Controlling construct-irrelevant factors through computer-based testing: disengagement, anxiety, & cheating

Steven L. Wise
NWEA, Portland, Oregon, USA

ABSTRACT
A decision of whether to move from paper-and-pencil to computer-based tests is based largely on a careful weighing of the potential benefits of a change against its costs, disadvantages, and challenges. This paper briefly discusses the trade-offs involved in making such a transition, and then focuses on a relatively unexplored benefit of computer-based tests – the control of construct-irrelevant factors that can threaten test score validity. Several unique advantages provided by computer-based tests are described, and how these advantages can be used to manage the effects of several common construct-irrelevant factors is discussed. Ultimately, the potential for expanded control may prove to be one of the most important benefits of computer-based tests.

KEYWORDS
Computer-based testing; construct-irrelevant factors; test-taking engagement; test anxiety

There has been a growing movement in large-scale educational assessment away from traditional paper-and-pencil tests (PPTs) towards computer-based tests (CBTs). As our societies and educational systems have become increasingly computerised, educators, students, and parents have witnessed many aspects of education becoming computer-based. It is therefore not unexpected that many should view CBTs as a manifestation of this trend. In this view, because CBTs represent the future of assessment, we should be drawn to developing and implementing them.

For an operational assessment program that originated as a PPT program, however, a decision to make a transition to CBT is potentially disruptive and requires a careful weighing of the costs and disadvantages of such a change against the potential benefits. In this paper, I will discuss the trade-offs involved in a decision to make such a transition, and then focus on a relatively unexplored advantage of CBTs – the control of construct-irrelevant factors – that may prove to be one of the greatest benefits to their use. I will argue that, ultimately, the attractiveness of CBTs will increase as they evolve into something quite different from what we see today.

Back in the 1980’s, computer-based testing was new, measurement practitioners were beginning to understand the potential benefits of CBTs, and the first computer-based assessment programs were emerging. Nearly 30 years ago, Barbara Plake and I provided an overview of the advantages and disadvantages of CBTs, and then speculated on their future use (Wise & Plake, 1989). It was instructive to go back and read that article with
the benefit of hindsight. Although most of the points we made remain valid today, we were clearly less prescient regarding some important issues. Probably the largest issue we did not anticipate was the enormous impact that the internet is having on educational measurement. Internet-based testing, which provides a strong basis for ubiquitous, on-demand testing (Luecht, 2016) has become increasingly common in the U.S.A. and is beginning to be used by international assessment programs.

**Costs of a CBT testing program**

One characteristic aspect of implementing CBT programs is that they are expensive. They typically require a substantial amount of initial resources to get the program up and running. There will be costs associated with infrastructure; if the test is to be delivered via the internet, adequate servers and other hardware will be needed to ensure adequate, reliable connectivity during testing. In addition, there need to suitable rooms, with sufficient computers or devices to provide a standardised testing environment. Computerised versions of test items will need to be developed, and test administration software will need to be acquired or developed. Often, the test administration software will need to be able to perform reliably on multiple computer operating systems and/or internet browsers.

Even after a CBT program has been initiated, maintaining the program will require substantial resources and personnel. Computer operating systems and internet browsers are frequently upgraded, which will require technical support to ensure that the testing program continues to function reliably. Additional test items will need to be developed and computerised.

Beyond the implementation and maintenance issues, transitioning from a PPT to a CBT program brings with it several psychometric issues that will require attention. The usability of the test administration software by test takers may pose a challenge, along with the related concern of test takers’ familiarity with its use. Moreover, if the number of test takers far exceeds the number of available computers, testing may need to be completed over a period of time. This practical issue carries with it the disadvantage that later test takers may learn about test content from those tested earlier, which threatens test security and, consequently, test score validity.

The most common challenge associated with a transition from PPT to CBT testing modes is score comparability. Although, ideally, test takers will perform equivalently under both testing modes, there are a number of factors (e.g. the way the CBT items appear or the test takers’ affective reactions to CBTs) that potentially can induce mode effects in performance. Consequently, score comparability should be investigated for each PPT-to-CBT transition. Useful discussions of this issue are provided by Leeson (2006) and Way, Davis, Keng, and Strain-Seymour (2016).

