*Article*

# Scoring Repeated Standardized Tests to Estimate Capacity, Not Just Current Ability

Policy Insights from the Behavioral and Brain Sciences
2019, Vol. 6(2) 218–224

DOI: 10.1177/2372732219862578
journals.sagepub.com/home/bbs
SAGE

**Daniel McNeish[1] and Denis G. Dumas[2]**

**Abstract**

Changes to educational policies have proliferated testing data to include multiple-administration assessments that repeatedly measure student performance over time. Psychometric models—extended for this type of data—estimate quantities typically associated with assessments that are given once, such as ability at a specific time point. This article considers how multiple-administration assessment offers the opportunity for models to estimate novel quantities that are not available from traditional single-administration assessments but may be of interest to educational researchers and stakeholders. Specifically, *dynamic measurement models* can directly estimate capacity—the expected future score once the construct of interest has fully developed. Preliminary evidence for this approach shows it may be less susceptible to effects of socioeconomic status and may improve predictions of future performance. An example with real-life operational assessment data is provided. Extensions and limitations for educational assessment are also discussed.

**Keywords**

standardized testing, repeated assessment, dynamic measurement, longitudinal psychometrics, multiple-administration assessment

## Tweet

Statistical models often do not consider useful quantities that longitudinal assessment can uniquely estimate. This article discusses directly estimating capacity (rather than extrapolating it); preliminary evidence shows how it may improve inferences from tests.

## Key Points

- Multiple-administration assessment is common, but most scoring methods are straightforward extensions of single-administration assessment.
- Multiple scores per student allow for statistical models to directly estimate different quantities, such as capacity (fully developed ability), rather than extrapolating from ability (current competence).
- Direct capacity estimates for reading and mathematics scores tend to be less affected by demographics.
- Direct capacity estimates tend to improve predictions of future performance with verbal test scores.
- If students are repeatedly tested, maximizing the information gleaned from these tests can provide better inferences, if long-term performance is a central interest.
- Testing agencies, state departments of education, and educators may consider incorporating direct measures of unique quantities as part of holistic evaluations of students.

## Introduction

Annually, millions of people in the United States take a standardized test, with the number worldwide likely exceeding one billion tests per year (Sternberg et al., 2002). Whether that test be a targeted cognitive battery administered one-on-one by a psychologist, a government-sponsored assessment administered across an entire school system, or a college admissions exam managed by a testing company, these tests are scored and interpreted using some type of psychometric model.

Most large-scale tests administer multiple items to the same students (Cai & Hansen, 2018). The responses to different questions from the same students presumably relate to one another because each response depends on the ability of the individual student. For example, mathematics item

[1]Arizona State University, Tempe, USA
[2]University of Denver, CO, USA

**Corresponding Author:**
Daniel McNeish, Department of Psychology, Arizona State University, P.O. Box 871104, Psychology Building, Tempe, AZ 85287, USA.
Email: dmcneish@asu.edu

responses from a given student theoretically will reflect their mathematics ability.

A *construct* (concept), such as mathematics ability, is not directly measurable in the same way that physical attributes are; consequently, *latent variable models* often model these physically unobservable, psychological constructs. Other factors aside, an unobservable, latent variable (e.g., mathematics ability) predicts responses to mathematics items. That is, mathematics ability is not physically observable, but can be captured indirectly through item responses. Each student has his or her own value for the latent variable, which differentially affects the probability of correct item responses.

Inferring ability assumes that the test measures one dimension. The constraint imposed by unidimensionality is that, after accounting for the latent variable, no residual correlation remains between item responses from the same student. When the test aims to assess a currently *developed* construct, such as how well an eighth-grade student knows eighth-grade mathematics, unidimensionality is a reasonable assumption (e.g., Carroll, 1993). This type of test is often *single administration*, whereby the test is given once to measure an ability that is fully formed at the time of testing (i.e., developed; Sternberg et al., 2002).

However, as psychometrics has expanded, interests became more ambitious and began to include *developing* constructs—abilities that are partially formed at the time of testing but that will not be fully realized until some point in the future. With developing constructs, multiple administrations are necessary to track progress toward the fully realized ability (Sternberg et al., 2002). However, today, many prominent testing programs—such as high-stakes admission tests in North America—continue with single-administration tests to assess developing constructs (e.g., Pfieffer, 2012). In these circumstances, the construct at the time of testing (e.g., eighth-grade mathematics ability on an eighth-grade test) is used to extrapolate what the fully realized ability will be (e.g., projected adult mathematics performance); however, it is the *meta-construct* that is of interest. That is, the test is capturing what the student has *achieved* so far (the construct the items are measuring), but a common interest is in the student's future *capacity* (the meta-construct behind the construct being measured). The two are related, but life circumstances could prevent otherwise capable students from acquiring the specific knowledge required for a given test.

