Practical Assessment, Research, and Evaluation

Volume 25 Article 8

2020

A Framework for Evaluating Stopping Rules for Fixed-Form Formative Assessments: Balancing Efficiency and Reliability

Deni L. Basaraba
Bethel School District #52

Paul Yovanoff
Southern Methodist University

Pooja Shivraj
American Board of Obstetrics & Gynecology

Leanne R. Ketterlin-Geller
Southern Methodist University

A Framework for Evaluating Stopping Rules for Fixed-Form Formative Assessments: Balancing Efficiency and Reliability

Cover Page Footnote

Funding: This work was supported by the Texas Education Agency (80-509001)

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 25 Number 8, October 2020 ISSN 1531-7714

Evaluating Stopping Rules for Fixed-Form Formative Assessments: Balancing Efficiency and Reliability

Deni L. Basaraba, *Bethel School District #52*
Paul Yovanoff, *Southern Methodist University*
Pooja Shivraj, *American Board of Obstetrics & Gynecology*
Leanne R. Ketterlin-Geller, *Southern Methodist University*

Stopping rules for fixed-form tests with graduated item difficulty are intended to stop administration of a test at the point where students are sufficiently unlikely to provide a correct response following a pattern of incorrect responses. Although widely employed in fixed-form tests in education, little research has been done to empirically evaluate the stopping rules in these tests that often have important instructional and/or placement implications for students. In this manuscript, we propose and research a framework for evaluating stopping rules with respect to two important and sometimes conflicting criteria: (1) efficiency, and (2) reliability. Using this framework, we provide an example in which we apply three increasingly complex methods for evaluating efficiency and two methods for examining reliability.

Many formative (and summative) assessments employ stopping rules (i.e., ceiling rules or discontinue rules) that specify the point at which test administration is discontinued but still provides sufficient information to support valid uses and interpretation of the results (Lonigan, Allan, & Lerner, 2011; Pearson, 2016). Examples of common stopping rules include specifying the number of items that either must be missed consecutively (e.g., zero words read correctly in the first 10 words of a reading passage; Good & Kaminski, 2002) or that can be missed within a set of given items (e.g., three items answered incorrectly within a set of five items) before administration of a test is discontinued. Stopping rules are implemented to limit the number of items that are administered, thereby gaining efficiency in administration. Equally important is that stopping rules maintain the reliability of student ability estimates. As such, key to setting an appropriate stopping rule is finding an acceptable balance between efficiency and reliability so as to support valid decision making. To do this, appropriately set stopping rules should identify the point at which a test will consistently discriminate between students with and without the ability to respond to the test items correctly without compromising the reliability of the ability estimates (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). In other words, a stopping rule should gain efficiency by limiting the number of items administered while also reliably estimating students' ability (Mather & Woodcock, 2001).

Despite the fact that multiple assessment systems (c.f., Key-Math 3 Diagnostic Assessment, Connolly; 2007; Kaufman Test of Educational Achievement [KTEA-3], Kaufman & Kaufman, 2014) employ stopping rules, little research has been conducted to investigate empirical procedures for establishing a stopping rule that simultaneously considers efficiency of test administration and reliability of score estimates. Moreover, data on stopping rules in technical documentation are not commonly reported (Clements, Sarama, & Liu, 2008;

Purpura, Reid, Eiland, & Baroody, 2015; Weiland, Wolfe, Horwitz, Sarama, & Yoshkawa, 2012). In this manuscript, we present a framework for evaluating stopping rules that explicitly considers efficiency and reliability. After presenting the framework we apply it to an operational testing program to illustrate how the proposed framework can be applied to extant data gathered from the administration of a fixed-form formative assessment (such as a universal screening or diagnostic assessment) in which items are ordered in increasing item difficulty. In doing so we provide considerations for test developers seeking to implement stopping rules in fixed-form tests as well as guidance for test users seeking to evaluate the stopping rules in the tests they are administering. Additionally, we identify relevant test design factors that impact the use and interpretation of results obtained when applying the framework to authentic data. Our goal is to offer the practical assessment community analytic procedures for examining stopping rules for fixed-form tests with increasing item difficulty.

Criteria for Evaluating Stopping Rules

The intent of a stopping rule is to limit the number of items students take on a test (efficiency) while simultaneously attempting to provide a reliable estimate of student ability (reliability). In this section, we provide conceptual definitions of these two criteria, consider how to find an acceptable balance between them, and then describe how these factors have been examined in the context of different types of educational tests.

Efficiency

Efficiency can be operationalized as an index of the amount of time required to complete the test (Anthony, DiPerna, & Lei, 2016; Jodoin, 2003; Weiss, 1982) and is particularly important in the context of classroom assessments. Increasing efficiency conserves classroom time for instructional activities (Parkes, 2013) and minimizes test-taker fatigue. Test-taker fatigue can be defined not only as a direct potential decrease in performance on a test as a result of its length, but also as a subjective quality (Ackerman & Kanfer, 2009). Excessive time on-task that results in test fatigue may also depend on other test characteristics, such as the degree of attention required to complete the task, the level of demand of intellectual functioning required, or lack of feedback about performance on the task (Ackerman et al., 2010). By seeking to limit the number of items a student sees that exceeds his/her ability level, stopping rules can help minimize test fatigue (Weiland et al., 2012). In this study, we operationalize efficiency as discontinuing test administration so that students do not respond to all items on the test, particularly those items that they have a higher probability of responding to incorrectly.

Reliability

Measurement reliability is a hallmark of technical adequacy (AERA, APA, & NCME, 2014) and should not be compromised for the sake of administration efficiency (i.e., administration of relatively few items). Reliability can be defined as consistency in the measurement of student ability from two perspectives: (a) internal consistency, and (b) standard error of ability score estimation. From a Classical Test Theory (CTT) perspective, internal consistency is the extent to which item responses are correlated (Haertel, 2006). Internal consistency tends to increase with test length, assuming items are sampled such that item responses are correlated. Using Item Response Theory (IRT), reliability can be measured by the magnitude of the standard error of measurement associated with the estimation of student ability scale score, which can serve as an index of confidence in the measurement of student ability (Haertel, 2006; Hays, Morales, & Reise, 2000). Stopping rules can be particularly useful for obtaining a reliable estimation of student ability by discontinuing administration of items that exceed the student's ability estimate (Anthony et al., 2016). As the difficulty of the items exceed a student's ability, the student is more likely to engage in guessing or random response behaviors. Consequently, their response choices may become less informative, and the reliability of the student ability estimate will decrease.

Balancing Efficiency and Reliability.

As previously noted, stopping rules need to balance efficiency and reliability to support valid decision making. Tension may exist between efficiency and reliability, in that efficiency focuses on minimizing test length while reliability, using CTT and item sampling, tends to increase test length. Stopping rules that prioritize efficiency seek to decrease the test-taking burden placed on students and any fatigue and/or frustration students may experience (Weiland et al., 2012) by administering the fewest number of

items. Stopping rules that prioritize reliability seek to maximize the information available from which students' ability can be estimated, and will discontinue administration of items when students' guessing tendencies compromise the accuracy of the ability estimate (AERA, APA, & NCME, 2014; Purpura et al., 2015; Watson & Peli, 1983). However, an inherent tension exists between these two objectives because increasing efficiency by reducing the number of items may have a negative impact on reliability. Conversely, increasing the number of items to improve reliability estimates may have a negative effect on efficiency by requiring students to take more items, items that are not necessarily informative because they likely exceed a student's ability level. As such, an appropriate balance between efficiency and reliability must be achieved to support the valid interpretations and uses associated with the purposes of the test (Schmeiser & Welch, 2006; Stiggins, 1992). Stopping rules can support this objective by identifying the threshold at which an assessment maximizes efficiency without compromising reliability. However, scant empirical research is available to support the specification and evaluation of employed stopping rules in the context of fixed-form tests with items ordered in increasing difficulty.