**The positive outcomes to moving to a CBT**

A decision to make the transition to a CBT depends, at least in part, on the strength of an argument that the benefits and advantages of CBTs outweigh the costs and limitations described above. Such an argument has two primary themes. The first is that testing can become more efficient and/or convenient. When selected response item
types are used, immediate scoring and reporting of results is possible. In addition, the capability to provide computer-based, automated scoring of answers to constructed-response items is steadily improving (Bennett & Zhang, 2016).

Furthermore, when an adequate-sized item pool is available, and immediate scoring is feasible, a computerised adaptive test (CAT) can be used. A CAT requires far fewer items to attain measurement precision comparable to that from a conventional fixed-form test or, equivalently, can attain a higher degree of precision in the same amount of testing time as a fixed-form test. In this way, CATs can provide more efficient measurement than traditional fixed-form tests (Wainer, 2000; Weiss & Kingsbury, 1984).

Some proponents have claimed that CATs are more motivating than PPTs or non-adaptive CBTs, because CAT test takers do not receive items that are far too easy or too difficult for them. While I have found some evidence for this claim (Wise, 2014), particularly for lower-achieving students, higher motivation has not typically been found to be accompanied by higher test performance. Interestingly, I noted one case in which a CAT did improve mean test performance over a conventional test, but this advantage was statistically equated away in order to attain score comparability between the two testing modes (Wise, 2014). That is, by focusing on adjusting for mode effects, an apparent performance advantage for a CBT was negated.

Beyond efficiency, CBTs may be considered more convenient than PPTs because they are both conducive to on-demand testing and because they can be taken from a variety of locations (if that is a desirable capability). Moreover, because many students have a particular computer they use at school, it is often desirable and convenient to test them on a computer with which they are familiar.

The second theme of the argument in favour of CBTs is that they can produce scores with higher validity (Hu & Sireci, 2001). The most commonly cited aspect of this theme is the greatly expanded types of items a CBT can administer. A CBT can, for example, require a test taker to use multiple senses to interact with tasks involving audio, video, and haptic elements. In addition, test takers can conduct simulations and experiments, which enhance our ability to measure their higher-order skills such as problem solving. This suggests that a transition from PPT to CBT brings with it enormous possibilities regarding the fidelity with which we test students and expands our ability to administer items that target a content domain more effectively. For example, a job analysis in a medical field might specify that the professional be able to competently listen to and interpret a patient’s breath sounds. With a PPT, we would be generally unable to assess this skill. With a CBT, in contrast, we could readily use its audio capabilities to assess the skill in multiple ways. This implies that when CBTs are used, we often can more comprehensively measure the essential elements of a content domain.

Improved coverage of a content domain, however, is not the only way that test score validity can be improved when a CBT is used. It can also be improved by reducing the effects of construct-irrelevant factors (Haladyna & Downing, 2004). Messick (1984) noted that “educational achievement tests, at best, reflect not only the psychological constructs of knowledge and skills that are intended to be measured, but invariably a number of contaminants” (p. 216). Such contaminants comprise influences on test scores that have nothing to do with the construct we are trying to measure. For example, if students’ performance on a mathematics test is too dependent on their level of reading comprehension, then reading comprehension becomes a construct-
irrelevant factor, because the definition of achievement in mathematics does not include reading comprehension (Haladyna & Downing, 2004).

Construct-irrelevant factors introduce systematic error variance into scores and thereby reduce validity. This suggests that if we can reduce the impact of construct-irrelevant factors, we can improve validity. This idea is analogous to how a signal-to-noise ratio might be improved. We can either enhance the signal (i.e. measure the domain better), or reduce the noise (i.e. reduce construct-irrelevant factors).

