Journal of Computerized Adaptive Testing

Volume 10 Number 2

May 2023

Expanding the Meaning of Adaptive Testing to Enhance Validity

Steven L. Wise

DOI 10.7333/2305-1002022

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

ISSN: 2165-65923

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

Journal of Computerized Adaptive Testing

Volume 10, Number 2, May 2023

DOI 10.7333/2305-1002022

ISSN 2165-6592

Expanding the Meaning of Adaptive Testing to Enhance Validity

Steven L. Wise
NWEA

Editor's note: This paper is based on a keynote presentation given at the 2019 International Association for Computerized Adaptive Testing (IACAT) conference in Minneapolis, Minnesota. The views expressed are solely those of the author and do not necessarily reflect the positions of NWEA.

The computerized adaptive test (CAT), which adjusts the difficulty levels of administered items to match the ability levels of the examinees, has a long history of providing efficient testing. After 50 years of research, however, CATs have evolved little beyond the initial idea of adapting item difficulty. In this paper, I suggest that we begin viewing adaptation in a more expanded manner, arguing that measurement will be further improved by a CAT's ability to detect and adapt to the presence of construct-irrelevant factors that threaten the validity of individual test scores. Several examples of innovative adaptations currently provided by an operational CAT program are discussed.

Keywords: validity; test-taking disengagement; adaptive testing; CAT; CBT

I was a doctoral student when I first heard about adaptive testing. I had learned about measurement from a classical test theory (CTT) perspective in my undergraduate and master's programs, and I was now learning about item response theory (IRT), which was about to become the dominant measurement theory used in educational measurement. IRT had several important advantages over CTT (e.g., item characteristics that were not group dependent, ability estimates that were not test dependent, easier test score equating, and estimation of reliability that did not require that test forms be parallel), which made easier many standard measurement tasks/processses, such as test form design, equating, ability estimation, or the assessment of item/test bias. And while I appreciated the value of these benefits, I did not find them overly exciting because they primarily improved our ability to do things we were already doing.

The big exception, however, was a special type of computer-based test (CBT). CBTs were just beginning to appear, and a computerized adaptive test (CAT) combined the invariance principle of IRT with a CBT's computing power to produce an innovative type of test that had not been seen before. Each examinee could receive a set of items whose difficulty was tailored to that examinee, yet examinee scores were on the same scale. As a result, testing efficiency was vastly improved, meaning that measurement precision could be attained that was comparable to that from a fixed-item test, or linear test, in about half as many items. Moreover, by dispensing with the one-size-fits-all constraint of linear tests, a CAT could yield even larger precision improvement for students in the tails of an ability distribution. This was an extremely attractive innovation that emerged just as IRT and CBT were each gaining traction in educational measurement. I was immediately enamored with the possibilities of this "smarter" form of testing.

A CAT is designed to adjust the level of item difficulty an examinee receives, based on the correctness of the item responses given earlier in the test event. The test administration algorithm seeks to administer items that are well-matched to the examinee's ability. Attaining a good match between difficulty and ability then leads to item responses that provide maximum item information and thereby efficiently increases test score precision. Thus, from its inception, the primary goal of a CAT has been to maximize the precision of ability estimates through difficulty-optimized item selection.

Of course, we soon learned that an operational CAT program is a bit more complicated than was suggested by the basic adaptive testing algorithm. In practice, item selection needed to be constrained to ensure that the set of items an examinee receives is adequately representative of the intended content domain or blueprint. Similarly, test item security demands required us to constrain the rate at which items were exposed from the item pool. These types of constraints, and others, reduced the efficiency of operational CATs somewhat, but it remained far more efficient than traditional linear tests.

