Journal of Computerized Adaptive Testing

Volume 8 Number 1

February 2020

Three Measures of Test Adaptation Based on Optimal Test Information

G. Gage Kingsbury and Steven L. Wise

DOI 10.7333/2002-0801001

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing

www.iacat.org/jcat

ISSN: 2165-6592

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing

Volume 8, Number 1, January 2020

DOI 10.7333/2002-0801001

ISSN 2165-6592

Three Measures of Test Adaptation Based on Optimal Test Information

G. Gage Kingsbury

Psychometric Consultant

Steven L. Wise

NWEA

This study extends the work of Reckase, Zu, and Kim (2019) by introducing three new measures of test adaptation. These measures are based on the amount of information that the test provides to individual examinees and are designed to provide theoretical, engineering, and operational views of the amount of adaptation that a particular test provides. The three measures are applied to simulations of adaptive, multistage, and fixed-form tests to describe how the measures could be used in an operational setting to identify difficulties with adaptation and to suggest processes for improving adaptivity.

Keywords: adaptive testing, adaptivity, test information, constrained testing

Computerized adaptive tests (CATs; Weiss, 1973) are commonly used in many fields, including educational assessment (Weiss & Kingsbury, 1984), placement testing (Sands, Waters, & McBride, 1997), and licensure testing (Zara, 1992). The purpose of an adaptive test is to provide efficient measurement by selecting items that are highly informative for each examinee from a bank of items with known characteristics. Such efficient accumulation of psychometric information results in more precise estimation of an examinee's trait level than could be obtained using a traditional fixed-item test of similar length.

The basic ideas underlying a CAT are relatively straightforward. From item response theory (IRT; Lord & Novick, 1968), it is known that items provide differing amounts of psychometric information across the trait scale, depending on their characteristics. Once a set of items is calibrated to a common measurement scale, the amount of information each item provides at each trait value can be calculated. An item provides maximum information at or near the item's estimated difficulty parameter value.

This implies that the most precise measurement can be obtained for a given examinee by selecting and administering from the item bank only items whose difficulties are well matched to the examinee's trait value. Of course, because an examinee's trait value is unknown prior to taking the test, the CAT must periodically calculate a provisional trait level estimate based on the examinee's responses to items administered earlier in the test session, and then select items based on this provisional estimate. In this fashion, a CAT "learns" where to target item difficulty and thereby attain efficient measurement. Thus, a CAT is designed to recursively refine its estimate of where an examinee is on the trait scale and focus item difficulty accordingly.

Applying the basic CAT ideas to an operational testing context brings with it a variety of practical constraints (Wise & Kingsbury, 2000) that are required to conform to the needs of the testing situation. These practical constraints can cause the test that is administered to a particular examinee to vary noticeably from the most efficient measurement possible. The extent to which a particular CAT design influences the efficiency of the measurement should be considered during the initial design of the test, during its subsequent implementation, and during its ongoing use as an operational test. As many factors influence the capability of a CAT to provide precise measurement for an individual, it is useful to define test adaptation as follows:

> *Test adaptation can be defined as the extent to which a test can provide examinees with an informative test, given the constraints imposed by the item bank and the test blueprint.*

The purpose of this paper is to discuss how constraints influence adaptation and to investigate three indices for quantifying a CAT's adaptation. These indices are designed for use in investigating theoretical test adaptivity, improving the engineering of test adaptivity, and identifying the operational test adaptivity of a CAT in use. This paper also considers the use of the indices with subgroups of examinees. First, the paper explicates some of the factors influencing test adaptivity. Second, it introduces the three new indices of adaptivity and compares them to current adaptivity measures. It also describes their potential for theoretical examination of adaptivity, engineering of better adaptivity, and operational adaptivity. Finally, it compares the characteristics of these indices using data from an operational CAT and discusses their potential utility.

Factors Affecting Adaptation

There are a variety of factors that influence a CAT's ability to adapt. Some of these factors are present in all types of CATs, while others are sometimes present depending on the needs of the testing program. Some of the primary factors influencing adaptation in a CAT are the following:

1. **Characteristics of the item bank.** In practice, there will always be limitations to item banks, which can vary widely in the degree to which they allow a CAT to adapt. This is largely a function of item bank size, but even very large banks can be deficient if they have few items providing information for a portion of the measurement scale.
2. **Item selection strategy.** The degree of adaptation can be influenced by **i**tem selection factors such as (1) which item parameters are used in selection, (2) whether Bayesian or maximum information selection methods are used, and (3) if there are limits on how much item difficulty can change from one item to the next. An especially important issue is

whether the CAT is item adaptive, cluster adaptive (Wainer & Kiely, 1987), or staged adaptive (Melican, Breithaupt, & Zhang, 2009).

3. **Test length.** Test length is critical to adaptation. A test that is too short might not have time to adapt to an individual's performance, while a test that is too long might exhaust the appro-priate items in the item bank.
4. **Content constraints.** In practice, there are frequently content constraints imposed that limit item selection. For example, a test blueprint might specify that 20% of the items must be drawn from each of five content categories. The CAT can satisfy this constraint by using any number of approaches (Kingsbury & Zara, 1991; van der Linden, 2007). The approach chosen, interacting with the characteristics of the item bank, will influence adaptation.
5. **Item exposure controls.** Exposure control procedures that can prevent items from being used too often (Hetter & Sympson, 1997) can also reduce the number and characteristics of the items in the item bank that are actually available for administration, which could in turn diminish adaptation.
6. **Item enemy avoidance.** Pairs of items that are too similar or that give clues to each other's answers are called item enemies, and some CATs avoid administering both items in the pair. As with item exposure controls, item enemy avoidance effectively decreases the size of the item bank, which could also diminish adaptation.
7. **Grade- or age-level constraints.** In K–12 achievement testing, CATs might be used across multiple grades, requiring item constraints by grade level for policy or instructional reasons. The imposition of such grade- or age-level constraints can decrease a CAT's ability to adapt by once again reducing the nominal size of the item bank, particularly for the highest and lowest performers in a grade.

