Research Report

Methods for Measuring Speededness: Chronology, Classification, and Ensuing Research and Development

ETS RR–21-22

Dakota W. Cintron

December 2021

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Methods for Measuring Speededness: Chronology, Classification, and Ensuing Research and Development

Dakota W. Cintron

ETS, Princeton, NJ

The extent to which a test's time limit alters a test taker's performance is known as speededness. The manifestation of speededness, or speeded behavior on a test, can be in the form of random guessing, leaving a substantial proportion of test items unanswered, or rushed test-taking behavior in general. Speeded responses do not depend solely on a test taker's ability and are therefore not appropriate for traditional item response theory. The literature on measuring the extent of speededness on a test is extensive and dates back over a half-century. Yet, simple rules of thumb for measuring speededness, dating back until at least Swineford in 1949, are still in operation—for example, 80% of the candidates reach the last item. The purpose of this research report is to provide a chronology and classification of methods for measuring speededness and to discuss ensuing research and development in measuring speededness.

Keywords Speededness; time limit; rapid guessing; unanswered items; aberrant behavior

doi:10.1002/ets2.12337

Theory of Speededness

In educational and psychological testing, there is a distinction between speed and power tests. A pure speed test is a test that contains a copious (but finite) number of easy items and a time limit. In a pure speed test, the number of items is so large that no test taker will finish all the items and so easy that no test taker will answer an item incorrectly. A test taker's score on a pure speed test is determined by the number of items the test taker completes within the time limit. Conversely, a pure power test is a test with no time limit and an attempt on every item. The score on a pure power test depends on the number of correct answers. The difference between speed and power tests is that speed tests reflect variations in speed of response, whereas power tests reflect variations in accuracy of response (Lu & Sireci, 2007). However, pure power and speed tests exist only in idealization (van der Linden & Hambleton, 1996, p. 166). Gulliksen (1950b) noted, "Most tests are a composite in unknown proportions of speed and power" (p. 230). Thus test takers' scores are determined by the number of items responded to (speed) and the number of correct responses (power). Lee and Chen (2011) classified this type of test as a time-limit test.[1] Davidson and Carroll (1945) characterized the issue with time-limit tests:

> The indiscriminate use of time-limit scores is one of the most unfortunate characteristics of current psychological testing since the time-limit score of a test frequently represents two relatively independent aspects of behavior: (a) the amount the subject knows or can perform (or in certain cases, the level of difficulty which he can reach) and (b) the rate at which the subject works. (p. 411).

The extent to which the time limit influences a test taker's test performance is known as speededness (Swineford, 1949a, 1949b). There exist several definitions of speededness. Yet, the commonality among these definitions is that they refer to the extent to which a test taker's behavior is potentially altered, as well as impaired, as a result of a time limit on a test. The roots of concern about speededness can be unearthed in the literature on the speed–accuracy trade-off (Heitz, 2014; Henmon, 1911; Luce, 1986). When a test taker finds that they need to work faster on a test because of the time limit, their score may not be identical to the score they would have received with a more ample time limit (van der Linden, 2011). The test taker may sacrifice accuracy for speed to complete the test within the time limit. Research on the speed–accuracy trade-off started with experimental psychology (Henmon, 1911; Lajoie & Shore, 1986; Lohman, 1986; McClelland, 1979;

Corresponding author: D. W. Cintron, E-mail: Dakota.Cintron@ucsf.edu

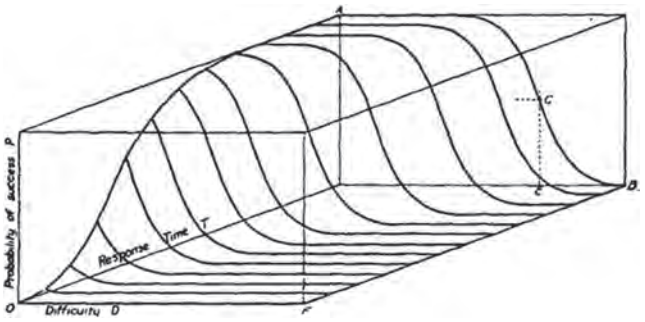

Figure 1 A graphical example of Thurstone's (1937) response surface.

Schouten & Bekker, 1967; Wickelgren, 1977), and the tradition of studying the relationship between ability and speed has persisted for more than a century (see Aikens et al., 1902; Courtis, 1924; Ellis, 1928; Freeman, 1931; Furneaux, 1952; Goldhammer, 2015; Jensen, 1982; Longstaff & Porter, 1928; McFarland, 1928; Mudge, 1921; Myers, 1915; Peak & Boring, 1926; Ruch & Koerth, 1923; Tate, 1950; Thorndike, 1914; Wissler, 1901).

Thorndike et al. (1926) and Thurstone (1937) were among the first scholars to explore the speed–accuracy trade-off within the context of examining effects on item difficulty (van der Linden, 2009). Thurstone proposed a notion of the speed–accuracy trade-off that accounted for item difficulty. Thurstone's depiction of the relationship between response time (speed), accuracy (probability of success), and difficulty can be seen in Figure 1, which demonstrates that (a) the probability of success decreases with increasing item difficulty; (b) with increasing response time, there is an increasing probability of success; and (c) with increasing item difficulty, there is an increasing need for more time to respond. Thurstone's depiction of the relationship between response time, probability of success, and difficulty summarizes the trade-offs that test takers evaluate when exhibiting speeded behavior, and these trade-offs are represented in different theories on speededness discussed herein.

In this research report, I examine the implications of speededness, provide a chronology and classification of methods for measuring speededness, and discuss directions for future research and development on speededness. This review is largely limited to methods developed for test-taking administrations that are high stakes, as opposed to low stakes (see Wise & DeMars, 2005, 2006; Wise & Kong, 2005); paper-and-pencil tests or computer-based tests, as opposed to computer-adaptive and multistage tests (see van der Linden et al., 2007); and multiple-choice responses, as opposed to constructed responses (see Kim, 2017). These constraints narrow the inclusion of methods to those specifically concerned with measuring speededness. For example, in low-stakes examinations, there are methods that could be suitable for measuring speededness but are explicitly designed to evaluate the extent of what has been termed low motivation (Lee & Chen, 2011). The assumption linking test speededness and low motivation is that the test takers who are either speeded or unmotivated exhibit *rapid-guessing behavior*. Rapid-guessing behavior is problematic for traditional item response theory (IRT) models because test takers' behavior deviates from what has been termed *solution behavior* (Lee & Chen, 2011; Wang & Xu, 2015). The resultant rapid guessing or rushed behavior from speeded test takers represents a form of construct-irrelevant variance, violating the IRT assumption of unidimensionality (Haladyna & Downing, 2004; Hambleton et al., 1991). The implications of this construct-irrelevant variance from speededness for both test theory and validity are discussed next.

Implications of Speededness

ETS (2014) defined validity as "the extent to which the interpretations of scores, the claims made about test takers, and inferences and actions made on the basis of a set of scores are appropriate and justified by evidence" (p. 63). Validity assumes that scores reflect the trait the test purports to measure and that differences in test scores between test takers represent differences in the degree to which the test takers possess or lack the trait measured by the test (Lu & Sireci, 2007). The literature on speededness demonstrates that there are several ways that speededness can erode validity (Lu & Sireci, 2007; Schnipke & Scrams, 1997). Consequently, the effects of speededness need to be accounted for, and removed, to alleviate test validity concerns (Lu & Sireci, 2007). The removal of the effects of test speededness can help establish test fairness, psychometric efficiency, and appropriate test administration (Donlon, 1980b).

One way in which speededness may violate the validity of test scores is through the effect of construct-irrelevant variance. Hambleton and Swaminathan (1985) noted that unidimensional IRT models assume that a test is nonspeeded. When speededness is not a direct component of the test design, the construct-irrelevant variance alters the construct the test intends to measure. For example, the construct-irrelevant variance could alter the factor structure of a test (Lu & Sireci, 2007). One instantiation of the issue of construct-irrelevant variance may be that items near the end of a test group together, as a result of rapid-guessing behavior due to the time limit, and represent a separate factor on the test (e.g., a speededness factor). The multidimensionality produced by the construct-irrelevant variance or speededness factor, essentially a nuisance dimension, may manifest in violations of the assumption of local independence (Douglas et al., 1998; Zenisky et al., 2002).

Another consequence of a separate speededness factor is that the speededness factor may give rise to differential item functioning (Dorans & Kulick, 1983, 1986; Holland & Wainer, 1993); that is, different subgroups may require additional time to complete a test. Thus it is possible that the time limit of a test could disproportionately disadvantage certain test taker populations. Differential speededness (i.e., one group is more likely to become speeded on a test)[2] has been evaluated on various ethnic (Dorans et al., 1992; Lawrence, 1993; Schmitt et al., 1991; Schmitt & Dorans, 1988, 1990), gender (Wild et al., 1982), native language (Talento-Miller et al., 2013; Younkin, 1986), and disability (Munger & Loyd, 1991; Sireci, 2005) populations.

Last, another consequence of speededness concerns the influence of speededness on the calibration of person and item statistics or estimates. The issue of speededness impacting item statistics dates back as far as Mollenkopf (1949, 1950). In particular, Mollenkopf (1950) found that "the proportion right of those attempting the item, the Δ index, and the biserial *r* were all found to have undesirable characteristics for items appearing late in a speeded test" (p. 291). More recently, Oshima (1994) evaluated the impact of speededness on item and ability estimates. Oshima found that the item discrimination and difficulties parameters were overestimated, the item guessing parameters were underestimated, and the ability estimates were influenced to a lesser degree by speededness. Bolt et al. (2002) also found that speededness resulted in overestimated item difficulties. Accordingly, the impact of speededness on item and person test calibration would have implications for IRT procedures, such as adaptive testing and equating (Kingston & Dorans, 1984; van der Linden et al., 2007).

Although a considerable body of literature has evaluated the impact of speededness, a widely accepted approach to measuring test speededness has yet to be embraced. For example, Talento-Miller et al. (2013) used the Swineford criterion (Swineford, 1949a, 1949b) to examine whether test takers differed in their degree of speeded behavior by native language. The differentiation among speededness methodologies leaves uncertainty around the impact of speededness with respect to important questions of fairness, psychometric efficiency, and test design. As Cronbach and Warrington (1951) stated, "test theory will be clarified if we can determine and measure degree of speeding. Then the false dichotomy between speeded and unspeeded tests can be discarded" (p. 184). In the next section, I detail different methods of measuring speededness to (a) provide an overview of the different methodologies and (b) highlight the strengths and shortcomings of the different methodologies relative to each other. Furthermore, I provide a classification of speededness methodologies to help future researchers understand the differences between the speededness methodologies discussed herein.

Measuring Speededness

I systematically reviewed the literature on speededness and identified 28 papers that developed a methodology for measuring speededness. These papers were found using a combination of the ETS ReSEARCHER database,[3] Internet search engines (Mahood et al., 2014), and snowballing (Greenhalgh & Peacock, 2005). In this section, I begin by discussing these papers through a chronology of methods for measuring speededness and follow with a classification of the methods for measuring speededness.

The chronology begins with methods for measuring speededness that were developed between the years of 1949 and 1994. In these years, researchers at ETS were behind the majority of the effort to develop methods for measuring and understanding speededness. The narrative then moves into methods developed in the years from 1995 to 2003. The methods developed in these years were characterized by the innovative use of mixture models to separate speeded and nonspeeded test takers into distinct classes. The three prominent methods of measuring speededness in this era were by Bolt et al. (2002), Schnipke and Scrams (1997), and Yamamoto (1995). These three papers would spur substantial research on mixture models to measure speededness over the next two decades. Next, I review methods for measuring speededness from the years 2004 to 2014. The defining characteristic of methods developed during this time is that the methods offered

Table 1 Chronology and Classification of Speededness Methodologies

Citation	RoT	DS	MB	Mx	RT	IR	HS	SIM	EMP	# citations
Gulliksen (1950a)[B]		✓								3,370
Cronbach and Warrington (1951)[A]		✓							✓	262
Stafford (1971)[A]		✓								21
Donlon (1973)[ETS]		✓					✓			10
Swineford (1974)[ETS]	✓									34
Donlon (1980b)[ETS]		✓							✓	9
Bejar (1985)[ETS]			✓			✓	✓	✓	✓	45
Secolsky (1989)[ETS]			✓						✓	12
Yamamoto (1995)[ETS]			✓	✓		✓		✓	✓	42
Schnipke and Scrams (1997)[A]			✓	✓	✓		✓		✓	165
Bontempo and Julian (1997)[CD]		✓							✓	3
Hadadi and Luecht (1998)[A]	✓								✓	9
Bontempo (2000)[CD]		✓							✓	5
Bolt et al. (2002)[A]			✓	✓		✓			✓	125
Mroch et al. (2005)[CD]			✓	✓		✓	✓	✓		10
Thomas (2006)[CD]			✓		✓				✓	5
Cao and Stokes (2008)[A]			✓	✓		✓	✓	✓	✓	52
Goegebeur et al. (2008)[A]			✓	✓		✓	✓	✓	✓	34
Meyer (2010)[A]			✓	✓	✓	✓		✓	✓	52
van der Linden (2011)[A]			✓		✓	✓	✓	✓	✓	25
Wang (2011)[CD]			✓	✓		✓	✓	✓	✓	1
Kahraman et al. (2013)[A]			✓		✓				✓	4
Chang et al. (2014)[A]			✓			✓	✓	✓	✓	13
Jin and Wang (2014)[A]			✓	✓		✓	✓	✓	✓	12
Wang and Xu (2015)[A]			✓	✓	✓	✓	✓	✓	✓	27
Shao et al. (2016)[A]			✓			✓	✓	✓	✓	4
Shao (2016)[CD]			✓		✓		✓	✓	✓	0
Williams (2017)[CD]			✓	✓		✓	✓	✓		0
Total	2	7	19	11	7	14	14	14	23	

Note. $N = 28$. Superscripts designate type as follows: A = journal article; B = book; CD = conference presentation or dissertation; ETS = ETS Research Report; DS = descriptive statistic; EMP = empirical study included in manuscript; HS = allowed for heterogeneity in speededness; IR = used a model-based approach for responses (e.g., two-parameter logistic [2PL] or three-parameter logistic [3PL]); MB = model-based; Mx = model-based with a mixture model component; RoT = rules of thumb; RT = used a model-based approach for response times; SIM = simulation study included in manuscript. Citation counts were collected on July 22, 2018.

new perspectives on how speededness manifests in tests (e.g., the "gradual process of change" or "leave the harder 'til later" models). Finally, I review methods developed from the years 2015 to 2018. These methods represent a combination of new (e.g., change-point analysis) and refined (e.g., mixture hierarchical model) methods for measuring speededness.

After discussing the chronology of methods for measuring speededness, I provide a classification or taxonomy of the methods for measuring speededness. This classification is meant to help clarify differences between methods for measuring speededness. I develop a classification taxonomy that I refer to as the *general classification of speededness methodologies*. Specifically, the methods were classified by whether they (a) were simple rules of thumb for measuring the extent of speededness, (b) were basic descriptive statistics for characterizing the extent of speededness, (c) were model-based, (d) were model-based with a mixture model component, (e) allowed for heterogeneity in speededness, (f) used a model-based approach for responses, and (g) used a model-based approach for response times. A complete breakdown of speededness methodologies, in terms of both their chronology and their classification, is presented in Table 1. Definitions of the classifications are provided in Table 2.

Chronology

The Years 1949–1994: ETS's Contribution

Aside from Cronbach and Warrington (1951) and Stafford (1971), the research on measuring speededness between the years 1949 and 1994 was conducted by ETS and its researchers. The earliest work I could find that directly assessed and

Table 2 Classification Description

Classification	Description
1. Rule of thumb	The rules-of-thumb classification was given to speededness methods that provided basic guidelines or rules of thumb on identifying or assessing speededness
2. Descriptive statistic	The descriptive statistic classification was given to speededness methods that were analytic but did not use a model-based approach (see later) to assess speededness
3. Model-based	The model-based classification was given to speededness methods that used a general modeling procedure to assess speededness. For example, the general modeling approaches covered in this classification include linear regression, hierarchical linear modeling, latent growth curve analysis, mixture modeling, change-point analysis, and item response theory
4. Model-based with a mixture component	The model-based with a mixture model component classification was given to speededness methods that used mixture modeling
5. Heterogeneity speededness	The speededness heterogeneity classification was given to speededness methods that allowed for speededness to manifest throughout a test at different points on a test or for different persons
6. Model-based for responses	The model-based for responses classification was given to speededness methods that used a model-based approach for the responses
7. Model-based for response time	The model-based for responses classification was given to speededness methods that used a model-based approach for the response times

reported speededness was by Swineford (1949a, 1949b). However, this is unlikely the provenance of the idea of measuring speededness. For example, Swineford (1949b) stated, "It is generally assumed that timing is satisfactory when it allows about 80 percent of the candidates to reach the last item" (p. 3). In the two Swineford reports, speededness was measured by what has come to be referred to as the Swineford criterion, and it is the same criterion used in operation at major testing organizations (Lu & Sireci, 2007; Swineford, 1956, 1974; van der Linden, 2011, 2017). The Swineford criterion states that a test is unspeeded if at least 80% of the test takers reach the last item on the test and if all the test takers reach at least 75% of the test items. If either criterion is violated, a test is regarded as speeded (i.e., test taker performance is impaired by speededness). There are a few issues with the Swineford criterion when it comes to measuring speededness. For example, the criterion would not work well in a number-right scored test in which the optimal strategy is not to leave any items unanswered (Lord, 1980). Or, if test takers arbitrarily decide to answer the last item but not the prior five items due to speededness, then it is possible that 80% of test takers could reach the last item on the test. In both scenarios, a test would appear unspeeded by the Swineford criterion.