Prior Empirical Evaluation of Stopping Rules

Recently, several efforts have been made to empirically establish and evaluate stopping rules when developing fixed-length tests with increasing item difficulty (Clements et al., 2008; Purpura et al., 2015; Weiland et al., 2012). In mathematics assessments specifically, stopping rules have been applied with assessments designed to (a) identify preschoolers who may be struggling with foundational mathematics concepts (Purpura et al., 2015) and (b) assess preschoolers' mathematical knowledge and skills (Clements et al., 2008; Weiland et al, 2012). Clements et al. (2008), for example, used data obtained from two pilot studies to establish a stopping rule of six consecutive incorrect responses on the Research-Based Early Maths Assessment (REMA). These data included the number of consecutive incorrect responses and Rasch probabilistic characteristics. More recently, Weiland and colleagues (2012) investigated the utility of a stopping rule of three consecutive incorrect responses for a shorter form of the REMA by not only comparing the item fit statistics and Rasch item difficulties from the two versions of the assessment (19 and 125 items, respectively) but also by examining the reliability of the student ability estimates. Most recently, Purpura et al. (2015) emphasized the importance of identifying a stopping rule for a mathematics screener that maximized efficiency by presenting the fewest number of items possible to mitigate the threat of test fatigue. The researchers examined reliability and validity of the stopping rule under three conditions: (a) three consecutive incorrect responses, (b) four consecutive incorrect responses, and (c) no stopping rule.

These studies represent important efforts to begin empirically examining stopping rules for fixed-form tests with increasing item difficulty. All studies used item response modeling procedures to estimate ability and item difficulty and reported various types of psychometric information – descriptive statistics, item fit statistics, and traditional indices of reliability and validity. In addition, Weiland et al. (2012) and Purpura et al. (2015) made an implicit effort to address the importance of efficiency by developing assessments with fewer items. Each of these studies has contributed to the available research on the specification and use of stopping rules. However, these studies have focused singularly on either efficiency or reliable estimation of student ability and item difficulty, without explicitly considering methods for simultaneously examining the influence of implemented stopping rules on test administration efficiency and the reliability of the student ability estimate. The current study aims to extend this work by framing stopping rules in terms of probability models and simultaneously considering efficiency of test administration and reliability of the student ability estimate and score estimate.

Purpose of the Study

The primary goal of this paper is to propose a framework for evaluating stopping rules for existing fixed-form tests with increasing item difficulty that explicitly considers efficiency and reliability. To do this, we apply alternative methods to an example of a fixed-form test of algebra readiness with graduated item difficulty. We describe (a) three methods for empirically evaluating efficiency and (b) two methods for examining reliability. Of particular note, our methods include procedures for estimating performance on items that have not been delivered using students' prior patterns of performance. Across

the procedures we propose, our intent is to take into account the various types of technical adequacy evidence that may be available. Ultimately, we anticipate that these procedures can inform the test development process (via application during pilot studies) and/or be applied to data obtained from existing tests.

Method

We conceptualize and apply procedures for examining the impact of efficiency and reliability on stopping rules using data collected during initial administration of a fixed-form diagnostic assessment of algebra readiness. In this section we describe the participants with whom our data were collected, the diagnostic assessment of algebra readiness used to gather the data for evaluating the stopping rule, our data preparation activities, and our analyses.

Participants

Two hundred seventy students from three middle schools in one southwestern state participated in the initial development research. The deidentified data set used for this study included 41 Grade 5 students, 195 Grade 6 students, and 34 Grade 7 students. Demographic data were available for approximately 90% of students in the school district recruited to participate in the study (n = 248). Approximately 66% of participating students were White, 24% were Hispanic/Latino, 4% were Black, and 4% were Asian. Fifty-three percent of the overall sample was male and 15% were English learners. Demographics for participating students were representative of student demographics for the school district.

Student outcome data from 48 students in Grades 5 and 6 who responded to one fixed-form test of algebra readiness piloted in the initial development research of a state assessment system were used to examine the proposed criteria for evaluating stopping rules in this paper. Because the data were deidentified it was not possible to identify the demographic data for this specific group of participating students who provided responses.

Measure

The fixed-form algebra-readiness test was designed for classroom use with students who have been previously identified as struggling with foundational algebra concepts. Multiple-choice items were written following detailed content specifications and included informative distractors so that students' selection of incorrect response options could provide teachers with information about why students may be struggling with the assessed content (Ketterlin-Geller, Shivraj, Basaraba, & Yovanoff, 2019). The test included three subtests, each with 10 to 11 items, for a total of 32 items in the final test form. All items were formatted for computer-based delivery. Administration of the test was not timed, and students could skip items and return to them before submitting their final responses, if desired. Items were organized within the test in two ways: (a) from least to most difficult within a subtest, and (b) from least to most difficult across subtests (Basaraba, Shivraj, Yovanoff, Bell, & Ketterlin-Geller, 2013).

The item parameters for the final, fixed-form test form were obtained using data from a pilot study of the items with approximately 10,000 students. During this pilot study, students in Grades 5-8 responded to one of 15 alternate forms comprised of 25-35 unique items assessing their algebra readiness knowledge and skills. Students were allowed to skip questions to minimize the effect of test fatigue and had the option to return to any unanswered items before submitting their test. As part of this study, a stopping rule of three consecutive incorrect responses within each subtest was implemented. For each subtest, students may not have responded to all items because this operational stopping rule required that administration of items *within* a subtest stop after students responded to three consecutive items incorrectly.

Item difficulties from the pilot study were estimated using a 2-parameter logistic (2PL) model. The 3-parameter logistic model was not used because the distractors were purposefully designed to elicit misconceptions and errors in student thinking related to the assessed content, thereby precluding the need to estimate a guessing parameter. Recent examination of data obtained from the test forms indicate that students' selection of a distractor was more systematic than it was random, as evidenced by statistically significant differences in the odds of a student selecting one distractor over other distractors (Ketterlin-Geller et al., 2019). These results suggest that students were purposefully selecting distractors that reflected their

misconceptions of the assessed content and lend empirical support to the use of the 2PL model that does not include the guessing parameter.

Data Preparation

Given that items were ordered within the test based on IRT modeling of item difficulty, we hypothesized that, within each subtest, students' responses to the items at or below their ability level would likely be correct and that their responses to items above their ability level would likely be incorrect. This hypothesis was grounded in two theoretical assumptions underlying IRT. For dichotomously scored items, (1) the probability of a correct response increases monotonically as student ability increases, and (2) because the item difficulty represents the location on the latent trait scale at which the probability of a correct response is equal to the probability of an incorrect response (0.50), students whose ability on the latent trait scale is greater than the item difficulty have a higher probability of responding correctly than responding to the item incorrectly (Embretson & Reise, 2000). Item responses were scored 0 if a student responded incorrectly or skipped the item and scored 1 if they selected the correct response. Items not administered after implementation of the stopping rule (three consecutive incorrect responses) were scored as missing.