Since the introduction in the 1970s of the theoretical ideas underlying adaptive testing, we have witnessed 50 years of research directed toward improving the applied science of CATs. This body of research has been extensive, focusing on numerous practical issues such as item pool development and maintenance, improved item selection algorithms, item exposure control, and test termination criteria. The result of these efforts has been the development of a variety of CATs, which can be applied to meet numerous measurement needs. At the same time, however, we might also view adaptive testing as having gone through limited evolution outside of its original goal of efficient testing. That is, after 50 years of evolution, the basic adaptive mechanism of a CAT has remained largely unchanged: CATs adapt through adjusting item difficulty.[1]

The limited evolution of CATs can be attributable, in part, to the fact that most CBTs emerged out of operational paper-and-pencil testing (PPT) programs. The transition was often done gradually, with CBT and PPT versions initially being used concurrently, followed by a move to a sole use of CBTs occurring later on (if ever). Whenever CBTs were introduced this way, questions naturally arose about the score comparability of the two test modes, and testing programs typically focused on the need to establish mode comparability as a key aspect of score validity. However, an unintended consequence of this emphasis on score comparability was that it constrained CBTs to be little more capable than the PPTs to which they were compared. As a result, CATs—when they appeared—often were simply more efficiently administered, computer-based versions of

[1] This general statement excludes CATs that utilize cognitive diagnostic models (CDMs). However, even CDM–CATs administer items with the goal of maximizing measurement precision of the latent attribute, without acting to minimize the impact of construct-irrelevant factors.

PPTs, with the primary advantage of reduced testing time. Item types other than multiple-choice were rarely seen and the capabilities of the computer to deliver tests providing more advanced tools and features were largely unused apart from adjusting item difficulty.

In recent years, more CBT testing programs have emerged without a legacy of PPT. And as CBTs have become untethered to PPT versions—rendering comparability no longer an issue—test developers have increasingly explored innovative test items and test administration methods. From the standpoint of adaptivity, this invites the question: What else can CBTs (and by implication, CATs) become?

Reconceptualizing Adaptivity More Broadly

Over the last half century, we have been conditioned to think of tailored item difficulty whenever we think of "adaptive testing." But a CBT can certainly adapt in other ways than simply adjusting item difficulty. To think more broadly about what that might mean, it is useful to reconsider our psychometric goal during a CAT test event. Traditional CATs prioritize reduction of the standard error of ability estimation, thereby enhancing score validity by improving score precision. A useful alternative goal, however, would be to focus on *maximizing the validity of individual test scores*. Moving the focus of a CAT to maximizing the validity of scores rather than their precision suggests an expanded definition:

> *A CAT is a type of computer-based test that adapts, during a test event, to examinee behavior in ways intended to improve the validity of individual scores.*

As noted above, this definition does not exclude traditional CATs that adjust item difficulty, because increased precision is expected to improve validity. But the definition opens the door to numerous other possibilities for improving validity during a test event.

Validity and the Psychology of Test Taking

There is a universal, tacit assumption underlying our contemporary IRT models: When we administer test items, we assume that the responses reflect what the examinee knows and can do. That is, we are assuming maximum performance from the examinee (Cronbach, 1960; Messick, 1989). What might threaten that assumption? Potentially many things. Examinees are vulnerable to myriad construct-irrelevant factors that might meaningfully affect test performance. For example, a student taking a reading comprehension test might be unmotivated. They might be affected by fatigue, anxiety, hunger, or illness. They might be worried about a sick pet or relative. Their testing environment might have noisy distractions. Factors such as these would generally be considered irrelevant to the measurement of the construct (i.e., the student's level of reading comprehension), yet they can meaningfully distort (typically negatively) test performance. Thus, the reality that construct-irrelevant factors are often present and can pose a serious threat to the validity of test score interpretations should motivate us to better understand the nature and potential impact of the most salient factors. We would therefore be well served to look beyond our traditional psychometric models and gain a better understanding of the psychology of test taking and the construct-irrelevant factors that can affect test performance.

A key challenge to test administrators is how to reduce the impact of construct-irrelevant factors. Such a reduction would improve test score validity. This "addition by subtraction" approach is complicated by the fact that (1) examinees are not equally affected by a given construct-irrelevant factor and (2) the impact of a construct-irrelevant factor will often change during a test event. This suggests, however, that if a CBT could detect, in real time, the presence of a particular construct-irrelevant factor, it might intelligently respond in some way. Or, said another way, the CBT could *adapt* to the examinee's behavior in ways other than simply adjusting item difficulty.

It should be noted that such an adaptation to the behavior of some, but not all examinees, runs counter to our long-standing traditions of standardized testing, which act to treat all examinees equally during a test event. Equal, however, is not the same as equitable. Test administrations that can adapt to the presence of a construct-irrelevant factor constitute a more individualized approach to measurement consistent with an "understandardization" perspective (Sireci, 2020). This underscores a key difference between a traditional CAT and CBT that adapts to construct irrelevance; in a traditional CAT, item difficulty adjustments are made for all examinees (albeit in unique ways) whereas construct-irrelevant factor adaptations are triggered only for examinees whose behavior warrants them.