After Swineford, the next ETS contributor to measure speededness was Gulliksen (1950a, 1950b). The Gulliksen measure for speededness compares the standard deviation of the number of not-reached items to the standard deviation of the number of items that were not given correct answers. Using the notation from Lu and Sireci (2007) to formalize the metric of speededness, let W, U, and X represent the number of items incorrectly answered or omitted, the number of items not reached, and the number of items that were not given correct answers, respectively. That is, $X = W + U$. Therefore, using basic variance algebra, the variance of X can be decomposed as

$$s_x^2 = s_w^2 + s_u^2 + 2r_{wu}s_w s_u. \tag{1}$$

Gulliksen's measure of speededness is based on the ratios of the standard deviations of U (i.e., the number of not-reached items) and X (i.e., items that were not given correct answers). Based on the theory of power and speed tests outlined by Gulliksen in the preceding "Theory of Speededness" section (see Gulliksen, 1950b, chapter 17), in a pure power test, s_u is equal to zero. In a pure speed test, s_w is zero. Based on these facts, Gulliksen noted that a test may be regarded as powered if the ratio s_u/s_x is small—0.1 or less. If the latter ratio is small, the variance of the number of incorrect items is mainly dependent on items incorrectly answered or omitted and not on items not reached. Similarly, Gulliksen noted that a test may be regarded as speeded if the ratio s_w/s_x is small (0.1 or less); that is, the variance of the number of incorrect items is primarily dependent on items not reached. A downside of the Gulliksen method is that the contribution of speed is not easily understood when the ratios are large (Lu & Sireci, 2007; Rindler, 1979). Moreover, the method would not work for a number-right scored test for which the test taker has an incentive not to leave any items unanswered.

Following the work of Gulliksen (1950a, 1950b), Cronbach and Warrington (1951) proposed a method to estimate the degree of speeding on a test. The approach by Cronbach and Warrington is considered a double administration procedure (Lu & Sireci, 2007); that is, the method requires multiple test administrations. The estimate of speededness proposed is referred to as τ, where τ compares test takers' performance on parallel tests administered using both power and speed conditions. Specifically, τ is the correlation between test scores obtained from the speed and power administrations corrected for attenuation, or

$$\tau = \frac{\rho\left[X_{PA}, X_{SB}\right]\rho\left[X_{PB}, X_{SA}\right]}{\rho\left[X_{PA}, X_{PB}\right]\rho\left[X_{SA}, X_{SB}\right]}. \tag{2}$$

X represents a test score, A and B represent the two different parallel tests, and P and S represent the power and speed administrations of a test (e.g., X_{PA} is the test score from the power administration of Test A). The stronger the correlation τ is, the more likely it is that the test is unspeeded, whereas the weaker the correlation τ is, the more likely it is that the test is impacted by speededness. A benefit of the Cronbach and Warrington (1951) approach is that it is free from any notions about dimensionality introduced by test time limits (e.g., multiple types of speededness could exist). Furthermore, τ is useful because it demonstrates how the score distributions change under timed and untimed conditions. The need for the double administration of a test could be considered a shortcoming of the method because speededness is not assessed in a single administration (Lu & Sireci, 2007). Yet, if methods for a single administration have false assumptions, the double administration procedure may be more advantageous, as it avoids the potentially false assumptions in measuring speededness on a single administration.

Between 1951 and 1970, there was not much work on developing methods of assessing speededness. A symposium was held on the effects of time limits on test scores in 1960 (see Myers et al., 1960), but no methods were proposed at the symposium. Stafford (1971) proposed the speededness quotient (SQ). The SQ is defined as the percentage of unattempted items in the total number of errors. Using the SQ, it is possible to assess speededness at both an individual and test level. Using the definitions of U and W from earlier, the formula for the SQ for individual i is

$$\mathrm{SQ}_i = \frac{U_i}{\mathrm{W}_i + U_i}. \tag{3}$$

Similarly, summing over all individuals taking a test, the estimate for the SQ for a test is

$$\mathrm{SQ} = \frac{\Sigma U}{\Sigma W + \Sigma U}. \tag{4}$$

The SQ would be equal to 100% for a purely speeded test due to the quantity W being 0 (i.e., no wrong items).[4] Likewise, the SQ would be equal to 0 for a purely power test due to the quantity U being 0 (i.e., no unattempted items). The SQ is easily computed and understood, making the method attractive due to its simplicity and ease of computation. However, as with the Gulliksen method, this speededness method would not be appropriate with number-right scoring if test takers tended to respond to all items.

The next two methods of speededness were developed by Donlon (1973, 1980b). Donlon (1980a) also provided a useful annotated bibliography of studies on test speededness. Donlon (1973) provided an estimate of speededness based on the "time needed" for a test. For those test takers who complete all the items on a test, the time-needed estimate is equivalent to the time limit of the test. However, for those test takers who fail to reach the last item, a time that would have been needed to reach the last item of the test is estimated. For example, if a test taker only completed one-half of a test in the allotted time limit, one could assume that the test taker would have finished the test in twice the time limit. The Donlon (1973) method is based on estimating the mean and standard deviation of the distribution of finishing times for a population of test takers if a given person has a consistent work rate (i.e., in items per minute). Under the assumption of normality of time-needed estimates, the author associates a z-score with each time-needed estimate by considering the proportion of the sample that exhibits estimates equal to or greater than the time-needed estimate under consideration. The equation for estimating time needed is as follows:

$$S = (Z)(\sigma) + M, \tag{5}$$

where S is the time-needed estimate, Z is the normal curve z-score linked to the percentage of test takers with a time needed equal to or less than S, and σ and M are parameters estimated from the group of test takers. Donlon's method has been referred to as the confidence interval method, as it is possible to use confidence intervals to estimate a range of

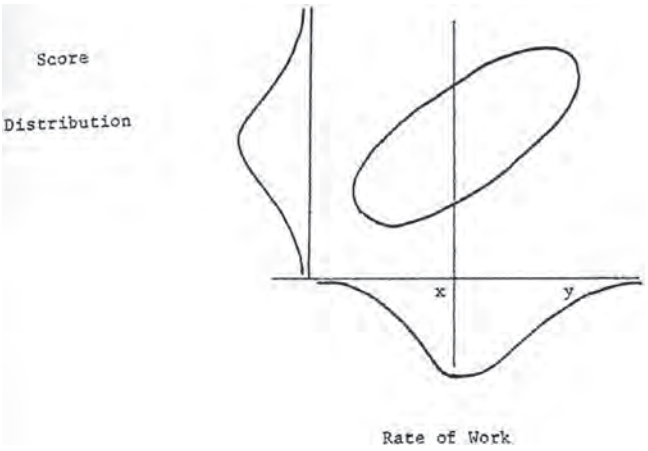

Figure 2 The bivariate normal distribution of score and rate of work from Donlon (1980b). Point x is a typical rate and point y a demanding rate.

time-needed estimates. Consequently, Donlon's method could be used to establish a time limit such that there is not a substantial number of test takers who require an amount of time to complete the test which is greater than the time limit.[5] A downside of the Donlon method is that it would incorrectly estimate the time needed for some test takers who finish a test earlier because it assumes that their time-needed estimate is equal to the time limit. Furthermore, as the time-needed estimate is equal to the time limit for those who finish a test, this method would not identify speededness in tests that use number-right scoring, as theoretically, no test taker would forgo answering any items. Thus the method may not provide an accurate understanding of time needed.

Donlon (1980b) proposed two methods for measuring speededness: the Reilly–Donlon approach and the biserial method. The methods provide correlations analogous to the τ by Cronbach and Warrington (1951) but can be derived from a single test administration. The methods build on Donlon's (1973) prior work by assuming normality in the rates of work, where the rate of work is defined as the number of items per minute a test taker can solve. In the Reilly–Donlon approach, the author attempts to "generalize from the observable correlation between speed and power, for those who fail to complete the test, to the correlation that would be observed for the total group" (Donlon, 1980b, p. 10). The model assumes that there is a bivariate normal distribution of score and rate of work. Furthermore, it is assumed that every test taker has a specific rate of work on a test and that the rates of work across test takers are normally distributed. To understand the generalization in Donlon (1980b), we can examine Figure 2, which shows two specific rates of work x (a typical rate) and y (a demanding rate) required to complete a test. Given the easier rate of work x, and the assumption that all test takers who complete the test are given time-needed estimates equal to the time limit, there would be a lack of variation in rates of work. The variation in rates of work would only come from those test takers who have rates of work less than x; that is, it is assumed that half the population has a rate of work that would enable them to complete the test with a required rate of work x. Accordingly, only half of the bivariate distribution in Figure 2 would be observed. However, with a rate such as that at point y, fewer test takers would finish the test, and there would be greater variation in rates of work, allowing the complete bivariate distribution in Figure 2 to be observed. For tests with time limits, such as x, the observed rate-of-work data must be generalized.

The first method for measuring speededness from Donlon (1980b) is computed as follows:

$$r_{xc} = \frac{s_c r'_{xc}}{\sqrt{s_c^2 \left(r'_{xc}\right)^2 + \left(s'_c\right)^2 \left(1 - \left(r'_{xc}\right)^2\right)}}, \tag{6}$$

where s_c "is the estimate of the standard deviation of items completed for the complete theoretical distribution (i.e., finishers and nonfinishers with an infinite number of additional parallel items)" (Donlon, 1980b, p. 13), s'_c "is the observed standard deviation of items completed for those not finishing the test" (p. 13), r'_{xc} is "the correlation between items completed and total score for nonfinishers" (p. 14), and r_{xc} is the correlation between items completed and total test score. Further details on how to derive the different estimates are provided in Donlon (1980b).

The second method proposed in Donlon (1980b) is the biserial estimation of the speed–power correlation. The biserial method is also based on the assumed bivariate normal distribution of score and rate of work (Figure 2). The formula for the biserial correlation coefficient is

$$r_{bis_{S-P}} = \frac{m_c - m_t}{s_t}\frac{P}{y}, \tag{7}$$

where

> $r_{bis_{S-P}}$ is the biserial estimate of the speed–power correlation, m_c is the mean test score of those completing, m_t is the mean test score of the total group, s_t is the standard deviation of the test score of the total group, p is the proportion who complete the test, and y is the ordinate of the normal curve above which p percent of the area lies. (Donlon, 1980b, p. 15)

There were issues with the two approaches presented in Donlon (1980b). Mainly, for the first method, the author was unable to carry out the computation of r_{xc}. Furthermore, the three methods across (Donlon, 1973, 1980b) are dependent on the assumption of normality in the rate at work. If this assumption is violated, the inferences based on the descriptive statistics would not be tenable.

After the methods proposed by Donlon (1973, 1980b), Bejar (1985), another ETS researcher, explored speededness on the *TOEFL*® test. Specifically, Bejar developed two measures to assess speededness (i.e., an item-level and a test-taker-level index), which accounted for number-right scoring and used IRT for scaling item responses. The item-level speededness index is based on the fit of each item. Bejar called the index Q, and Q is based on the goodness-of-fit index developed by Yen (1981) and modified by Cook and Douglass (1982). The formula for Q is, where test takers are grouped into 15 equally spaced intervals from −3 to 3 based on estimated ability,

$$Q = \sum_{j=1}^{15} \frac{\left[N_j\left(O_{ij} - E_{ij}\right)\right]}{E_{ij}\left(1 - E_{ij}\right)}. \tag{8}$$

N_j is the number of test takers in cell j, O_{ij} is the observed proportion of test takers in cell j that answer the ith item correctly, and E_{ij} is the predicted proportion of test takers in cell j that answer the ith item correctly based on the 3PL. The test-taker-level speededness index contrasts the observed performance on the test's most difficult items against the performance expected by the model for the same items. The goal of the test-taker-level index is to identify students of low ability who get spuriously high scores. For the test-taker-level index, the test is split into two parts: the easy part (75% easiest items) and the difficult part (25% hardest items). The expected performance on the hard items is computed using the ability estimate based on the easy items:

$$E\left(\theta_e\right) = \sum_i P_i\left(\theta_e\right). \tag{9}$$

P_i is the probability of a test taker getting item i correct based on the 3PL, the sum is over the difficult items, and θ_e is the estimated ability based on the easy items. The observed performance on the difficult items is

$$O = \sum_i u_i, \tag{10}$$

where u_i is equal to 1 if the ith item is answered correctly, and 0 otherwise. The sum is taken over the hard items only. A fundamental assumption in Bejar's study is that when speed is not a factor, student response behavior can be modeled by the three-parameter logistic (3PL) model. Another assumption is that more difficult items would be left for last and therefore would be the subject of random responding or guessing. This assumption makes the 3PL a less accurate model for students with low ability and can lead to nonmonotonicity. Both the test-taker-level and item-level index attempt to identify evidence of model misfit as a result of speededness. However, it is possible that model misfit could be confounded with something other than speededness (Lu & Sireci, 2007). Yamamoto and Everson (1995) pointed out that Bejar's (1985) method is circular or subject to a tautology because

> the IRT parameters were estimated using the items that were assumed to be affected by speededness. More important, the speededness index did not reflect the fit of the 3PL IRT model. He reported, for example, model misfit

> in the extreme high and low ability regions where the IRT model parameters were least accurately estimated. As a result, Bejar's index failed to detect speededness in the more important regions of the ability distribution. (p. 4)

The last method in the period from 1949 to 1994 is the method proposed by Secolsky (1989) at ETS. Secolsky, like Bejar (1985), examined speededness in the TOEFL. Secolsky used a regression-based technique to measure speededness. The technique rests on the assumption that under power (nonspeeded) test conditions, scores on the early items of the test should correlate highly with scores on the test's later items (i.e., early scores should predict later scores). If speededness on a test is present, this correlation would not be present. Secolsky (as cited in Yamamoto & Everson, 1995) determined that the TOEFL was speeded because "the observed scores on a subset of examinees on the last 4 to 6 items were significantly lower than the scores predicted by the performance on the first 4 to 6 items" (p. 4). Limitations of the method, as noted by Yamamoto and Everson, include that with such a small number of items, the regression-based method is unreliable and subject to errors of classification based on uncertainty and that the conclusion that test speededness could be confounded with the unreliability inherent in regression methods. Furthermore, the point at which Secolsky (1989) examined speededness was arbitrary and did not attend to individual differences in speed of responding.

The Years 1995–2003: Landmark Mixture Models

Three influential papers, published between 1995 and 2003, marked a shift in the measurement of speededness. The three papers were Bolt et al. (2002), Schnipke and Scrams (1997), and Yamamoto (1995). All three papers involved the use of mixture models, but each had a different take on the manifestation of speededness. Three additional articles offered suggestions on how to measure speededness with response times (Bontempo, 2000; Bontempo & Julian, 1997; Hadadi & Luecht, 1998). Notably, the methods of Bontempo (2000), Bontempo and Julian (1997), and Hadadi and Luecht (1998) involved either descriptive statistics or rules of thumb and did not use a model-based procedure to evaluate response times.