Arguments against single-administration tests for such purposes have been a magnet for criticism from the very beginnings of psychometrics as a field of study. W.E.B. Du Bois (1920/2013) criticized this practice directly in his 1920 essay *Race Intelligence*, arguing that current ability (the construct, eighth-grade math ability) could not be logically equated to future potential (the meta-construct, adult math ability). He predicted that confounding still-developing constructs and fully developed constructs would inevitably be a tool for the continued oppression of those who have historically had fewer opportunities to develop their abilities early in life.

This situation is especially true if those scoring poorly on single-administration tests never receive the instruction necessary to develop their ability (e.g., they are denied college admission or placed in a lower academic track), leading to a circular affirmation of the assessment's underprediction of their potential. A similar opinion was provided by the pioneering psychometrician Edward Thorndike (1921) who stated, "Some of us, I fear, claimed a generality for our measures of status and a surety of inference from them to original capacity which it would be very hard to justify" (p. 125). In psychometric terminology, no matter how reliable scores may be, the value of the scores is difficult to justify if they are used improperly (i.e., if they have low *validity* for the intended purpose). These concerns appear to have manifested in the current testing environment as research continues to explore the achievement gap in standardized test scores (Cohen, Garcia, Apfel, & Master, 2006; Lee, 2002; Lee & Bowen, 2006; Reardon, 2013).

## Rise in Multiple-Administration Data

Today, the benefits of multiple-administration assessment have not gone unnoticed (though the focus has been on assessing teachers rather than students; Betebenner, 2009; Castellano & Ho, 2013; Monroe & Cai, 2015). In the United States, legislation has required states to collect student test scores longitudinally (e.g., No Child Left Behind, Every Student Succeeds Act). These data are collected at multiple levels of government (U.S. Department of Education, 2011) and increasingly by testing agencies such as ACT, Inc.'s Aspire test that follows students from Grade 3 to Grade 10 (ACT, 2014a, 2014b) or NWEA's MAP (Measures of Academic Progress) Growth assessments that test students 3 times per year beginning in kindergarten and following them through eighth grade, and in some cases through high school (NWEA, 2019; Thum & Hauser, 2015). Despite these policy changes emphasizing multiple-administration testing, popular psychometric models continue to focus on quantities produced by single-administration assessments.

Multiple-administration tests scored with traditional methods may violate unidimensionality—meaning they measure more than one thing—because items may relate to the latent variable (current ability) *and* responses at a previous administration (the cumulative prior base). Recent research has extended the traditional framework with feasible methods such as longitudinal item response models (Cai, 2010; von Davier, Xu, & Carstensen, 2011).

However, multiple-administration data can create opportunities to directly estimate the meta-constructs (e.g., college-level capacity) that can only be inferred with single-administration testing. That is, rather than generalizing the traditional framework to produce multiple construct scores per person (e.g., yearly ability test scores),

the multiple administrations can be used to directly estimate different quantities—such as ultimate capacity—that may be of more central psychological or educational interest.

Rather than feature the construct score as the focus of the analysis, the approach we will discuss suggests that the construct scores are an instantaneous measure that, although useful, should be a secondary focus. This is much like the items in a standard psychometric analysis—the items are evidence of the latent variable and are analyzed collectively to obtain a single score for the unobservable construct of interest. An individual item is the piece of the puzzle, but a single item is not sufficient by itself. Here, we argue that the single time point construct scores are partial evidence of capacity but are not enough by themselves. Rather, construct scores are pieces of a larger puzzle that can be analyzed collectively to obtain a score of each individual student's unobservable capacity (the estimate of the ability once it has been fully developed). From a more formal statistical perspective, rather than accounting for dependence between scores from different administrations, we explore explicitly modeling these dependencies.

The next section discusses how computational advances have facilitated estimation of statistical models needed to obtain capacity information on a large scale. This recent approach may be able to provide scores that can be more valid than the traditional approaches; estimating construct scores multiple times per student or extrapolating to capacity from single-administration construct score often assume that students' ranks are consistent from the time of administration to the time of inferences (McNeish & Dumas, 2018). New techniques avoid this.

