI is therefore more closely matched to an integrative composite likelihood ratio (Thompson & Ro, 2007; Thompson, 2009; Weitzman, 1982).

MIMI also enforces its own stopping rule; it will not select a next item when all imputed response patterns produce the same score. At that point, mutual information of any further item is estimated at zero. The nominal error rate should then be the inverse of the number of imputations, but that is not the error rate in practice, as the studies that follow demonstrate.

Study 1: Service Reliability Assessment

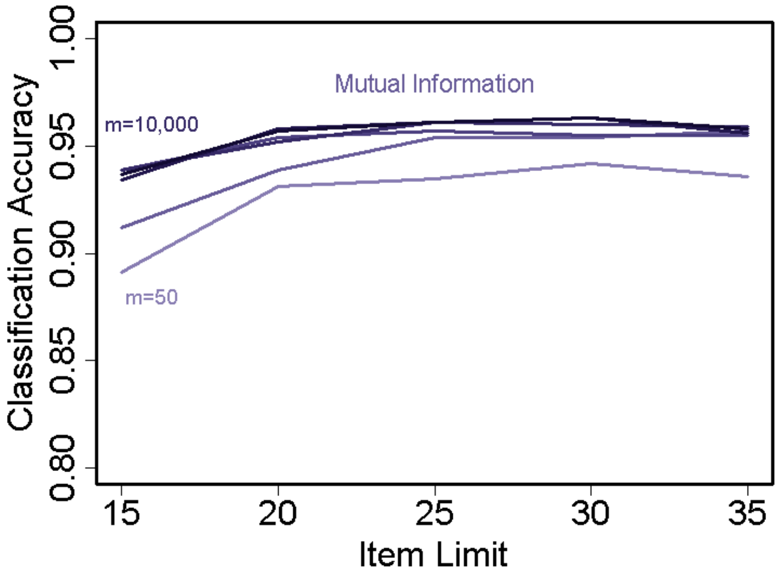
Method. One thousand applications including responses to all 65 Service Reliability Assessment items were selected at random from a large population. Applications represented the full range of possible scores; most did not result in hires. These applicant responses were resampled according to the MIMI CAT algorithm; a subroutine simulated an applicant taking the CAT form of the assessment by providing each given answer when the corresponding question appeared. Although each real applicant had actually responded to every item, only responses to those items that would have been administered were provided to the scoring algorithm. Responses to "unasked" items were masked. The resulting final scores were compared to scores produced by the full 65-item scoring algorithm in order to determine classification accuracy.

The MIMI-enforced stopping rule was used in combination with a "hard" item limit. Repeated simulations varied the item limit and imputation count used, as well as substituting Segall's credible region method (1996; 2001) for comparison to MIMI.

Results. Figure 3 shows the relationship observed between item limit, imputation count, and classification accuracy. For at least 20 allowed items and 500 imputations, a performance asymptote around 95-96% classification accuracy was observed. Lower item limits and lower imputation counts both resulted in decreased classification accuracy.

Low imputation counts (such as 50 imputations per item administered) most likely reduce classification accuracy through raising the nominal error rate of the stopping rule; 50 of the same wrong score might occur by chance. However, the observed error rate is in all cases higher than the nominal error rate, suggesting an additional source of error.

Figure 3. Item Limit, Imputation Count, and Classification Accuracy on the MIMI CAT Version of the Service Reliability Assessment. Imputation Count is Shown as *m*={50, 100, 500, 1,000, 5,000, 10,000}, With Darker Lines Representing Higher Imputation Counts