In 1995, Yamamoto used the HYBRID model (Yamamoto, 1989) to assess the degree of speededness of a test and examine the effects of speededness on test taker ability and item parameter estimates in TOEFL. The HYBRID model assumes that test takers switch from an ability-based response strategy (e.g., solution behavior) to a random response strategy (e.g., rapid guessing) at some point during the test. The HYBRID model represents this distinction using distinct latent classes (Clogg, 1995). When the test takers use their abilities to answer a given item, their response behavior can be characterized using IRT-based models (i.e., a latent class with variation in individual ability). Conversely, when the test taker's response behavior is characterized by guessing, the probability of a correct response is modeled using a guessing-class multinomial model. From Yamamoto (1995), the model is defined as

$$P\left(x_i = 1|\theta, \beta_i, k\right) = \left\{1 + \exp\left[-Da_i\left(\theta - b_i\right)\right]\right\}^m c_i^{m+1}, \tag{11}$$

where $m = -1$ when $i \leq k$ and $m = 0$ when $i > k$, x_i is a dichotomous response (0 = wrong and 1 = right) on item i, β_i represents item parameters (a,b), θ is ability, c_i is the expected proportion correct under a patterned response strategy (i.e., random guessing after switch point), and k is the strategy switch point. In the preceding model, the 2PL IRT model holds until the switch point, and then a constant conditional probability for random responses holds for the remainder of the items. Lu and Sireci (2007) noted, "The model [HYBRID], however, limits the strategy switch to occur once, which makes the model useful mostly in cases where random or omitted responses are found at the end of the test" (p. 35). An example of the switch point resulting from the HYBRID model can be found in Yamamoto's (1995) analysis of 48 items from TOEFL, where 1 switch point was depicted. The example from Yamamoto is shown in Figure 3, which shows how the proportion in the rapid-guessing class becomes larger after the switch point (Item 28). The graph's different lines represent three different timing conditions: 50, 55, and 60 minutes. Figure 3 shows that those in the 50-minute condition had higher proportions of rapid guessing after the switch point. Shortcomings of this model include that the model does not include response time; the model fixes the location at which speededness manifests or occurs — that is, the model does not allow for different test takers to become speeded at different points in the tests (i.e., the model assumes speededness homogeneity — see the Theories on the Manifestation of Speededness section in this report), and last, the model assumes that only items at the end of the test are impacted by speededness. The latter shortcoming could be an issue, for example, if harder items are left until later and are the items that are subject to speeded behavior.[6]

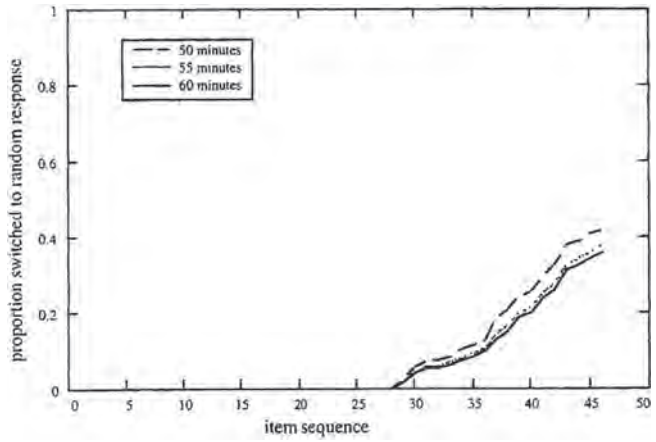

Figure 3 Example of switch point at item 28 in the HYBRID model (Yamamoto, 1995). The figure depicts the cumulative distributions of examinees affected by speededness under three time limits. We can see that after the point of speededness (Item 28), the proportion of examinees affected by speededness increases, with the 50-minute test having the most rapid increase.

Schnipke and Scrams (1997) proposed a second influential method of measuring speededness, and their method was the first to include response times to measure speededness.[7] The authors hypothesized that an accurate assessment of speededness should reflect the tendency of test takers "to rapidly guess on items as time expires" (p. 213). Like Yamamoto (1995), Schnipke and Scrams (1997) hypothesized that there was a rapid-guessing component of speededness. Schnipke and Scrams aimed to model the rapid-guessing component using a two-state mixture model. The two-state mixture model assumes that for each item, a test taker is either engaged in a rapid-guessing or solution behavior on a given item and that these two types of behaviors have different response-time distributions. Using these ideas, an observed response-time distribution can be described as a mixture of rapid guessing and solution-behavior response-time distributions. Specifically, the two-state mixture model is defined as:

$$F_{Oi} = \rho_i F_{Gi} + \left(1 - \rho_i\right) F_{Si}, \tag{12}$$

where F_{Oi} is the observed response-time distribution for item i, ρ_i is the proportion of rapid guesses on item i, F_{Gi} is the rapid-guessing response-time distribution for item i, and F_{Si} is the solution-behavior response-time distribution for item i. The parameters of the model are identified if the shapes of the underlying distributions can be specified reasonably and ρ_i is not too close to 0 or 1(too few rapid-guessing observations or too few solution-behavior observations). When $\rho_i = 0$ (no rapid guessing), the preceding mixture model simplifies to (single-state model)

$$F_{Oi} = F_{Si}. \tag{13}$$

To finish, Schnipke and Scrams considered a final model that assumed the parameters of the rapid-guessing distribution in the standard mixture model to be constant across items, where

$$F_{Oi} = \rho_i F_{CG} + \left(1 - \rho_i\right) F_{Si}, \tag{14}$$

and F_{CG} is the common-guessing mixture distribution and is the same for all items. The shapes of the underlying distributions were assumed to be unimodal and positively skewed, thus approximating log-normal distributions. Schnipke and Scrams compared the three mixture models' results on the items from a computer-administered reasoning test. Schnipke and Scrams found that a single-state model fit the response-time distributions for most of the items on the first half of the test and that the standard mixture and common-guessing models fit the response-time distributions on the latter portion of the test (i.e., when final estimates were available). Schnipke and Scrams noted that the mixture model did not have final parameter estimates for some items early in the test because there were not enough rapid guesses. Also, Schnipke and Scrams concluded that rapid-guessing behavior is the same across all items, that guessing increases toward the end of the test, and that accuracy increases as a function of response time to a point at which accuracy is constant. Response time was coded as total time spent on the item during all attempts (measured within 1-second precision). Speededness in the model was determined by combining the proportion of rapid-guessing behavior (ρ_i) with the proportion of test takers

who did not reach each item. This was accomplished by adjusting the estimate of ρ_i (from the common-guessing mixture) so that the measure is the proportion of rapid-guessing behavior in the entire test taker population rather than just the proportion of rapid-guessing behavior among those who reached the item.

Between 1997 and 2000, three papers were published that provided methods and rules of thumb or descriptive statistics for measuring speededness (Bontempo, 2000; Bontempo & Julian, 1997; Hadadi & Luecht, 1998). The first paper, Bontempo and Julian (1997), outlined three such methods of analyzing speededness: The first method examines test taker response rate, the second method examines increases in test taker response rates within individual examinations, and the third method examines increases in test taker response rates across repeated examinations taken by the same test taker. The first method is like the Swineford (1949a, 1949b, 1956, 1974) criterion. First, the total time that each test taker utilized is divided by the number of items completed; that is, a test taker's response rate, or rate of work in Donlon vernacular, is determined. Then, the mean, 80th percentile, and 99th percentile of all test takers' response rates were used to extrapolate an expected completion time. The distribution of mean response rates for test takers could then be used to determine the degree of speededness based on cutoffs like the Swineford criterion (e.g., the percentage of test takers who would finish less than a certain percentage of the items given their response rate). The second method examined whether test takers rush with some of their later items. The method calculates a test taker's relative response rate to a group of items:

$$\overline{S_c} = \frac{\sum_{i=1}^{I}\left(T_{ic} - \overline{T_i}\right)}{I}, \tag{15}$$

where T_{ic} is test taker c's response time to item i, $\overline{T_i}$ is the item length for item i (the average time spent on the items by all test takers), I represents the total number of items (or a subset of items) taken by test taker c, and S_c is the relative response rate of test taker c. To measure speededness, the relative response rate of a subset of items (e.g., first half) would be compared to another subset's relative response rate (e.g., second half). If speededness were not present, there would be no statistically significant difference between the relative response rate between the two halves of the test. The last method by Bontempo and Julian requires repeated administration in which the relative response rate from above is compared across the repeated measurements. Note that the relative response rate can be computed for each test taker as well as the group.

Hadadi and Luecht (1998) provided "rules of thumb" based on their analysis of two empirical studies—all involved medical school students. The authors provided a set of three necessary conditions, or rules of thumb, to infer rapid-guessing behavior, and these conditions form their operational definition and method for detecting test speededness. The three conditions are (a) a substantial change in the response latencies (i.e., response time) during the latter portion of the test (assuming that the majority of test takers take the test primarily in the item presentation sequence), (b) a corresponding negative change in performance, and (c) the ruling out of alternative explanations for the observed changes in response latencies or performances (e.g., increased item difficulty, substantial increases in the reading load per item).

Bontempo (2000) proposed an experimental procedure for measuring speededness requiring double administration procedures: one with a time limit and one without. Speededness using the Bontempo method is measured by comparing the speed rating scale scores between different administrations. The speed rating scale scores are a specific method used in Bontempo to convert the natural logarithm of response times into speed scores on a scale of 0–9, with faster responses receiving higher scores. Figure 4 provides an example of how the response-time data were converted to the speed rating scale. Figure 4 shows that rapid response times of 0.0–7.4 seconds were coded as 9 and slower responses from 403.4 seconds to infinity as 0.

The last paper in the landmark mixture model era was the paper by Bolt et al. (2002), whose specific aim was to reduce the contamination in the item difficulty estimates resulting from speededness (see Oshima, 1994). The proposed mixture model relied on a set of ordinal constraints. Specifically, ordinal constraints were applied to a mixture Rasch model to classify test takers into latent classes of speeded test takers (i.e., did not have sufficient time to answer end-of-test items) and nonspeeded test takers (i.e., had sufficient time to answer all items). The outcome measure in the model was item scores for which test takers are distinguished according to whether their item response patterns indicate a disproportionately large number of incorrect responses on end-of-test items. The mixed Rasch model with ordinal constraints model is

$$P\left(U = 1|g, \theta_g\right) = \frac{\exp\left(\theta_g - b_{ig}\right)}{1 + \exp\left(\theta_g - b_{ig}\right)}, \tag{16}$$

RESPONSE TIME DATA CONVERSION TABLE

Time (Seconds)		Ln(Time)		Speed
Lower	Upper	Lower	Upper	Rating Scale
0.0	7.4	0.0	2.0	9
7.4	12.2	2.0	2.5	8
12.2	20.1	2.5	3.0	7
20.1	33.1	3.0	3.5	6
33.1	54.6	3.5	4.0	5
54.6	90.0	4.0	4.5	4
90.0	148.4	4.5	5.0	3
148.4	244.7	5.0	5.5	2
244.7	403.4	5.5	6.0	1
403.4	α	6.0	α	0

Figure 4 The response-time data conversion table from Bontempo (2000).

with

$$b_{i0} = b_{i1}, \text{for } i < k, \tag{17}$$

and

$$b_{i0} < b_{i1}, \text{for } i \geq k. \tag{18}$$

Variable g is a class membership parameter that determines the relative difficulty ordering of the items for the test taker, θ_g is a continuous latent ability parameter in class g (the parameter impacts the number of items the test taker is expected to answer correctly), and b_{ig} is the Rasch difficulty parameter of item i for class g; k denotes the first item for which the test takers experience the effects of speededness. A norming condition $\sum_i b_{ig} = 0$ is required for identification. Assumptions of the model include that the Rasch model holds within each class, but the item difficulty parameters are different between classes, Specifically, speeded test takers are expected to perform worse than nonspeeded test takers on items located at the end of tests; this is accomplished by constraining the item difficulties to be higher in the speeded class than in the nonspeeded class. Estimation is conducted using the Markov chain Monte Carlo method to handle the ordinal constraints on model parameters—

> for example, if the difficulty parameter for item i in Class 1, b_{i1}, were constrained to be greater than its difficulty parameter in Class 2, b_{i2}, then we can elect to sample values for b_{i2} with respect to its posterior only over the domain of values less than that sampled for b_{i1}. (Bolt et al., 2002, p. 335)

The priors for the model parameters can be found in the article. The size of the speeded class indicates the speededness of a test. A drawback of this model, like Yamamoto's (1995), is that it does not include response time, and the authors themselves noted that the model would likely find its primary application in paper-and-pencil tests.

The Years 2004–2014: Theories on the Manifestation of Speededness

After the development of the landmark mixture models for assessing test speededness, new methods appeared that offered alternative theories on the manifestation of speededness on a test. However, they were, at times, simple refinements of the prior mixture models of speededness. The major theories on the manifestation of speededness that arose during this time were the theory of the existence of *speededness heterogeneity* or multiple points of speededness on a test for different individuals or items (Goegebeur et al., 2008; Mroch et al., 2005), that is, the methods allowed for the onset of speededness at different points on a test for different test takers;[8] the theory that speededness could be a function of item difficulty (Cao & Stokes, 2008; Chang et al., 2014); the theory of a gradual process change that hypothesizes that the probability of success gradually decreases and eventually reduces to the probability of success under random guessing (Goegebeur et al., 2008; Jin & Wang, 2014); the theory that speededness was a function of pacing behavior (Kahraman

et al., 2013; Thomas, 2006); and the theory that models for speededness should incorporate both responses and response times (Meyer, 2010; van der Linden, 2011).

The first alternative theory of speededness proposed was by Mroch et al. (2005). The authors defined their theory building on Wollack and Cohen (2004), who proposed a model for simulating test data with speededness. Specifically, the authors proposed the multiclass mixture Rasch model (MMRM), which allowed for speededness effects to occur at different locations on a test for different test takers (i.e., speededness heterogeneity). The flexibility of the model allowed for rushed responses as opposed to strictly guessing. Under the MMRM, the probability that a test taker j answers an item i correctly is given as follows:

$$P\left(u_{ij}=1|\theta_{gj},b_{ig},g\right)=\frac{\exp\left(\theta_{gj}-b_{ig}\right)}{1+\exp\left(\theta_{gj}-b_{ig}\right)}, \tag{19}$$

where u_{ij} is the 0/1 response of test taker j to item i, θ_{gj} is the latent ability parameter of test taker j in class g ($g = 1,2, \ldots, k$ latent classes), and b_{ig} is the difficulty parameter for item i in class g. The model allows the Rasch model to be estimated for each latent class. The model proposed by Bolt et al. (2002) is a special case of the MMRM. The MMRM model assumes that only end-of-test item parameters differ across classes, and Mroch et al. applied constraints that forced the classes to represent classes distinguished by speededness (for a discussion of these constraints, see Mroch et al., 2005). Furthermore, a certain number of items at the beginning of the test are assumed not speeded (the authors described the procedures to identify the number of items at the beginning of a test that are assumed not speeded). In the model, the point of the test at which speeded responses begin distinguishes the latent class. The model assumes that items are answered in order, so the only items that are impacted by speededness, for a given latent class, are items occurring after the speededness point. Within speeded and unspeeded classes, item difficulties are assumed to be equal. Across classes, ordinal constraints are applied so each speeded item is defined in each class such that the speeded item difficulty is always higher than the difficulty for the same item in the nonspeeded case (these are the same constraints used by Bolt et al., 2002). The model retains the shortcomings from Yamamoto (1995) and Bolt et al. (2002) yet does enable the examination of different classes of speededness, which are defined by the point at which speededness occurred on the test.

The next paper on speededness from this era used a general modeling approach that had not been used until this point in time. Thomas (2006) aimed to model the *trajectory* of response times using both hierarchical linear modeling and latent growth curve (LGC) modeling. The paper was only the second to directly use a model-based procedure on the response times. The model aimed not to measure speededness directly but rather to recognize that pacing behavior may change over the course of an examination (e.g., speededness occurs when one's pacing strategy fails). Thomas stated, "A new definition of speededness may be in order. 'Speededness' should be operationally redefined as a significant within-individual change in pacing trajectory, accompanied by a corresponding change in performance" (p. 87). The sequential item response time was treated as the dependent variable for the hierarchical linear modeling and LGC models. The details of hierarchical linear modeling and LGC modeling are not discussed here; the reader is referred to Thomas for further detail. A shortcoming of the model is that it does not incorporate item responses or item response accuracy.

In this era, the next paper, Cao and Stokes (2008), proposed three models that could be used to detect speededness. The three models for speededness were the IRT threshold-guessing model (IRT-TG); the IRT difficulty-based guessing model (IRT-DG); and the IRT continuous-guessing model (IRT-CG). The models the authors proposed assume latent classes of students with both high and low motivation (these could also be speeded and unspeeded groups, and the authors explicitly noted this). The estimation of each of the three models is carried out with Bayesian techniques. Furthermore, each model is a mixture model extended from the 2PL IRT model. Cao and Stokes noted that the IRT-TG "is the same as the HYBRID model proposed by Yamamoto (1995). … Yamamoto implemented the marginal maximum likelihood method to estimate the parameters in the model, while our estimation is Bayesian" (p. 212). The IRT-DG model is based on the theory of test-taking motivation by Wise and DeMars (2005); that is, effort decreases as a function of item difficulty. The model assumes that a guesser resorts to guessing behavior for items that are difficult for them, where the difficulty threshold is related to the test taker's ability parameter. In the IRT-DG model, the probability of a correct response from test taker i to item j is given by

$$P\left(x_{ij}=1|\theta_i,\delta_j,\gamma_j,\eta,c_j\right)\frac{\exp\left[\gamma_j\left(\theta_i-\delta_j\right)-\beta_i I\left(\delta_j-\theta_i-\eta\right)\left(\gamma_j\left(\theta_i-\delta_j\right)-c_j\right)\right]}{1+\exp\left[\gamma_j\left(\theta_i-\delta_j\right)-\beta_i I\left(\delta_j-\theta_i-\eta\right)\left(\gamma_j\left(\theta_i-\delta_j\right)-c_j\right)\right]}, \tag{20}$$

where

> $\beta_i = 1$ if person i is a guesser and 0 otherwise and η is a parameter that measures the difficulty differential threshold that would entice a guesser to guess. If the relative difficulty of item j $(\theta_i - \delta_j)$ is higher than η, the guesser will take a random guess on the item. … Otherwise, the answer to the item reflects the examinee's true ability. (Cao & Stokes, 2008, p. 214)

Parameters γ_j and δ_j are the item parameters. The model allows for guessers to switch multiple times between solution behavior and guessing behavior. The IRT-CG assumes that both motivated and unmotivated test takers are taking the test. The model assumes that "the motivated examinees answer all items using their knowledge and the unmotivated examinees expend less effort as the test progresses" (Cao & Stokes, 2008, p. 215). Furthermore, because the test takers expend less effort, it is assumed that their probability of answering items correctly decreases over the duration of the test. The model for the IRT-CG is given by

$$P\left(x_{ij} = 1 | \theta_i, \delta_j, \gamma_j, \beta_i, \phi_j\right) = \frac{\exp\left[\gamma_j\left(\theta_i - \delta_j - \beta_i\phi_j\right)\right]}{1 + \exp\left[\gamma_j\left(\theta_i - \delta_j - \beta_i\phi_j\right)\right]}, \tag{21}$$

where $\beta_i = 1$ if person i is a guesser, and 0 otherwise, and ϕ_j is a parameter that represents a motivation factor (or likelihood of guessing) associated with item j. The parameter ϕ_j is also described by Cao and Stokes as the "extent of effort held back on item j" (p. 215). Because the authors assume a gradual change in the motivation factor, the parameter ϕ_j can also be considered a function (smooth) of the item location j. Furthermore, if $\beta_i = 0$ (i.e., the ith test taker is a nonguesser), then the IRT-CG reduces to the 2PL IRT model. The new theory of speededness proposed by the authors was related to the IRT-DG model. That is, the model allowed speededness to be manifest in more difficult items, as these were the items on which test takers were most likely to guess. A drawback of the model is that it does not allow for the incorporation of response times.