Our analyses focused on the relation between a student's response to a 'current' item and the probability of their response to the 'next' item in the fixed test form. For each of the 48 students, we created two categorical variables for use in our analyses. The first variable (var1) was a dichotomous variable for the scored response to the next sequential item in the test; this variable was scored 0 or 1 based on whether the student's response to the next item was correct (scored as 1) or incorrect (scored as 0). The second variable (var2) was an ordered categorical variable representing the number of consecutive incorrect responses obtained by each student. Values for this variable ranged from 0 to 3, indicating if the item response was correct (a value of 0, or not a consecutive incorrect response), or if it was the student's first, second, or third consecutive incorrect response. Depending on a student's response patterns within a subtest it was possible for a student to have multiple values for this variable (e.g., a student could respond to two consecutive items incorrectly, followed by a correct response, followed by one incorrect response). We describe how we used var1 and var2 in our analyses in more detail in the next section.

Analyses

In the sections that follow we first describe our proposed methods for evaluating efficiency, followed by our proposed methods for evaluating reliability.

Evaluating Efficiency

We have conceptualized efficiency as minimizing the number of item responses required while still obtaining a reliable estimate of student ability. Keeping in mind that our proposed framework focuses specifically on fixed-form tests with items sequenced from easy to difficult, operationally an efficient test is one for which administration is discontinued before students are required to respond to items that exceed their ability level. We present three increasingly complex procedures for evaluating when a test should be stopped such that the probability of a student responding to future items correctly is less than 0.50. The procedures vary in complexity depending on the assumptions underlying the test design and the availability of item-level psychometrics.

Observed probability of responding to the next item (cross-tabulation)

Using cross-tabulation of the number of consecutive incorrect responses with the scored response to the next test item, we computed the proportion of examinees with one or two consecutive incorrect items (var1). The observed proportions can be interpreted as probabilities of a correct response to the next item conditional on having one or two consecutive incorrect responses. Our data are based on a stopping rule of three consecutive incorrect responses within each subtest. Therefore, one noteworthy limitation when using cross-tabulation procedures is that it is not possible to condition observation of the next item response on three consecutive incorrect responses because of the stopping rule that was implemented. A second limitation is that the probabilities are sample-dependent; conducting the same analyses with another sample of data may result in very different observed probabilities. This procedure is appropriate when limited item-level data are available. As noted above, without estimated item characteristics, such as item difficulty, the probability of a correct response to

future items cannot be formally estimated by a psychometric model. This implies that the actual item difficulties are unknown, and therefore that the sequencing of items from least to most difficult item sequence is at least a reasonable guess and hopefully based on evaluation of items by content experts.

Estimated probability of responding correctly to the next item (logistic regression)

To address the limitations of the cross-tabulation analyses, we used hierarchical generalized linear modeling (HGLM) to evaluate efficiency by estimating a student's probability of selecting a correct response for an item based on his/her pattern of previously correct responses. This model also accounted for the conditional dependence among responses created by the nesting of item responses within students. In these nested logistic regression models, the scored response to the next item was dependent on the number of consecutive incorrect responses (var2; 0, 1, 2, 3) was the independent variable, with item responses as the Level-1 variable and student-IDs as the Level-2 variable. The model specified is shown below, with n.seq.incorrect representing the number of consecutive incorrect responses:

Level 1:

$$\text{Scored response to the next item} = \beta_{00} + \beta_{10} * n.seq.incorrect + e_{ij}$$

Level 2:

$$\beta_{00} = \gamma_{00} + \mu_{0j}$$

$$\beta_{10} = \gamma_{10} + \mu_{1j}$$

Mixed Model:

$$Scored\ response\ to\ the\ next\ item = \gamma_{00} + (\gamma_{10} + \mu_{1j}) * n.seq.incorrect + e_{ij} + \mu_{0j} + \mu_{1j}$$

Although advantages to this procedure include accounting for the nested structure of the data and being able to account for each student's pattern of incorrect responses, this procedure is not without limitations, Namely, this procedure was conducted using the raw data, which means that estimating the probability of a response to a subsequent item was possibly only to the point at which the stopping rule was implemented.. Obviously, it would be ideal to have data in which all students responded to all items as this would allow us to explore the trade-offs between efficiency and reliability for stopping rules across a range of consecutive incorrect responses.

Estimated probability of responding correctly to the next item response (item response modeling)

To address the limitations of using the raw data, we used BILOG (Zimowksi, Muraki, Mislevy, & Bock, 1996) to estimate student ability after each consecutive item response conditional on the item characteristics of administered items obtained from the IRT modeling. This may constitute a best-case scenario in which carefully estimated item parameters are available. With the estimated student ability and the known item characteristics of the next item (i.e., the 2-PL item difficulty and item discrimination parameter estimates), the probability of a correct response on the next item was estimated conditional on whether the student responded to one, two, or three consecutive items incorrectly. The dichotomous 2-PL model (shown below) provides the probability of student i responding correctly to item j, with a difficulty of b and a discrimination of a, conditional on their ability θ_i.

$$P(X_{ij} = k | \theta_i) = \frac{\exp(b_{jk} + a_{jk}\theta_i)}{1 + \exp(b_{jk} + a_{jk}\theta_i)}$$

Evaluating reliability

We have conceptualized reliability in two ways: (a) the minimum number of items required to have confidence in the correlation between the individual item responses, and (b) the level of precision associated with the student ability estimate. To evaluate reliability we propose two procedures, each addressing an important aspect of reliability: (a) the inter-item and item-total score correlations, and (b) the standard error of estimation of the student ability estimate.

Internal consistency of the items on the test form. Cronbach's alpha was used to summarize the inter-item correlations as an index of construct measurement reliability. Although a high value for Cronbach's alpha is not an indicator of unidimensionality, it is grounded in Classical Test Theory (CTT) that describes the extent to which all of the items on a test measure the same construct (Kline, 2000). Cronbach's alpha is directly influenced by test

length, such that decreasing the number of items in the test tends to decrease the reliability estimate (alpha). Although Cronbach's alpha is a common index of reliability for all types of assessment, it may be a more appropriate index of reliability for tests that are designed to distribute student scores (e.g., norm-referenced tests, such as summative state assessments) and may be less appropriate for classroom assessments that are criterion-referenced and designed to measure the achievement of learning objectives that are specific to a course (Parkes, 2013). For the purposes of this study, we computed Cronbach's alpha by subtest conditional on the addition of items delivered to determine the minimum number of item responses required within each subtest to reach acceptable reliability.

As noted above, there are many important test design features that merit consideration, of which many (if not most) are beyond the scope of this paper. One, however, that we do consider is the test assembly. We chose to focus on the reliability of each subtest rather than the overall score because the score from the subtests were designed to provide educators with instructionally useful information about the algebra-readiness concepts with which students were struggling. Calculating the reliability of each subtest also aligns with the structure of the test in which items were ordered by increasing item difficulty within and across subtests. Because subtests will, by definition, have fewer items than the overall test, the reliability estimates may be relatively low (depending on construct dimensionality).

Mean reliability of student ability estimate. Development of the test form using the 2PL IRT model provided us with additional information to estimate reliability beyond that which is available using CTT approaches. Specifically, IRT modeling allows for the estimation of student ability and the standard error of the estimated ability at each iteration of an item response. The standard error of the ability estimate is transformed into an index of reliability conditional on a sequence of consecutive incorrect responses. For each student we computed the reliability of the student ability estimate conditional on stopping the test after 0, 1, 2, or 3 consecutive incorrect responses. To evaluate reliability for each stopping rule, we estimated the mean reliability of the student ability estimate across students in our sample.

Results

The primary goal of this study is to propose a framework for evaluating stopping rules for fixed-form tests with items sequenced from easy to difficult. Our analyses focused on the accumulation of item responses that allowed for estimation of student ability and prediction of subsequent responses (should a next item be presented). We emphasize that our application of the proposed framework (e.g., efficiency and reliability) is specifically for fixed-form tests with graduated item difficulty. Furthermore, the stopping rule specified on the test described in our application is in terms of a sequence of observed incorrect responses. Our results proceed from a series of descriptive statistics detailing the number of correct and incorrect responses prior to reporting results from analyses of the efficiency and reliability when applying the stopping rule.