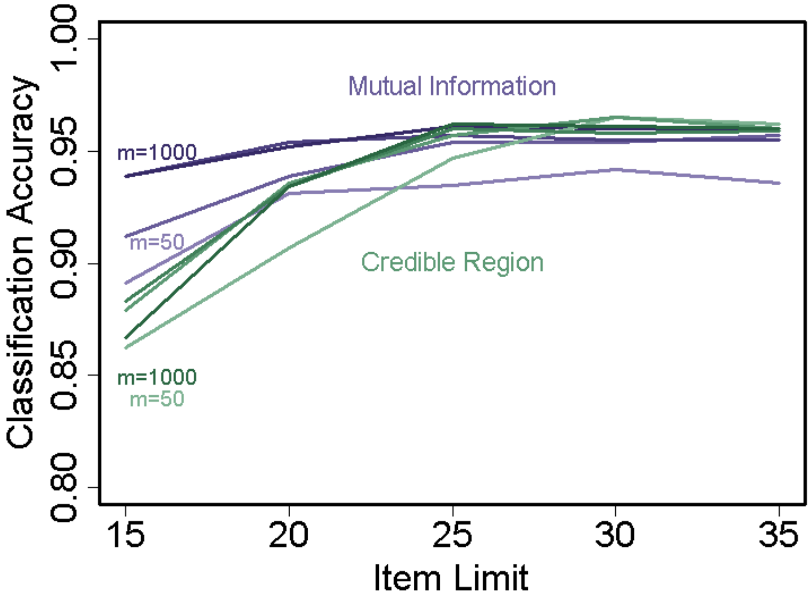


For comparison, Figure 4 shows a comparison between MIMI CAT (in blue) and Segall's credible region method (in green), over several item limits and imputation counts. When the credible region method is adapted to use empirical posterior distributions based on the method of multiple imputations, rather than a parametric form, its distinction from MIMI amounts to using the estimated determinant of the posterior covariance matrix of scores, rather than mutual information, to select the next item.

The credible region method exhibited the same 95-96% classification accuracy asymptote as MIMI CAT; for item limits at or above 25 and imputation counts above 100, the methods were equal. At 50 imputations and above 25 allowed items, the credible region method outperformed

MIMI; the credible region method is used in fixed-length CAT and is not subject to premature stopping when few imputations are used. MIMI, on the other hand, is less susceptible to low item limits; it showed 3-9% greater classification accuracy when allowed only 15 items, depending on the number of imputations.

Figure 4. Item Limit, Imputation Count, and Classification Accuracy on the MIMI CAT Version of the Service Reliability Assessment, Compared to a Credible Region CAT Version of the Same Test. Imputation Count is Shown as *m*={50, 100, 500, 1,000}, With Darker Lines Representing Higher Imputation Counts



Study 2: Reliability Assessment

The methods and outcomes of Study 1 were replicated on a second assessment, the Reliability Assessment. The Reliability Assessment is structurally similar to the Service Reliability Assessment, using a neural network core algorithm to predict the same two outcomes, termination type, and days on payroll. However, the Reliability Assessment is shorter and applicable to different jobs.

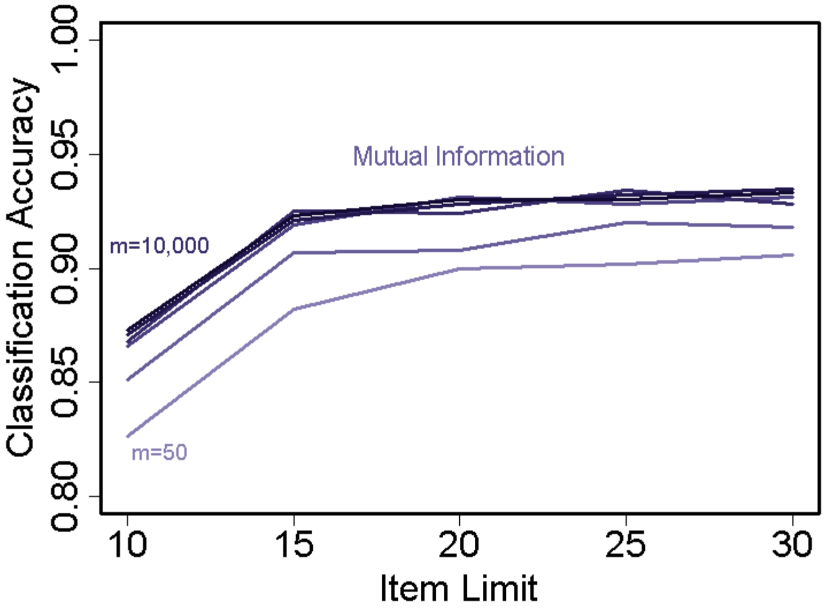
Method. One thousand applications including responses to all 35 Reliability Assessment items were selected at random from a large population. As in Study 1, applications represented the full range of possible scores; most did not result in hires. These applicant responses were resampled according to the MIMI CAT algorithm following the Study 1 design. The resulting final scores were compared to scores produced by the full 35-item scoring algorithm in order to determine classification accuracy.

The MIMI-enforced stopping rule was used in combination with a "hard" item limit. Repeated simulations varied only the item limit and imputation count used, not the method of item selection.