In the same year as the preceding paper, Goegebeur et al. (2008) proposed a speededness model that consisted of two random processes—a problem-solving process and a random-guessing process. The theory of speededness that the authors proposed was referred to as the *gradual process of change*. The gradual process of change model hypothesized that the random-guessing process gradually takes over the problem-solving process (this is opposed to, e.g., Yamamoto, 1995, who proposed a fixed switch point). The change point and change rate are treated as random parameters in the authors' model to model the gradual change between behaviors. The parameters enable the model to capture test taker differences in change point and rate of change. Under the problem-solving process, item responses are determined by a 3PL model. If a test taker is speeded, then at some point in the model, the problem-solving process gradually decreases, eventually reducing to the probability of success under random guessing. The point at which the change occurs and the rate of change are test taker specific. From Goegebeur et al. (2008), let p be a test taker index, $p = 1, \ldots, P$, and i be an item index, $i = 1, \ldots, I$. The model is

$$Y_{pi} \mid \theta_p, \eta_p, \lambda_p \sim \text{Bern}\left(\pi_{pi}\right), \tag{22}$$

with

$$\pi_{pi} = c_i + \left(1 - c_i\right) P\left(\theta_p\right) \min\left\{1, \left[1 - \left(\frac{1}{I} - \eta_p\right)\right]^{\lambda_p}\right\}, \tag{23}$$

where c_i is a parameter that represents random guessing, η_p ($\eta_p \in [0,1]$) represents the point that a test taker becomes speeded, and λ_p ($\lambda_p \geq 0$) is the speededness rate (change rate) of test taker p. The $P_i\left(\theta_p\right)$ is determined using the 2PL model. The parameter η_p detects the point in the test at which test taker p first experiences an effect due to speeding. The speededness point parameter is expressed as a fraction of the number of items (Goegebeur et al., 2008, p. 68). There is not an effect of speedings for items with $i < \eta_p I$. Once a test taker passes their speededness point, the term $i/I - \eta_p$ becomes positive and results in a decrease of the term $\min\left\{1, \left[1 - \left(\frac{1}{I} - \eta_p\right)\right]^{\lambda_p}\right\}$. The rate of decrease in the term $\min\left\{1, \left[1 - \left(\frac{1}{I} - \eta_p\right)\right]^{\lambda_p}\right\}$ is governed by λ_p. For larger values of λ_p, there is a faster process of change or change rate. Two visual summaries of the model can be found in Figures 5 and 6, which depict the decision tree and gradual process of change of the Goegebeur et al. speededness model. The rationale for the model is as follows:

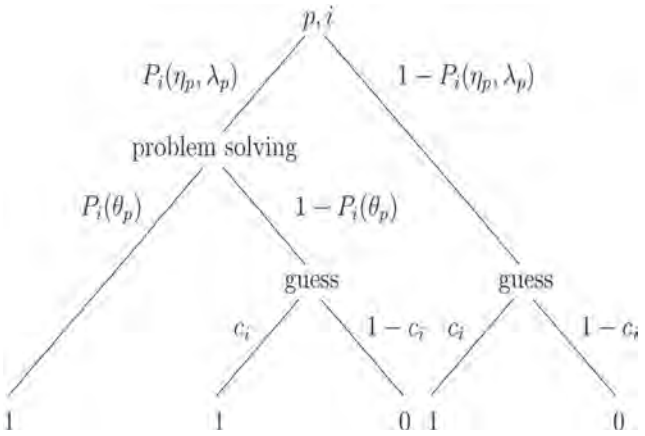

Figure 5 Decision tree depicting the process of the model proposed by Goegebeur et al. (2008). $P_i\left(\eta_p, \lambda_p\right)$ is the probability that an examinee is using a solution process to solve the item (in Goegebeur et al., 2008, the solution process was represented by the three-parameter logistic). When $1 - P_i\left(\eta_p, \lambda_p\right)$ is larger, the probability of random guessing is high.

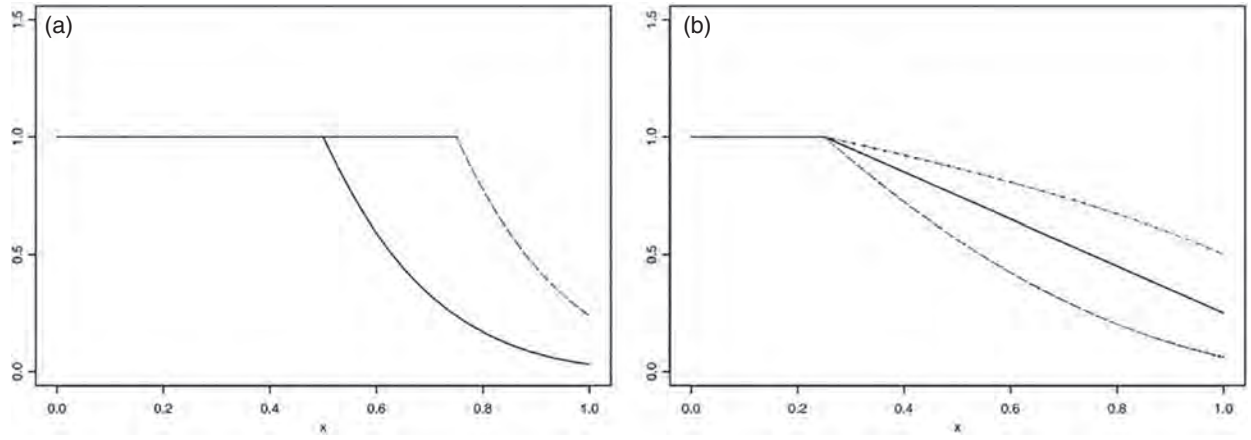

Figure 6 The decay function $\min\left\{1, [1 - (x - \eta)]^{\lambda}\right\}$ depicted as (a) a function of η_p, the point of speededness, and as (b) λ_p, the rate of speededness. In Figure 6a, $\lambda = 5$, $\eta = 0.5$ (solid line) and $\eta = 0.75$ (dashed line). In Figure 6b, $\eta = 0.25$, $\lambda_1 = 1$ (solid line), $\lambda_2 = 2$ (broken line), and $\lambda_3 = 0.5$ (dash-dotted line; Goegebeur et al., 2008). We can see how a lower λ leads to a lower rate of speededness.

> Denote $P_i\left(\eta_p, \lambda_p\right) = \min\left\{1, \left[1 - \left(\frac{1}{I} - \eta_p\right)\right]^{\lambda_p}\right\}$. When examinee p encounters item i he/she answers according to either a 3PL process or a random guessing process, with probabilities $P_i\left(\eta_p, \lambda_p\right)$ and $1 - P_i\left(\eta_p, \lambda_p\right)$, respectively. Under random guessing the answer is correct with probability c_i. Under the problem-solving process the examinee knows the answer with probability $P_i\left(\theta_p\right)$; if ignorant, the examinee guesses at random. (Goegebeur et al., 2008, p. 68)

The standard assumptions regarding the IRT model parameters are made. The marginal distributions of θ_p, η_p, and λ_p are denoted by G1, G2, and G3, respectively. The marginal distributions' dependencies are modeled using a copula function C, and the joint distribution of G1, G2, and G3 is defined using Sklar's (1959) formula. The model allows each test taker to become speeded at a different point on the test (speededness heterogeneity). This aspect of the model is unlike the Yamamoto (1995) or Bolt et al. (2002) model but like that of Mroch et al. (2005). Also, the model allows for the potential of a gradual process of change, which may be a more realistic assumption regarding the manifestation of speededness. A shortcoming of this model is that it does not incorporate response times.

Meyer (2010) proposed a mixture Rasch model with item response time components. The model combined the two-class mixture Rasch model (MRM) of Bolt et al. (2002) and the log-normal mixture model of item response time

(LNMMT) of Schnipke and Scrams (1997) to identify rapid guessing and solution behavior with both item responses *and* response times. Given $i = 1, \ldots, N$ test takers, $j = 1, \ldots, J$ items, and $g = 1, \ldots, G$ latent classes, the item response model, following the MRM, is as follows:

$$U_{ij} \sim f\left(u_{ij}|\theta_{gi}, b_{gi}, g_i\right), \tag{24}$$

with success parameter

$$p\left(U_{ij} = 1|\theta_{gi}, b_{gi}, g_i\right) = \frac{\exp\left(\theta_{gi} - b_{gi}\right)}{1 + \exp\left(\theta_{gi} - b_{gi}\right)}. \tag{25}$$

The item difficulty and ability within each class g are specified as b_{gi} and θ_{gi}, respectively. The average ability, u_g, can vary between classes. The model assumes that the Rasch model fits for each class. The constraint of $\Sigma_{j=1}^{J} b_{gj} = 0$ on the item difficulties is required within each class. The LNMMT models the response time T_{ij} using the log-normal distribution (Schnipke & Scrams, 1997) as follows:

$$T_{ij} \sim f\left(t_{ij}|\lambda_{gj}, \tau_{gj}, g_i\right), \tag{26}$$

where

$$f\left(t_{ij}|\lambda_{gj}, \tau_{gj}, g_i\right) = \sqrt{\frac{\tau_{gj}}{2\pi}}\frac{1}{t_{ij}}\exp\left[-\frac{\tau_{gj}}{2}\left(\log t_{ij} - \lambda_{gj}\right)^2\right], \tag{27}$$

and λ_{gj} and τ_{gj} are the mean and precision for class g, respectively. The class-specific mean reflects the average time taken by members of class g to respond to item j. The precision reflects response-time standard deviation for each class, where $\sigma_{gj} = 1/\tau_{gj}$—a small τ indicates a larger standard deviation of item response times. In the Meyer model, π is a mathematical constant, not a mixing proportion. An identifying assumption of the Meyer model is that item response times and item response accuracy are independent given the latent class. With the assumption of conditional independence between item responses and response times, the joint probability of the model is defined by the product of Equations 24 and 26, or

$$f\left(u_{ij}, t_{ij}|\theta_{gi}, b_{gi}, \lambda_{gj}, \tau_{gj}, g_i\right) = f\left(u_{ij}|\theta_{gi}, b_{gi}, g_i\right)f\left(t_{ij}|\lambda_{gj}, \tau_{gj}, g_i\right). \tag{28}$$

Meyer noted that van der Linden (2007) justifies the same assumptions by assuming independence given ability and a speed parameter. Meyer argued that while the model contains no person speed parameter, the latent class plays a similar role in justifying the independence between responses and response times. Specifically, Meyer argued, "for the present model, conditioning on latent class and theta has the same result as conditioning on time and theta" (p. 524). A benefit of this model is that it uses both response and response times. Although the model appears flexible enough to handle multiple latent classes, it was defined using two classes, and in carrying out data analysis, the authors only compared the fit of one-class and two-class models. However, it is possible that there may be more than two classes of speeded or aberrant behavior.

A method of measuring speededness that has an intuitive interpretation and meaning was proposed by van der Linden (2011). The definition and measurement of speededness are developed by extending the hierarchical model for item responses and response times (van der Linden, 2007). Specifically, van der Linden defined speededness as the probability of a test taker running out of time on a test as a function of the time limit. The idea is similar to Donlon's definition of rate-of-work and time-needed estimates. Let T_i be the random variable denoting the test taker's response time on item i. Therefore the total time on a test with items $i, \ldots, n$ is simply the sum over the n items on the test, or $T_{\text{tot}} = \Sigma_{i=1}^{n} T_i$. In addition, if t_{lim} is defined as the time limit for the test, the degree of speededness for a test taker working at speed t is defined as the probability of passing the time limit without completing the test; that is,

$$\pi_\theta = \Pr\left(\sum_{i=1}^{n} T_i > t_{\text{lim}}|\tau, \boldsymbol{\alpha}, \beta\right) = 1 - F_{\text{tot}}\left(t_{\text{lim}}|\tau, \boldsymbol{\alpha}, \beta\right), \tag{29}$$

where $\beta = (\beta_1, \ldots, \beta_n)$ is a vector of item parameters representing the amount of labor required by the items (i.e., time intensities in van der Linden, 2007); $\boldsymbol{\alpha}$ is the vector of item discrimination parameters from van der Linden (2007); and $F_{\text{tot}}\left(t_{\text{lim}}|\tau, \boldsymbol{\alpha}, \beta\right)$ is the cumulative distribution function for T_{tot} given τ (see van der Linden, 2011, p. 45). The latter equation is for an individual test taker. A simple extension of the model was provided by van der Linden (2007) to calculate

the extent of speededness on a test by determining the proportion of test takers running out of time. A benefit of the van der Linden model is that it uses both response and response times. The model could further be thought of as allowing for speededness heterogeneity as test takers could have different probabilities of running out of time on a test. However, we would not be able to identify the point at which a test taker becomes speeded, which could be useful if one wanted to explain variation in points of speededness between test takers.

Wang (2011) proposed a mixture cross-classification IRT model for test speededness. The multilevel mixture IRT model identifies speededness in terms of latent classes. The model is formulated by Wang to allow for covariates. The unconditional multilevel mixture model for test speededness is expressed as follows:

$$\eta_{ij} = \theta_j + \beta_{ijg}, \tag{30}$$

where

$$\theta_j \sim N\left(\mu_g, \sigma_1^2\right), \tag{31}$$

and

$$\beta_{ijg} \sim N\left(\overline{\beta_g}, \sigma_2^2\right), \tag{32}$$

with $\overline{\beta_1} = \overline{\beta_2}$ (item parameters between the two classes) for items unaffected by speededness and $\overline{\beta_1} > \overline{\beta_2}$ for items affected by speededness—ordinal constraints from Bolt et al. (2002). To help identify speeded and nonspeeded groups, person covariates are included to help predict latent classes. The latter is done in a two-stage procedure in which the unconditional model is fit, and then variables with significant associations with latent group membership are included in a latent variable regression model. The likelihood of response vector X_j based on the mixture cross-classified IRT model is

$$X_j = \sum_{g=1}^{G} \pi_g \prod_{i=1}^{k} p^{x_{ij}} \left(1-p\right)^{1-x_{ij}}, \tag{33}$$

and the probabilities of the mixtures with covariates are expressed as

$$\pi_{jg} \mid X_j = \frac{\exp\left(\lambda_{0g} + \sum_{c=1}^{C} \lambda_{cg} X_{jc}\right)}{\sum_{g=1}^{G} \exp\left(\lambda_{0g} + \sum_{c=1}^{C} \lambda_{cg} X_{jc}\right)}, \tag{34}$$

where π_{jg} is the probability for test taker j to be in latent class g, λ_{0g} is the class-specific intercept when the covariate X_{jc} has no effect on the probability of group membership, and λ_{cg} is the class-specific effects of covariate X_{jc} on the probability of group membership; λ_{0g} and λ_{cg} are constrained to zero for identification. The model assumptions include that items at earlier locations on the test are not affected by test speededness, nonspeeded items are fixed to be equal in both the speeded and nonspeeded groups, items in the speeded portion of the test (i.e., end-of-test items) are assumed to have smaller difficulties for the nonspeeded group than the speeded group, and the item variances are constrained to be equal across the whole test (but allowed to be different between latent classes), and it is assumed that there is no covariation among items due to local independence. A shortcoming of this model is that it does not incorporate response times. Furthermore, like Meyer (2010), this method assumes there are two latent classes when it is possible that there may be more than two classes of speeded or aberrant behavior.