Quantifying Adaptation

Each of the constraints described above has the potential to reduce a CAT's ability to adapt to examinees. Given the variability in the ways that CATs can be designed and implemented, it is useful to have a metric for assessing the degree of adaptation associated with a particular CAT.

There are at least two advantages associated with having such a metric. First, the metric could provide a more accurate view of the degree to which a particular CAT design with its unique constraints truly improves measurement efficiency over a fixed-item test. Second, it could help test developers better understand the benefits or costs that would result from a particular change to the CAT design.

Current Approaches

There are several existing approaches to quantifying the adaptive nature of a CAT. Reckase, Ju, and Kim (2019) proposed three adaptation indices based on the variance of the difficulty (*b*) parameters of administered items:

1. $r\left(\bar{b}_j, \hat{\theta}_j\right)$ The correlation between the examinee's final trait level estimate and the mean difficulty of the items the examinee was administered, across examinees.

2. $s\left(\bar{b}_j\right)\Big/ s\left(\hat{\theta}_j\right)$. The ratio of the standard deviation (SD) of the average item difficulties administered to an examinee and the SD of the final trait level estimates, across examinees.
3. $\left[s^2\left(b\right) - \text{ pooled } s^2\left(b_j\right)\right]\Big/ s^2\left(b\right)$ The proportion reduction of variance (PRV) from the variance of item difficulty estimates in the item bank to the items administered to an examinee, across examinees.

In their initial study, Reckase et al. (2018) showed the utility of these indices in identifying and comparing the adaptivity of different approaches to item exposure control in simulated and operational CATs. Reckase's indices are useful for identifying the degree of adaptivity in a test or set of tests. This is consistent with the reason for their development, which was to show the difference in adaptivity for different test designs. Taken together, the three indices developed by Reckase and his co-authors do an excellent job of this, as "they provide objective information about the matching of the item selection to the level of performance for everyone taking the test" (Reckase et al., 2019).

At the same time, Reckase's (2019) indices do not facilitate the following three useful functions related to adaptivity:

1. Identifying the theoretical limits of adaptivity for a CAT and comparing an existing test to this theoretically optimal test.
2. Identifying how the engineering characteristics of a CAT influence its adaptivity for all examinees and for subgroups.
3. Identifying the adaptivity of an operational CAT, including its adaptivity for subgroups of the population.

Using Reckase's original indices (Reckase et al., 2018), it would be difficult to use a difference in correlations or a ratio of variances to accomplish these three functions, which would probably be the next issues facing a test developer who found differences in the adaptivity of test designs. In order to facilitate these functions, this study introduces three new indices of adaptivity and describes how they can be used in the development, maintenance, and improvement of a CAT. These three indices—the Theoretical Optimal Information index (TOI), the Engineering Optimal Information index (EOI), and the Operational Optimal Information index (OOI)—are designed to answer specific questions concerning test adaptivity and should improve the robustness of available indicators.

Throughout this study, item information was calculated using Birnbaum's formula (Lord & Novick, 1968):

$$I(\theta) = p'(\theta)^2 \big/ p(\theta)q(\theta)\,, \tag{1}$$

where θ is the true trait level of the examinee; $p'(\theta)$ is the first derivative of the item response function evaluated at θ; p is the probability of a correct response as a function of the trait level; and q is the probability of an incorrect response as a function of the trait level.

Test information was calculated as the sum of the item information values from all items in a test given to a particular examinee. For some IRT models, an alternative value of information is the response pattern information function (Samejima, 1969); but since the two formulations are

equivalent for the one-parameter logistic (1PL) model used in this study, the Birnbaum formulation was used here.

Note that Birnbaum's formulation requires knowledge of the true trait level for evaluation; but as the true trait level is unknown in practice, this formula is commonly evaluated using the provisional trait level ($\hat{\theta}$) based on items administered up to that point during a CAT. Therefore, to determine the "optimal" information values for this study, the true value of θ or the estimated value of θ can be used in association with an existing item bank (with known item parameter estimates) or a perfect item bank (with all item parameter values available). These varying methods for evaluating optimal test information lead to the measures of adaptation discussed below.

The Theoretical Optimal Information Index (TOI)

TOI examines how well the test information observed at the final estimated θ compares to the information obtained at the final estimated θ from a theoretical test consisting of items equal in difficulty to the true θ level.

The TOI index identifies the percentage of test information that a CAT provides for a group of examinees, given the maximum amount of information that could be obtained at each examinee's true trait level. For this index, imagine that each examinee is administered a set of items with item difficulty equal to their true trait level. The particular responses that the examinee gives will cause the final trait level estimate to differ from the true trait level somewhat. TOI uses the information observed at the final trait level estimate based on the information available in a theoretically perfect test administering items at the true trait level.

This index is theoretical in two ways. It assumes that a set of items exists that perfectly match each examinee's trait level, and it assumes that these items are chosen and administered to each examinee. Since the true trait level is unknown, and the perfect item bank does not exist, this is only a theoretical maximum. However, it can serve as a useful measure of the adaptive nature of any particular test design in the context of a simulation study.

It should be noted that the index is designed for use with the 1PL IRT model, which constrains the maximum amount of information that could be obtained from any particular item to a known value. For the 1PL model, the response to a dichotomously scored item will provide a given amount of information with a fixed maximum. The specific numerical values will depend on the measurement scale chosen for use.