Results. Figure 5 shows the relationship observed between item limit, imputation count, and classification accuracy. For at least 15 allowed items and 500 imputations, a performance asymptote around 92-93% classification accuracy was observed. As was observed for the Service

Reliability Assessment, lower item limits and lower imputation counts both resulted in decreased classification accuracy.

Figure 5. Item Limit, Imputation Count, and Classification Accuracy on the MIMI CAT Version of the Reliability Assessment. Imputation Count is Shown as *m*={50, 100, 500, 1,000, 5,000, 10,000}, With Darker Lines Representing Higher Imputation Counts



Design and Performance Considerations

There are several possible variations on the MIMI CAT system shown in Figure 2. Some of these affect the relationship between imputation counts and classification accuracy, the median number of items administered before the stopping rule is reached, or computational resource requirements. Two of these will be explored, within the context of Kronos's hourly hiring solutions.

Although classification accuracy is of primary concern, usage of computational resources greatly influences the cost of a MIMI CAT system. In our example, Kronos presents and scores assessments over the Internet, using a client-server architecture where most processing occurs on the server side. Under MIMI CAT, that processing includes scoring and next-item selection in between items, as well as at the end of the assessment. Sufficient computational resources must be allocated that delays do not occur during periods of high application volume. The computational performance differences described in the following section span nearly three orders of magnitude; Variation 2 might have an even greater effect. For the commonly used Service Reliability Assessment and Reliability Assessment, three orders of magnitude might well mean the difference between "one process on a shared server" and "a dedicated server farm." The cost of the latter is better avoided.

The CAT systems of previous decades commonly ran on standalone personal computers. If client-server systems were used, item selection and scoring depended on computations done on the client side, lest network delays impair system responsiveness. Today, test-takers understand that Web pages can take more than a second or two to load. Nonetheless, responsiveness and efficiency remain worthy of consideration.

The basic MIMI CAT architecture (Figure 2) has features that are well-suited to Internet administration within a client-server model:

1. *As noted previously, all item selection and scoring computation can be done on the server.* In Internet testing, it is risky to depend on the client system to run any program code. Client systems vary; hardware platforms, operating systems, Web browsers and configurations, security firewalls, parental filter systems, and inadvertently interfering third applications are many and diverse. The more complex the task the client system must perform, the more likely that some users cannot take the test at all. Therefore, it is beneficial to aggregate complex computations, such as item selection and scoring, on the known server architecture.
2. *Between item selection cycles, only the observed item-response pairs need be stored.* It is useful in engineering a CAT server application to think of an update—one scoring and item selection cycle— as the scope and unit of the CAT program, rather than an entire test. While the server is waiting for the client to return an item response, the CAT program need neither run continuously nor "remember" that the user is there. Preserving state variables of mid-test examinees, such as a representation of a posterior distribution, can drive up memory usage on the server. If some examinees never complete the assessment, their states might remain semi-permanently, resulting in a "memory leak." Mid-test states should be recorded as simply as possible in a permanent database, and derived variables should be regenerated as necessary.
3. *Data transmitted between client and server is limited to item content and the response identifier.* Network transmissions are costly and slow, and can dwarf local computations in terms of both efficiency and responsiveness. Efficient network transmissions, containing the bare minimum of information needed, both improve responsiveness and reduce the total cost of the system.

For these reasons, variations on the MIMI architecture will be considered with the secondary goal of maintaining the existing client-server separation.

Design Variation 1: Imputation Within Items

One possible variation pertains to the sampling of plausible response patterns across the distribution of responses to one particular unasked item. If all unasked item responses are imputed at random according to their modeled probabilities, and one or more item responses is rarely given, few plausible response patterns will include that item response. The distribution of plausible scores associated with that item response, then, will be sparse, and its parameters poorly estimated. Although the formula for mutual information has a built-in scheme to give more weight to frequently sampled cells, it was considered that deliberately oversampling rare responses might improve item selection performance.

Rare response oversampling was achieved by generating plausible response patterns and plausible scores separately for each response to each unasked item, with that response held fixed as if it had been observed. All remaining item responses varied according to the imputation model as before. The same number of imputations was used for each unasked item response. When the plausible score distributions for a particular unasked item were combined, they were uniformly distributed across possible responses to the target item. Mutual information was calculated according to the plausible score distributions generated for that item only.