Descriptive Statistics

In Table 1, we present two sources of descriptive statistics from our illustrative diagnostic test of algebra readiness.

The first panel of Table 1 shows that the average number of correct responses within each subtest ranged from 8.52 – 9.10, indicating that students typically responded to only 1-2 items incorrectly in a subtest. The other panels within the Table present frequencies for the patterns of correct responses followed by a given number of consecutive incorrect responses. Because students can exhibit a pattern of one incorrect (1, 0) or two consecutive incorrect (1, 0, 0) or three consecutive incorrect (1, 0, 0) responses within each subtest, we accounted for the number of instances the pattern of each was observed, as opposed to the number of students who exhibited these patterns. The means reported here represent the average number of times that each pattern of responses was observed within each subtest and indicate that, on average, the pattern of a correct response followed by an incorrect response was observed most frequently and more than once per subtest. The minimum and maximum values represent the minimum and maximum number of times the patterns were observed within each subtest and indicate, for example, that the pattern of a correct response followed by one incorrect response was observed as many as six times within a subtest. Collectively these data indicate that the pattern

of a correct response followed by one incorrect response was observed the most often while the patterns of a correct response followed by two or three consecutive incorrect responses were observed far less frequently.

Table 1. Descriptive statistics and frequency of response patterns observed for algebra-readiness assessment (n = 48)

	Descriptive Statistics		1 Incorrect (1,0)				2 Consecutive Incorrect (1,0,0)				3 Consecutive Incorrect (1,0,0,0)	
Subtest	N Items in form	*M* (*SD*) Total correct	*n* instances pattern observed	*M* (*SD*)	Min	Max	*n* instances pattern observed	*M* (*SD*)	Min	Max	*n* instances pattern observed	Proportion for whom stopping rule implemented
1	11	8.52 (2.56)	56	1.17 (1.28)	0	6	15	0.31 (0.51)	0	2	33	0.69
2	10	8.83 (1.67)	58	1.21 (0.99)	0	3	10	0.21 (0.46)	0	2	25	0.52
3	11	9.10 (2.36)	60	1.25 (1.23)	1	6	27	0.56 (0.68)	0	2	28	0.58

It is important to note that the stopping rule was implemented for 52% to 69% of students (i.e., students who provided three consecutive incorrect responses). This suggests that although students responded to the majority of items correctly, the majority of these correct responses may have been to the least difficult items within each subtest. This pattern of responding is not only consistent with the design of the test (items presented from easy to difficult), but would also help explain the apparent dissonance between the relatively high mean total correct scores for each subtest and the relatively high proportion of students for whom the stopping rule was implemented within each subtest. Collectively, these data indicate that the stopping rule was implemented for a nontrivial proportion of students (i.e., at least 50% of students within each subtest) who had a pattern of three inconsecutive incorrect responses. Moreover, these data indicate that, commensurate with the design of the test, students provided a sequence of increasing numbers of incorrect responses as the test items increased in difficulty and eventually surpassed students' ability level, at which point the probability of an incorrect response was greater than the probability of a correct response.

To further illustrate the design and implementation of the stopping rule in the context of the algebra readiness assessment in which items were presented in a fixed item sequence from least to most difficult, we present item-level statistics in Table 2.

As described previously, the items within this algebra readiness test were presented from least to most difficult within subtests as well as across subtests; the items presented in this table are ordered by empirical item difficulty within subtest. We present the number and percentage of total respondents (n = 48) who responded to each item, the proportion of students who responded correctly, the point biserial correlation, and the IRT parameters (i.e. item difficulty, standard error of item difficulty, and item discrimination) estimated from the 2PL model. Although there were many students for whom the stopping rule was not implemented (i.e., they responded to all items within a subtest; 15, 26, and 22 students for Subtests 1, 2, and 3, respectively), there was a nontrivial proportion of students (0.52 – 0.69) for whom the stopping rule was implemented. Moreover, with the exception of one or two anomalies (e.g., Item 11 in Subtest 1, Item 10 in Subtest 3), the proportion of students who responded to the items correctly decreased with each subsequent item in that subtest. These data, in conjunction with the item difficulties presented in the far right panel of Table 2, indicate that as items increased in difficulty the proportion of students who responded correctly to the items decreased, as did the number of students who responded to the items.

Evaluating Efficiency

In order of statistical rigor, our three procedures for evaluating administration efficiency are: (a) the observed probabilities for responding correctly to the next item using cross-tabulation of observed frequencies, (b) the estimated probability of responding correctly to the next item using HGLM

analyses, and (c) the estimated probability of responding correctly to the next item using the IRT mean estimates.

Observed probability of responding correctly to the next item

In Table 3, we present the probabilities of a correct response conditional on the observed sequence of incorrect responses. For Subtests 1 and 2, the probability of responding correctly to the next item decreases rapidly after two consecutive incorrect responses are observed. Using this procedure, the probability of selecting a correct response conditional on three consecutive incorrect responses cannot be computed because administration of the items within a Subtest was discontinued after three consecutive incorrect responses.

Table 2. Item-Level CTT and IRT Statistics

Item	Item Order in Test	Subtest	Responded to Items		CTT Statistics		IRT 2PL Statistics		
			N	%	Proportion Correct	Point Biserial	*b*	SE of *b*	α
1	2	1	48	100.00	0.73	0.50	-2.060	0.589	0.49
2	7	1	48	100.00	0.79	0.32	-1.326	0.363	0.65
3	8	1	48	100.00	0.71	0.60	-1.262	0.331	0.68
4	14	1	45	93.75	0.69	0.37	-0.122	0.303	0.70
5	17	1	44	91.67	0.43	0.30	-0.058	0.267	0.77
6	20	1	41	85.42	0.34	0.18	0.110	0.281	0.79
7	21	1	36	75.00	0.64	-0.12	0.274	0.300	0.76
8	24	1	31	64.58	0.23	0.04	0.604	0.440	0.59
9	25	1	28	58.33	0.29	0.14	0.810	0.530	0.58
10	27	1	25	52.08	0.28	0.18	1.025	0.486	0.63
11	28	1	15	31.25	0.53	0.21	1.131	0.540	0.63
12	1	2	48	100.00	0.90	-0.13	-2.148	0.535	0.56
13	4	2	48	100.00	0.83	0.52	-1.685	0.272	0.83
14	6	2	48	100.00	0.75	0.03	-1.348	0.281	0.75
15	9	2	47	97.92	0.72	0.32	-1.239	0.368	0.61
16	13	2	46	95.83	0.65	0.04	-0.160	0.299	0.73
17	16	2	46	95.83	0.57	0.15	-0.081	0.387	0.57
18	19	2	44	91.67	0.43	0.42	0.037	0.371	0.64
19	22	2	39	81.25	0.44	0.35	0.304	0.295	0.72
20	23	2	32	66.67	0.50	0.47	0.519	0.305	0.74
21	31	2	26	54.17	0.42	0.16	2.199	0.808	0.81
22	3	3	48	100.00	0.81	0.07	-1.794	0.360	0.24
23	5	3	48	100.00	0.79	-0.25	-1.483	0.370	0.20
24	10	3	48	100.00	0.46	0.25	-0.788	0.162	0.17
25	11	3	48	100.00	0.56	0.13	-0.745	0.344	0.19
26	13	3	46	95.83	0.52	0.34	-0.157	0.284	0.23
27	15	3	40	83.33	0.55	0.02	-0.088	0.438	0.15
28	18	3	39	81.25	0.62	0.19	0.025	0.217	0.33
29	26	3	36	75.00	0.39	0.28	0.938	0.597	0.16
30	29	3	34	70.83	0.21	0.09	1.422	0.648	0.16
31	30	3	28	58.33	0.57	0.01	1.423	0.658	0.15
32	32	3	22	45.83	0.27	0.11	3.123	1.256	0.13

Table 3. Observed Probability of Responding Correctly to the Next Test Item Conditional on a Sequence of Consecutive Incorrect Responses

Subtest	Number of Consecutive Incorrect Responses	
	1	2
1	0.58	0.27
2	0.58	0.19
3	0.47	0.46

Note. Observed probability of a correct response was not possible when 3 consecutive incorrect responses obtained, at which point item administration stopped.