For example, if the scale used has a mean of 0.0 and an SD of 1.0, the maximum information available from a single item is 0.25. That is the measurement scale used in this study. The use of TOI with other IRT models would require the measurement practitioner to make assumptions about an item's maximum information by specifying additional parameter values (e.g., discrimination, lower asymptote) corresponding to "typical" or "high quality" items. TOI can be calculated as

$$\text{TOI} = 100 \times \frac{\sum_{j=1}^{J} \left(IA_j / IM_j \right)}{J}, \tag{2}$$

where IA_j is the actual test information observed for examinee j at their final trait level estimate, IM_j is the amount of information provided at the final trait level estimate by a set of items with

difficulty equal to the true trait level, and J is the number of examinees in the sample.

It should be noted that TOI can take on values greater than 100 for small groups of examinees. This occurs when the final trait level estimate differs noticeably from the true trait level. These high values indicate that the test is adapting well to observed responses, but those observed responses differ somewhat from expectation.

Uses of the TOI Index

A CAT is designed to choose the items administered based on an examinee's performance during the CAT. It is done, however, with the intent of estimating the examinee's true standing on the trait. Therefore, the TOI index is designed to answer questions concerning how well the CAT adapts to the true trait levels of a group of examinees. Values of TOI below 100 indicate that the test is providing less information at the final trait level estimates than would be obtained from a test made up of items at the true trait levels. This index could be used to answer questions concerning the quality of the adaptation relative to the true characteristics of the group of examinees. Thus, if the CAT results in a TOI value that is close to 100, it would indicate that the test is adapting well to the true trait level of the group of examinees, even if this performance deviates from the true trait levels. For instances in which the index exceeds 100, it might indicate that the test is adapting well to the estimated trait levels but that the estimated trait levels are not close to the actual trait levels. The TOI index can be used to answer questions such as the following:

1. How close does the current test design, with the current item bank and required constraints, come to the best adaptivity possible?
2. How good is the adaptivity of the current test for the subpopulations of examinees that it is designed to measure, compared to the best possible measure?

Because TOI requires knowledge of each examinee's true θ level, it can only be used in simulation. It is designed for use in the developmental stage of a CAT, when the item bank and test constraints can still be modified.

It should be noted that TOI can be calculated for a full group of simulated examinees, a subgroup with a different overall distribution of trait levels, or even for a single simulated examinee. This is also true for the other two indices to be introduced below. Although the results from a single examinee might not say much about the nature of adaptivity of a test design, patterns of change in the index calculated at the examinee level and observed as a function of the true trait level might be quite useful. This is explained in more detail below.

The Engineering Optimal Information Index (EOI)

EOI examines how well the test information observed at the final estimated θ compares to the information obtained from a theoretical test consisting of items equal in difficulty to the final estimated θ level.

EOI calculates the percentage of the maximum information at the final trait level estimate that is actually obtained by the test administered (rather than the true trait level, which is used for the TOI index). This index assumes a perfect item bank and administration of items with difficulty equal to the final trait estimate—a theoretical value that quantifies the most informative test that

could be administered, and compares it to the observed information value. The EOI index for the 1PL model can be computed as

$$\text{EOI} = 100 \times \frac{\sum_{j=1}^{J} \left(IA_j / 0.25K \right)}{J}, \tag{3}$$

where K is the number of items administered to examinee *j*; 0.25 is the maximum information available from an item for the 1PL using a measurement scale with a mean of 0.0 and an SD of 1.0; and EOI has a maximum value of 100 for a small group or an individual. This value indicates that the test has administered the set of items that provides the most information at the final trait level estimate for each individual in the group.

Uses of the EOI Index

EOI is designed to provide information concerning how well the test administered maximizes the information obtained at the final trait levels by a group of examinees. For many examinees, the maximum information is not available due to deficiencies in the item bank, the constraints imposed on item selection, and the test design. The term "engineering" is used here to describe the potential for this index to be used to improve the characteristics of the adaptivity of the test by adjusting these factors.

The EOI index could help evaluate, for example, the benefits of doubling the item bank size or the costs of decreasing test length by 30%. It is important to consider, however, that the impact of a change will often vary across values of the trait scale. For instance, adding 100 difficult items to a bank might improve a CAT's adaptation for higher performing examinees, while having no impact on adaptation for lower performers.

As EOI can be calculated for an entire testing sample or for select subgroups of examinees, its variability suggests areas on the measurement scale in which the CAT is not measuring as well as it could. Therefore, EOI can provide clues to future engineering of the item bank and test administration approaches that will improve the test being administered. It can also be used to identify differential adaptivity for important subgroups of examinees. EOI can be used in simulation or with operational test data. It can be used to examine questions such as:

1. If the item bank or the item selection procedure were modified, how would this change the adaptivity of the test?
2. If the techniques to handle constraints or some of the constraints themselves were modified, how would this change the adaptivity of the test?

Because EOI does not require knowledge of an individual's true θ levels, it can be used in simulation but might prove most useful with operational data from a test of interest. As most operational CAT programs seek to improve or evolve over time, EOI is designed to be used to improve a CAT from year to year. This can be particularly useful in settings that have samples of examinees with characteristics that change over time (for instance, certification tests with differing populations, depending on employment needs.)

The Operational Optimal Information Index (OOI)

OOI examines how well the test information observed at the final estimated θ compares to the information obtained at the final estimated θ from a test consisting of available items with difficulty closest to the true θ level.