It might be apparent that more plausible response patterns are generated and scored during each item selection cycle under this design than under the standard MIMI architecture. Oversampling was tested for effects on the classification accuracy, median stopping count, and computational performance of the Service Reliability Assessment and Reliability Assessment; its computational performance cost was also projected for a third assessment, the Frontline Healthcare with Retention Assessment, to explore the effect of a larger item pool. Comparisons were done at 100, 500, and 1000 imputations, and item caps of 15, 25, 35 and 45 for the Service Reliability Assessment; 100, 500, 1000 and 5000 imputations, and item caps of 15 and 25 for the Reliability Assessment.

For both of the assessments tested, the impact of imputation within items on classification accuracy was small. The classification accuracy of the Service Reliability Assessment ranged from 91% to 96% for standard MIMI CAT, and 90% to 97% for the oversampled design. The Reliability Assessment made greater gains; its accuracy ranged from 90% to 94% for standard MIMI CAT, and 94% to 96% for the oversampled design. For both assessments, the median examinee reached the stopping rule two to three items earlier under the oversampled design (see Table 1 for ranges).

Table 1. Performance Effects of a Design That Oversamples Rare Responses

	Assessment	Impute once	Impute within items
Median items administered	Service Reliability	14-16 (22-25%)	12-13 (18-19%)
	Reliability	10-11 (29-31%)	7-10 (20-29%)
Classification accuracy	Service Reliability	91-96%	90-97%
	Reliability	90-94%	94-96%
Plausible score generation cycles between first and second items	Service Reliability	1	256
	Reliability	1	136
	Frontline Healthcare with Retention	1	498-501
Plausible score generation cycles between 24th and 25th items	Service Reliability	1	164
	Reliability	1	44
	Frontline Healthcare with Retention	1	383-455

Both improvements, however, are overshadowed by the large effect of this design change on computational performance. For a complex nonlinear scoring model, the scoring of many plausible response patterns is by far the most computationally expensive phase of an item selection cycle. To score plausible response patterns separately for each remaining item response multiplies the computational requirements by the number of remaining item responses. The multiplier is largest at the beginning of the assessment, smallest at the end, and scales with the size of the item pool. For the Reliability Assessment, with an item pool size of 35 and four responses per item, the delay between the submission of the first item response and the presentation of the second item—when 34 items remained—was increased by a factor of 136. For the Service Reliability Assessment, with an item pool size of 65 and four responses per item, the delay increased by a factor of 256. Finally, for the Frontline Healthcare with Retention Assessment, with 124 items ranging from two to five responses per item, the delay increased by a factor of 498 to 501 depending on the first item selected.

In sum, imputation within items delivers a small improvement in number of items administered before the stopping rule is reached, and an improvement of up to four percent in classification accuracy. However, it imposes a high practical cost on larger item pools, which might be needed to increase the validity of the full-length assessment. It does not make sense to increase classification accuracy at the expense of potential validity, as the purpose of seeking high classification accuracy is to retain validity. Imputation within items is therefore of dubious value.

Design Variation 2: More Than One Unrelated Item Per Screen

In Internet testing, some users prefer that more than one item be displayed at a time. If items are not grouped into testlets or pages in advance of administration, an ideal CAT system will select the *n* jointly most informative items on each update cycle.

MIMI CAT is readily extended to *n*-item selection. Following Chambless and Scarborough (2001), mutual information can be calculated exhaustively for pairs, triples, or *n*-tuples of items in order to select those that are most informative in combination. Additional imputations are not necessarily required for this calculation; the existing plausible response patterns and plausible scores can be partitioned according to the full pattern of responses to the *n* items. If *n* is large, more imputations might be required to adequately sample all response *n*-tuples.

No significant change in classification accuracy was observed for the Service Reliability Assessment for $n = 2$, using an item cap of 20 and conditions of 500 and 1,000 imputations per update. Median test length increased to 16 when $n = 2$ from 14 (500 imputations) or 15 (1,000 imputations) when $n = 1$.

Computation time scales exponentially with *n* greater than about 3, if *n*-tuples are tested exhaustively for information. For small *n*, scoring the plausible response patterns, which is not affected, takes much longer than computing mutual information, which is. The exact value of *n* for which the computational requirements of the mutual information calculation exceed those of scoring depends on the complexity of the scoring algorithm.