HGLM estimated probability of responding correctly to the next item

Results of the HGLM analyses are summarized in Table 4. For Subtests 1 and 2, the probability of responding correctly decreased considerably as a function of the number of consecutive incorrect responses and is less than 0.50, meaning that there is less than a 50% chance that a student of average ability (i.e., ability estimate of 0.50) will provide a correct response immediately after two consecutive incorrect responses. For Subtest 3, only after three consecutive incorrect responses did we observe the probability of selecting a correct response that is less than 0.50. In this case, for a student providing one or two consecutive incorrect responses, it is likely that they will respond correctly to the next item. These results indicate that, for Subtests 1 and 2, after two consecutive incorrect responses are observed there may be little value in administering additional items. For Subtest 3, however, a stopping rule of three consecutive incorrect responses is recommended.

Table 4. Hierarchical Generalized Linear Regression Mean Probability Estimates of Responding Correctly to the Next Test Item Conditional on a Sequence of Consecutive Incorrect Responses

Subtest	Number of Consecutive Incorrect Responses	
	1	2
1	0.46	0.22
2	0.49	0.18
3	0.55	0.46

IRT estimated probability of responding correctly to the next item.

We present the results of the IRT analyses in Table 5. In comparison to the HGLM estimates reported in Table 4, which are based solely on the pattern of students' responses, the estimates in Table 5 include the use of item parameters. The data reported in Table 5 indicated that with each increase in the number of consecutive incorrect responses the probability of responding correctly to the next item decreases dramatically once you also take into account the empirical item difficulties of each item in the fixed-order sequence. For each of the three subtests, after two consecutive incorrect responses a student is about two times more likely to provide an incorrect response than a correct response. The difference between the HGLM estimates (Table 4) and the IRT estimates (Table 5) is most notable for Level 3. For example, in Table 4, the probability of an incorrect response conditional on three incorrect responses is 0.46, compared to the corresponding mean probability estimate of 0.18 in Table 5, indicating a significant decrease in the likelihood of responding correctly to a fourth item after three consecutive incorrect responses results when the item difficulty is considered.

Table 5. IRT Mean Probability Estimates of Responding Correctly to the Next Test Item Conditional on a Sequence of Consecutive Incorrect Responses

Subtest	Number of Consecutive Incorrect Responses		
	1	2	3
1	0.48	0.29	0.19
2	0.50	0.28	0.15
3	0.49	0.29	0.18

Note. Estimation of probability is a function of the examinee ability estimate and the next item IRT 2PL parameter estimates. Therefore, probability of a correct response varies across examinees.

Evaluating Reliability

The results obtained from evaluating the reliability of the implemented stopping rule using both the CTT-based approach and the IRT-based approach are presented next. As described previously, the purpose of each approach is different. Consideration of both types of reliability evidence is important as one procedure serves as an index of the confidence that the items are measuring the construct of interest consistently (CTT-based approach) while the other

procedure (IRT-based approach) serves as an index of the confidence with which the items on the test are able to accurately and consistently measure student ability.

Internal consistency of the items on the test form

The internal consistency estimates presented in Table 6 (conditional on the number of items delivered on the test) show that a minimum of eight items need to be delivered in each subtest to reach a reliability estimate between 0.67 (Subtests 2 and 3) to 0.71 (Subtest 1), which is considered acceptable (Kline, 2000). A student would need to respond to at least eight items before any stopping rule was implemented to be reasonably confident in the estimation of the students' algebra readiness. These estimates are not markedly lower than the internal consistency estimates for each subtest when all items in the subtest were delivered, which ranged from 0.72 – 0.83.

Table 6. Cronbach's Alpha Reliability Estimates Conditional on Additional Items Delivered on Test

Number of Items	Number of Consecutive Incorrect Responses		
	Subtest 1	Subtest 2	Subtest 3
2	0.06	0.12	-0.59
3	0.46	0.09	-0.18
4	-0.84	-0.87	-0.13
5	0.31	0.07	-0.45
6	0.53	0.56	-0.02
7	0.64	0.66	0.51
8	0.71	0.67	0.67
9	0.77	0.68	0.76
10	0.82	0.72	0.80
11	0.83	-	0.82

Mean reliability of the student ability estimate

We present the mean reliability of the ability estimates conditional on the number of consecutive incorrect responses and the empirical item difficulties associated with those items in Table 7. As the number of consecutive incorrect responses increases from one to three, the mean reliability of the student ability estimate increases while the standard errors decrease. This is expected, up to a certain number of consecutive incorrect responses. Examination of the results in Table 7 reveal that the reliability of the mean ability estimates associated with stopping administration of items after one incorrect response is low and, for each subtest, has a standard deviation ranging from a third to almost half the size in magnitude as the mean reliability of the ability estimate. Examination of the mean reliability of the ability estimates conditional on a sequence of two consecutive incorrect responses, however, reveals reliabilities that are larger in magnitude and that are associated with appreciably smaller standard deviations (SD = 0.07 – 0.12).

Though the increase in the reliability and corresponding decrease in the standard error of the ability estimate from one to two consecutive incorrect responses seems considerable, that does not seem to be the case when moving from two to three consecutive incorrect responses. Multiple sources of evidence gathered from these analyses support stopping administration of items within a subtest after two consecutive incorrect responses: (1) mean reliabilities of the student ability estimates (Table 7), (2) mean probability estimates obtained from the HGLM analyses (Table 4), and (3) the results of the IRT-based analyses (Table 5). This conclusion is supported n

Table 7. IRT Estimated Mean Reliability of Ability Estimates Conditional on a Sequence of Consecutive Incorrect Responses

	Number of Consecutive Incorrect Responses								
	1			2			3		
Subtest	*M*	*SD*	SE	*M*	*SD*	SE	*M*	*SD*	SE
1	0.39	0.20	0.020	0.49	0.12	0.013	0.52	0.10	0.016
2	0.37	0.13	0.018	0.47	0.07	0.012	0.51	0.06	0.016
3	0.36	0.17	0.017	0.44	0.12	0.011	0.50	0.09	0.017

Note. Estimation of probability is a function of the examinee ability estimate and the next item IRT 2PL parameter estimates. Therefore, probability of a correct response varies across examinees.

only by the observation that the probability of responding correctly to an item after two consecutive incorrect responses is relatively low, but also because the reliability and standard error of the student ability estimate at that point are arguably acceptable (depending, of course, on the measurement purpose). It is important to note that unlike a CAT, in which the delivery of the items is designed to provide students with an approximately 50% probability of responding correctly (conditional on their ability level), the items in the assessment described here are presented in a fixed-order and are not conditional on ability. Consequently, with the administration of each item, the item difficulty will eventually diverge from the student ability estimates because the items are ordered in increasing difficulty, which attenuates the mean reliability of the student ability estimates.