OOI calculates the information that the test administered provides at the examinee's final trait level estimate as a percentage of the information that would be provided at the final trait level estimates by administering a hypothetical test composed of the items in the bank that would provide the most information at the examinee's true trait level. This calculation assumes that only items in the item bank are available, and that the length of the test is the same for the hypothetical test and the actual test administered. The OOI index can be computed by assuming:

1. That the optimal amount of test information that can be obtained for an examinee from a given item bank is IO_j, obtained by administering the most informative items for examinee j for the entire length of the test. Because this requires knowing the true trait level for the examinee, it is impossible to do with operational data but serves as a reasonable upper bound for any given item bank and test length.
2. That the test information observed for examinee j is IA_j, as defined in Equation 1. This value will be influenced by the actual items chosen for the examinee. In turn, it will be influenced by the test length, the item selection procedure, and other factors not held constant by the test requirements and will yield the OOI index, which can be calculated as

$$\text{OOI} = 100 \times \frac{\sum_{j=1}^{J} \left(IA_j / IO_j \right)}{J}. \tag{4}$$

It should be noted that OOI can take on values greater than 100 for small groups or individuals. As with TOI, this will occur when the final trait level estimate differs noticeably from the true trait level. These high values indicate that the test is adapting well to observed responses and that the best items available at the true trait level provide less information at the observed trait level estimate than those that were administered.

Uses of the OOI Index

The OOI index can be applied to the full group taking an assessment or subgroups conditional on ability level. It is designed to answer questions concerning test design, given an existing item bank and set of constraints. It can be used to identify relative areas of strength or weakness in the adaptation of the test, given the item bank that is available. It can also be used to consider changes in the item administration strategy or the various constraints on item selection. The OOI index might be used to address questions such as the following:

1. Which adaptive testing design will allow the test to provide the most adaptivity, given the item bank that is available and the constraints that are required?
2. How good is the adaptivity of the current test for the entire population of examinees that it is designed to measure, compared to what the item bank could provide?
3. How good is the adaptivity of the current test for important subgroups in the population of examinees that it is designed to measure, given the current item bank and constraints?

Because OOI requires knowledge of each individual's true θ level, it can only be used in simulation. It is designed to be used either in development or with operational tests to identify how well the test provides information compared to what might be obtained at the true trait level, given constraints and the existing item bank. The index is designed to give a snapshot of adaptivity that could be used in technical reports or other documentation.

Empirical Evaluation

This study compared the three new indices to the procedures developed by Reckase et al. (2018) in a simulation using a fixed-form test, multistage test (MST), and an item level CAT. The outcomes from the various procedures were considered in light of the types of questions that might be asked of a CAT program. As the new indices are designed to work with important subgroups of the testing population, this use was also examined with one of the indices.

The indices developed in this study can also serve as a tool for more precise engineering of adaptive tests. Different test types and test designs have specific costs in terms of adaptivity and test accuracy across the breadth of the testing population. The current study considered these types of costs and suggests approaches to better adaptive test engineering.

Methodology

This study computed the three adaptation indices described above and those developed by Reckase et al. (2018) for a constant sample of 1,000 simulated examinees drawn from a normal (0,1) distribution. It computed all indices for the same simulated sample. Each index was computed for three different test types (a fixed-form test, an MST, and a fully adaptive test). Each test was drawn from a bank of 200 items calibrated to the 1PL IRT model with item difficulty drawn from a normal (0,1) distribution. All momentary achievement level estimates and final achievement level estimates were obtained using maximum-likelihood estimation.

Fixed Form

The fixed-form test consisted of 45 items drawn systematically from the bank of 200 items. To choose the items for the test, the bank was ordered by item difficulty. Forty-five items were chosen by starting with the easiest item and selecting each fourth or fifth item to obtain a total of 45 items. This approach resulted in a test form with item difficulty distributed much like the entire item bank.

Multistage (MST)

The MST used in this study was a three-stage design with one subform in the first stage, two subforms in the second stage, and three subforms in the third stage. Each subform consisted of 15 unique items drawn systematically from the item bank. The subforms had characteristics as follows:

1. ***Subform 1***. 15 items drawn from a difficulty range of −1.0 to 1.0. All examinees took this subform first.
2. ***Subform 2***. 15 items drawn from a difficulty range of −1.5 to 0.0. Examinees with a trait level estimate less than 0.0 after the first subform took this subform second.

3. ***Subform 3***. 15 items drawn from a difficulty range of 0.0 to 1.5. Examinees with a trait level estimate greater than or equal to 0.0 after the first subform took this subform second.
4. ***Subform 4***. 15 items drawn from a difficulty range of −3.0 to −1.0. Examinees with a trait level estimate less than −1.0 after the first two subforms took this subform third.
5. ***Subform 5.*** 15 items drawn from a difficulty range of −1.0 to +1.0. Examinees with a trait level estimate greater than or equal to −1.0 and less than 1.0 after the first two subforms took this subform third.
6. ***Subform 6***. 15 items drawn from a difficulty range of 1.0 to 3.0. Examinees with a trait level estimate greater than or equal to 1.0 after the first two subforms took this subform third.

Therefore, as in the fixed-form test, each simulated examinee was administered a total of 45 unique items during the MST.

CAT

Each examinee was administered 45 items during the CAT drawn from the bank of 200 items. After each item was administered, the examinee responses were scored using maximum-likelihood estimation, and the next item was chosen to maximize the item information at this momentary trait level. No exposure control was used for this study, and all items were treated as equivalent to the test blueprint. The first item administered assumed a trait level estimate of 0.0. Until the maximum-likelihood estimate was finite, a Bayesian modal estimate [N(0,1)] was used.

Results

Table 1 provides the value of each of the adaptation indices for each test type. Each of the indices shows the same pattern of adaptation, with the fully adaptive test being the most adaptive, the fixed-form test being the least adaptive, and the MST being somewhere in between. These results are expected, and in line with results seen in Reckase et al. (2018).