Example: Disengagement on the MAP Growth Assessment

This idea of expanding the meaning of test adaptation to include methods for managing construct-irrelevant factors might be unfamiliar to many readers. To make this more concrete, I will provide examples drawn from the operational CAT of NWEA's *Measures of Academic Progress* (*MAP Growth)* assessment. For this assessment, I will present three different ways that this CAT could, in addition to adjusting item difficulty, adapt to examinee behavior. The first two adaptations have been in operational use with MAP Growth since 2017, while the third has been researched and might be considered for future operational use.

MAP Growth is an online multiple-choice testing system that administers CATs to measure the academic achievement of millions of U.S. students in grades K-12. MAP Growth scores are expressed as scale scores (termed Rasch unit, or RIT) on a vertical scale that permits a student's growth to be assessed when they are tested at different grades and at different times during a school year. MAP Growth is based on the Rasch IRT model, with maximum likelihood estimation (MLE) used to calculate student scores. The RIT scale is centered at 200 with each logit equal to 10 points. MAP Growth assessments are administered with liberal time limits, resulting in test events that are generally unspeeded.

The purpose of MAP Growth is to provide educators information about the academic progress and instructional needs of individual students. However, MAP Growth should be considered low-stakes from a student's perspective because test performance does not typically count toward their school grades. This low-stakes nature invites questions about the engagement levels of students when taking MAP Growth. As a result, test-taking disengagement is considered a major construct-irrelevant factor threatening the validity of score interpretations of MAP Growth. More specifically, disengagement is likely to diminish test performance, leading to RIT scores that underestimate the student's achievement level and thereby mislead educators about that student's instructional needs.

Before joining NWEA, I had previously researched disengagement on low-stakes university general educational assessments. When CBTs were used, item response time was recorded and available, and we had measured disengagement through identification of instances of *rapid-guessing behavior* (Wise & Kong, 2005). Rapid guesses were found to reflect disengagement, as they resembled essentially random responses, with accuracy rates that were usually near chance level. However, the most important and distinctive characteristic of rapid guesses was that they tended to be *psychometrically uninformative* in that their correctness tends to be unrelated to examinee ability (Wise, 2017). This indicated that rapid guesses are not contributing to measurement of an examinee's achievement level.

When I joined NWEA and began studying test event data from MAP Growth, I discovered that rapid guessing was not uncommon. Furthermore, we learned a lot about the dynamics of rapid guessing. For example, it showed a consistent pattern: boys rapid guessed at about twice the rate of girls, and the prevalence of rapid guessing steadily increased with grade. In addition, treating rapid guesses as missing during scoring (because they are uninformative) resulted in RIT scores that were less distorted by disengagement, though with higher standard errors (Wise & Kingsbury, 2016).

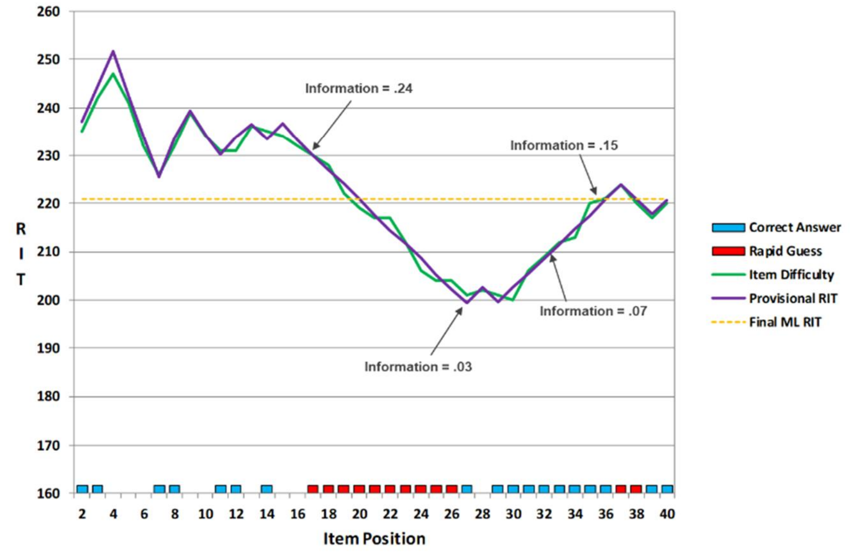
Rapid Guessing Can Distort Item Selection