How can using a CBT reduce the effects of construct-irrelevant factors? CBTs have three key advantages over PPTs: they can provide greater control over a test event, they can measure more types of information, and they can adapt the test event to test-taker behaviour. These advantages can work singly, or in combination, to improve validity. In this section, I will discuss these advantages, and in the following section I will discuss how these advantages can be used to reduce construct-irrelevant factors.

**Greater control**

The use of CBTs allows test givers to exert a degree of administrative control over a test event that is not practically possible with a PPT. When PPTs are used, particularly when given in a typical group administration, test takers are permitted a great deal of control over how they interact with the test items. A test taker might review all of the items before answering any, answer items in any order they choose, review/change answers multiple times, and omit answering altogether if they choose.

In contrast, when a CBT is used, much more control can be imposed. The test giver can control many aspects of a test event, including (a) the order with which items are seen, (b) whether an answer must be given before moving on to the next item, (c) whether answers can be reviewed/changed, or (d) the difficulty of the items given during a CAT. The amount of time a test taker is allowed to interact with each item could be controlled, if desired. In addition, if accommodations are provided, the test giver can control which types are provided to which test takers. A CBT could also detect and notify the test giver of incomplete or inadvertently omitted answers and give the test taker an opportunity to complete these items prior to completing their test.

Note that the degree of control a test giver exerts over a CBT test event is inversely related to the amount of control the test taker is permitted. In this way, the test giver controls the degree of control that the test taker is allowed. Beyond this relationship, however, a CBT could also provide a test taker with additional control not available with PPTs. An example of this would be if test takers are permitted to choose the difficulty levels of the items they receive, with test performance calculated using item response theory (IRT) methods to account for difficulty levels chosen (Rocklin & O’Donnell, 1987).

**More information can be gathered**

At the completion of a PPT test event, the only information typically available to the test giver is the responses made by the test taker to the items on the test. On a CBT, in
contrast, multiple types of information about the behaviour of the test taker are potentially available. The most commonly studied information has been the amount of time a test taker spent on each item (termed *item response time*). But numerous other actions taken by a test taker might be collected, including (a) the computer location from which the test was taken, (b) the order with which items were completed, (c) exactly when answers were given, (d) which items were viewed versus not reached, (e) which answers were reviewed, and (f) how answers were changed. In addition, a variety of biometric data could be collected, examples of which include eye tracking, skin conductance, pulse rate, fingerprints, and brain wave data.

**Adapting to test-taker behaviour**

CBTs could adapt in a variety of ways to a test taker’s behaviour exhibited during a test event. A CAT – which adapts the difficulty of items administered to each test taker – is the most commonly seen example of this. But a CAT is not the only way a CBT could be adaptive. There are potentially many ways in which a CBT could adapt to test taker behaviour during a test event and improve test score validity (Wise, in press). Examples of this will be described below.

**Reducing the impact of construct-irrelevant factors**

In any measurement context there are likely to be clearly identifiable construct-irrelevant factors that are most likely to adversely threaten score validity. For example, in low-stakes educational assessments, test-taker disengagement tends to be the most significant issue, whereas in high-stakes assessments, test anxiety or cheating are more of a concern. Managing the validity threat posed by a particular construct-irrelevant factor requires one or both of the following components: the development of measures for detecting its presence, and the adoption of testing methods that can reduce its impact. Each of the advantages provided by CBTs (greater control, more information, adaptation) enhances our ability to establish the two components.

In the following sections, I discuss several construct-irrelevant test-taking factors – disengagement, anxiety, and, cheating – that I feel are particularly amenable to management through adoption of a CBT. These factors, along with a brief description of the role that CBT could play in their management, are shown in Table 1.

**Test-taking disengagement**

Attaining valid scores on achievement tests requires engaged test takers who direct effort towards applying their knowledge, skills, and abilities to solving the challenges posed by the test items they receive (Eklöf, 2010; Wise, 2015). Without sufficient effort, it is difficult to understand the degree to which poor test performance is attributable to lack of knowledge, absence of motivation, or both. An interpretation of a poor test score as valid therefore requires a discounting of plausible construct-irrelevant factors, such as disengaged test taking, that can produce poor performance. The impact of disengaged test taking on test score validity has been well established. It constitutes a person-specific, construct-irrelevant behaviour that tends to negatively distort
achievement test scores (Haladyna & Downing, 2004). The amount of distortion can be sizable; it has been found that unmotivated test takers tend to underperform their motivated peers by more than one-half standard deviation (Wise & DeMars, 2005).