Table 1. Value of Each Adaptation Index for Each Test Type Across All Examinees

Index	CAT	Fixed Form	MST
$r\left(\bar{b}_j, \hat{\theta}_j\right)$	0.92	0.00	0.90
$s\left(\bar{b}_j\right) / s\left(\hat{\theta}_j\right)$	0.81	0.00	0.43
PRV	0.88	0.43	0.74
TOI	94.99	76.74	87.41
EOI	92.74	74.66	85.24
OOI	98.27	79.11	90.32

Although the pattern seen in these results is consistent across all indices, the interpretation of the values is somewhat different. For instance, the correlation between item difficulty and the trait level $r\left(\bar{b}_j, \hat{\theta}_j\right)$ will always be 0.00 for a fixed-form test. However, if the item bank is no larger

than the fixed-form test, then using the fixed-form test will be the best that can be done given the size of the item bank. Clearly, the indices that emphasize variability in item difficulty, $s\left(\bar{b}_j\right)\Big/ s\left(\hat{\theta}_j\right)$ and PRV, will not reflect whether the item selection makes best use of the available item bank. Although this might be a good measure of adaptivity, it might not reflect appropriate adaptivity, given the constraints that the item bank imposes.

Another aspect of interpretation that differs among the indices is their relationship with test design. Three of the indices—$r\left(\bar{b}_j, \hat{\theta}_j\right)$, $s\left(\bar{b}_j\right)\Big/ s\left(\hat{\theta}_j\right)$, and PRV—describe the order of adaptivity in the different tests quite well but do not point toward strategies that might be used to improve a particular test. For instance, it is not clear how increasing the test length or size of the item bank might influence these indices. The other three indices (TOI, EOI, and OOI) can give the developer guidance concerning the impact of specific changes on the adaptivity of the test.

Because the TOI index is designed to identify how well a test adapts to examinees compared to the best possible theoretical test, it can reflect how much damage is done by administering the fixed-form test or the MST rather than the CAT. From Table 1, it can be seen that 13% of the information would be lost by administering the MSTs and 23% of the information by administering the fixed-form test. This is useful information in situations in which some examinees need to take an alternate test form.

The EOI allows examination of test engineering questions. In this case, it might be asked how the MST could be modified to provide as much adaptivity as the CAT. Because the MST provided 7% less information than the CAT, it might be useful to examine the characteristics of an MST that was 7% longer than the CAT. This probably will not work perfectly, as the item bank will have been depleted by drawing the original tests; but as this index can be computed from simulated data, several alternative lengths could be tried. Although it is unlikely that the MST will ever have the measurement characteristics of a fully adaptive test, this approach could improve their similarity.

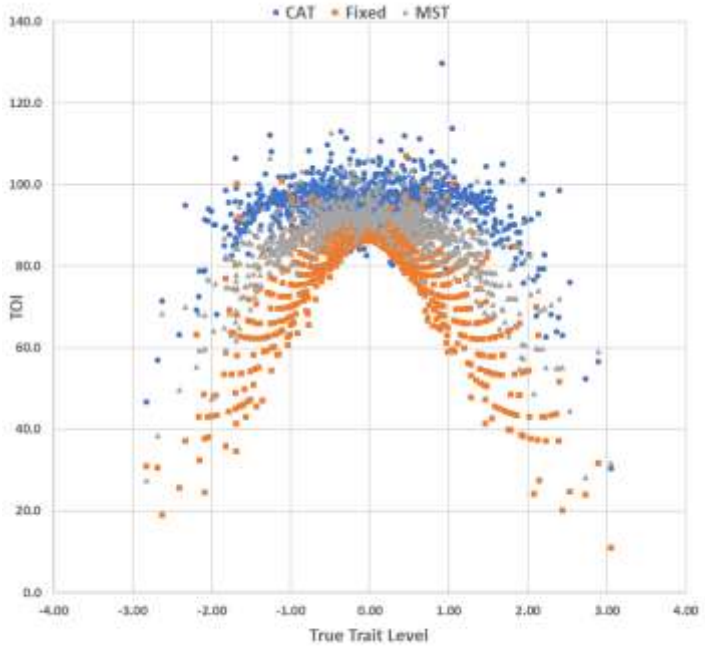
The OOI index indicates that the CAT is providing about as much adaptivity as it can, given the item bank and constraints. The data in Table 1 indicates that if 2% more equivalent items (slightly less than one item) were added to the length of the CAT, overall results could be expected that would be similar to those from the optimal test with the current test length. In a similar fashion, 21% more items would be needed for the fixed-form test, and 10% more items would need to be added to the MST to obtain the optimal information value. Of course, this assumes that the items added are equivalent to those already included in the test. This might not be possible due to limitations of the content being assessed or limitations in item development.

Figures 1 to 3, respectively, show the TOI, EOI, and OOI calculated for each simulated examinee (simulee) for each test type, as a function of the true trait level. It is useful to consider the OOI values in Figure 3 as an example of the ways in which the index information can be used. It can be seen from Figure 3 that the fully adaptive test has a consistent OOI index value across the entire trait range, with most values near 100. For the MST, the OOI index values are reasonably consistent throughout the trait range from −1 to +1 but are slightly lower than the fully adaptive test, and much lower for the more extreme trait values. For the fixed-form test, the OOI values peak at a trait level of 0.0 and drop off as the trait level increases or decreases. It is useful to note

that the OOI index has values higher than 100 for some simulees for whom the items administered were closer to the final trait level estimate than those at the true trait level would be. Results from the TOI and the EOI can be interpreted in the same manner.