The characteristics of rapid guessing described above have generally been found with CBTs. However, we discovered an additional problem that uniquely affects a CAT, which uses performance on earlier items to calculate the provisional ability estimates for selected subsequent items during a test event. MAP Growth selects items that should be correctly answered roughly 50% of the time, but the accuracy of rapid guesses is much lower—around chance level (i.e., 20–25%). This implies that when a student exhibits rapid-guessing behavior, the provisional ability estimates will tend to drift below the student's true ability level—resulting in mistargeted item selection. In essence, the item selection algorithm is "confused" by the low-accuracy rapid guesses, which appear consistent with a student of much lower ability. In this way, rapid guesses degrade the CAT's efficiency.

But the problems caused by mistargeting can be even worse. We found that, even though the prevalence of rapid guessing generally increased across item position, at the individual test event level rapid guessing showed many different patterns (Wise & Kingsbury, 2016). It was not uncommon for students to alternate between engagement and disengagement in idiosyncratic ways. For example, a student might become disengaged—exhibiting a string of rapid guesses, but then appear to re-engage—exhibiting a string of solution behaviors (i.e., the engaged counterpart of rapid guesses). This type of pattern poses a special type of challenge for a CAT's efficiency.

As an illustration, Figure 1 shows an actual 40-item MAP Growth test event in reading. The body of the graph shows the provisional ability estimates as they were updated at each item position, along with the closely tracked item difficulty values. Along the horizontal axis, each item response is indicated as correct/incorrect along with whether the response was classified as a rapid guess. In this example, the student initially behaved as we normally would expect on a CAT, showing solution behaviors for the first 16 items, correctly responding to seven of them, with provisional ability estimates beginning to converge around a 235 RIT score. Beginning at the 17th item, however, they exhibited 10 consecutive rapid guesses—each of which were incorrect. Clearly, the student's test-taking behavior had changed, and their provisional ability estimates showed a steady decline down to around a 200 RIT. If we believe that the student's "true" ability level was around 235, the selected item difficulties became increasingly mistargeted during the period of rapid guessing. This mistargeting is reflected in diminished item information values. Again, assuming that true ability equaled 235, the Rasch model theoretical information ($p \times q$) provided by the 16th item response equaled 0.24, whereas by the 26th item it had decreased by nearly 90% to 0.03. This illustrates the effects of rapid guessing on CAT item selection and score precision.

Figure 1. Mistargeted Item Difficulty Selection Caused by Rapid Guessing



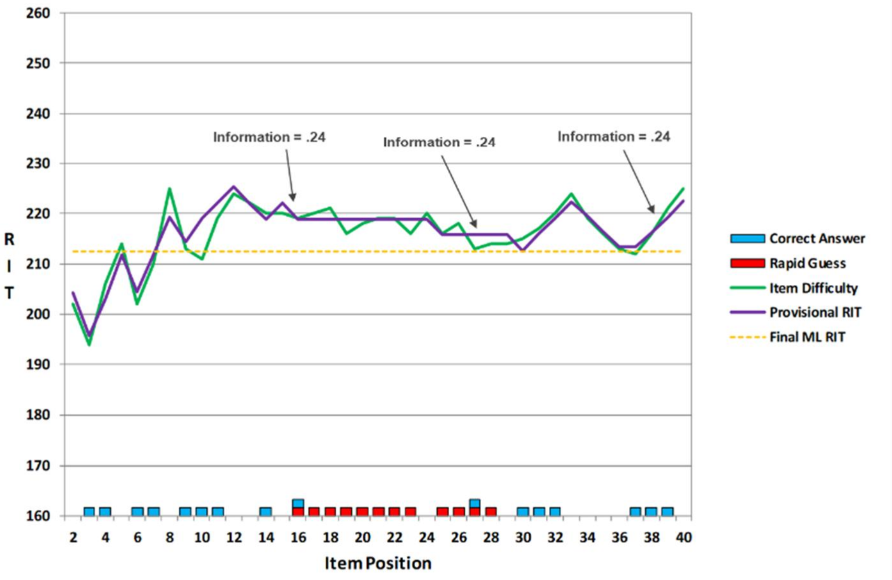
But the most interesting part of the example is what happened next. The student apparently re-engaged, correctly answering 11 of the final 14 items (with only two rapid guesses). This would seemingly be good news; the student had re-engaged and was again providing psychometrically informative responses. But notice that after re-engagement, the responses continued to provide little item information, due to mistargeting, with this deficit persisting through the rest of the test event. That is, the effects of rapid guessing continued long after the disengagement had ceased!