Kahraman et al. (2013) proposed a LGC model to measure speededness, roughly equivalent in conception to Thomas (2006) but specifically using a piecewise LGC model. Kahraman et al. explored the use of LGC modeling to study pacing behavior and test speededness. Test taker response times from a test used in making high-stakes decisions were analyzed using LGC models to detect identifiable patterns of test taker pacing behavior. The authors used LGC modeling, rather than hierarchical linear modeling (HLM), to "take advantage of greater flexibility in terms of modeling the nonlinear relationship between the order in which items were presented during the examination" (p. 344). The LGC models enable the evaluation of item response-time trajectories, the variability around the initial status and overall rate of change in response time, and the impact of covariates on trajectories and variance around initial status and trajectories (Kahraman et al., 2013). The model assumes that test takers start with the first item and continue through to the last. The model for change used in the analysis was a three-piece piecewise growth model in which the three

intervals were the beginning of the test, the middle of the test, and the end of the test. The Level 1 model from Kahraman et al. is

$$y_{it} = \eta_{0i} + \eta_{1i}a_{1it} + \eta_{2i}a_{2it} + \eta_{3i}a_{3it} + \epsilon_{it}, \tag{35}$$

and the Level 2 model is

$$\eta_{0i} = \beta_0 + \zeta_{0i}; \eta_{1i} = \beta_1 + \zeta_{1i}; \eta_{2i} = \beta_2 + \zeta_{2i}; \eta_{3i} = \beta_3 + \zeta_{3i}. \tag{36}$$

Occasions are at Level 1, and test takers are at Level 2; η_{0i} is the intercept, or initial status, and η_{1i} through η_{3i} are the slopes for the three different pieces of the LGC model. The β_0 through β_3 are fixed effects, and ζ_{0i} through ζ_{3i} are random effects. Kahraman et al. did not provide a substantive discussion of how to interpret speededness from the growth model. In an empirical application, it was found that, in the first phase, test takers appeared to move at a relatively slow pace, and there was not much decline in response times. In the second phase, test takers were somewhat faster, but no drastic changes in their speed occurred. In the final stage, test takers moved from one item to the next at a faster rate. Yet, two observations of speededness could be conjectured about the growth model using the results from Thomas (2006): (a) if there was a *meaningful* change in the third slope (larger negative slope indicating a rapid decline in response time) or (b) speededness could result in *meaningful* variation in speed trajectories. A shortcoming of this model is that it does not incorporate item responses. The random effects in the model by Kahraman et al., however, due allow for possible speededness heterogeneity based on response times.

Chang et al. (2014) proposed an alternative theory of speededness like the IRT-DG model (Cao & Stokes, 2008). The theory is that speededness manifests in the harder items on a test. Chang et al. (2014) proposed a speededness model that assumes that harder items are left until later in the test and are thereby the items likely impacted by speededness. The model by Chang et al. (2014) is referred to as the "leave the harder 'til later" model. Basically, in the Chang et al. model, a speeded-effect term is incorporated into the 2PL model. Chang et al. noted the model proposed by Bolt et al. (2002) and the IRT-CG model by Cao and Stokes (2008). Chang et al. (2014) stated that both of the latter models assume that test takers answer items according to item ordering (i.e., speededness effects end-of-test items only)—a concept dating back to the HYBRID model. Instead of assuming that test takers answer items according to item ordering, Chang et al. proposed a model like Bejar (1985) and Cao and Stokes's (2008) IRT-DG. The assumption is that harder items are left until later and thus are the items impacted by speededness. The model compares test taker–specific thresholds to location parameters (the thresholds are all free to be estimated, which is more flexible than in the IRT-DG). Furthermore, the probability of entering into the problem-solving process for an item from Goegebeur et al. (2008) is adopted; that is, the gradual process of change model is assumed. Chang et al. (2014) hypothesized that test takers,

> while facing tests within an allocated time. … answer easier items first and leave items with a certain level of difficulty till later. After answering easier items, the examinee would attempt some of the first-skipped items, and eventually some of the first-skipped and attempted ones may be correctly answered. Probabilities of the first-skipped items being attended later are modeled, and the harder the item is, the smaller the probability will be. If an examinee does not have enough time to attend to all the first-skipped items due to the time limit, some items will be left blank at the end of the test. (p. 257)

The model formulation is as follows: Let Y_{pj} be the dichotomous response of test taker p on item j, where $p = 1, 2, \ldots, P$, and $j = 1, 2, \ldots, J$. The typical 2PL item and ability parameters are assumed. The probability of test taker p obtaining a correct response on item j is a product of two terms, the probability of getting a correct response under the 2PL and the probability of getting into the solving process affected by test speededness:

$$P\left(Y_{pj} = 1 | a_j, b_j, \theta_p, \tau_p, \lambda\right) = \pi_{pj}, \tag{37}$$

with

$$\pi_{pj} = \frac{1}{1 + e^{-a_j\left(\theta_p - b_j\right)}} e^{-\lambda\left(b_j - \tau_p\right) I b_j > \tau_p}, \tag{38}$$

where τ_p is the test taker–specific threshold parameter for the speededness effect of test taker p (referred to as the speededness point for test taker p), λ ($\lambda > 0$) is the overall speededness rate, and I is an indicator function. Chang et al. referred to

the $e^{-\lambda\left(b_j-\tau_p\right)Ib_j>\tau_p}$ term as the speeded term. The interpretation of the location parameter b_j is no longer a direct indicator of item difficulty. The authors referred to difficult/easy instead of large/small location parameters. Items with a location parameter that exceeds the test taker threshold are regarded as not-so-easy items requiring more time to be solved. The leave the harder 'til later model assumes that the latter items would not be put into the problem-solving process until all easier items are answered (i.e., test taker p would first try to answer items with location parameters smaller than τ_p). The probability of harder or retained items (i.e., first-skipped items) being in the problem-solving process is governed by $e^{-\lambda\left(b_j-\tau_p\right)}$, where harder items have a smaller probability (this piece of the speeded term is referred to as the decay rate). According to Chang et al., the decay rate is determined by the speededness rate λ. When b_j exceeds τ_p, a larger speededness rate λ implies that a test taker p has a lower probability of answering item j. Conversely, when the location parameter b_j does not exceed the test taker–specific threshold τ_p, the probability that test taker p answers item j equals 1; otherwise, the probability is $e^{-\lambda\left(b_j-\tau_p\right)}$ (see Chang et al., 2014, p. 258).

Lastly, the leave the harder 'til later model by Chang et al. assumes that, given b, a, θ_p, τ_p, and λ, responses over all items for test taker p are conditionally independent and that the distribution of ability and τ_p, the test taker–specific threshold parameter for the speededness effect, is assumed to be bivariate normally distributed with a correlation of ρ. A shortcoming of this model is that it does not incorporate response times. However, the model does allow for the exploration of speededness heterogeneity given test taker–specific thresholds.

Jin and Wang (2014) proposed a model that aims to account for aberrant testing behaviors (e.g., low motivation, speededness) in IRT models. The authors proposed a mixture IRT model in which it is assumed that test takers are in multiple classes, and a decrement parameter is added to each latent class. The decrement parameter describes a decline in performance. Assumptions of the model are that (a) test takers respond to items in the sequence of item order; (b) test takers attempt items with full effort until specific item locations, and then begin attempting the remaining items with less effort; (c) different test takers may begin to attempt items with less effort at different item locations; and (d) the decrement in effort is irreversible. The decrement parameter is incorporated for each test taker into the IRT framework to adjust for performance decline, where the probability of success on item i for person n is defined as

$$P\left(X_{ni}=1\right)=c_i+\left(1-c_i\right)\frac{\exp\left[a_i\left(\theta_n-b_i-\eta_{ni}\right)\right]}{1+\exp\left[a_i\left(\theta_n-b_i-\eta_{ni}\right)\right]}. \tag{39}$$

The new parameter in the model is η_{ni} (where $\eta_{ni}\geq 0$); η_{ni} is the decrement parameter for person n at item location i. The decrement parameter adjusts for performance decline due to loss of motivation, speededness, or whatever factor is involved. The ability parameter and the decrement parameter are linked linearly. Furthermore, the model assumes that

$$\eta_{ni}=\begin{cases}0;\text{ if } i\leq\delta_n,\\ \gamma_{\delta_n};\text{ if } i>\delta_n;\end{cases} \tag{40}$$

δ_n is a test taker effort threshold and is an integer between 1 and the test length I, and γ_{δ_n} is the decrement and a positive number. The model assumes there are I latent classes and that test takers belonging to latent class I have the same decrement parameter γ_{δ_n}. A shortcoming of this model is that it does not incorporate response times. However, through an test taker effort threshold, the model does allow for speededness heterogeneity by allowing different test takers to begin to attempt items with less effort at different item locations.

The Year 2015 to the Present: New and Refined Methods for Measuring Speededness

The last four papers discussed in this research report describe a mix of new methods (Shao, 2016; Shao et al., 2016) and refinements of existing methods (Wang & Xu, 2015; Williams, 2017). The new methods for modeling speededness include change-point analysis (Barry & Hartigan, 1993) with and without response times. The refined methods are extensions of Meyer (2010) and Jin and Wang (2014) for Wang and Xu (2015) and Williams (2017), respectively.

The first paper discussed is Wang and Xu (2015). The model is the second to combine responses and response times to measure speededness. Wang and Xu note about Meyer (2010) that

> examinees are assigned to one of the latent classes (speeded or nonspeeded) and each latent class has class-specific item parameters as well as population mean/variances of person parameters. However, assigning an examinee

to a speeded class still provides no hint regarding on which item (or from which item onwards) rapid guessing happens. (p. 3)

The authors extended Meyer's (2010) work by formulating a model that allows for multiple switch points (or heterogeneity in the point of speededness). The model presented is like Schnipke and Scrams's (1997) in that only two behaviors, or latent classes, are modeled: a solution behavior and a rapid-guessing behavior. A mixture hierarchical model accounts for differences among item responses and response-time patterns arising from the two behaviors. The mixture model framework is as follows: Let i index persons and j index items. The observed response time, T_{ij}^{obs}, is assumed to take the form

$$T_{ij}^{\text{obs}} = \left(1 - \Delta_{ij}\right) T_{ij} + \Delta_{ij} C_{ij}, \tag{41}$$

where delta is a latent variable indicating response behavior; that is, $\Delta_{ij} = 1$ if test taker i answers item j by guessing, and $\Delta_{ij} = 0$ otherwise. T_{ij} represents the time required for the test taker to respond to item j using a solution behavior, and C_{ij} denotes the time required for test taker i to respond to item j using a rapid-guessing behavior. Therefore, conditioning on the latent index Δ, the density function of observed RT (T_{ij}^{obs}) in Equation 41, $p_{ij}(t)$, where t is a realization of T_{ij}^{obs}, can be written as.

$$p_{ij}\left(t|\Delta_{ij}\right) = \left(1 - \Delta_{ij}\right) f_{ij}(t) + \Delta_{ij} h_{ij}(t); \tag{42}$$

f and h denote the density functions of T_{ij} and C_{ij}, respectively. In parallel to the density function of observed RT, the probability that the ith test taker gives the correct response to the jth item can be decomposed as.

$$P\left(Y_{ij} = 1|\Delta_{ij}\right) = \left(1 - \Delta_{ij}\right) P\left(Y_{ij} = 1|\Delta_{ij} = 0\right) + \Delta_{ij} P\left(Y_{ij} = 1|\Delta_{ij} = 1\right); \tag{43}$$

$P(Y_{ij} = 1|\Delta_{ij} = 0)$ is the probability of a correct response from the solution behavior, whereas $P(Y_{ij} = 1|\Delta_{ij} = 1)$ captures the probability of a correct response from the guessing behavior. The model assumes that T_{ij} follows a log-normal distribution as in van der Linden (2007), where.

$$\log\left(T_{ij}\right) \sim N\left(\beta_j - \tau_i, \sigma_j^2\right); \tag{44}$$

τ_i is a random coefficient denoting the latent speed of test taker i, where test takers with higher speed tend to answer items faster. β_j is the time-intensity parameter for item j; items with higher β_j are more time consuming. Variable σ_j quantifies the dispersion of the log-normal distribution; this serves as the time discrimination power of item j. Likewise, Wang and Xu (2015) assumed that the distribution of response times in the guessing class is $\log(C_{ij})$ with a common mean and variance of μ_c and ${\sigma_c}^2$, respectively, for all i and j. If $\Delta_{ij} = 0$, the 3PL model is used for solution behavior. When $\Delta_{ij} = 1$, the authors assumed that test taker i responds to item j using rapid guessing with item-specific guessing probability g_j, or

$$P\left(Y_{ij} = 1|\Delta_{ij} = 1\right) = g_j. \tag{45}$$

The authors note that a 2PL should suffice in place of using g_j. As in the van der Linden (2007) hierarchical model, person parameters of the measurement model $\xi_i = (\theta_i, \tau_i)$ are assumed to follow a bivariate normal distribution

$$\xi_i \sim N\left(\mu_p, \Sigma_p\right) \tag{46}$$

with mean vector

$$\mu_p = \left(\mu_\theta, \mu_\tau\right)' \tag{47}$$

and covariance matrix

$$\Sigma_p = \begin{bmatrix} \sigma_\theta^2 & \rho\sigma_\theta\sigma_\tau \\ \rho\sigma_\theta\sigma_\tau & \sigma_\tau^2 \end{bmatrix}. \tag{48}$$

The authors noted that the guessing indicator, Δ, can be dependent on many factors. For instance, rapid-guessing behavior could depend on the remaining time or elapsed time of a test. The latter requires modeling the entire test as a dependent process, for example, using counting process models to capture the dependence. Rapid-guessing behavior could be dependent on item position: In a speeded test, test takers could be more likely to guess on those items given at the end of a test. Furthermore, another method would be to allow rapid-guessing behavior to depend on the latent ability

and speed via a person-specific guessing proportion parameter, $\pi_i = P(\Delta_{ij} = 1)$. It is expected that faster-responding and high-ability test takers will be less likely to guess, resulting in a smaller π. In the current model, Wang and Xu (2015) treated π as a fixed effect. Wang and Xu noted that the probability of rapid guessing could be modeled with a parametric function of latent ability and speed, in which case, marginal maximum likelihood estimation (Bock & Aitkin, 1981) could be used. To end, Wang and Xu (2015) noted that it would be of interest to allow for a potential interaction between item types and the probability of rapidly guessing and allowing the distribution of time intensity to differ by item type. However, according to Schnipke and Scrams (1997), we might not expect the distribution of guessing times to vary. Wang and Xu (2015) found that the proportion of items with rapid-guessing behavior is greater when the time intensity of a test is greater. They referred to the time intensity also as the level of speededness; that is, tests with higher time-intensity parameters have a higher level of speededness. The latter assumes a fixed testing time. The van der Linden (2011) method could also be used to calculate speededness in this model.