## Dynamic Measurement

### *Precursor:* Dynamic Assessment *(DA)*

Following World War II, an unenviable task facing some Israeli psychometricians was to assess the cognitive abilities of child Holocaust survivors and place them into grade levels. Traditional single-administration assessment was found to systematically underestimate the appropriate grade level because these students' traumatic circumstances meant they knew less than would be otherwise expected for their age at the time of testing, despite the fact that their capacity remained mostly intact and they were capable of learning more complex topics.

To address this problem, DA tests a student multiple times, conducting targeted learning opportunities with a clinician between assessments (Feuerstein, Rand, & Hoffman, 1979). The implicit model of DA appears in Figure 1: Student improvement over time typically follows a nonlinear, decelerating trajectory toward some asymptotic *capacity* (Feuerstein, Feuerstein, & Falik, 2015). In DA, this capacity is the quantity of interest rather than any single-administration score used to calculate it—similar to how the interest in a single-administration test is on the latent ability, not the individual items used to calculate it.

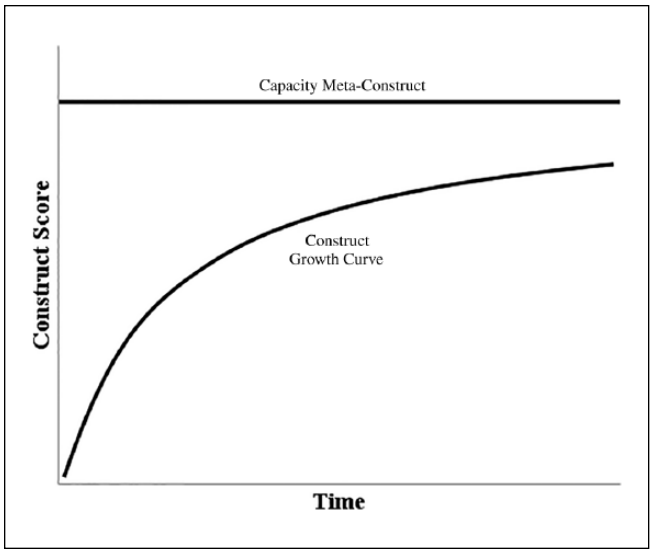


**Figure 1.** Conceptual diagram of dynamic assessment.

DA methods continue to develop as researchers have worked to identify the most appropriate tasks, modes of instruction, and timescale of the test administrations, to produce the most meaningful capacity scores (Elliott, Resing, & Beckmann, 2018; Resing, Bakker, Pronk, & Elliott, 2017). And, in some countries (e.g., the Netherlands, Israel), DA estimates capacity for test respondents from historically marginalized populations or those who have an intellectual disability (Haywood & Lidz, 2007).

Although the idea cleverly circumvents real-world challenges, difficulties can prevent DA from generalizing to the large-scale testing programs that currently operate in North America and around the world. For instance, DA is resource intensive in terms of time and finances, given the necessarily close correspondence between a clinician and a student during the targeted learning opportunities between test occasions. This one-on-one approach to learning also makes standardization more difficult. In addition, DA has historically not relied on formalized statistical models, but, in its most common instantiation, features the descriptive plots of student growth without underlying psychometric frameworks to evaluate the quality of student scores. Nonetheless, dynamic measurement modeling (DMM) generalizes the conceptual framework into a large-scale assessment framework.

### *DMM*

DMM aims to formalize the goal of DA to measure capacity with a statistical model using multiple-administration scores but without requiring the one-on-one interventions between test administrations as in DA. After vertically scaling

scores across time so that they are on a common scale, the conceptual diagram in Figure 1 resembles a nonlinear mixed effect model commonly used for growth modeling (e.g., Cudeck & Harring, 2007; Grimm, Ram, & Hamagami, 2011). These models have increased in popularity in the last decade after advances in statistical computing have made estimation far less arduous (e.g., Proc Nlmixed in SAS, Hamiltonian Markov Chain Monte Carlo as implemented in the Stan software). The general idea of these models is that—with random effects—each student in the data receives a unique monotonic, decelerating growth curve for his or her test scores.