Several characteristics of disengaged test taking are important to the current discussion. First, disengaged test taking occurs most frequently with low-stakes tests, for which test takers are likely to perceive an absence of personal consequences associated with their test performance (Knekta, 2017; Wolf & Smith, 1995). Second, it appears that disengagement rarely occurs over an entire test event (Wise & Kong, 2005; Wolf, Smith, & Birnbaum, 1995). Third, during a test event, test takers can move in and out of engagement in an idiosyncratic fashion (Wise & Kingsbury, 2016), driven largely by variation in how mentally taxing the test items are (Wolf et al., 1995).

There are several methods by which a test taker’s engagement can be assessed. First, the test proctor could observe and provide a rating of the test taker’s engagement. Second, the test taker could be asked to complete a self-report instrument immediately after a test event (Eklöf, 2006; Sundre & Moore, 2002). Third, the test taker’s item responses could be statistically analysed to assess the degree to which they fit the measurement model being used (Meijer, 2003). Fourth, test takers’ behaviour (beyond simply the item responses) could be studied to assess the degree to which they behaved in a manner consistent with that expected under engaged test taking.

The last method for assessing engagement – analysing test-taker behaviour – is particularly compatible with a CBT’s capability for collecting additional information about a test event. This method has the unique advantage that it can allow engagement to be assessed down to the level of individual item responses. The behaviour most frequently studied is item response time. Schnipke (1995) studied the data from timed, high-stakes, multiple-choice CBTs and showed that as time was running out, some test takers would begin quickly filling in answers to remaining items in hopes of getting some of them correct through lucky guessing. This rapid-guessing behaviour can be characterised as a test taker submitting an answer before they had a chance to fully read and understand the challenge posed by the item. Schnipke concluded that the presence

| Construct-Irrelevant Factor | Additional Information | CBT Advantage |
|-----------------------------|------------------------|---------------|
| Disengaged Test Taking      | Response time used to identify rapid guessing | Effort-moderated scoring |
| Test Anxiety                | Biometric information  | Self-adapted testing |
| Cheating Behaviour          | Response time plus response accuracy | Proctor notification |

Table 1. Examples of using CBT advantages to identify and control construct-irrelevant factors.
of rapid guessing at the end of a high-stakes test event indicated that the test was speeded for that test taker.

A decade later, when analysing data from untimed, low-stakes CBTs, a graduate student and I discovered rapid guessing occurring throughout test events, and not just at the end as Schnipke (1995) had found (Wise & Kong, 2005). We demonstrated that in low-stakes contexts, rapid guessing indicated unmotivated, disengaged test taking. We proposed an index based on rapid guessing – termed response time effort (RTE) – for measuring the overall level of effort a test taker exhibited during a test event.

Drawn from this initial research on rapid guessing, two central ideas formed the foundation of a multifaceted strategy for addressing the construct-irrelevant factor of disengaged test taking. The first was that each item response could be classified as either a rapid guess or its effortful solution behaviour counterpart. The second was that these classifications could be aggregated to produce an overall measure of engagement. For example, the classifications are aggregated across the items a test taker received to calculate RTE, which equals the proportion of item responses for which a test taker exhibited solution behaviour. An analogous index for items, response time fidelity (RTF), aggregates the classifications across test takers and indicates the amount of engagement the item received (Wise, 2006).