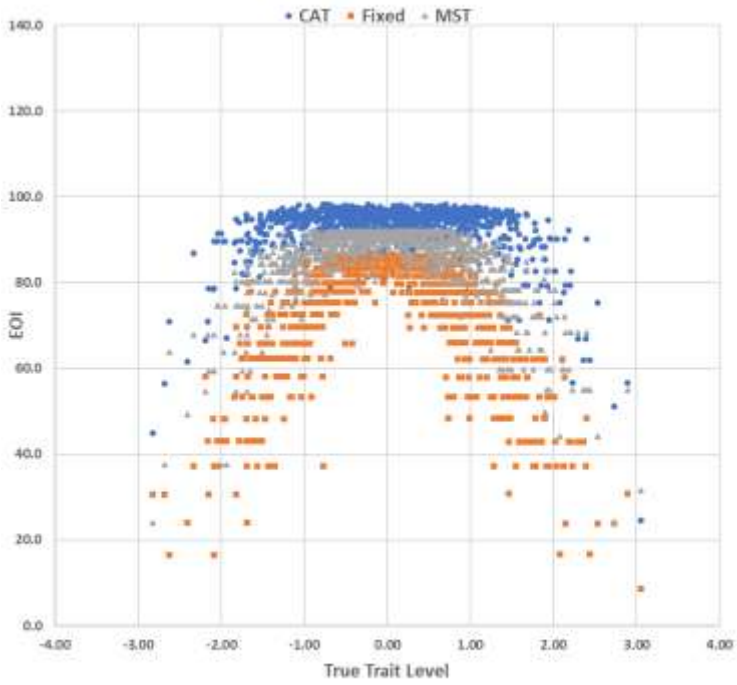
Figure 1 shows that the same overall trends observed in the full-group analysis tend to be seen in this investigation of TOI at the simulee level as a function of the true trait level. For most simulees, the CAT results in a higher TOI than the MST and the fixed-form test. Likewise, for most simulees, the MST resulted in a higher TOI than the fixed-form test.

Figure 1. Theoretical Optimal Information (TOI) Index for Each Simulee for Each Test Type as a Function of True Trait Level



The patterns of TOI across the true trait levels are also useful to consider. For the CAT, TOI tends to hover between 90 and 100 for most simulees with a true trait level between −1.5 and +1.5, and begins to drop for most simulees outside that range. This is probably due to the small number of items in the item bank (less than 10) beyond this range. For these simulees, the bulk of the 45 items administered were further from the true trait level, and therefore provided less information than the perfect theoretical test. For the MST, simulees between −1.0 and +1.0 tended to have TOI values above 90; and for the fixed-form test, simulees rarely obtained TOI values greater than 90. This indicates that the fixed-form test and the MST had noticeably less adaptivity for the extreme simulees. If important decisions are to be made for examinees outside that range, the test design probably should be fully adaptive.

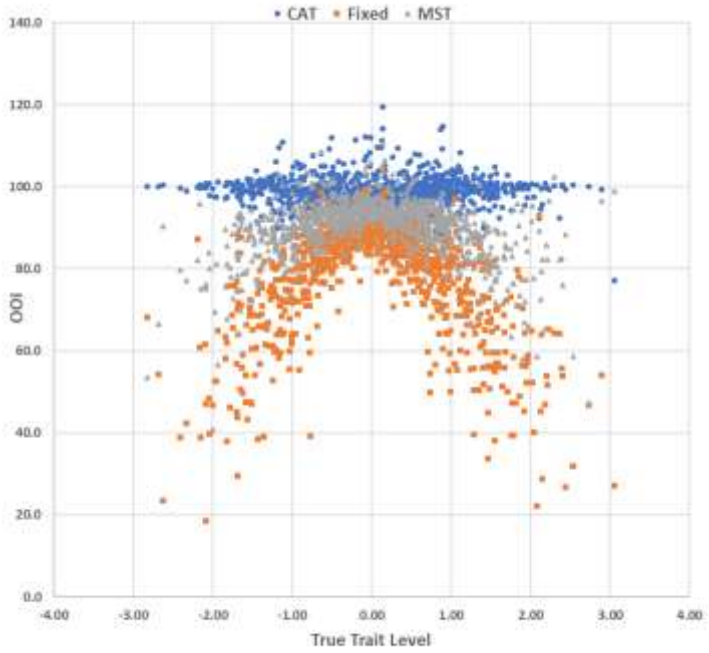
Figure 2. Engineering Optimal Information (EOI) Index for Each Simulee for Each Test Type as a Function of True Trait Level



One final trend to take note of in Figure 1 is the scallop pattern seen for the fixed-form test and to a lesser extent for the MST, with a curvilinear relationship being observed between TOI and bands of true trait levels. This is caused by the limited number of observed scores that are available for these tests. For the 1PL model, the number correct within a specific item set will result in the same trait level estimate. Although this does not have much impact on test design, it is an interesting pattern to note.

In Figure 2, the EOI values show the same overall trends, with the CAT providing the best adaptivity, the MST showing the next best adaptivity, and the fixed-form test showing the least adaptivity. Inspecting EOI at the individual simulee level as a function of the true trait level tells a more complete story. As with the TOI, the EOI reveals a broader range of the trait in which the CAT delivered values above 90 (approximately −1.5 to 1.5), while the MST never delivered the adaptivity of the CAT and delivers EOI values above 90 for most simulees with true trait levels in the range of −0.8 to 0.8. The fixed-form test did not result in EOI values above 90 for the majority of examinees at any trait level,and dropped off rapidly beyond the true trait range between −0.6 and 0.6. In order to create a fixed-form or MST that could provide the adaptivity of the CAT, it is clear that just increasing the number of items administered would not overcome the deficit in adaptivity observed for the more extreme simulees. Figure 3 shows that the OOI values for the overall group were highest for most examinees for the CAT, next highest for the MST, and lowest for the fixed-form test. This is in keeping with the earlier findings and the results from the Reckase indices (Reckase et al., 2018).

Figure 3. Operational Optimal Information (OOI) Index for Each Simulee for Each Test Type as a Function of True Trait Level



Examining the individual results as a function of the true trait level, one finding concerning the CAT stands out: There was no consistent drop off in the OOI values at the extremes of the range of the trait. This does not mean there was no drop off in information for those simulees. Instead, according to the definition of OOI, it means that the CAT achieved as high a level of information as is possible given the item bank. As a snapshot of the performance of the test, this finding that the CAT did about as well as it could in the circumstances is very positive, and useful for the stakeholders to know.