The model directly estimates an upper asymptote (ceiling) on the same scale as the original test scores. Given that each student receives a unique growth curve, this means that each student has a potentially unique upper asymptote representing the ability level he or she would be estimated to obtain once the construct were fully developed (e.g., as time approaches infinity), as well as a unique learning trajectory over time. As in DA, this estimate of the upper asymptote serves as the capacity score. So, rather than extrapolating an ability score into the future, DMM directly estimates the future score, given past scores and a functional form of learning (Dumas & McNeish, 2017).

Various functional forms can be parameterized to include an upper asymptote, each characterizing a different type of growth that may be differentially appropriate, depending on students' age and researcher knowledge about how the construct of interest changes over time. These include S-shaped curves that emphasize inflection points (Gompertz, Richards, and logistic curves; Gompertz, 1825; Richards, 1959) or J-shaped curves that decelerate smoothly (Michaelis-Menten and von Bertalanffy curves; Michaelis & Menten, 1913; von Bertalanffy, 1934). Figure 2 shows the different types of curves that can be parameterized with an upper asymptote within the DMM framework (for parameterizations of these curves with DMM, see McNeish & Dumas, 2017).

Preliminary studies have assessed whether DMM's generalization of the DA framework might improve validity of test scores—given the theoretical argument that it directly estimates the quantity of interest in high-stakes tests (i.e., capacity), rather than requiring extrapolation.

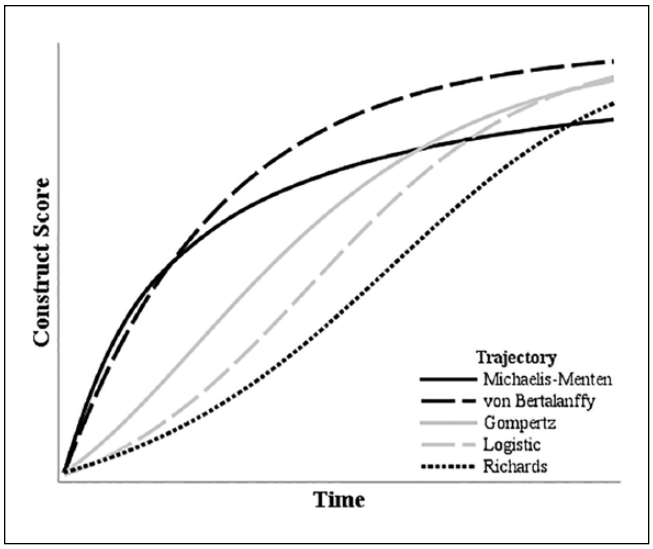


**Figure 2.** Conceptual plot of different nonlinear growth trajectories parameterized with an upper asymptote.
*Note.* All curves demonstrated here have the same intercept and upper asymptote but characterize how change occurs over time differently.

## Preliminary Evidence for Validity

### *Consequential Validity Evidence*

The original conception of the model was demonstrated on data from the Early Childhood Longitudinal Study Kindergarten (Tourangeau et al., 2009). Follow-up studies demonstrating the *consequential validity* of scores from this model similarly use mathematics and reading scores from these data (Dumas & McNeish, 2017, 2018). Briefly, these data contain seven assessment scores from kindergarten to Grade 8 for the same students (the scores are vertically scaled to place all seven scores on a common scale, to directly compare scores over time). Consequential validity refers to the social consequences of a test (Messick, 1995), so these studies looked at the effect of demographic characteristics (socioeconomic status [SES], in particular) on test scores considered in two different ways: (a) on each of the seven test scores individually and (b) the capacity characterized from all seven scores concurrently estimated from a DMM.

Effect sizes of SES on each of the seven reading scores nears the medium cutoff, as do the seven mathematics scores. However, the capacity from the DMM for both reading and math are well below the threshold for a small effect. That is, DMM minimizes the SES effect on test scores. The reliability for DMM capacities is moderate, whereas the reliability for each single-administration score is high. (DMM reliabilities are lower because the future quantity being estimated is still developing. These data also featured few measurement occasions near the bend in the curve, which negatively affects reliability.) Consequently, a possible explanation may be that the DMM effect sizes are smaller because they are attenuated. However, even after correcting for attenuation (i.e., assuming perfect reliability; Osborne, 2003), DMM effect sizes are still negligible and below the small cutoff.

Although far from conclusive, taken together, these studies suggest some preliminary evidence that DMM may have improved consequential validity by providing test scores that are less susceptible to the effects of demographics and may possess some effectiveness for measuring a characteristic such as capacity that should not be broadly affected by demographics. With DMM, students from lower socioeconomic backgrounds appear to be less disadvantaged than they are

with single-administration assessment. Nevertheless, another major question is whether the capacities from DMM are able to predict future performance.