Shao et al. (2016) and Shao (2016) published two related analyses. The first paper, Shao et al. (2016), used a change-point analysis (CPA) to detect the point at which a test taker becomes speeded using item responses. The model classifies test takers into speeded and nonspeeded groups, for which the speededness among speeded test takers can occur at different locations on a test. CPA detects whether there is an abrupt change in a sequence of data. The method proposed by Shao et al. assumed that the 2PL fits for regular responses that are not speeded and that the items are answered in sequence. The CPA works by defining s_i as the last set of items on which a test taker i is speeded, where $s_i = 0, 1, 2, \ldots, (J-1)$; that is, test taker i starts speeding on item $(J - s_i + 1)$. When $s_i = 0$, a test taker is not speeded. The model assumes that when s_i is positive, test takers' abilities begin to drop as they start to speed up, where the decrease in latent ability is defined by d_i. Pulling the aforementioned together, the logistic model for the CPA is.

$$P_{ij}(\theta) = \frac{\exp\left[a_j\left(\theta_i - b_j - d_i \cdot I\left(j > J - s_i\right)\right)\right]}{\exp\left[a_j\left(\theta_i - b_j - d_i \cdot I\left(j > J - s_i\right)\right)\right]}; \tag{49}$$

here $I(j > J - s_i)$ is an indicator function. $I(j > J - s_i)$ equals 1 when $j > (J - s_i)$, and 0 otherwise. The equation demonstrates how a test taker's ability drops by d_i for the last s_i questions, which is like the decrement parameter of Jin and Wang (2014). The logistic model reduces to the 2PL when $s_i = 0$. Shao et al. (2016) noted that the model is different from that of Goegebeur et al. (2008) in that "the probability of answering an item correctly drops sharply after a certain point, instead of gradually" (p. 1121). Ultimately, the goal of the CPA is to detect a change point. Specifically, Shao et al. wanted to test the null hypothesis that $s_i = 0$ against an alternative hypothesis that $s_i > 0$. Shao et al. used a test statistic based on log-likelihood ratio instead of the probability of getting an item correct before and after the change point. Last, Shao et al. controlled for multiple comparisons using the false discovery rate (Benjamini & Hochberg, 1995). The model did not use response time to model speededness, but Shao et al. (2016) noted,

> With the increasing popularity of computer-based testing, response-time data become an increasingly important source of information. Wang and Xu (2015) recently proposed the mixture hierarchical model to detect speededness, which takes advantage of both response and response-time data. Speeded responses towards the end of a test are expected to have a shorter response time than typical. Note that response time is a continuous variable, making it possible to use more conventional CPA techniques than the procedure proposed in this paper. It will be very interesting to apply CPA to both response and response time data. (p. 1139)

The second paper, Shao (2016), in addition to examining whether the CPA by Shao et al. (2016) can help improve item parameter estimation, proposed a CPA using item response-time information for speededness detection but not item responses. Shao (2016) noted,

> Using CPA on response time data has its potential because the CPA method can help to find the estimated speeding point, which makes it much easier for practitioners to investigate the speededness effect rather than just checking the response time patterns in a plot. (p. 62)

A gradual-change log-normal model is proposed to model better the real-life change of response time affected by speededness. The model used by Shao (2016) builds on that of van der Linden (2007). The model assumes that τ (i.e.,

working speed) is constant if a test taker is nonspeeded but has a change point if a test taker is speeded. To test for a change in τ, a Wald statistic (Lafontaine & White, 1986) and a likelihood ratio test statistic are used. If α_j and β_j are known, and local independence holds, the likelihood for the model takes the form (the following narrative of equations is largely repurposed from Shao, 2016)

$$L\left(\tau_i; t_i\right) = \prod_{j=1}^{j} \frac{\alpha_j}{t_{ij}\sqrt{2\pi}} \exp -\frac{1}{2}\left\{\alpha_j\left[\ln t_{ij} - \left(\beta_j - \tau_i\right)\right]\right\}^2. \tag{50}$$

Thus the log-likelihood function can be obtained as

$$l\left(\tau_i; t_i\right) = \sum_{j=1}^{j} \ln \frac{\alpha_j}{t_{ij}\sqrt{2\pi}} - \frac{1}{2}\sum_{j=1}^{j}\left\{\alpha_j\left[\ln t_{ij} - \left(\beta_j - \tau_i\right)\right]\right\}^2. \tag{51}$$

The likelihood statistic is then

$$\Delta l_i = 2\left(l_i^{H_a} - l_i^{H_0}\right); \tag{52}$$

$l_i^{H_0}$ is obtained by plugging $\widehat{\tau_{i,0}}$ (which is the speed parameter estimate using all J items) into Equation 51 and

$$l_i^{(k)} = l\left(\widehat{\tau_{i,1}}; t_i\right) + l\left(\widehat{\tau_{i,2}}; t_i\right), \tag{53}$$

where

$$l_i^{(k)} = \sum_{j=1}^{J} \ln \frac{\alpha_j}{t_{ij}\sqrt{2\pi}} - \frac{1}{2}\sum_{j=1}^{k}\alpha_j\left(\ln t_{ij} - \left(\beta_j - \widehat{\tau_{i,1}}\right)\right)^2 - \sum_{j=k+1}^{k}\alpha_j\left(\ln t_{ij} - \left(\beta_j - \widehat{\tau_{i,2}}\right)\right)^2; \tag{54}$$

$\widehat{\tau_{i,1}}$ is the speed parameter estimate using the first k items, and $\widehat{\tau_{i,2}}$ is the speed parameter estimate using items $k+1$ to J. Then,

$$l_i^{H_\alpha} = \max_{1,2,\ldots,(J-1)} l_i^{(k)}. \tag{55}$$

The estimated number of speeded responses is then

$$\widehat{s_i} = J - \operatorname*{argmax}_{1,2,\ldots,(J-1)} l_i^{(k)}. \tag{56}$$

The Wald statistic tests whether the speed parameter of the first k items is the same as the speed parameter including the last $j-k$ items. The form of the Wald statistic is.

$$W_i^{(k)} = \frac{\left(\widehat{\tau_{i,1}} - \widehat{\tau_{i,2}}\right)}{\frac{1}{I_1\left(\widehat{\tau_{i,0}}\right)} + \frac{1}{I_2\left(\widehat{\tau_{i,0}}\right)}}. \tag{57}$$

$I_1\left(\widehat{\tau_{i,0}}\right)$ is the estimated Fisher information for items 1 to k of test taker i, and $I_2\left(\widehat{\tau_{i,0}}\right)$ is the estimated Fisher information for items $(k+1)$ to J. The $W_{\max,i}^{(k)}$ and corresponding $\widehat{s_i}$ take the following forms:

$$W_{\max,i}^{(k)} = \max_{1,2,\ldots,(J-1)} W_i^{(k)}, \tag{58}$$

and

$$\widehat{s_i} = J - \operatorname*{argmax}_{1,2,\ldots,(J-1)} W_i^{(k)}. \tag{59}$$

(For information on the Fisher information calculations, the reader is referred to Shao, 2016.) Last, the speeded response-time model for simulating data proposed by Shao is

$$\log\left(t_{ij}\right) = \left(\beta_j - \tau_{ij} + \epsilon_{ij}\right) \cdot \min\left\{1, \left[1 - \left(\frac{j}{J} - \eta_i\right)\right]^{\lambda_i}\right\}. \tag{60}$$

If a test taker is speeded after a certain item, the minimum term will be smaller than 1, indicating that the test taker will spend less time on the item than had they not been affected by speededness. The speeded response-time model can be used to simulate response time affected by speededness. A downside of the model by Shao is that it does not incorporate

item responses. However, a benefit of this model is that it does allow for different test takers to have different change points or speededness heterogeneity.

The last paper discussed in this chronology is Williams (2017). Williams noted that no existing model for measuring speededness allowed for speededness to be a function of a test taker's ability; that is, the manifestation of speededness is dependent on the test taker's ability. Williams argued that existing models of speededness do not account for the association between speededness and ability. To demonstrate the utility of modeling the relationship between speededness and ability, the author extended the method of Jin and Wang (2014), which Jin and Wang referred to as the 2PLMix model, to what they referred to as the modified 2PLMix (M2PLMix). The model modification includes the assumption that the probability of being speeded depends on ability. As in Jin and Wang, the probability of a correct item response is

$$P\left(X_{ni}=1|\alpha_j,\beta_j,\theta_i,\eta_{ij}\right)=\frac{\exp\left[\alpha_j\left(\theta_i-b_j-\eta_{ij}\right)\right]}{1+\exp\left[\alpha_j\left(\theta_i-b_j-\eta_{ij}\right)\right]}, \tag{61}$$

where α_j is the discrimination parameter, β_j is the difficulty parameter, θ_i is the ability parameter, and η_{ij} is the decrement parameter. Again, the decrement parameter can be represented as

$$\eta_{ij}=\begin{cases}0;\text{if } j\le\delta_i,\\ \gamma_{\delta_i};\text{if } j>\delta_i;\end{cases} \tag{62}$$

here γ_{δ_i} is the speededness effect parameter, which can be written as

$$\gamma_{\delta_i}=\kappa\left(J-\delta_i\right). \tag{63}$$

Up until this point, the model presentation is identical to Jin and Wang (2014). To extend the model, Williams allowed the location of speededness δ_i to depend on ability level θ_i. In Jin and Wang, π_j denotes the probability of not being speeded—which is the same for the entire test taker population. Williams (2017) argued,

> This assumption not only makes the parameter not applicable under all testing contexts (i.e., low- and high-stakes assessments), but also makes the use of this parameter unrealistic. The probability of not being speeded on a test in the 2PLMix is a property of the test and not a characteristic of the test-taking population. (p. 28)

In the M2PLMix, the function that describes the probability of not being speeded, as a function of ability, is defined as

$$\pi\left(\theta_i\right)=\lambda+(1-\lambda)\frac{1}{1+\exp\left(-\theta_i\right)}, \tag{64}$$

where λ is the threshold probability of not being speeded. λ represents the baseline probability of not being speeded for the test taker population. Therefore the model assumes that every test taker in the population has the same baseline probability of not being speeded—Williams argued that if the assessment is low stakes, we can assume that λ is small and that λ may be small if the time limit is properly defined on an assessment used to make high-stakes decisions. The equation for the probability of not being speeded works as follows: The higher one's ability is, the larger is the logit, and therefore the higher is the probability of not being speeded. Figure 7 provides a graphical depiction of how the M2PLMix influences the probability of being speeded. The figure additionally uses the graphical depiction to compare the M2PLMix to the 2PLMix and 2PL. The method is limited by the fact that it does not incorporate response times. However, like Jin and Wang (2014), the model does allow for speededness heterogeneity by allowing different test takers to begin to attempt items with less effort at different item locations.

Classification

The current section aims to provide an overview of the different ways in which the 28 different methods for measuring speededness reviewed herein vary. A classification system or taxonomy is developed based on the review of the studies. I develop a classification taxonomy that I refer to as the *general classification of speededness methodologies*. Specifically, the articles were classified by whether they (a) were simple rules of thumb for measuring the extent of speededness, (b) were

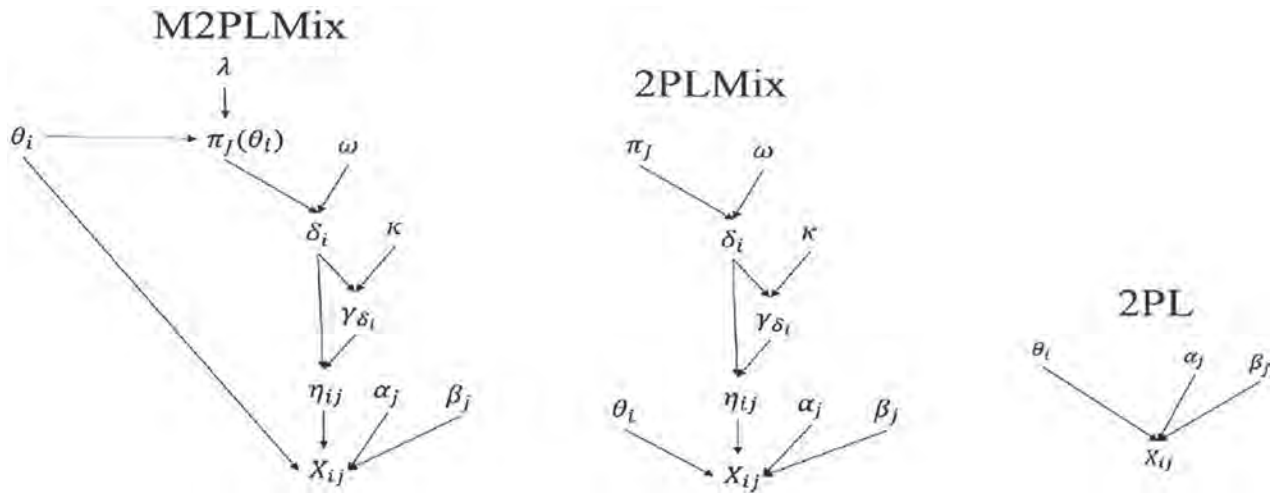

Figure 7 Williams (2017) demonstrated how the modified 2PLMix (M2PLMix) and the two-parameter logistic (2PL) mixture model (2PLMix) relate to the standard 2PL model. The 2PLMix displays a decrement parameter η_{ij}. In the 2PLMix, the decrement parameter is a function of item location of speededness Δ_{ij} and the speededness impact or effect parameter γ_{Δ_i}. In the 2PLMix, there is hyperparameter ω and the probability of being speeded π_j. The M2PLMIX extends the 2PLMix to allow the location of speededness to depend on the probability of being speeded as a function of ability level θ_i and the baseline probability of not being speeded λ. Note that the 2PLMix is the model of Jin and Wang (2014).

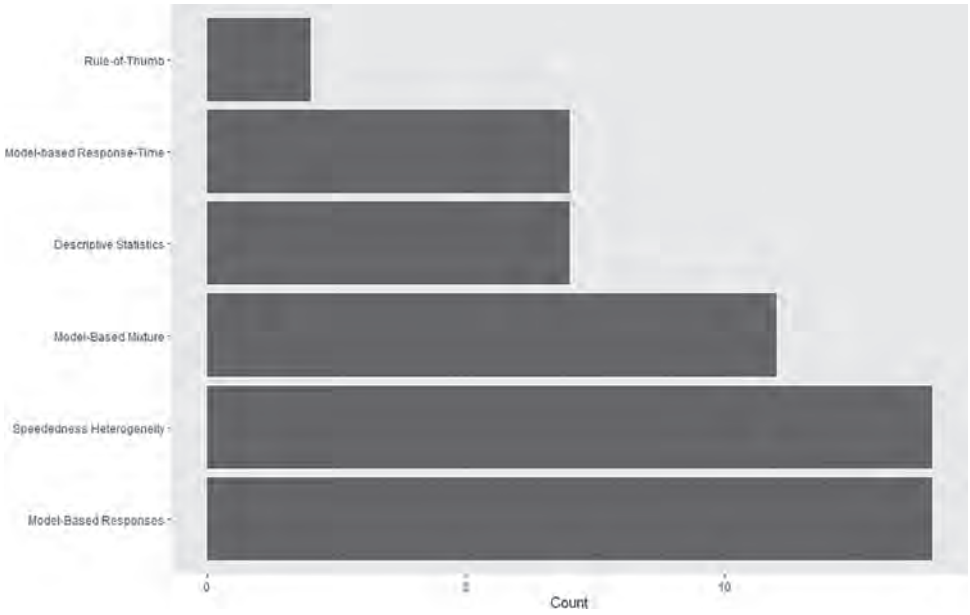

Figure 8 Counts of methods of measuring speededness by general classification of speededness methods. The counts are sorted ascendingly. Model-based is excluded because of its overlap with the other model-based classifications.

basic descriptive statistics for characterizing the extent of speededness, (c) were model-based, (d) were model-based with a mixture model component, (e) allowed for heterogeneity in speededness, (f) used a model-based approach for responses, or (g) used a model-based approach for response times. Figure 8 provides a graph that represents the breakdown of the 28 articles in terms of the number in each category of the general classification of methodologies. Notable takeaways from Figure 8 are that (a) rules of thumb are lacking ($N = 2$) and (b) methods with a model-based approach for responses and methods that allowed for speededness heterogeneity have proved popular for measuring speededness ($N = 14$).

This section ends with a discussion of what each element of the classification taxonomy means and provides examples of how different methods of measuring speededness satisfy a given element of the classification taxonomy. The complete

breakdown of speededness methodologies, in terms of both their chronology and their classification, is presented in Table 1. Column 1 contains the names of the different manuscripts and descriptors about the manuscripts (e.g., whether the manuscript was a dissertation). Columns 2–10 detail characteristics of the models, which can cut across different classes of methodologies. Column 11 provides a count of the citations. Table 2 provides definitions of the different elements of the classification taxonomy.

Rules of Thumb

The rules-of-thumb classification was given to speededness methods that provided basic guidelines or rules of thumb on identifying or assessing speededness. The most straightforward and notable rule of thumb is the Swineford criterion (Swineford, 1956, 1974). The other methods that were classified as rules of thumb included Hadadi and Luecht (1998) and Bontempo (2000). The general rules-of-thumb classification is an area that has had little development in the speededness literature.

Descriptive Statistic

The descriptive statistic classification was given to speededness methods that were analytic but did not use a model-based approach to assess speededness. A prime example of a descriptive statistic was the response rate developed by Bontempo and Julian (1997). Notably, the descriptive statistics for measuring speededness primarily occurred before the instantiation of the landmark mixture models. For example, the methods proposed by Cronbach and Warrington (1951), Donlon (1973, 1980b), Gulliksen (1950a, 1950b), Stafford (1971), and Bejar (1985) all essentially included descriptive statistics of the extent of test speededness, rather than a model-based process to assess speededness on a test.

Model-Based

The model-based classification was given to speededness methods that used some general modeling procedure to assess speededness; that is, the general modeling approaches covered in this classification include linear regression, HLM, LGC analysis, mixture modeling, linear regression, CPA, and IRT. Besides van der Linden (2007), the general modeling approaches have not found much use in the literature on speededness. For example, the relative citation count (Column 8 of Table 1), excluding van der Linden (2007), of the different general modeling approaches is relatively low compared to the mixture modeling approaches with an average of six citations per article. The model-based procedures are notable from basic descriptive statistics because they attempt to model the speededness process in some form directly rather than using basic counting and/or rank ordering.

Model-Based With a Mixture Model Component

The model-based with a mixture model component classification was given to speededness methods that used mixture modeling. The mixture model approach to measuring speededness burgeoned considerably after Yamamoto (1995), Schnipke and Scrams (1997), and Bolt et al. (2002). For instance, eight different mixture models were proposed after the three landmark mixture models for measuring speededness (Cao & Stokes, 2008; Goegebeur et al., 2008; Jin & Wang, 2014; Meyer, 2010; Mroch et al., 2005; Wang, 2011; Wang & Xu, 2015; Williams, 2017). Mixture modeling methods are characterized by an attempt to separate individuals into classes (for speededness research, nonspeeded vs. speeded class). Two points to note about the mixture models are as follows: First, the use of a decrement or decay parameter (e.g., Jin & Wang, 2014) was a notably different way of creating a mixture of responders than, say, Schnipke and Scrams (1997), in which mixtures were defined item by item. The decay parameter allows for individual speededness processes to occur, whereas the item-by-item mixtures evaluate the effects of speededness item by item. Second, none of the mixture models extensively evaluated the appropriateness of class solutions greater than two. Two papers evaluated the appropriateness of a two-class solution compared to a one-class solution (Wang, 2011; Wang & Xu, 2015). Research in mixture modeling has been developed for identifying the appropriate number of classes for a set of dependent variables (Lo et al., 2001; Nylund et al., 2007). Consequently, research on the appropriateness of the assumption of two classes may be required. For example, more than one class of aberrant responders may exist in addition to a solution-response class (e.g., a

solution-response class, a speeded class, and a cheating class). For more detail on the differentiating between aberrant response classes, the reader is referred to Wang et al. (2018).