The solution to this problem, while relatively simple, required a new type of adaptation. The test administration software can identify rapid guesses as they occur, and the solution is to treat them as missing when computing provisional ability estimates (Wise & Kingsbury, 2016). This means that the item selection algorithm would not be confused by the low-accuracy rapid guesses, as a provisional ability estimate would be "locked" after a rapid guess. Figure 2 shows a different student's test event after the software had been modified to provide the difficulty locking adaptation. This student showed an engagement pattern similar to that in Figure 1. The first 15 item responses were solution behaviors (eight were correct), followed by 12 rapid guesses over a 13-item span (two were correct), and ending with 12 solution behaviors. If it was assumed that this student's true ability was around 220, the Rasch model item information values were about 0.24 just prior to the first rapid guess at the 16th item and remained at this level for the remainder of the test event. The set of rapid guesses did not cause item difficulty mistargeting, resulting in item selection remaining well-targeted when re-engagement occurred at the 29th item. Thus, Figure 2 shows how a test event could adapt to instances of rapid guessing to preserve the CAT's efficiency.

Effort Monitoring to Curtail Disengagement

A second adaptation implemented in MAP Growth is designed to reduce rapid guessing. When I worked with university general educational assessments, we did the initial research on an *effort monitoring CBT*, which can monitor engagement during a test event in real time and then would intervene by displaying warning messages to a student clearly exhibiting rapid guessing (Kong et al., 2006; Wise et al., 2006). This research found that effort monitoring held promise for curtailing subsequent rapid-guessing behavior and improving test score validity.

Figure 2. Item Difficulty Mistargeting Is Mitigated by Difficulty Locking

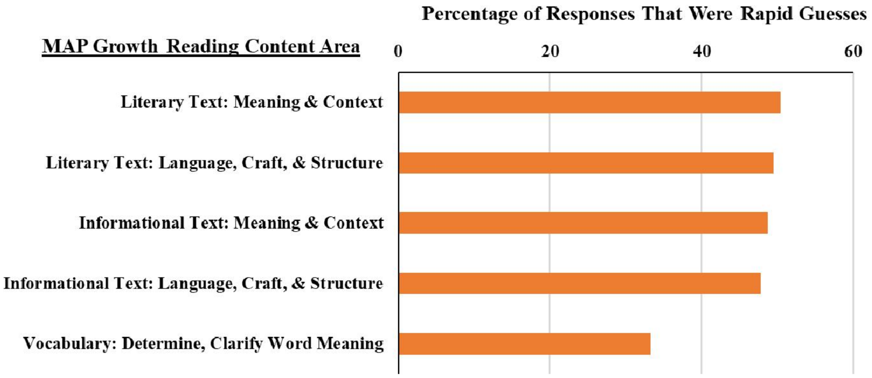


In 2017, MAP Growth became the first large-scale CBT program to operationally implement effort monitoring. Its disengagement intervention, however, was new. Whenever a student displays a predetermined amount of rapid guessing, their test event is automatically paused, and they are encouraged to slow down and identify themselves to the test proctor. At the same time, the test proctor is notified via their computer that that student has been "auto-paused," and they must physically visit the student at their computer to (presumably) encourage the student's reengagement and to resume their test event. A study of the impact of effort monitoring on MAP Growth (Wise et al., 2019) found that, for students who triggered proctor notifications, after the notification occurred (1) item response time increased by 61%, (2) mean rapid guessing rates decreased by 0.34 standard deviations, and (3) mean test performance increased by 0.38 standard deviations. The significant increase in post-notification test performance was especially important because it supports the claim that disengaged students will give rapid guesses to items that they could answer correctly if engaged, rather than rapid guesses indicating the student did not know how to solve the item's challenge. Thus, effort monitoring represents an additional CAT adaptation that can improve the validity of score interpretations for disengaged examinees by curtailing their subsequent rapid-guessing behavior.