### *Predictive Validity Evidence*

Most test score data collection terminates near the end of secondary school, and large-scale assessment programs are relatively recent creations, so few data resources follow test scores on the same people across the life span. Standardized IQ data (McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009) linked three separate longitudinal studies of verbal IQ (using the Stanford–Binet, Wechsler, or revisions thereof) to create a data set that followed hundreds of people from age 3 to age 72. To recreate how high-stakes standardized testing currently operates—whereby a single test at the end of secondary school is used to make inferences about students' college or work readiness—the data were split at age 20, such that the score at (or nearest to) age 20 for each person was used predict scores later in the life span (McNeish, Dumas, & Grimm, 2019). Dynamic measurement models were also fit to all the scores from age 20 or younger, and the resulting capacity estimates were used to predict scores later in life. The predictive ability of DMM capacities was compared with the predictive ability of the score at age 20.

The DMM capacities increased the predictive validity by about 30 absolute percentage points (from about $R^2 = 20\%$ for age 20 scores to about $R^2 = 50\%$ with DMM capacities; values varied slightly depending on how outliers were treated). On the metric of percent relative change, this equates to a relative increase of more than 100% in favor of DMM capacities. This is the difference between about a .70 correlation with DMM capacities from data up to age 20 and scores later in the life span—compared with about a .45 correlation with age 20 score and scores later in the life span. Furthermore, the correlation between the DMM capacities and the age 20 score was about .60, showing that although the scores are related, the two quantities are distinct, and the DMM capacities are not simply an ornamental transformation of the age 20 score.

Overall, these results seem to indicate that—as one might expect—incorporating multiple test scores vastly improves prediction of future performance compared with a single score. The contribution of DMM is not the insight that more information leads to better predictions, but rather in *how* the multiple pieces of information can be combined into a single, interpretable score to capitalize on this axiom. That is, rather than using all the information to produce an ability score at age 20 (via longitudinal item response theory) whose interpretation is similar to if the test was only given once at age 20, DMM threads the multiple administrations together to yield a quantity not estimated by other multiple-administration models.

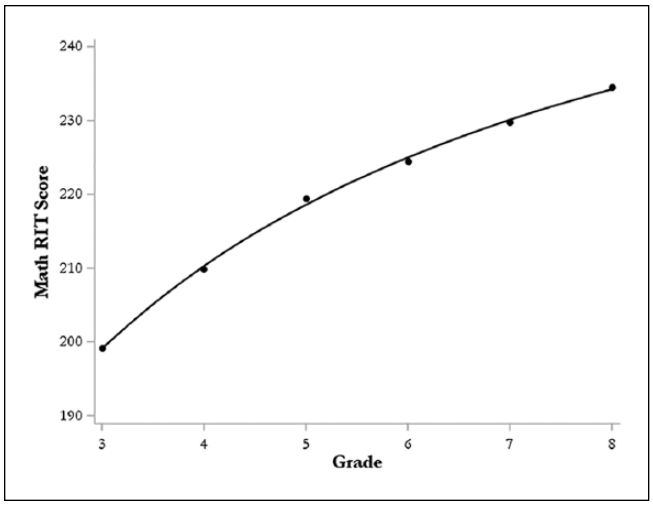


**Figure 3.** Fitted trajectory plot against empirical means for NWEA MAP Growth Mathematics test data.
*Note.* NWEA = Northwest Evaluation Association.

## Example With Operational Standardized Test Data

Using advanced statistical techniques (such as DMM) to model multiple-administration data has the allure that the data already exist. That is, testing agencies and state departments of education currently possess multiple-administration data for which models such as DMM are appropriate. To demonstrate, a DMM can be fit to student longitudinal data from the NWEA's MAP Growth assessments.

NWEA has contracts with school systems in nearly every U.S. state and longitudinally administers tests to millions of students each year across a number of academic domains (mathematics, reading, language usage, and science) to provide a description of growth for both students and schools. The data are vertically scaled Rasch unit (RIT) scores taken from NWEA's MAP Growth Mathematics test. For simplicity, we will demonstrate using data from students test data from at least one administration in each year from Grade 3 to Grade 8. This resulted in an analytic sample of 3,647,034 tests administered to 607,839 students for this demonstration.