In conjunction with the reliability values presented in Table 6, it seems appropriate to conclude that, for this test, delivering a minimum of eight items before implementing a stopping rule of two consecutive incorrect responses would be appropriate to obtain a balance between efficiency and reliability. Although delivering more items on the test could increase the reliability (Table 6), doing so would compromise the efficiency of the test administration. On the other hand, while implementing a stopping rule of three consecutive incorrect responses could increase the mean reliability estimates slightly (Table 7), the probability of a student responding correctly to that third item may be so low (Table 5) that it is inefficient to do so.

Discussion

The purpose of this study was to propose a framework for evaluating stopping rules for fixed-form tests in which items are presented from least to most difficult. The proposed framework simultaneously considers efficiency (by limiting the number of items a student needs to take that exceeds his/her ability level) and reliability (by having the student respond to appropriately sampled items to obtain a reliable estimate of his/her ability level). We then presented an example of the application of the proposed framework to illustrate how efficiency and reliability information can be considered simultaneously when evaluating a stopping rule for a fixed-form test in which items are sequenced from least to most difficult.

Although we recognize that CATs, by their very design, incorporate stopping rules that balance efficiency and reliability while providing an estimate of student ability, many tests administered currently to students are fixed-form tests with items ordered from least to most difficult (Rueter et al., 2018). The Woodcock Reading Mastery Test, Third Edition (WRMT-III; Pearson, 2011) (a standardized test of reading achievement) for example, requires that administration of a subtest be discontinued if a student misses four consecutive items, the Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4; Dunn & Dunn, 2007) is discontinued if a student makes eight or more consecutive errors in a set of 12 items (Capp, Ethridge, & Odland, 2018), and administration of the Expressive Vocabulary Test is discontinued if a student provides five consecutive incorrect responses (Moyle & Long, 2013). Similarly, many widely-used universal screening tests of foundational literacy and mathematics skills, such as the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002), AIMSweb (NCS Pearson, 2012), and easyCBM (Riverside, n.d.) incorporate stopping rules that specify administration of any given subtest should be discontinued if a student responds to a given number of items incorrectly. Consequently, we feel that providing a framework that allows test developers and test users to empirically evaluate the reliability of estimated ability and administration efficiency when using stopping rules incorporated in these types of tests is not only critical but may lead to some important changes to test administration. The results of applying the proposed framework to our sample algebra-readiness test, for example, suggest that administration of items within a subtest could discontinue after two (instead of three) consecutive incorrect responses and tests users could still have confidence that students' algebra-readiness skills are being measured reliably.

However, in our interpretation of these results we would like to emphasize that our focus is on the utility of the proposed methods for evaluating efficiency and reliability of a stopping rule for fixed-form tests, not the actual results of the applied example. In other words, we are not arguing that a stopping rule of two consecutive incorrect responses will be appropriate for all fixed-form tests and recognize that application of the proposed framework to other fixed-form tests with items ordered from least to most difficult will likely yield a different stopping

rule. It is our hope that including the applied example helps illustrate how the proposed framework can be applied to data collected from other fixed-form tests with items sequenced from least to most difficult to evaluate the efficiency and reliability of the stopping rules for those tests in an effort to maximize instructional time, minimize test fatigue for students, and still have confidence that the construct of interest is being measured reliably.

In the present study, we evaluated the efficiency of an operational stopping rule of three consecutive incorrect items in three ways: (1) examining students' observed patterns of responses, (2) estimating the probability of a correct response based on the pattern of observed responses, and (3) estimating the probability of a correct response conditional on the IRT estimate of item difficulty. These increasingly statistically technical methods for examining efficiency that take into consideration increasing amounts of information produced consistent results (i.e., the stopping rule in the context of our illustrative example could be implemented after two consecutive incorrect responses).

Although these results indicate that application of any of the proposed methods for evaluating efficiency in the context of fixed-form tests with items sequenced from least to most difficult may provide useful information, the probabilistic approach and use of IRT estimates may provide the most robust results, as these procedures take into consideration item and student information when available (Anthony et al., 2016). While the IRT procedures apply only when ample data (specifically item parameter estimates) are available, they are preferable because they can mitigate sample dependence and explicitly account for variation in item difficulties and differences in student ability. Regarding variation in item difficulty, our proposed framework assumes items are sequenced from easy to difficult. Obviously, this is a strong assumption when actual item difficulty parameters are unavailable and less informative evidence is the basis for the assumed sequencing of items. The advantages of known item parameters stem from the impact that a broad range of item difficulties will likely provide a more informative sample of responses, as evidenced in IRT-based adaptive testing models. For the relatively extreme examinees, a narrow sample of item difficulties will fail to provide a reliable estimate. Though not adaptive administratively, the linear sequence of item responses will benefit from a suitably broad span of item difficulties in which case a stopping rule can be implemented reasonably.

Efficiency, however, comes with a tradeoff. Two consecutive incorrect responses will, of course, occur before observing a sequence of three consecutive incorrect responses. Classically, observing fewer items tends to result in a less reliable estimate than would be obtained with administration of additional items (Tavakol & Dennick, 2011). Therefore, it becomes important to consider the tradeoff between efficiency and reliability.

Although multiple methods for consideration of estimating reliability in the context of classroom assessment have been proposed (Brookhart, 2003; Parkes, 2013), in this study we examined reliability in two specific ways: (1) internal consistency, or the extent to which items on the assessment measure the same construct, and (2) precision in the estimate of student ability. Examination of Cronbach's alpha values indicated that a student would need to respond to at least eight items to achieve an acceptable level of reliability (Kline, 2000) and to be reasonably confident that performance on the items was related to overall performance on the test. However, this definition of reliability may not be most appropriate in the context of classroom assessment where the primary goal is not often to understand how one student performs relative to other students but rather is to obtain more information about the student's current level of knowledge and skills and mastery of the assessed content (Parkes, 2013). To address this potential shortcoming of Cronbach's alpha (in the context of classroom assessment) and the lower levels of Cronbach's alpha obtained in this study with our applied example, we also calculated the mean reliability of each student's ability estimate. Doing so revealed that implementing the stopping rule after a student responded incorrectly to two consecutive items within a subtest produced a sufficiently reliable mean ability estimate and that waiting until a student responded incorrectly to three consecutive items did not result in appreciable improvement in the reliability of students' mean ability estimates. Given the importance of reliability within the context of implementing a stopping rule and the low levels of reliability observed in our applied example, we opted to include multiple

indicators of reliability in our proposed framework; doing so is intended to mitigate the lack of dependence we could place on any one reliability measure to provide us with confidence that that construct of interest was measured adequately and that the estimate of student ability accurately reflects students' knowledge and skills before administration of items is discontinued.

Contributions of the Proposed Methods for Evaluating Stopping Rules

To date, several studies (Clements et al., 2008; Purpura et al., 2015; Weiland et al., 2012) have empirically examined stopping rules when developing fixed-form tests in which items are organized from least to most difficult. Although there are some similarities in our proposed methods with the work these researchers have also completed, there are also several notable differences. First, the approaches to establish a stopping rule published previously have focused on either efficiency or reliability but have not attempted to address the trade-off between them. We have attempted to do so by attending not only to the point at which an assessment can be discontinued when the probability of responding incorrectly is greater than the probability of responding correctly (i.e., efficiency), but also by attending to the reliability of the ability estimate at that point in the assessment. Second, we have conducted empirical examinations of efficiency with reliability in mind by exploring whether administering fewer items compromised the reliability of the ability estimate; this approach allowed us to consider the potential trade-offs between efficiency and reliability. Third, we employed three different, increasingly complex analyses that allowed us to predict a students' response to a future item even after the stopping rule of three consecutive incorrect items had been implemented. These procedures advance our conceptualization and applied frameworks for evaluating implementation of specific test administration stopping rules when items are administered from least to most difficult.