Rapid Guessing Varies Across Content Areas

A recent research study of MAP Growth data revealed that the prevalence of rapid guessing varies across content areas (Wise, 2020). That is, some content areas are found more likely to elicit rapid guessing. For example, Figure 3 shows that when MAP Growth reading is administered, students who gave 10% or more rapid guesses did not do so uniformly across the five content areas. Vocabulary items receive rapid guesses at a much lower rate compared to the other content areas (which use reading passages). That this occurs is not particularly surprising, as vocabulary items tend to be much shorter than items containing reading passages. Item length has been shown to be a primary correlate of rapid guessing (e.g., Wise et al., 2009) probably because students perceive that shorter items require less effort to interact with.

Figure 3. The Prevalence of Rapid Guessing Varies Across Content Areas for MAP Growth Reading



Note. The percentages in the graph indicate the prevalence of rapid guessing by students for whom at least 10% of their item responses were rapid guesses.

To understand the importance of differential rapid-guessing rates across content areas, keep in mind that achievement test events are typically built around a planned blueprint, which specifies the degree to which different content areas are to be represented during a test event. If a disengaged student shows differential rapid guessing, the content distribution of their psychometrically informative item responses (i.e., solution behaviors) might deviate significantly from the intended distribution—thereby introducing uncertainty into interpretations made about their test score. Thus, when students rapid guess, the realized content balance might be markedly different from that specified by the test blueprint.

This problem suggests a third adaptation that might be adopted. If a student has shown differential rapid guessing, at the end of a test event the testing software could administer a small number of additional items to make up for the missing informative item responses in an attempt to rebalance the content representation. Such an adaptation remains speculative, however, because its effectiveness has not yet been investigated.

What About Other Construct-Irrelevant Factors?

As the example has shown, difficulty locking, effort monitoring, and content rebalancing are three representative ways that a CAT can adapt to student behavior exhibited during a test event. They each adapt to the presence of test-taking disengagement and are based on item response time rather than the correctness of the student's responses. In these ways, MAP Growth represents a step away from the traditional CAT adaptation of adjusting item difficulty.

How do we identify other potential adaptations? One key requirement is establishing a valid indicator of a particular construct-irrelevant factor. Such an indicator could come in the form of process data (i.e., log files) that can be unobtrusively collected. Response time is the most studied type of process data. Alternatively, some researchers have begun asking examinees to periodically self-report during a test event on some factor of interest, such as motivation or anxiety (e.g., Finney et al., 2020). Finally, we could collect additional biometric measures during test events, such as heart rate, eye-tracking, or scanned cognitive activity, if the use of such measures did not raise disqualifying privacy/intrusiveness or data security/retention concerns. It is essential, however, that an indicator's use be both practical and validated, by either a clear theoretical or empirical link between the indicator and the presence of the target construct-irrelevant factor.

Once a validated indicator has been established, its use would require monitoring by the CAT to detect the presence of the construct-irrelevant factor(s), and incorporation of an intervention designed to preserve test score validity. As a hypothetical example, imagine that highly test-anxious examinees had been found to exhibit a particular pattern of eye movements when viewing an item, and that eye-tracking methodology could be used to detect this pattern in real time during a test event. Anxiety is a construct-irrelevant factor that can negatively affect test performance during high-stakes tests. If a CAT could detect examinee anxiety, it might adapt by switching from a traditional CAT format to a self-adapted test (S-AT)—in which the examinee is allowed to select the difficulty level prior to each item being administered. An S-AT, which provides the examinee some control over their test event, has been found to reduce anxiety and improve performance (Rocklin & O'Donnell, 1987; Wise et al., 1992). Moving from a CAT to an S-AT would reduce testing efficiency somewhat, but it would be an adaptation designed to preserve the validity of that student's score. Note that the adaptations used with disengagement (e.g., proctor notification) would likely be of limited use for anxious examinees, illustrating that different construct-irrelevant factors would require unique types of tailored adaptations.