Could a page of *n* items be assembled stepwise, rather than by exhaustive comparison of all possible combinations? The most informative item could be selected, followed by the next most informative item given the first, and so on. Computation time then scales linearly rather than exponentially with *n*. However, plausible response patterns must be partitioned separately for each step, resulting in a more complex data model, the memory requirement of which still scales exponentially. For small *n*, stepwise assembly is probably not practical; for larger *n* such as 4 or 5, it might be.

Summary

CAT by mutual information and multiple imputations (MIMI CAT) permits the administration of reduced-length classification tests with minimal loss of classification accuracy, regardless of the dimensionality or mathematical form of the underlying scoring algorithm. An item selection procedure envelops an arbitrarily defined scoring algorithm. Plausible responses to unasked items are imputed, the completed response patterns are scored, and the resulting dataset is used to calculate the mutual information contributed by each item about the score.

MIMI CAT versions of two neural network-scored assessments, each resulting in three possible classifications, were examined. Classification accuracies of 95% and 93% resulted from assessments reduced in length by 60-70% for all examinees, or up to 80% for the median examinee. The MIMI CAT design can be easily broken into client and server responsibilities, suitable for an Internet testing context, and is readily extended to the selection of multiple unrelated items for simultaneous presentation.

While specific results such as classification accuracy figures were not identical even between two similarly structured and similarly purposed assessments, the MIMI CAT architecture is generally applicable, and the general pattern of results observed might be replicable in other contexts.

References

Chambless, B. & Scarborough, D. (2001). Information theoretic feature selection for a neural behavioral model. *Proceedings of the International Joint Conference on Neural Networks of the IEEE,* Washington, D.C.

Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20,* 213-229.

Cheng, Y. (2009). Computerized adaptive testing for cognitive diagnosis. In D. J. Weiss (Ed.). *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* http://www.psych.umn.edu/psylabs/CATCentral/

Diao, Q. & Reckase, M. (2009). Comparison of ability estimation and item selection methods in multidimensional computerized adaptive testing. In D. J. Weiss (Ed.). *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* http://www.psych.umn.edu/psylabs/CATCentral/

Fisher, R.A. (1925) Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society, 22,* 700-725.

Hunt, E. (1995) *Will we be smart enough? A cognitive analysis of the coming workforce.* New York, NY: Russell Sage Foundation.

Kullback, S. (1959). *Information theory and statistics.* New York: Wiley.

Kullback, S. & Leibler, R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22,* 79-86.

Mislevy, R.J.; Johnson, E.G. & Muraki, E. (1992) Scaling procedures in NAEP. *Journal of*

Educational Statistics, 17, 131-154.

Rubin, D.B. (1987) *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement, 1,* 233-247.

Scarborough, D. & Somers, M. (2006). Using neural networks in employee selection. In *Neural Networks in Organizational Research: Applying pattern recognition to the analysis of organizational behavior.* APA Books Inc., Washington, D.C.

Segall, D.O. (1996). Multidimensional adaptive testing. *Psychometrika, 61,* 331-354.

Segall, D.O. (2000). Principles of multidimensional adaptive testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 53-73). Norwell MA: Kluwer.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27,* 379-423.

Thissen-Roe, A. (2005). *Adaptive selection of personality items to inform a neural network predicting job performance.* (Doctoral dissertation, University of Washington, 2005). Dissertation Abstracts International, 66(06), 3460B.

Thompson, N.A. (2009). Utilizing the Generalized Likelihood Ratio as a termination criterion. In D. J. Weiss (Ed.). *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* http://www.psych.umn.edu/psylabs/CATCentral/

Thompson, N.A. & Ro, S. (2007). Computerized classification testing with composite hypotheses. In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* http://www.psych.umn.edu/psylabs/CATCentral/

Wald, A. (1947). *Sequential analysis.* New York: Wiley.

Wang, C. & Chang, H. (2009). Multidimensional adaptive test: the application of Kullback-Leibler information. In D. J. Weiss (Ed.). *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* http://www.psych.umn.edu/psylabs/CATCentral/

Weissman, A. (2003). *Information theoretic approaches to item selection.* Paper presented at the 13th international meeting of the Psychometric Society, Sardinia, Italy.

Weissman, A. (2007). Mutual information item selection in adaptive classification testing. *Educational and Psychological Measurement, 67,* 41-58.

Weitzman, R. A. (1982). Sequential testing for selection. *Applied Psychological Measurement, 6,* 337-351.

Xu, X.; Chang, H. & Douglas, J. (2003). *A simulation study to compare CAT strategies for cognitive diagnosis.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago IL.