Based on the response time information that CBTs can collect and use to identify rapid-guessing behaviour, several solutions have been developed to reduce the impact of disengaged test taking. One solution is to include information about test taker engagement on score reports. Such information can help persons interpreting test scores (i.e. educators, policy makers, and parents) draw more valid inferences about student test performance. A second solution is to statistically adjust scoring to account for disengagement. Rapid guesses tend to be psychometrically uninformative, as their correctness has been found to have little, if any, relationship to a test taker’s achievement level (Wise, 2017; Wise & Gao, 2017; Wise & Kingsbury, 2016). This finding, coupled with the fact that rapid guessing tends to negatively distort scores, suggests that if rapid guesses are excluded from scoring, validity will increase. This is what has been found with effort-moderated scoring (Wise & DeMars, 2006), which produces IRT-based scores based only a test taker’s solution behaviours.

Although providing engagement information on score reports and using effort-moderated scoring are useful ways to manage the problems caused by disengaged test taking, they represent remedies that are applied after a test event has been completed and disengagement has occurred. Use of a CBT can also be used with a different type of remedy that can be applied in real time during a test event. In effort monitoring, the CBT software classifies each item response as solution behaviour or a rapid guess immediately after it happens. If a test taker begins to exhibit multiple rapid guesses, the CBT initiates some type of intervention intended to re-engage the test taker and eliminate subsequent rapid guessing. Note that this utilises a different advantage of CBTs – the ability to adapt to the behaviour of a test taker during a test event.

In the initial implementation of effort monitoring, messages were sent directly to the computer screens of disengaged test takers. These messages noted that the test taker appeared to be disengaged (while making no mention of rapid guessing) and encouraged the test taker to give more effort. In two experimental studies (Kong, Wise, Harmes, & Yang, 2006; Wise, Bhol, & Yang, 2006) this effort-monitoring CBT was
found, for test takers deserving messages, to (a) increase RTE by about one standard deviation, (b) increase test performance by nearly one-half standard deviation, and (c) improve the correlations of test scores with external variables. These studies illustrate how a CBT can proactively change test taker behaviour and improve test score validity.

Recently, my own organisation (NWEA) has implemented a different type of effort monitoring with its MAP® Growth™ assessment, which is a group-administered, low-stakes CAT administered in many U.S. school districts. During a MAP Growth administration, the proctor has a computer console from which to track the progress of each student. If a particular student begins to exhibit rapid-guessing behaviour, the proctor is notified (rather than the student) and encouraged to intervene with the student to help them re-engage with the test. Moreover, the CAT item selection algorithm in MAP Growth has been modified to ignore rapid guesses, whose presence can confuse item selection and result in mis-targeted items being administered. These testing features have been found to have positive effects on test-taking engagement, test performance, and test score validity (Wise, Kuhfeld, & Soland, 2018).

Although the previous examples have all focused on rapid guessing to identify disengaged test taking in multiple-choice tests, these ideas have recently been extended. Wise and Gao (2017) showed that omitting answers or entering very brief responses to constructed-response items also indicate disengagement, if these behaviours also occur rapidly. In addition, Harmes and Wise (2016) investigated engagement on complex multiple-choice items that required test takers to perform certain actions when presented items, such as viewing tabbed information or playing video clips. Harmes and Wise found that the degree to which a test taker performed all of the actions expected from an engaged test taker provided a novel measure of item-level engagement based on behaviours other than response time.

Test anxiety

In low-stakes tests, disengagement constitutes a major construct-irrelevant factor that threatens validity. In contrast, with high-stakes testing an important validity threat comes not from disengagement, but from the anxiety test takers can feel during a test event about their performance. Zeidner (1998) stated that “test-anxious students are characterised by a particularly low-response threshold for anxiety in evaluative situations, tending to view evaluative situations, in general, and test situations, in particular, as personally threatening” (p. 18). The prevalence of test anxiety can be sizable; Hill (1984) estimated that over a quarter of U.S. students suffer the effects of debilitating stress in evaluative situations. Such stress can have a systematic negative effect on test performance. Based on meta-analytic results, Hembree (1988) found correlations between test anxiety and performance to generally be around -.25.