Concluding Comments

After 50 years of research and development on CATs, I believe we have reached a firm understanding of traditional adaptive testing, which selects item difficulty in pursuit of efficient testing. Furthermore, I believe major advances in this type of traditional adaptation is unlikely in the near future. However, the future of CATs that adapt in other ways is bright.

The beauty of CATs that I saw as a graduate student lay in the way that principles of IRT could be exploited by CBTs to tailor test events for examinees differing in ability. Despite its profound impact on modern measurement, IRT has a serious limitation. It models a world in which the only important examinee characteristic is ability. In this world, examinees are assumed to be always motivated to attain maximum performance, never get tired, and are never anxious. But although this standard IRT model applies reasonably well to many examinees, for some it does not fully capture the reality of ability measurement. We should strive for CATs to be efficient, but not forget that we are measuring people—many of whom will have their test performance influenced by construct-irrelevant factors. Therefore, we should direct efforts toward a new generation of CATs that can detect the presence of these factors and adapt to them in ways that mitigate their impact.

Developing a new generation of CATs will require a (r)evolutionary refocus away from the long-standing traditional CAT goal of reducing the standard error of ability estimation and toward the broader goal of maximizing the validity of each examinee's score. Attaining this goal will move us toward more individualized testing in which the test adapts, as needed, to elicit maximally valid scores from each examinee. It is worth noting that a traditional CAT—which administers a unique set of items to each examinee—already represents a form of individualized testing. As we expand the ways in which a CAT can adapt, individualized testing promises to improve test score validity by becoming *caring assessments* (Zapata-Rivera et al., 2020) that actively pursue maximum validity by adapting to the presence of construct-irrelevant factors. This evolution of what a CAT can become, however, brings with it an important challenge to measurement researchers: to cultivate new, validated, real-time indicators of construct-irrelevant factors and to develop effective means of adapting test events to manage their presence.

References

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). Harper & Row.

Finney, S. J., Perkins, B. A., & Satkus, P. (2020). Examining the simultaneous change in emotions during a test: Relations with expended effort and test performance. *International Journal of Testing, 20*(4), 274–298. CrossRef

Kong, X. J., Wise, S. L., Harmes, J. C., & Yang, S.-T. (2006, April 8–10). *Motivational effects of praise in response-time based feedback: A follow-up study of the effort-monitoring CBT* [Paper presentation]. Annual meeting of the National Council on Measurement in Education, San Francisco, CA, United States.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11. CrossRef

Rocklin, T. R., & O'Donnell, A. M. (1987). Self-adapted testing: A performance-improving variant of computerized adaptive testing. *Journal of Educational Psychology*, *79*(3), 315–319. CrossRef

Sireci, S. G. (2020). Standardization and understandardization in educational assessment. *Educational Measurement: Issues and Practice*, *39*(3), 100–105. CrossRef

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretations, and implications. *Educational Measurement: Issues and Practice, 36*(4), 52–61. CrossRef

Wise, S. L. (2020). The impact of test-taking disengagement on item content representation. *Applied Measurement in Education*, *33*(2), 83–94. CrossRef

Wise, S. L., Bhola, D. S., & Yang, S.-T. (2006). Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. *Educational Measurement: Issues and Practice 25*(2)*,* 21–30. CrossRef

Wise, S. L., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement, 53*(1), 86–105. CrossRef

Wise. S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183. CrossRef

Wise, S. L., Kuhfeld, M. R., & Soland, J. (2019). The effects of effort monitoring with proctor notification on test-taking engagement, test performance, and validity. *Applied Measurement in Education*, *32*(2), 183–192. CrossRef

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*(2), 185–205. CrossRef

Wise, S. L., Plake, B. S., Johnson, P. L., & Roos, L. L. (1992). A comparison of self-adapted and computerized adaptive tests. *Journal of Educational Measurement*, *29*(4), 329–339. CrossRef

Zapata-Rivera, D., Lehman, B., & Sparks, J. R. (2020). Learner modeling in the context of caring assessments. In R. A. Sottilare & J. Schwarz (Eds.), *Adaptive instructional systems: International Conference on Human-Computer Interaction* (pp. 422–431). Springer. CrossRef

Author's Address

Steven L. Wise, NWEA, 121 NW Everett St., Portland, Oregon 97209, U.S.A.
Email: steven.wise@hmhco.com