The Michaelis–Menten trajectory (see Figure 2) may be the most useful for DMM based on its parsimonious specification and interpretable parameters. We similarly fit this trajectory to these data for the demonstration here. Figure 3 plots the fitted Michaelis–Menten trajectory against the empirical means of the test scores across all 607,839 students. The model-implied trajectory fits closely to the empirical means from the data, suggesting that Michaelis–Menten appears appropriate for these data as well. The average capacity estimate across all students is 274.68, but as described in previous sections, a unique DMM capacity score would be estimated for each student.

This could be potentially useful if long-term mathematics capacity were of interest because this quantity is directly estimated rather than needing to be extrapolated from the last assessment at Grade 8.

## Discussion

Over the past century, the statistics and psychometrics literatures have expanded with complex models that are appropriate for nearly any type of data or research question related to single-administration testing. However, multiple-administration test score data are now routinely collected by various entities, but the psychometric and statistical modeling literature has focused on extending single-administration quantities to multiple-administration data, rather than exploring new quantities afforded by multiple administrations. Statistical modeling commonly asserts that having multiple observations per person is almost always more desirable than having only one observation. However, the advantages to be gained from looking at all test administrations as a cohesive set, rather than as multiple individual assessments has not yet been fully realized in the psychometric modeling literature. High-stakes assessments that are often used as de facto determiners of who is granted access to scarce academic resources are most typically based on one administration, even though other previous test data could strengthen scores, and ultimately predictions they produce about future performance.

As outlined here, dynamic measurement models are one possible way to score multiple-administration test data to directly estimate capacity by blending concepts from the statistical literature on growth modeling into the psychometric literature. Preliminary evidence shows potential advantages in prediction and equity when taking this approach. Seeing as dynamic measurement is a nascent concept; further work is undoubtedly required to fully demonstrate its capabilities and ultimately uncover its weaknesses to more completely evaluate its potential contribution and whether capacity is a useful quantity. But, as testing policy and practice continue to operate with multiple administrations per student, some type of model that incorporates new quantities available from longitudinal data will be indispensable for improving the inferences we make about students, whether that future model follows the dynamic measurement framework presented here or not.

In essence, the main takeaway is that orienting psychometric analysis to incorporate multiple administrations and to expand to new quantities is not an idealist call that would require massive overhaul: Policies producing multiple-administration test data are already in place and these data already exist for use by testing agencies, state departments of education, and educators when forming evaluations of students. By embellishing psychometric models so that they capitalize on additional features of multiple-administration data—rather than treating them as a multivariate embodiment of the historical single-administration paradigm—predictions from test scores can be more accurate, and the interpretations of these scores can be more refined, which is especially desirable, given the substantial personal and societal consequences hanging in the balance.

### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### References

ACT. (2014a). *ACT Aspire® summative assessment technical bulletin #1*. Iowa City, IA: Author.

ACT. (2014b). *ACT Aspire® summative assessment technical bulletin #2*. Iowa City, IA: Author.

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*, 42-51.

Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*, 307-335.

Cai, L., & Hansen, M. (2018). Improving educational assessment: Multivariate statistical methods. *Policy Insights From the Behavioral and Brain Sciences*, *5*, 19-24.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York, NY: Cambridge University Press.

Castellano, K. E., & Ho, A. D. (2013). Contrasting OLS and quantile regression approaches to student "growth" percentiles. *Journal of Educational and Behavioral Statistics*, *38*, 190-215.

Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. *Science*, *313*, 1307-1310.

Cudeck, R., & Harring, J. R. (2007). Analysis of nonlinear patterns of change with random coefficient models. *Annual Review of Psychology*, *58*, 615-637.

Du Bois, W. E. B. (2013). *W. E. B. DuBois on sociology and the Black community*. Chicago, IL: The University of Chicago Press. (Original essay published 1920).

Dumas, D. G., & McNeish, D. M. (2017). Dynamic measurement modeling: Using nonlinear growth models to estimate student learning capacity. *Educational Researcher*, *46*, 284-292.

Dumas, D. G., & McNeish, D. M. (2018). Increasing the consequential validity of reading assessment using dynamic measurement modeling: A comment on Dumas and McNeish (2017). *Educational Researcher*, *47*, 612-614.

Elliott, J. G., Resing, W. C. M., & Beckmann, J. F. (2018). Dynamic assessment: A case of unfulfilled potential? *Educational Review*, *70*, 7-17.