Speededness Heterogeneity

The speededness heterogeneity classification was given to speededness methods that allowed for speededness to be manifest throughout a test at different points on a test or for different persons. The methods allow individuals to become speeded at different points on a test, and the methods characterize individual speededness behavior. The methods of measuring speededness that allowed for speededness heterogeneity can be found in Column 8 of Table 1. There were slight differences between the methods in how they incorporated speededness heterogeneity. The existence of speededness heterogeneity could be thought of as occurring at the item level or the person level. For example, Bejar (1985) was included in this classification because of his item-level index. In particular, the item-level index would enable one to identify the items or points on a test where speededness had occurred. Conversely, Mroch et al. (2005) allowed for speededness heterogeneity to emerge at different test locations for different test takers. The distinction of speededness heterogeneity is important because it identifies methods that can handle either item-level speededness heterogeneity or person-level speededness heterogeneity. Methods that can handle an interaction between these two types of heterogeneities should be preferred.

Model-Based Approach for Responses or Response Times

The model-based approach for responses or response-time classification was given to speededness methods that used a model for the responses or the response times. Methods for measuring speededness are classified by whether they modeled item responses or response times to distinguish those methods that used both. Specifically, methods for measuring speededness are classified by whether they modeled item responses using a model-based approach (e.g., IRT or CPA). Using a model-based approach for the item responses allows for the identification of speededness based on response accuracy patterns. Similarly, using a model-based approach for the item response times to identify speededness allows for the identification of speededness through a model of response time. A few different methods used response times, yet they varied drastically in their use, ranging from CPA to mixture modeling to LGC modeling. Fourteen articles used a model for the responses. Only three manuscripts used a model for both the responses and response times: These manuscripts were Meyer (2010), van der Linden (2011), and Wang and Xu (2015). The use of joint models for responses and response times in measuring speededness offers a major advantage over methods that only use responses. Response times are of interest because they provide information over and beyond response accuracy, and that information helps explain how a test taker is processing a given test item.

Ensuing Research and Development

The current section details ensuing research and development in the area of speededness. Four areas appear to require additional research and clarification regarding the measurement of speededness: (a) the joint modeling of responses and response times, (b) a theory on the manifestation of speededness, (c) validity issues in test speededness, and (d) experimental research for measuring speededness.

A major takeaway from the review of the literature on measuring speededness is that ensuing research and development in the area of speededness will need to focus on incorporating response times and responses into their methods. The existing research has already demonstrated the utility of using response time in the measurement of speededness. Yet, the author was only able to identify three papers that incorporated responses and response times (Meyer, 2010; van der Linden, 2011; Wang & Xu, 2015) into their methods for measuring speededness. Notably, these papers each relied on a joint model for responses and response times (van der Linden, 2007). Alternative measurement models for incorporating responses and response time have been proposed but have not found use in the measurement of speededness (e.g., Bolsinova et al., 2017; Entink et al., 2009; Fox & Marianti, 2016; Ingrisone, 2008; Lee & Ying, 2015; Loeys et al., 2014; Meng et al., 2015; Molenaar et al., 2016; Molenaar et al., 2015; Roskam, 1997; Tatsuoka & Tatsuoka, 1979; Thissen, 1983; Verhelst et al., 1997; Wang et al., 2013; Wang & Hanson, 2005). Fortunately, the advent of computer-based testing makes the collection of how test takers progress on tests and how much time test takers spend on test items possible and should lead to encouraging research and development in this area (Parshall et al., 2002). Another important takeaway from this

review is that there are multiple theories on the manifestation of speededness. For example, the four articles that used both responses and response times used different theories to define how speededness manifests in a test. Last, given the lack of consensus on a method for measuring speededness, and the focus of research on developing speededness methods, less research has been conducted on how speededness impacts test fairness, psychometric efficiency, and test administration procedures. Furthermore, less experimental research has been conducted on measuring speededness, with only a handful of studies conducting experiments to measure speededness.

Joint Models for Responses and Response Time

The use of joint models for responses and response times in measuring speededness offers a major advantage over methods that only use responses. Response times provide information over and beyond response accuracy. We can use timing data to understand cognitive processes or task-taking behavior exhibited on a test. For example, the use of joint models for responses and response time has further clarified our understanding of the speed–accuracy trade-off. However, one critical assumption in the joint model for responses and response time is the conditional independence of response accuracy and response time (van der Linden, 2007). Notably, recent research has suggested that the conditional independence between response times and ability (response accuracy) may be violated in practice (Bolsinova et al., 2017; Bolsinova & Tijmstra, 2018); that is, response accuracy and response time may be conditionally dependent. Bolsinova et al. (2017) suggested that the conditional dependence between response accuracy and response time is present in the residuals of the response times. One area for ensuing research and development is to examine how speededness results in residual differences by first modeling the conditional dependence without accounting for speededness and then evaluating the conditional dependence after accounting for speededness. For example, does accounting for speededness remove residual dependence in time-limit tests? Does any specific method for measuring speededness result in less conditional dependence between response accuracy and response time? Another assumption of the joint model for responses and response times is constant speed. Though all individuals have estimates of their latent speed, these speeds are assumed to be constant across test items. The manifestation of speededness would seem to contradict this (i.e., a test taker begins to *guess rapidly*), and models that allow for differential speed across a test, or account for differences in speed across items, may be necessary (see Fox & Marianti, 2016; Molenaar et al., 2016; Wang & Xu, 2015).

Manifestation of Speededness

As this review demonstrates, there are multiple theories on the manifestation of speededness: Some suggest that speededness is manifest in omitted and not-reached responses; some suggest that speededness is manifest at a point on a test at which there is no recovery; some suggest that speededness is manifest at a point on a test, but the influence of speededness only gradually takes over, and at different rates for different individuals; some suggest that speededness is manifest when one's pacing strategy fails; some suggest that speededness occurs on every item, but with differing proportions; and some suggest that speededness is only manifest on harder items. There is a great deal of theory on the manifestation of speededness, yet the manifestation of speededness may be context dependent. Research that elucidates when different theories of speededness may be useful would be helpful. Research that compares the impact of assuming different theories of the manifestation of speededness on item and parameter estimates would also be helpful. Each of the theories on speededness appears reasonable, and it may be the case that speededness is manifest with respect to the test context (e.g., low stakes vs. high stakes) or test taker populations (see Williams, 2017). Given the heterogeneity in how speededness may be manifest in a test or test taker population, it would be worthwhile to understand what factors contribute to speededness (e.g., ability). Therefore models that allow speededness to vary as a function of ability or additional covariates would be useful (see Williams, 2017). The only papers I identified that attempted to answer these questions were Wang (2011) and Williams (2017), yet neither of these incorporated response times. Therefore models that incorporate response times and allow covariates (including ability) to explain variation in speededness manifestation would be of use.

Test Fairness, Psychometric Efficiency, and Test Administration Procedures

Test fairness, in terms of the effects of speededness, is one area in which further research is needed. Even though previous studies examined the impact of speededness on different groups, this research has not kept pace with the research

on speededness methods. Further research on group differences, using the new and innovative methods for measuring speededness, is required. The research must clarify whether speededness has a differential impact on different groups (e.g., ethnicity, gender, language). Another area that has had a shortage of research concerns psychometric efficiency. For example, as Williams (2017) stated, "no research has shown how speededness-reducing models have an influence on the estimation of linking coefficients and equating functions" (p. 118). Moreover, even with the advent of a multiplicity of methods for measuring speededness, little research has been conducted on understanding the implications of speededness on item and person estimates using the various methods for measuring speededness. For example, it may be instructive to consider how the different methods, which employ different theories of speededness, subsequently remove any bias in item and person estimates.

Experimental Research for Measuring Speededness

Last, there is a critical need for research and discussion about what is an appropriate time limit or time limit per item. The existing experimental research evidence has had conflicting results and has been mostly context dependent. For example, some experimental research evidence suggests that greater time limits benefit everyone on the *GRE*® test (Wild et al., 1982) and Law School Admissions Test (Evans & Reilly, 1972). However, other experimental research evidence suggests that greater time limits are most beneficial for middle-ability examinees and of no benefit for examinees of low ability (Bridgeman et al., 2004) or, possibly, only beneficial for students of low proficiency (Harik et al., 2018). There are conflicting results, and we are no closer to defining what appropriate time limits are for test takers. Likewise, although some research has evaluated how to control for differential speededness in computer-adaptive tests (see van der Linden et al., 2007), there has been less research and discussion on how items should be administered in both computer-based and paper-and-pencil tests, to reduce the influence of speededness. True experimental studies (e.g., randomly assigning test takers to tests with different time limits) may help produce the clearest evidence of speededness at both the test level and the item level (e.g., Bridgeman et al., 2004).

Concluding Remarks

In current practice, standardized tests are designed to be power tests but are administered with a time limit. Brigham (1932), while evaluating the 1927 administration of the *SAT*® test, stated,

> Perhaps the most frequent inquiry addressed to members of the Committee concerns the matter of accuracy of timing. … It is hoped that time limits may eventually be adjusted so that every candidate will have ample time to answer all problems which he is able to answer. (p. 342)

The hopes of measuring the extent of speededness on a test today may not be far from Brigham's hopes in 1927. The issue that we are concerned with today is that, when performing under the pressure of a time limit is not part of the construct we intend to measure, how do we ensure that each test taker is given ample time to perform at their best and demonstrate their true ability? When test takers do not have enough time to perform at their best and demonstrate their true ability, the extent of speededness is too large. Measuring and removing the effects of speededness moves us closer to ensuring that every candidate will have ample time to answer all problems that they are able to answer. The current review of speededness methods helps clarify potential methods of alleviating the effects of speededness and identifying appropriate time limits. However, a question that may still be open for debate is, how much speededness is acceptable? Given the practical constraints of testing organizations (e.g., the cost of test administration time) and possible fatigue factors with long testing times, there is a trade-off between the time limit and the cost of the time limit. Identifying an optimal time limit and the extent of speededness requires considering both the needs of the test taker and the testing organization.

Acknowledgments

Dakota Cintron conducted this research while participating in ETS's Summer Internship Program. Dr. Cintron is currently a postdoctoral scholar at UCSF's Center for Health & Community.

Notes

1 In the mid-20th century, time-limit tests were referred to as speeded tests (cf. Mollenkopf, 1950).
2 Differential speededness is defined at the individual level by van der Linden et al. (2007). Nonetheless, group differences in differential speededness are also possible.
3 http://search.ets.org/researcher/
4 Equation 4 would need to be multiplied by 100 to make it a percentage. Stafford (1971) did make the multiplication by 100 explicit in his formulas for the SQ but noted that "percentages are used rather than decimals to distinguish the SQ from reliability, validity, and difficulty coefficients" (p. 276).
5 What determines substantial here is a matter of merit, worth, or significance (Scriven, 1991).
6 This is a reasonable assumption in large-scale testing. For example, the 2020 edition of *The Official SAT Study Guide* by the College Board (2019) stated, "Don't dwell on questions that stump you. Circle ones you decide to skip so that you can return to them quickly later. Remember that a question you answer easily and quickly is worth as much as a question that you struggle with or take a lot of time on" (p. 26).
7 Note that item response theory measurement models, or model-based procedures, for response times were proposed as early as Tatsuoka and Tatsuoka (1979) and Thissen (1983).
8 The different points on a test for an item or individual are where the switch points occurred on a test from solution behavior to rapid-guessing behavior. Also, it should be noted that Schnipke and Scrams (1997) did not discuss the existence of a switching behavior (i.e., onset of rapid-guessing behavior); rather, their model allows for the existence of rapid-guessing behavior on each item, which may be more flexible than the heterogeneous item "switching" models and could also be characterized as allowing for speededness heterogeneity. This is reflected in the classification (see Table 1).

References

Aikens, H. A., Thorndike, E. L., & Hubbell, E. (1902). Correlations among perceptive and associative processes. *Psychological Review*, *9*(4), 374–382. https://doi.org/10.1037/h0070072, 382

Barry, D., & Hartigan, J. A. (1993). A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, *88*(421), 309–319. https://doi.org/10.1080/01621459.1993.10594323

Bejar, I. I. (1985). *Test speededness under number-right scoring: An analysis of the Test of English as a Foreign Language* (Research Report No. RR-85-11). ETS. https://doi.org/10.1002/j.2330-8516.1985.tb00096.x

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459. https://doi.org/10.1007/BF02293801

Bolsinova, M., De Boeck, P., & Tijmstra, J. (2017). Modelling conditional dependence between response time and accuracy. *Psychometrika*, *82*, 1126–1148. https://doi.org/10.1007/s11336-016-9537-6

Bolsinova, M., & Tijmstra, J. (2018). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology*, *71*(1), 13–38. https://doi.org/10.1111/bmsp.12104

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*(4), 331–348. https://doi.org/10.1111/j.1745-3984.2002.tb01146.x

Bontempo, B. D. (2000). *Assessing speededness using probabilistic models* [Unpublished doctoral dissertation]. University of Chicago.

Bontempo, B. D., & Julian, E. R. (1997, March 25–27). *Assessing speededness in variable-length computer adaptive testing* [Paper presentation]. Annual meeting of the National Council on Measurement in Education, Chicago, IL, United States.

Bridgeman, B., Trapani, C., & Curley, E. (2004). Impact of fewer questions per section on SAT I scores. *Journal of Educational Measurement*, *41*(4), 291–310. https://doi.org/10.1111/j.1745-3984.2004.tb01167.x

Brigham, C. C. (1932). *A study of error: A summary and evaluation of methods used in six years of study of the scholastic aptitude test of the college entrance examination board*. College Entrance Examination Board.

Cao, J., & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika*, *73*, 209–230. https://doi.org/10.1007/s11336-007-9045-9

Chang, Y.-W., Tsai, R.-C., & Hsu, N. J. (2014). A speeded item response model: Leave the harder till later. *Psychometrika*, *79*, 255–274. https://doi.org/10.1007/s11336-013-9336-2

Clogg, C. C. (1995). Latent class models. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 311–359). Plenum Press. https://doi.org/10.1007/978-1-4899-1292-3_6

College Board. (2019). *The official SAT study guide* (2020 ed.).

Cook, L. L., & Douglass, J. B. (1982, March 18–23). *Analysis of fit and vertical equating with the three-parameter model* [Paper presentation]. Annual meeting of the American Educational Research Association, New York, NY, United States.

Courtis, S. (1924). The relation between rate and quality in educational measurement. *Journal of Educational Research*, *10*(2), 110–131. https://doi.org/10.1080/00220671.1924.10879486

Cronbach, L. J., & Warrington, W. G. (1951). Time-limit tests: Estimating their reliability and degree of speeding. *Psychometrika*, *16*, 167–188. https://doi.org/10.1007/BF02289113

Davidson, W. M., & Carroll, J. B. (1945). Speed and level components in time-limit scores: A factor analysis. *Educational and Psychological Measurement*, *5*(4), 411–427. https://doi.org/10.1177/001316444500500408

Donlon, T. F. (1973, November). *Establishing appropriate time limits for tests* [Paper presentation]. Annual meeting of the Northeast Educational Research Association, Ellenville, NY, United States.

Donlon, T. F. (1980a). *An annotated bibliography of tests of speededness* (GRE Board Report No. GREB-76-09R). ETS.

Donlon, T. F. (1980b). *An exploratory study of the implications of test speededness* (GRE Board Report No. GREB-76-9P). ETS.

Dorans, N. J., & Kulick, E. (1983). *Assessing unexpected differential item performance of female candidates on SAT and TSWE forms administered in December 1977: An application of the standardization approach* (Research Report No. RR-83-09). ETS. https://doi.org/10.1002/j.2330-8516.1983.tb00009.x

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement*, *23*(4), 355–368. https://doi.org/10.1111/j.1745-3984.1986.tb00255.x

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement*, *29*(4), 309–319. https://doi.org/10.1111/j.1745-3984.1992.tb00379.x

Douglas, J., Kim, H. R., Habing, B., & Gao, F. (1998). Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics*, *23*(2), 129–151. https://doi.org/10.3102/10769986023002129

Ellis, R. S. (1928). A method of constructing the scoring test given with time limits to eliminate or weigh the effect of speed. *School and Society*, *28*, 205–207.

Entink, R. K., Fox, J. P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*, 21–48. https://doi.org/10.1007/s11336-008-9075-y

ETS. (2014). *ETS standards for quality and fairness*.