How can the use of a CBT help address the debilitating effects of test anxiety? Although the research on this issue is much less developed than that found for test-taking disengagement, the studies conducted to date have been encouraging. One promising set of research efforts is based on the idea of perceived control – that people can tolerate a stressful situation better if they perceive they have some control over their situation (Lazarus & Folkman, 1984; Wise, 1994). Applied to a CBT, this idea has been studied using a self-adapted test (S-AT; Rocklin & O’Donnell, 1987). In a S-AT, the test taker is sequentially
asked to choose the difficulty level of each item (though not the item itself) from a set of difficulty levels. The overall set of item responses is then scored using IRT-based methods, which take into account the difficulty levels of the items that had been administered.

An experimental study conducted by Rocklin and O'Donnell (1987) found that a S-AT yielded scores that were both significantly higher and less related to self-reported anxiety level than scores from a fixed-form CBT. A follow-up study (Wise, Plake, Johnson, & Roos, 1992) found that test takers randomly assigned to take a S-AT exhibited higher test performance and lower posttest state anxiety than those assigned to take a CAT, while the mean item difficulty levels administered under the two test types were virtually the same. In a related study, when test takers were allowed to choose between receiving a CAT or a S-AT, those reporting higher levels of pretest state anxiety were much more likely to choose the S-AT (Wise, Roos, Plake, & Nebelsick-Gullett, 1994).

Use of a S-AT does not come without disadvantages. A S-AT tends to take more time to complete than a CAT or a fixed-form test, due to the extra time needed for item difficulty choices. In addition, though on S-AT most test takers appear to choose difficulties well matched to their achievement levels (similar to that from a CAT), some make poorly matched choices, resulting in test scores with high standard errors. Moreover, Pitkin and Vispoel (2001) performed a meta-analysis of the effects of a S-AT on test performance, and found only a small mean effect size. It is important to note, however, that the studies they synthesised had all been done in low-stakes contexts, such as with university volunteer research subjects. In consequential, high-stakes testing contexts, it seems reasonable to expect that test anxiety levels would be higher, and that a S-AT would likely have a greater positive impact.

Beyond the studies of S-ATs, limited research has been reported on the relationships between CBTs and test anxiety. Ortner and Caspers (2011) found that CAT can have a negative impact on the performance of highly anxious test takers. This finding is likely due to the fact that, in a CAT, test takers tend to pass only about half of their test items, which could be anxiety-provoking for test takers who are accustomed to passing a much higher percentage of the items they receive on fixed-form tests. This suggests that it might be useful if a CAT could adapt to anxious test takers by administering easier items that they have a much higher probability of passing. While such a practice would decrease the efficiency benefits of a CAT, this might be more than offset by measurement benefits resulting from reduced anxiety. This idea is speculative, however, and research should be directed towards better understanding its net effect on score validity.

The abovementioned idea requires, however, that we be able to measure a test taker’s level of anxiety during a test. There are several possibilities for how to do this, though I am unaware of research into their efficacy. One might periodically ask a test taker during a test event to self-report their current anxiety level. Alternatively, it may be found that some behaviour-based information, such as response time or person fit statistics can validly indicate test anxiety levels. In addition, the types of biometric information described earlier might be useful to collect during a test event, such as heart rate, skin conductance, or eye tracking. And in the future, brainwaves or other neuropsychological information that could identify the presence of test anxiety may become unobtrusively collectable in operational testing settings. But until we can validly measure test anxiety in real time, it will be difficult to develop CBTs that can adapt to its presence.
Cheating behaviour