Limitations and Directions for Future Research

We recognize that this study is subject to at least four important limitations. First, the proposed methods were applied to a diagnostic test of algebra readiness during its pilot phase before the test became operational. Consequently, the proposed criteria have been applied to only one set of items administered to a relatively small number of students and we recognize that the results obtained for our applied example are sample dependent, particularly due to our small sample size and the specificity of item difficulties. As noted above, the item sample with respect to difficulty is a critically important. Moreover, because of the practical constraints of releasing the assessment as part of the statewide initiative, empirically evaluating the implemented stopping rule of three consecutive incorrect items before the assessment was in its final, operationalized form was not possible. We aim to address this limitation in our future research with simulation studies in which we will apply the proposed methods to other, similar fixed-form tests with graduated item difficulty that have been administered to significantly larger and more diverse samples of students.

Second, because data were collected within the context of the operationalized stopping rule of three consecutive incorrect responses, we were only able to consider stopping rules that were more lenient, but unable to investigate stopping rules that may have been more conservative. Third, one procedure (and perhaps the most informative) relied on IRT modeling of item parameters. While IRT-based test development is a standard today, it is not necessarily the case that item parameters are available and, even if they are, application of IRT methods requires that that data meet certain assumptions (e.g., unidimensionality, local independence of responses, and invariance of item parameters and latent trait across different sample characteristics) and a relatively large sample size. Fourth, for the purposes of this study we were not able to take into consideration the instructional utility of the data provided to end-users of the assessment when the stopping rule was implemented. That is to say, at this point (without feedback from teachers) we are unclear as to whether the data teachers obtained from the assessment when the stopping rule of three consecutive incorrect items was implemented was useful for helping teachers plan their instruction. We realize that the instructional utility of the data is an essential element of the assessment-instruction cycle that warrants further investigation.

In an effort to address some of these limitations, our future research efforts include conducting simulation studies using datasets that

include all of the factors our applied example of algebra readiness did here, as well as those factors that were missing. One critical component of a future simulation study would be to have all students respond to all items in the fixed-form test so that effects of various stopping rules (i.e., patterns of consecutive incorrect responses) could be explored. Similar to our applied example, a simulation study would require item-level responses to generate the IRT parameter estimates for items with a range of item difficulties. An important factor for simulation studies is the sample of item characteristics, namely item difficulty, although including item discrimination and/or the guessing parameters would also be of value, as would evidence of construct dimensionality. We would hypothesize that a broad and uniformly distributed sequence of item difficulty will enable more defensible evaluation of stopping rules. Additional features to be considered include the item response format and scoring, as well as other psychometrics (apart from item characteristics), such as test assembly and the sampling model used to collect data for evaluating the stopping rule. Some of these topics are discussed in the applied testing and psychometric literature (e.g., Buyske, 2005; van der Linden, 1998), but they have yet to be explored in the applied context we are addressing.

While three subtests comprised the fixed-form test of algebra readiness in our applied example, a simulation study would include potentially more subtests with varying numbers of items that would allow us to better explore the relation between internal consistency reliability and reliability of the student ability estimate as indices of reliability for a stopping rule. A simulation study would also allow us to work with a dataset in which all students responded to all items within the fixed-form test; this design would allow us to investigate stopping rules with varying levels of leniency (e.g., 2 or 3 consecutive items incorrect vs. 5 or 6 consecutive items incorrect). Although this design wouldn't necessarily require the application of the three analytic approaches for efficiency described here (i.e., cross-tabulation analyses, HGLM analyses using raw data, and IRT analyses) because all of the data needed to calculate the IRT estimated probability of responding correctly to the next item would be included the dataset, a simulation study would allow us to more systematically and rigorously compare these three methods for evaluating efficiency. Lastly, a simulation study would include a significantly larger and more diverse sample than that which was available with our applied example, thus giving us more data to work with when exploring the proposed methods for evaluating efficiency and reliability.

Implications for Test Development

The primary goal of this study was to introduce a framework for evaluating stopping rules for fixed-form tests with items sequenced from least to most difficult that can be applied and evaluated in the context of other classroom assessments. In doing so, our aim was to inform future considerations of the establishment of stopping rules during the test development process, as well as the level of information that is shared with test users. It may be possible, for example, that a test developer could identify a stopping rule, pilot their assessment, evaluate the proposed stopping rules using the framework outlined here for efficiency and reliability, and modify the stopping rule (if necessary). Also, as demonstrated by our illustrative application of the proposed framework to a pre-existing test of algebra readiness, test developers and end-users of tests could apply these criteria to data they have already collected to evaluate the stopping rule of an assessment. Engaging in this process may prompt test developers to reconsider the stopping rule implemented in their assessment or prompt test users to consider trying another assessment that meets the same intended purposes.

We also recognize that the purpose of a test (and the information it is intended to provide) is likely to influence which criteria – efficiency or reliability - receives greater weighting when establishing a stopping rule. Consequently, identifying an appropriate stopping rule for a test may require balancing the tradeoffs between efficiency and reliability. Universal screening assessments, for example, that are designed to be administered to all students to identify those who would benefit from additional instructional support by their very nature need to be time and resource efficient (Clemens, Keller-Marguilis, Scholten, & Yoon, 2016; Kettler, Glover, Albers, & Feeney-Kettler, 2014) and, consequently, may benefit from stopping rules that place greater emphasis on efficiency. This is not to say that stopping rules for screening assessments disregard reliability of the student ability estimate but rather that, in the interest of efficiency, it may make more sense to establish a simple, easy-to-apply stopping rule that is

almost certain to identify students who are in intensive need of additional instructional support (e.g., discontinuing administration if a student responds incorrectly to all items in the first row of the test). Conversely, diagnostic assessments that are designed to help educators identify why students may be struggling to learn key content by eliciting misconceptions and errors in students' thinking may benefit from stopping rules that allow students to respond incorrectly to more items because the incorrect responses can provide instructionally relevant information (Ketterlin-Geller et al., 2019). Allowing students to respond incorrectly to additional items may increase the reliability of the student ability estimate (at a slight cost to efficiency) which may be appropriate, given the purpose of the assessment.

Practical Considerations for Test Users

We recognize that the methods we have described not only increase the requirements for data reporting and analyses, but also raise important questions for future consideration. With respect to the data reporting and analyses requirements, analyses such as those that we have described here require item-level data and are not possible using the students' overall total score. For educational assessments that are delivered using a paper-pencil format, it may not be reasonable to collect-item level data. Moreover, we also realize that many of the instructional and placement decisions made using educational assessments are based on the students' total score and, therefore, item-level data may not be available. Additionally, the analyses we described here were conducted using item-level data and may require additional training and practice beyond the descriptive and comparative analyses using total scores typically used to make instructional and placement decisions.

References

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. Journal of Experimental Psychology: Applied, 15, 163-181. DOI 10.1037/a0015719.

Ackerman, P. L., Kanfer, R., Shapiro, S. W., & Newton, S. (2010). Cognitive fatigue during testing: An examination of trait, time on-task, and strategy influences. Human Performance, 23, 381-402. DOI 10.1080/08959285.2010.517720.

American Educational Research Association [AERA], National Council on Measurement in Education [NCME], & American Psychological Association [APA]. (2014). Standards for Educational and Psychological Testing. Washington, DC: AERA.

Anthony, C. J., DiPerna, J. C., & Lei, P-W. (2016). Maximizing measurement efficiency of behavior rating scales using Item Response Theory: An example with the Social Skills Improvement System – Teacher Rating Scale. Journal of School Psychology, 55, 57-69. DOI 10.1016/j.jsp.2015.12.005.