Feuerstein, R., Feuerstein, R., & Falik, L. H. (2015). *Beyond smarter: Mediated learning and the brain's capacity for change*. New York, NY: Teachers College Press.

Feuerstein, R., Rand, Y., & Hoffman, M. B. (1979). *The dynamic assessment of retarded performers: The learning potential assessment device, theory, instruments, and techniques*. Baltimore, MD: University Park Press.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. F. R. S. &c. *Philosophical Transactions of the Royal Society of London*, *2*, 513-583.

Grimm, K. J., Ram, N., & Hamagami, F. (2011). Nonlinear growth curves in developmental research. *Child Development*, *82*, 1357-1371.

Haywood, C., & Lidz, C. (2007). *Dynamic assessment in practice: Clinical and educational applications*. New York, NY: Cambridge University Press.

Lee, J. S. (2002). Racial and ethnic achievement gap trends: Reversing the progress toward equity? *Educational Researcher*, *31*, 3-12.

Lee, J. S., & Bowen, N. K. (2006). Parent involvement, cultural capital, and the achievement gap among elementary school children. *American Educational Research Journal*, *43*, 193-218.

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods*, *14*, 126-149.

McNeish, D., & Dumas, D. G. (2017). Nonlinear growth models as measurement models: A second-order growth curve model for measuring potential. *Multivariate Behavioral Research*, *52*, 61-85.

McNeish, D., & Dumas, D. G. (2018). Calculating conditional reliability for dynamic measurement model capacity estimates. *Journal of Educational Measurement*, *55*, 614-634.

McNeish, D., Dumas, D. G., & Grimm, K. J. (2019). *Estimating new quantities from longitudinal test scores to improve forecasts of future performance*. doi: 10.31234/osf.io/s6p5f.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, *14*, 5-8.

Michaelis, L., & Menten, M. L. (1913). Die Kinetik der Invertinwirkung. *Biochem. Z*, *49*, 333-369.

Monroe, S., & Cai, L. (2015). Examining the reliability of student growth percentiles using multidimensional IRT. *Educational Measurement: Issues and Practice*, *34*, 21-30.

NWEA. (2019). *MAP® growth™ technical report*. Portland, OR: Author.

Osborne, J. W. (2003). Effect sizes and the disattenuation of correlation and regression coefficients: Lessons from educational psychology. *Practical Assessment, Research & Evaluation*, *8*, 1-7.

Pfeiffer, S. I. (2012). Current perspectives on the identification and assessment of gifted students. *Journal of Psychoeducational Assessment*, *30*, 3-9.

Reardon, S. F. (2013). The widening income achievement gap. *Educational Leadership*, *70*, 10-16.

Resing, W. C. M., Bakker, M., Pronk, C. M. E., & Elliott, J. G. (2017). Progression paths in children's problem solving: The influence of dynamic testing, initial variability, and working memory. *Journal of Experimental Child Psychology*, *153*, 83-109.

Richards, F. J. (1959). A flexible growth function for empirical use. *Journal of Experimental Botany*, *10*, 290-301.

Sternberg, R. J., Grigorenko, E. L., Ngorosho, D., Tantufuye, E., Mbise, A., Nokes, C., . . . Bundy, D. A. (2002). Assessing intellectual potential in rural Tanzanian school children. *Intelligence*, *30*, 141-162.

Thorndike, E. L. (1921). Intelligence and its measurement: A symposium. *Journal of Educational Psychology*, *12*, 124-127.

Thum, Y. M., & Hauser, C. H. (2015). *NWEA 2015 MAP norms for student and school achievement status and growth* (NWEA research report). Portland, OR: Northwest Evaluation Association.

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., Najarian, M., & Hausken, E. G. (2009). *Early Childhood Longitudinal Study, Kindergarten class of 1998–99 (ECLS-K): Combined user's manual for the ECLS-K eighth-grade and K–8 full sample data files and electronic codebook*. Washington DC. Retrieved from https://nces.ed.gov/ecls/data/ECLSK_K8_Manual_part1.pdf

U.S. Department of Education. (2011). *Final report on the evaluation of the growth model pilot project*. Retrieved from https://www2.ed.gov/rschstat/eval/disadv/growth-model-pilot/gmpp-final.pdf

von Bertalanffy, L. (1934). Untersuchungen über die Gesetzlichkeit des Wachstums [Inquiries on Growth Laws]. *Development Genes and Evolution*, *131*, 613-652.

von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika*, *76*, 318-336.