For the MST, the OOI values on the individual level indicate a loss of 10 to 15% of the information that might be obtained from the item bank given the test length and constraints. The OOI values for the MST started to drop off beyond the range of −1.0 to 1.0 on the true trait metric, indicating more loss of available information. For the fixed-form test, OOI values indicated a loss of 15 to 20% in the middle of the trait distribution. The OOI values dropped off almost immediately for simulees that differed from the mean of the distribution. These results consistently indicate that these three test designs differ in adaptivity with each index. But more importantly, the differences depend on the range of examinee performance that is being considered. This leads to a direct example of the use of the indices for a salient subgroup of examinees.

Subgroup Example

Figures 1 through 3 suggest an opportunity for an additional analysis of adaptivity within subgroups of the examinee population, using the OOI index as an example of this type of analysis for examinees with a true trait level below −1.0. If the tests examined here were used in a classroom of students, the students who scored below −1.0 would be those who would often struggle with the course material and might be eligible for special education services. These values would differ across situations, but the example was designed to show that the adaptivity of the overall test does not tell the whole story when important subgroups of the testing population exist. Differential adaptivity then becomes extremely important.

Figure 4. OOI Index for Each Low Performing Simulee for Each Test Type as a Function of True Trait Level Less than –1.0

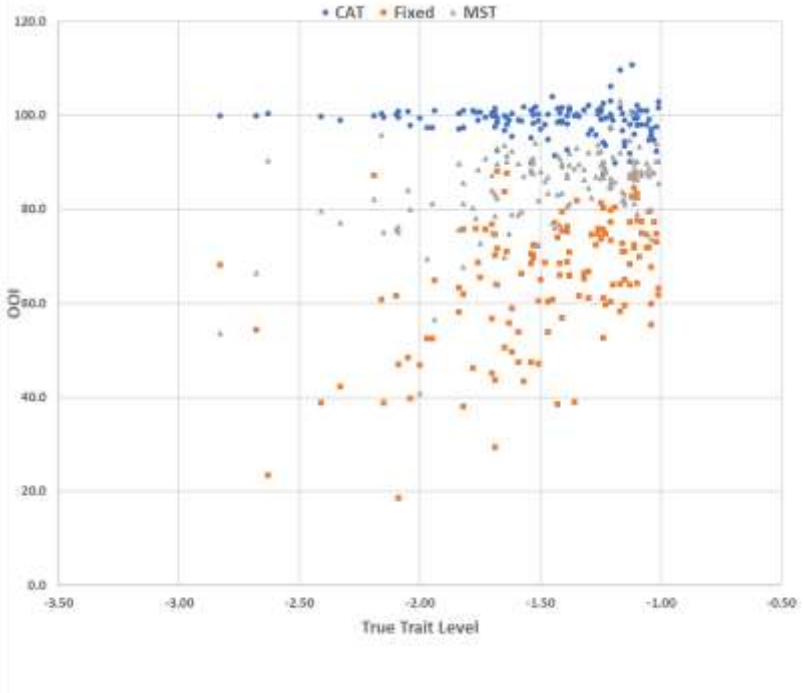


Figure 4 shows the OOI values for each examinee with a true trait level lower than −1.0 for each of the test types. For this low subgroup of simulees, the observed OOI was 98.86 for the CAT, 65.48 for the fixed-form test, and 84.62 for the MST. Comparing these values to those from the full group of examinees, it can be seen that the OOI for the CAT was virtually identical, while the OOI values dropped noticeably for the fixed-form test and the MST.

Given this group of examinees, the CAT could be improved to the optimal information level by adding 2% more equivalent items to the test length. The MST would require 15% more equivalent items to reach the optimal information level (approximately 50% more additional items than indicated by the full-sample analysis) and the fixed-form test would require 35% more equivalent items to obtain the optimal information level (approximately 66% more equivalent

items than indicated by the full-sample analysis). It appears from this analysis that the CAT is the only test design that maintains its relationship to the optimal information values.

Conclusions

This study demonstrated that the TOI, EOI, and OOI indices can be used to closely examine the adaptivity of a test for the full group of examinees and for subgroups of examinees. These indices also show promise for pre-engineering tests to obtain the optimal information by adjusting the length of tests. This information can also be used in a cost analysis, given the cost of developing equivalent items and the cost of administering a longer test.

For the three test types used in this study, the OOI index indicated that the fully adaptive test consistently provided a level of information that approached the optimal information level, while the other test types provided less information compared to the optimal level. Further, these other test types provided a smaller percentage of the optimal information for the simulees in the subgroup analysis, and an even smaller percentage when visually examining the more extreme simulees.

Is it possible to create a multistage assessment that will mimic the EOI characteristics of the fully adaptive test used in this study? Perhaps, although the increased length of the resulting test might exhaust the item bank. However, using EOI as a tool to suggest how the MST could be modified—conceivably through more stages, or longer subtests, or different difficulty distributions—is a step in the direction of designing tests using appropriate information.

It is useful to consider how these indices might be used in a common situation such as the development of a new test for use in a K–12 educational setting to measure reading achievement in the elementary grades. Commonly, a decision about the test design would be made by rough rules of thumb, such as "We don't have very many items in our item bank, so we should use an MST rather than a test that adapts after every item." Consider how the confidence in this decision (and possibly the decision itself) might change with the use of the adaptation indices described above. The new decision process might be as follows:

1. Draw up a rough draft of the potential item selection designs that might be used for the new test, including draft constraints, using the existing item bank.
2. Select items from the bank according to the constraints for those designs that will not use the entire item bank.
3. Simulate the potential designs with a distribution of true θ values that roughly approximates that expected of the actual examinees. Note that this will almost never be a normal distribution.
4. Calculate the TOI for the full group of simulees. (You will also collect other important information about the proposed designs at this point, but this discussion focuses on the adaptivity indices.)
5. Calculate the TOI for each important subgroup of simulees, again approximating the distributions from what is expected of actual examinees. For an achievement test, these subgroups would probably include specific grade levels and students who might receive special services as a result of the test outcome—talented and gifted services, or an Individual Educational Plan (IEP), for instance. From the TOI values for the total group and the subgroups, determine which draft design meets the needs of the testing organization

most completely. In this example, a single test that could identify proficiency in reading and need for special services for the most able students and the most struggling students in multiple grades might have great value for a K–12 school system.