A completely different type of construct-irrelevant factor commonly associated with high-stakes tests is cheating behaviour, in which an attempt is made to illicitly attain a test score that overstates what a test taker knows and can do. Although there are numerous ways that test takers (and test givers) can cheat, CBTs could generically help in the detect of these types of test taking behaviour through the additional information it can collect, greater control, and adaptation by the CBT to test-taker behaviour. For example, one way in which test takers could cheat is by acquiring advance knowledge of test items. Such advance knowledge could be detected using item response time information if the test takers give rapid, correct responses (Meijer & Sotaridona, 2006; Qian, Staniewska, Reckase, & Woo, 2016). A second example would be the discrete-option multiple-choice item (Foster & Miller, 2009), in which the CBT controls the administration of each item’s response options in a way that renders the item more difficult to cheat on and steal from. A third example would be if one test taker copied answers from another during a test event. Such test takers would have highly similar response patterns. Observing a similar response pattern, in and of itself, would not provide strong evidence of copying, but if the CBT system could show that (a) the test takers’ computers were in close proximity during the test event and (b) one test taker’s responses were consistently submitted shortly after those from the other test taker, the evidence would be much more compelling. Moreover, we would have useful information regarding who had copied from whom.

Research to date on the use of CBTs to manage cheating behaviour has tended to focus primarily on both post-test detection and on testing methods that reduce the prevalence of cheating. A good overview of this research is provided by Foster (2016). It is likely, however, that some of the methods developed could be used to adapt to a test taker’s behaviour. For instance, in the copying example above, a CBT might perform real-time cheating monitoring and provide warnings to test takers or notifications to proctors when illicit behaviour was detected.

Closing comments

Nearly three decades ago, Bunderson, Inouye, and Olsen (1989) described their vision of what computer-based measurement would become, through their four generations of computerised testing. They discussed an evolution from standalone fixed-form CBTs towards advanced assessments that integrate assessment and instruction. In some ways, computer-based assessment has progressed in this fashion. The progression of Bunderson et al., however, is based on the same fundamental assumption that is inherent in our measurement models: that if you administer an item to a test taker, the response will reflect only what the test taker knows and can do. But we know that there are frequent instances in which that assumption is invalid. Test takers sometimes will not demonstrate what they know and can do, because of construct-irrelevant factors.

Because construct-irrelevant factors such as disengagement, test anxiety, and cheating threaten the validity of inferences we make based on educational assessments, it is essential that we be able to effectively manage their presence. CBTs, with their capability to both collect an array of valuable information about test events and adapt to test taker
behaviour, will play an increasingly important role in the managing these types of validity threats. Future developments in artificial intelligence and other types of machine learning will inevitably become incorporated in our educational assessments, and as we learn more about construct-irrelevant factors, CBTs will enhance the collection of increasingly valid measures of a student’s achievement.

Notes on contributor

Steven L. Wise is a Senior Research Fellow at NWEA. Dr. Wise has published extensively during the past three decades in applied measurement, with particular emphases in computer-based testing and the psychology of test taking. In recent years, Dr. Wise’s research has focused primarily on methods for effectively dealing with the measurement problems posed by examinee disengagement on achievement tests.

References

Bennett, R. E., & Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.), Technology and testing: Improving educational and psychological measurement (pp. 142–173). New York: Routledge.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerised educational measurement. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 367–408). New York: Macmillan.

Eklöf, H. (2006). Development and validation of scores from an instrument measuring student test-taking motivation. Educational and Psychological Measurement, 66, 643–656.

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. Assessment in Education: Principles, Policy & Practice, 17, 345–356.

Foster, D. (2016). Testing technology and its effects on test security. In F. Drasgow (Ed.), Technology and testing: Improving educational and psychological measurement (pp. 235–255). New York: Routledge.

Foster, D., & Miller, H. L., Jr. (2009). A new format for multiple-choice testing: Discrete-option multiple-choice. Results from early studies. Psychology Science Quarterly, 51, 355.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. Educational Measurement: Issues and Practice, 23(1), 17–27.

Harmes, J. C., & Wise, S. L. (2016). Assessing engagement during the online assessment of real-world skills. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), Handbook of research on technology tools for real-world skill development (pp. 804–823). Hershey, PA: IGI Global.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. Review of Educational Research, 58, 47–77.

Hill, K. T. (1984). Debilitating motivation and testing: A major educational problem, possible solutions, and policy applications. Research on Motivation in Education: Student Motivation, 1, 245–274.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. Educational Measurement: Issues and Practice, 20(3), 16–25.