Basaraba, D. L., Shivraj, P., Yovanoff, P., Bell, J., & Ketterlin-Geller, L. (2013). Middle School Students in Texas Algebra Ready (MSTAR): Diagnostic Assessment Pilot Study for Grades 3-8 (Tech. Rep. No. 13-10). Dallas, TX: Southern Methodist University, Research in Mathematics Education.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. Educational Measurement: Issues and Practice, 22(4), 5-12. DOI 10.1111/j.1745-3992.2003.tb00139.x.

Buyske, S. (2005). Optimal design in educational testing. In M. P. F. Berger & W. K. Wong (Eds) Applied optimal designs (pp. 1-19). West Sussex, England: John Wiley & Sons.

Capp, K., Ethridge, K., & Odland, A. (2018). Peabody Picture Vocabulary Test. In B. A. Frey (Ed.) The Sage encyclopedia of educational research, measurement, and evaluation (pp. 1226-1228). Thousand Oaks, CA: Sage.

Clemens, N. H., Keller-Marguilis, M. A., Scholten, T., & Yoon, M. (2016). Screening assessment within a multi-tiered system of support: Current practices, advances, and next steps. In S. R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.) Handbook of response to intervention: The sciences and practices of multi-tiered systems of instructional support (2nd ed.) (pp. 187-214). New York, NY: Springer.

Clements, D. H., Sarama, J. C., & Liu, X. (2008). Development of a measure of early mathematics achievement using the Rasch model: The Research-based Early Maths Assessment. Educational Psychology, 28, 457-482. DOI 10.1080/01443410701777272.

Connolly, A. J. (2007). Keymath -3 diagnostic assessment: Manual forms A and B. Minneapolis, MN: Pearson.

Dunn, L. M., & Dunn, D. M. (2007). Peabody picture vocabulary test (4th ed). Minneapolis, MN: Pearson.

Embretson, S. E., & Reise, S. P. (2000). Item Response Theory for Psychologists. Mahwah, NJ: Lawrence Erlbaum Associates.

Good, R. H., & Kaminski, R. A. (2002). Dynamic Indicators of Basic Early Literacy Skills (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement.

Gordon, B., & Elliott, C. D. (2002). Assessment with differential ability scales. In D. H. Saklofske, J. J. Andrews, H. L. Janzen, & G. D. Phye (Eds.) Handbook of Psychoeducational Assessment: Ability, Achievement, and Behavior in Children (1st ed.) (pp. 65-102). San Diego, CA: Academic Press.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), Educational measurement, (4th ed. pp. 65-110), New York, NY: American Council on Education & Praegar.

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. Medical Care, 38, 28-42. DOI 10.1097/00005650- 200009002-00007.

Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. Journal of Educational Measurement, 40, 1-15. https://www.jstor.org/stable/1435051.

Kaufman, A. S., & Kaufman, N. L. (2014). Technical and interpretive manual: Kaufman Test of Educational Achievement (3rd edition). Bloomington, MN: Pearson.

Ketterlin-Geller, L., Shivraj, P., Basaraba, D. L., & Yovanoff, P. (2019). Using mathematical learning progressions to design diagnostic assessments. Measurement: Interdisciplinary Research and Perspectives, 17 (1-22). DOI 10.1080/15366367.2018.1479087.

Kettler, R. J., Glover, T. A., Albers, C. A., & Feeney-Kettler, K. A. (Eds.) (2014). An introduction to universal screening in educational settings. In R. J. Kettler, T. A. Glover, C. A. Albers, & K. A. Feeney-Kettler (Eds.) Universal screening in educational settings: Evidence-based decision making for schools (pp. 3-16). Washington, DC: American Psychological Association.

Kline, P. (2000). The handbook of psychological testing. London, UK: Routledge. Linacre, J. M. (1998). CAT: Maximum possible ability. Rasch Measurement Transactions, 12(3), 657-658.

Lonigan, C. J., Allan, N. P., & Lerner, M. D. (2011). Assessment of preschool early literacy skills: Linking children's educational needs with empirically supported instructional activities. Psychology in the Schools, 48, 488-501. DOI 10.1002/pits.20569.

Mather, N., & Woodcock, R. W. (2001). Examiner's Manual: Woodcock-Johnson III Tests of Achievement. Itasca, IL: Riverside Publishing.

Moyle, M., & Long, S. (2013). Expressive Vocabulary Test II. In F. R. Vollmer (Ed.) Encyclopedia of Autism Spectrum Disorders. New York, NY: Springer.

NCS Pearson (2012). AIMSweb Test of early literacy administration and scoring guide. Bloomington, MN: Pearson.

Parkes, J. (2013) Reliability in classroom assessment. In J.H. McMillan (Ed.) Sage Handbook of Research on Classroom Assessment (pp. 107-123). Thousand Oaks, CA: Sage.

Purpura, D. J., Reid, E. E., Eiland, M. D., & Baroody, A. J. (2015). Using a brief preschool early numeracy skills screener to identify young children with mathematics difficulties. School Psychology Review, 44, 41-59. DOI 10.17105/SPR44-1.41-59.

Riverside (Producer). (n.d.). easyCBM video tutorial series [Video]. Rolling Meadows, IL. https://secure2.easycbm.com/training/?action=measures

Rueter, J. A., McWhirter, R., & Delello, J. (2019). Decision-making practices during the instrument selection process: The choices we make. Assessment for Effective Intervention, 44, 281-291. DOI 10.1177/1534508418758370.

Schmeiser, C. B., & Welch, C. N. (2006). Test development. In R. L. Brennan (Ed.) Educational measurement (4th ed.) (pp. 307-354). Westport, CT: Praeger.

Stiggins, R. J. (1992). High quality classroom assessment: What does it really mean? Educational Measurement: Issues and Practice, 11(2), 35-39. DOI 10.1111/j.1745-3992.1992.tb00241.x.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International Journal of Medical Education, 2, 53-55. DOI 10.5116/ijme.4dfb.8dfd.

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. Applied Psychological Measurement, 22, 195-211. DOI 10.1177.01466216980223001.

Watson, A. B., & Peli, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. Perception & Psychophysics, 33, 113-120. DOI 10.3758/BF03202828.

Weiland, C., Wolfe, C. B., Horwitz, M. D., Clements, D. H., Sarama, J. H., & Yoshikawa, H. (2012). Early mathematics assessment: Validation of a short form of a prekindergarten and kindergarten mathematics measure. Educational Psychology, 32, 311-333. DOI 10.1080/01443410.2011.654190.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement, 6, 473-492. DOI 10.1177/014662168200600408. Weiss, L. G. (June 2016). Standardized assessment for clinical practitioners: A primer. CLINA29061 07/16. San Antonio, TX: Pearson.

Yen, W. M. & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), Educational Measurement (4th ed.) (pp. 111-154), New York, NY: American Council on Education & Praegar.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multiple Group IRT analysis and test maintenance for binary items. Chicago, IL: Scientific Software International.

Citation:

Basaraba, D. L., Yavanoff, P., Shivraj, P., & Ketterlin-Geller, L.R. (2020). Evaluating Stopping Rules for Fixed-Form Formative Assessments: Balancing Efficiency and Reliability. *Practical Assessment, Research & Evaluation*, 25(8). Available online: https://scholarworks.umass.edu/pare/vol25/iss1/8/

Corresponding Author

Deni L. Basaraba
Bethel School District #52
4640 Barger Drive
Eugene, Oregon 97402

email: deni.basaraba [at] bethel.k12.or.us