Evans, F. R., & Reilly, R. R. (1972). A study of speededness as a source of test bias. *Journal of Educational Measurement*, *9*(2), 123–131. https://doi.org/10.1111/j.1745-3984.1972.tb00767.x

Fox, J. P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research*, *51*(4), 540–553. https://doi.org/10.1080/00273171.2016.1171128

Freeman, F. S. (1931). The factors of speed and power in tests of intelligence. *Journal of Experimental Psychology*, *14*(1), 83–90. https://doi.org/10.1037/h0073577

Furneaux, W. (1952). Some speed, error and difficulty relationships within a problem-solving situation. *Nature*, *170*(4314), 37–38. https://doi.org/10.1038/170037a0

Goegebeur, Y., De Boeck, P., Wollack, J. A., & Cohen, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika*, *73*, 65–87. https://doi.org/10.1007/s11336-007-9031-2

Goldhammer, F. (2015). Measuring ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental control. *Measurement: Interdisciplinary Research and Perspectives*, *13*(3–4), 133–164. https://doi.org/10.1080/15366367.2015.1100020

Greenhalgh, T., & Peacock, R. (2005). Effectiveness and efficiency of search methods in systematic reviews of complex evidence: Audit of primary sources. *BMJ*, *331*(7524), 1064–1065. https://doi.org/10.1136/bmj.38636.593461.68

Gulliksen, H. (1950a). The reliability of speeded tests. *Psychometrika*, *15*, 259–269. https://doi.org/10.1007/BF02289042

Gulliksen, H. (1950b). *Theory of mental tests*. Erlbaum. https://doi.org/10.1037/13240-000

Hadadi, A., & Luecht, R. M. (1998). Some methods for detecting and understanding test speededness on timed multiple-choice tests. *Academic Medicine-Philadelphia*, *73*(10), S47–S50. https://doi.org/10.1097/00001888-199810000-00042

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17–27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer-Nijhoff. https://doi.org/10.1007/978-94-017-1988-9

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. *2*). Sage.

Harik, P., Clauser, B. E., Grabovsky, I., Baldwin, P., Margolis, M. J., Bucak, D., Jodoin, M., Walsh, W., & Haist, S. (2018). A comparison of experimental and observational approaches to assessing the effects of time constraints in a medical licensing examination. *Journal of Educational Measurement*, *55*(2), 308–327. https://doi.org/10.1111/jedm.12177

Heitz, R. P. (2014). The speed–accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8* Article No. 150, 1–19. https://doi.org/10.3389/fnins.2014.00150

Henmon, V. A. C. (1911). The relation of the time of a judgment to its accuracy. *Psychological Review*, *18*(3), 186–201. https://doi.org/10.1037/h0074579

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Addison-Wesley.

Ingrisone, S. J. (2008). *An extended item response theory model incorporating item response time*. Florida State University.

Jensen, A. R. (1982). Reaction time and psychometric g. In H. J. Eysenck (Ed.), *A model for intelligence* (pp. 93–132). Springer. https://doi.org/10.1007/978-3-642-68664-1_4

Jin, K. Y., & Wang, W. C. (2014). Item response theory models for performance decline during testing. *Journal of Educational Measurement*, *51*(2), 178–200. https://doi.org/10.1111/jedm.12041

Kahraman, N., Cuddy, M. M., & Clauser, B. E. (2013). Modeling pacing behavior and test speededness using latent growth curve models. *Applied Psychological Measurement*, *37*(5), 343–360. https://doi.org/10.1177/0146621613477236

Kim, M. (2017). *Detection of speededness in constructed response items using mixture IRT models* [Unpublished doctoral dissertation]. University of Georgia.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, *8*(2), 147–154. https://doi.org/10.1177/014662168400800202

Lafontaine, F., & White, K. J. (1986). Obtaining any Wald statistic you want. *Economics Letters*, *21*(1), 35–40. https://doi.org/10.1016/0165-1765(86)90117-5

Lajoie, S. P., & Shore, B. M. (1986). Intelligence: The speed and accuracy tradeoff in high aptitude individuals. *Journal for the Education of the Gifted*, *9*(2), 85–104. https://doi.org/10.1177/016235328600900201

Lawrence, I. M. (1993). *The effect of test speededness on subgroup performance* (Research Report No. RR-93-49). ETS. https://doi.org/10.1002/j.2333-8504.1993.tb01560.x

Lee, Y. H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*(3), 359–379.

Lee, Y. H., & Ying, Z. (2015). A mixture cure-rate model for responses and response times in time-limit tests. *Psychometrika*, *80*, 748–775. https://doi.org/10.1007/s11336-014-9419-8

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*(3), 767–778. https://doi.org/10.1093/biomet/88.3.767

Loeys, T., Legrand, C., Schettino, A., & Pourtois, G. (2014). Semi-parametric proportional hazards models with crossed random effects for psychometric response times. *British Journal of Mathematical and Statistical Psychology*, *67*(2), 304–327. https://doi.org/10.1111/bmsp.12020

Lohman, D. F. (1986). The effect of speed–accuracy tradeoff on sex differences in mental rotation. *Perception & Psychophysics*, *39*, 427–436. https://doi.org/10.3758/BF03207071

Longstaff, H. P., & Porter, J. P. (1928). Speed and accuracy as factors in objective tests in general psychology. *Journal of Applied Psychology*, *12*(6), 636–642. https://doi.org/10.1037/h0071593

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Erlbaum.

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice*, *26*(4), 29–37. https://doi.org/10.1111/j.1745-3992.2007.00106.x

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press.

Mahood, Q., Van Eerd, D., & Irvin, E. (2014). Searching for grey literature for systematic reviews: Challenges and benefits. *Research Synthesis Methods*, *5*(3), 221–234. https://doi.org/10.1002/jrsm.1106

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review*, *86*(4), 287–330. https://doi.org/10.1037/0033-295X.86.4.287

McFarland, R. A. (1928). The role of speed in mental ability. *Psychological Bulletin*, *25*(10), 595–612. https://doi.org/10.1037/h0074228

Meng, X.-B., Tao, J., & Chang, H.-H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *Journal of Educational Measurement*, *52*(1), 1–27. https://doi.org/10.1111/jedm.12060

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement*, *34*(7), 521–538. https://doi.org/10.1177/0146621609355451

Molenaar, D., Oberski, D., Vermunt, J., & De Boeck, P. (2016). Hidden Markov item response theory models for responses and response times. *Multivariate Behavioral Research*, *51*(5), 606–626. https://doi.org/10.1080/00273171.2016.1192983

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research*, *50*(1), 56–74. https://doi.org/10.1080/00273171.2014.962684

Mollenkopf, W. G. (1949). *Item-analysis data from an experimental study of the effects on item-analysis data of changing item placement and test time limit* (Research Bulletin No.RB-49-10). ETS. https://doi.org/10.1002/j.2333-8504.1949.tb00914.x

Mollenkopf, W. G. (1950). An experimental study of the effects on item-analysis data of changing item placement and test time limit. *Psychometrika*, *15*, 291–315. https://doi.org/10.1007/BF02289044

Mroch, A. A., Bolt, D. M., & Wollack, J. A. (2005, April 12–14). *A new multi-class mixture Rasch model for test speededness* [Paper presentation]. Annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Mudge, E. L. (1921). Time and accuracy as related to mental tests. *Journal of Educational Psychology*, *12*(3), 159–161. https://doi.org/10.1037/h0069817

Munger, G. F., & Loyd, B. H. (1991). Effect of speededness on test performance of handicapped and nonhandicapped examinees. *Journal of Educational Research*, *85*(1), 53–57. https://doi.org/10.1080/00220671.1991.10702812

Myers, C. T., Mollenkopf, W. G., Morrison, E. J., Toops, H. A., & Wesman, A. G. (1960). Symposium: The effects of time limits on test scores. *Educational and Psychological Measurement*, *20*(2), 219. https://doi.org/10.1177/001316446002000201

Myers, G. (1915). Communications and discussions: Learning against time. *Journal of Educational Psychology*, *6*(2), 115–116. https://doi.org/10.1037/h0071330

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, *14*(4), 535–569. https://doi.org/10.1080/10705510701575396

Oshima, T. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, *31*(3), 200–219. https://doi.org/10.1111/j.1745-3984.1994.tb00443.x

Parshall, C. G., Spray, J. A., Kalohn, J., & Davey, T. (2002). *Practical considerations in computer-based testing*. Springer. https://doi.org/10.1007/978-1-4613-0083-0

Peak, H., & Boring, E. G. (1926). The factor of speed in intelligence. *Journal of Experimental Psychology*, *9*(2), 71–94. https://doi.org/10.1037/h0071020

Rindler, S. E. (1979). Pitfalls in assessing test speededness. *Journal of Educational Measurement*, *16*(4), 261–270. https://doi.org/10.1111/j.1745-3984.1979.tb00107.x

Roskam, E. E. (1997). Models for speed and time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187–208). Springer. https://doi.org/10.1007/978-1-4757-2691-6_11

Ruch, G., & Koerth, W. (1923). Power vs. speed in army alpha. *Journal of Educational Psychology*, *14*(4), 193–208. https://doi.org/10.1037/h0074006

Schmitt, A. P., & Dorans, N. J. (1988). *Differential item functioning for minority examinees on the SAT* (Research Report No. RR-88-32). ETS. https://doi.org/10.1002/j.2330-8516.1988.tb00288.x

Schmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement*, *27*(1), 67–81. https://doi.org/10.1111/j.1745-3984.1990.tb00735.x

Schmitt, A. P., Dorans, N. J., Crone, C. R., & Maneckshana, B. T. (1991). *Differential speededness and item omit patterns on the SAT* (Research Report No. RR-91-50). ETS. https://doi.org/10.1002/j.2333-8504.1991.tb01417.x

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, *34*(3), 213–232. https://doi.org/10.1111/j.1745-3984.1997.tb00516.x

Schouten, J., & Bekker, J. (1967). Reaction time and accuracy. *Acta Psychologica*, *27*, 143–153. https://doi.org/10.1016/0001-6918(67)90054-6

Scriven, M. (1991). *Evaluation thesaurus*. Sage.

Secolsky, C. (1989). *Accounting for random responding at the end of the test in assessing speededness* (Research Report No. RR-89-11). ETS. https://doi.org/10.1002/j.2330-8516.1989.tb00337.x

Shao, C. (2016). *Aberrant response detection using change-point analysis* [Unpublished doctoral dissertation]. University of Notre Dame.

Shao, C., Li, J., & Cheng, Y. (2016). Detection of test speededness using change-point analysis. *Psychometrika*, *81*, 1118–1141. https://doi.org/10.1007/s11336-015-9476-7

Sireci, S. G. (2005). Unlabeling the disabled: A perspective on flagging scores from accommodated test administrations. *Educational Researcher*, *34*(1), 3–12. https://doi.org/10.3102/0013189X034001003

Sklar, A. (1959). Fonctions de repartition a *n* dimensions et leurs marges. *Publications de l'Institut de Statistique del'Université de Paris*, *8*, 229–231.

Stafford, R. E. (1971). The speededness quotient: A new descriptive statistic for tests. *Journal of Educational Measurement*, *8*(4), 275–277. https://doi.org/10.1111/j.1745-3984.1971.tb00937.x

Swineford, F. (1949a). *Fellowship record examination* (Research Bulletin No. RB-49-07). ETS. https://doi.org/10.1002/j.2333-8504.1949.tb00010.x

Swineford, F. (1949b). *Law school admissions test — WLS* (Research Bulletin No. RB-49-12). ETS. https://doi.org/10.1002/j.2333-8504.1949.tb00915.x

Swineford, F. (1956). *Technical manual for users of test analyses* (Statistical Report No. SR-56-42). ETS.

Swineford, F. (1974). *The test analysis manual* (Statistical Report No. SR-74-06). ETS.

Talento-Miller, E., Guo, F., & Han, K. T. (2013). Examining test speededness by native language. *International Journal of Testing*, *13*(2), 89–104. https://doi.org/10.1080/15305058.2011.653021

Tate, M. W. (1950). Notes on the measurement of mental speed. *Journal of Educational Psychology*, *41*(4), 219. https://doi.org/10.1037/h0053990, 236

Tatsuoka, K., & Tatsuoka, M. (1979). *A model for incorporating response-time data in scoring achievement tests* (Report No. CERL-E-7). University of Illinois at Urbana-Champaign, Computer-Based Education Research Lab.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 179–203). Academic Press. https://doi.org/10.1016/B978-0-12-742780-5.50019-6

Thomas, M. H. (2006). *Modeling differential pacing trajectories in high-stakes computer adaptive testing using hierarchical linear modeling and structural equation modeling* [Unpublished doctoral dissertation]. University of North Carolina, Greensboro.

Thorndike, E. L. (1914). On the relation between speed and accuracy in addition. *Journal of Educational Psychology*, *5*(9), 537–541. https://doi.org/10.1037/h0075308

Thorndike, E. L., Bregman, E. O., Cobb, M. V., & Woodyard, E. (1926). *The measurement of intelligence*. Teachers College Bureau of Publications. https://doi.org/10.1037/11240-000

Thurstone, L. L. (1937). Ability, motivation, and speed. *Psychometrika*, *2*, 249–254. https://doi.org/10.1007/BF02287896

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287. https://doi.org/10.1007/s11336-006-1478-z

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*(3), 247–272. https://doi.org/10.1111/j.1745-3984.2009.00080.x

van der Linden, W. J. (2011). Test design and speededness. *Journal of Educational Measurement*, *48*(1), 44–60. https://doi.org/10.1111/j.1745-3984.2010.00130.x

van der Linden, W. J. (2017). Test speededness and time limits. In W. J. van der Linden (Ed.), *Handbook of item response theory: Vol. 3. Applications* (pp. 249–265). Chapman and Hall/CRC Press. https://doi.org/10.1201/9781315119144

van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, *44*(2), 117–130. https://doi.org/10.1111/j.1745-3984.2007.00030.x

van der Linden, W. J., & Hambleton, R. K. (1996). *Handbook of modern item response theory*. Springer. https://doi.org/10.1007/978-1-4757-2691-6

Verhelst, N. D., Verstralen, H. H., & Jansen, M. G. H. (1997). A logistic model for time-limit tests. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 169–185). Springer. https://doi.org/10.1007/978-1-4757-2691-6_10

Wang, A. (2011). *A mixture cross-classification IRT model for test speededness* [Unpublished doctoral dissertation]. University of Georgia, Athens.

Wang, C., Fan, Z., Chang, H.-H., & Douglas, J. A. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics*, *38*(4), 381–417. https://doi.org/10.3102/1076998612461831

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456–477. https://doi.org/10.1111/bmsp.12054

Wang, C., Xu, G., & Shang, Z. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika*, *83*, 223–254. https://doi.org/10.1007/s11336-016-9525-x

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*(5), 323–339. https://doi.org/10.1177/0146621605275984

Wickelgren, W. A. (1977). Speed–accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*(1), 67–85. https://doi.org/10.1016/0001-6918(77)90012-9

Wild, C. L., Durso, R., & Rubin, D. B. (1982). Effect of increased test-taking time on test scores by ethnic group, years out of school, and sex. *Journal of Educational Measurement*, *19*(1), 19–28. https://doi.org/10.1111/j.1745-3984.1982.tb00111.x

Williams, I. J. (2017). *A speededness item response model for associating ability and speededness parameters* [Unpublished doctoral dissertation]. Rutgers University.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19–38. https://doi.org/10.1111/j.1745-3984.2006.00002.x

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

Wissler, C. (1901). The correlation of mental and physical tests. *Psychological Review: Monograph Supplements*, *3*, i-62. https://doi.org/10.1037/h0092995

Wollack, J., & Cohen, A. (2004, April 12–16). *A model for simulating speeded test data* [Paper presentation]. Annual meeting of the American Educational Research Association, San Diego, CA, United States.

Yamamoto, K. (1989). *Hybrid model of IRT and latent class models* (Research Report No. RR-89-41). ETS. https://doi.org/10.1002/j.2333-8504.1982.tb01326.x

Yamamoto, K. (1995). *Estimating the effects of test length and test time on parameter estimation using the hybrid model* (Research Report No. RR-95-02). ETS. https://doi.org/10.1002/j.2333-8504.1995.tb01637.x

Yamamoto, K., & Everson, H. T. (1995). *Modeling the mixture of IRT and pattern responses by a modified hybrid model* (Research Report No. RR-95-16). ETS. https://doi.org/10.1002/j.2333-8504.1995.tb01651.x

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*(2), 245–262. https://doi.org/10.1177/014662168100500212

Younkin, W. F. (1986). *Speededness as a source of test bias for non-native English speakers on the college level academic skills test* [Unpublished doctoral dissertation]. University of Miami.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the medical college admissions test. *Journal of Educational Measurement*, *39*(4), 291–309. https://doi.org/10.1111/j.1745-3984.2002.tb01144.x

Suggested citation:

Cintron, D. W. (2021). *Methods for measuring speededness: Chronology, classification, and ensuing research and development* (Research Report No. RR-21-22). ETS. https://doi.org/10.1002/ets2.12337

Action Editor: Jim Carlson

Reviewers: Brent Bridgeman