Knekta, E. (2017). Are all pupils equally motivated to do their best on all tests? Differences in reported test-taking motivation within and between tests with different stakes. Scandinavian Journal of Educational Research, 61(1), 95–111.

Kong, X. J., Wise, S. L., Harmes, J. C., & Yang, S. (2006, April). Motivational effects of praise in response-time based feedback: A follow-up study of the effort-monitoring CBT. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Lazarus, R. S., & Folkman, S. (1984). Stress, coping and appraisal. New York: Springer.
Leeson, H. V. (2006). The mode effect: A literature review of human and technological issues in computerised testing. *International Journal of Testing*, 6, 1–24.

Luecht, R. M. (2016). Computer-based test delivery models, data, and operational implementation issues. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 179–205). New York: Routledge.

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods*, 8, 72–87.

Meijer, R. R., & Sotaridona, L. S. (2006). *Detection of advance item knowledge using response times in computer adaptive testing*. Law School Admission Council Computerised Testing Report 03-03. Newtown, PA: Law School Admission Council.

Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement*, 21, 215–237.

Ortner, T. M., & Caspers, J. (2011). Consequences of test anxiety on adaptive versus fixed item testing. *European Journal of Psychological Assessment*, 27(3), 157–163.

Pitkin, A. K., & Vispoel, W. P. (2001). Differences between self-adapted and computerised adaptive tests: A meta-analysis. *Journal of Educational Measurement*, 38, 235–247.

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement: Issues and Practice*, 35(1), 38–47.

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

Schnipke, D. L. (1995). *Assessing speededness in computer-based tests using item response times* (Unpublished doctoral dissertation). Johns Hopkins University.

Sundre, D. L., & Moore, D. L. (2002). The student opinion scale: A measure of examinee motivation. *Assessment Update*, 14(1), 8–9.

Wainer, H. (2000). Introduction and history. In H. Wainer (Ed.), *Computerised adaptive testing: A primer* (pp. 1–21). Mahwah, NJ: Lawrence Erlbaum Associates.

Way, W. D., Davis, L. L., Keng, L., & Strain-Seymour, E. (2016). From standardization to personalization: The comparability of scores based on different testing conditions, modes, and devices. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 260–284). New York: Routledge.

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

Wise, S. L. (1994). Understanding self-adapted testing: The perceived control hypothesis. *Applied Measurement in Education*, 7, 3–14.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes, computer-based test. *Applied Measurement in Education*, 19, 93–112.

Wise, S. L. (2014). The utility of adaptive testing in addressing the problem of unmotivated examinees. *Journal of Computerised Adaptive Testing*, 2(1), 1–17.

Wise, S. L. (2015). Effort analysis: individual score validation of achievement test data. *Applied Measurement in Education*, 28, 237–252. doi:10.1080/08957347.2015.1042155

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

Wise, S. L. (in press). An intelligent CAT that can deal with disengaged test taking. In H. Jiao & R. W. Lissitz (Eds.), *Applications of artificial intelligence (AI) to assessment*. Charlotte, NC: Information Age Publishing, Inc.

Wise, S. L., Bhola, D., & Yang, S. (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.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10, 1–17.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, 43, 19–38.

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*, 30, 343–354.
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*, 86–105.

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

Wise, S. L., Kuhsfeld, M. R., & Soland, J. (2018, April). *The effects of effort monitoring with proctor notification on test-taking engagement, test performance, and validity*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Wise, S. L., & Plake, B. S. (1989). Research on the effects of administering tests via computers. *Educational Measurement: Issues and Practice, 8*, 5–10.

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*, 329–339.

Wise, S. L., Roos, L. L., Plake, B. S., & Nebelsick-Gullett, L. J. (1994). The relationship between examinee anxiety and preference for self-adapted testing. *Applied Measurement in Education, 7*, 81–91.

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8*, 227–242.

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test motivation, and mentally taxing items. *Applied Measurement in Education, 8*, 341–351.

Zeidner, M. (1998). *Test anxiety: The state of the art*. New York: Plenum Press.