6. Begin development of the test that was identified in the previous step.
7. If none of the draft designs completely meets the needs of the organization based on TOI, consider revising the most successful draft, potentially using different values for needed constraints and different approaches to handling issues such as item exposure and repeated testing.
8. Simulate these new drafts again with the same groups and subgroups.
9. Calculate the EOI for each of the drafts for total groups and subgroups. At this point, consider evaluating EOI for the simulees closest to cutoff values that are important to the testing organization, such as the dividing line between students eligible and not eligible for special services.
10. If the draft tests do not meet the needs of the organization, use the EOI values to estimate how to add items to the bank and the distribution of difficulty that these items should have.
11. Consider the possibility of phasing in constraints as the test matures, to allow needed accuracy in the initial use of the test.
12. Once the test is in the field (and, hopefully, successful) calculate the OOI values for appropriate groups and subgroups and use these values as a portion of the technical documentation for the test.
13. As the test matures and needs change, use EOI values to continuously re-engineer the item bank to meet these new circumstances.

The example above suggests moving from "test development by expert opinion" to "test engineering" using the best available information to provide the test design that meets the needs of the testing organization. It is hoped that, as information ratios, these indices will be readily interpretable by practitioners. It should be noted that this study was limited by using just simulated data. It will be important to expand the analysis to include operational test data to identify how the indices work in practice. Additionally, the study was limited in the constraints that were used for the analysis. Examining the new indices with tests using a wider variety of constraints—such as fixed and variable content constraints, and a variety of exposure control constraints—should improve understanding of the operational utility of the indices in test analysis and in test engineering.

In addition, a key extension of the current research will be to use the indices with other IRT models. Although the OOI index can be used as defined in this study, the EOI and TOI require the definition of an optimal item so that the information available from the optimal item can be used in the computation of the indices. As mentioned above, the optimal item for the dichotomous Rasch model is simply an item with difficulty equal to the true trait level of the item. For a given measurement scale, this also yields a consistent relationship between the amount of information at the true trait level and the amount of information at any estimated trait level.

For other response models that do not have a maximum information value, an appropriate optimal item needs to be described. Consider the three-parameter model, which has an information function that is asymmetric. One approach would be to establish a value for the *a* (discrimination) parameter for the optimal item as the highest observed value in the item bank, and to establish a

value for the *c* (pseudo-guessing) parameter as the lowest observed value in the item bank. Then, the optimal value for the *b* parameter would be the value that maximizes the information for a particular true trait level. For example, if the optimal item had $a = 1.0$ and $c = 0.20$, it would need a *b* value that was 0.27 θ units below the individual's true θ level to provide the maximum information (where the offset from the true θ level would vary with optimal values chosen for *a* and *c*). This could be done using Equation 2 from Reckase et al. (2019). This type of approach would create an "optimal" value for the item parameters for the theoretical items, which in turn would create the information values needed to calculate the EOI and TOI indices. However, the interpretation of the indices would also differ somewhat from those described in this study.

References

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 145-149). Washington, DC: American Psychological Association. CrossRef

Kingsbury, G. G., & Zara, A. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education, 4*(3), 241-261. CrossRef

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

Melican G. J., Breithaupt, K., & Zhang Y. (2009). Designing and implementing a multistage adaptive test: The uniform CPA exam. In W. J. van der Linden & C. A. W. Glas, (Eds.), *Elements of adaptive testing* (pp. 167-189). New York, NY: Springer. CrossRef

Reckase, M. D., Ju, U., & Kim, S. (2019). How adaptive is an adaptive test: Are all adaptive tests adaptive? *Journal of Computerized Adaptive Testing, 7*(1), 1-14. CrossRef

Reckase M.D., Ju, U., & Kim, S. (2018). Some measures of the amount of adaptation for computerized adaptive tests. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.). *Quantitative Psychology IMPS 2018: Springer Proceedings in Mathematics & Statistics, Vol. 233* (pp. 25-40). Cham, Switzerland: Springer International Publishing. CrossRef

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. CrossRef

Sands, W. A., Waters, B. K., & McBride, J. R. (1997). *Computerized adaptive testing: From inquiry to operation.* Washington, DC: American Psychological Association. CrossRef

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*(3), 185-201. CrossRef

Weiss, D. J. (1973). *The stratified adaptive computerized ability test* (Research Report 73-3). Minneapolis, MN: University of Minnesota, Department of Psychology, Psychometric Methods Program. Available at www.iacat.org/biblio.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*(4), 361-375. CrossRef

Wise, S., & Kingsbury, G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica, 21*(1), 135-155. Available at www.iacat.org/biblio.

van der Linden, W. J. (2007). The shadow-test approach: A universal framework for implementing adaptive testing. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* Available at www.iacat.org/biblio.

Zara, A. (1992, April). *A comparison of computerized adaptive and paper-and-pencil versions of the national registered nurse licensure examination.* Paper presented at the annual meeting of the American Educational Research Association. San Francisco, CA. Available at www.iacat.org/biblio.

Author's Address

Steven L. Wise, NWEA, 121 NW Everett St., Portland, Oregon 97209, U.S.A.
Email: steve.wise@nwea.org