4-2016

The Effects of Scaling on Trends of Development: Classical Test Theory and Item Response Theory

Weldon Z. Smith
University of Nebraska-Lincoln, weldon@huskers.unl.edu

Spring 4-21-2016

The Effects of Scaling on Trends of Development: Classical Test Theory and Item Response Theory

Weldon Zane Smith

The Effects of Scaling on Trends of Development: Classical Test Theory and Item Response Theory

by

Weldon Zane Smith

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of the Requirements

For the Degree of Master of Arts

Major: Educational Psychology

Under the Supervision of Professor James A. Bovaird

Lincoln, Nebraska

May, 2016

The Effects of Scaling on Trends of Development: Classical Test Theory and Item Response Theory

Weldon Zane Smith, M.A.

University of Nebraska, 2016

Advisor: James A. Bovaird

The scale metrics used in educational testing are often arbitrary, and this can impact interpretation of scores on measurements. Both classical test theory sum scores and item response theory estimates measure the same underlying dimension, but differences in the two scales may lead one to be more preferential than the other in interpreting data. Mismatch between individual ability and test difficulty can further result in difficulties in correctly interpreting trends of development in longitudinal data. A previous limited simulation by Embretson (2007) demonstrated that classical test theory sum scores result in misinterpretation of linear trends of development, and that item response theory estimates improve upon the problem. This study replicates the results from the previous literature, as well as extends the results to include simulation of development in both quadratic and cubic trends. Results indicate that while item response theory scaling does improve estimates for the linear, quadratic, and cubic trends simulated, ultimately the two methods perform very similar to one another. Item response theory estimates resulted in marginally fewer Type I and Type II errors, especially when

investigating interaction effects. The mismatch between test difficulty and ability level of test takers has the strongest impact on correctly interpreting how individuals develop over time.

Acknowledgements

I would like to express gratitude to my advisor Dr. Bovaird for his help in completing my master thesis, as well as for introducing me to my topic. I would also like to thank my reader Dr. De Ayala for his course on IRT that helped me learn how to do the simulation necessary for my thesis.

I would like to thank my family, Zane, Wendy, and Alyssa, for always being there. Your love and warmth has meant the world to me and always helps me get through any problems I face. Thank you to all my extended family members, who are too numerous to name here but all of whom have had a lasting impact on my life. All of you never stopped believing in me, and that has always kept me going.

Special thanks to my friends here at UNL, including HyeSun Lee, the rest of the birthday club, and the guys that manned the NEAR Center with me. I could not have made it through without all of you making my graduate career fun. To my friends back in Blanco and the surrounding areas, thanks for always believing in me. Last but not least, to my buddy James, thanks for always being there you big galoot.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Classical test theory (CTT) and item response theory (IRT) offer different approaches to scaling latent constructs. The differences between these methods may impact their usefulness in different situations. Embretson (2007) investigated how both CTT and IRT estimates of ability influence linear trends of development. Results indicated that IRT scaling may help to observe the correct linear trend of development. This study aims to replicate previous findings and extend this line of inquiry to investigate ether the same results occur for quadratic and cubic trends of development. Factors such as Type I and II error rates, and effect sizes are reported in order to outline the differences between these two scaling methods. By understanding issues of measurement and investigating CTT and IRT scaling in a longitudinal development this research aims to help researchers and practitioners select the correct scale of measurement to use for their data.

Arbitrary Scales of Measurement

Measurement plays an important role in all types of research. Educational research is no exception to this as test development, evaluation of students, evaluation of interventions, evaluating teacher performance, and many other testing related issues are deeply ingrained in education. Reliability of data and the validity of inferences from a measure are important aspects of measurement that can help us to understand the measures we use. Both reliability and validity are properties of the data resulting from a scale, not the actual scale itself (Messick, 1995). All inferences made from scales rely on the context in which the data were collected, so the individuals observed by a measure and the setting in which an observation takes place can influence the meaning placed on a measure. Therefore, it is important to consider many aspects of a measure as well as the

situation surrounding it in order to give meaning to what the measure actually tells us. One very important aspect of measurement next to reliability and validity is that of the actual metric of the measurement. This metric can be clearly grounded in reality, like height or weight, or something more abstract.

Generally, the aspects of behavior we are interested in measuring with a metric are not directly observable, especially in education in which we may be interested in measuring something like a student's math skill. Since we cannot test the entire domain of math knowledge we instead test a sample of the universe of math items a student might be expected to know, and from that sample we hope to ascertain the student's actual math skill. Relating a sample of math items to a construct like math skill is intuitive, however, as other constructs of interest such as student motivation or self-efficacy may not map as easily to something directly observable. Depending on how a measure is created and how behaviors are defined we may end up with many measures in an attempt to make inferences about a construct. Each of these measures will likely be on a different metric, but purportedly measure the same construct.

Blanton and Jaccard (2006a) discuss the concept of arbitrary metrics in measurement, defining a measure as having an arbitrary metric when the actual location of observed scores on the underlying dimension of a construct is unknown, or when the meaning of a change in unit on a scale is unknown. This inability to understand how exactly an observed metric maps onto an unobserved metric can cause problems for understanding how reliable the measure is or how valid inferences made from the measure are. Blanton and Jaccard discuss a strategy for addressing arbitrary metrics, suggesting researchers should identify relevant meaningful events, make a case for the

importance of the events and position them on the underlying psychological dimension in relation to one another, build consensus with other researchers about the ordering and placement of such events, link test scores to events in order to help add meaning to the metric of the test, and build evidence and consensus for values used in diagnostic statements based on test scores. Blanton and Jaccard concluded that individual scores and changes between scores may not be meaningfully interpreted, but that research findings do not suffer for involving arbitrary metrics.

Embretson (2006) addressed the issues brought up in Blanton and Jaccard (2006a), stating arbitrary metrics have a direct impact on statistics for group comparisons and trend analysis, suggesting that inferences may be impacted by scale as significance levels and power will be impacted. Embretson demonstrated this claim by showing that data generated with mean effects but no interactions can indeed result in an interaction effect although one should not be present. Further, these differences seem to relate to test difficulty and that longitudinal studies on trend will we affected by such false interactions. Schulz and Nicewander (1997) suggest that depending on measurement scale different growth functions across time as well as different patterns of variances will change. Embretson suggested that IRT may be useful in mapping constructs to items. Blanton and Jaccard (2006b) responded to Embretson's rebuttal, stating that while it did not capture the complex issues surrounding mapping psychological behaviors onto underlying dimensions, the claim that that IRT scaling and similar models are of potential use for addressing issues with arbitrary scaling.

The four scales of measurement: nominal, ordinal, interval and ratio, popularized by Stevens (1946), play an integral role in scaling issues. While raw scores are typically

treated as being at the interval or ratio scale for analyses, these scores are inherently ordinal in nature. According to Morse, Johanson, and Griffeth (2012), typical parametric statistical models may not be appropriate for ordinal scaled variables due to limitations of the scale. Davison and Sharma (1990) demonstrated that using ordinal scale variables in regression analyses can result in the detection of interaction effects which are actually not present in the data. Using raw scores in analyses such as ANOVA may result in higher rates of Type I and Type II errors due to this problem (Embretson, 1996). Kang and Waller (2005) suggest that spurious interactions are more likely to occur when there is a mismatch between test difficulty and average ability level of test takers.

Another issue related to arbitrary metrics is the issue of floor and ceiling effects in measurement. A floor effect can occur when individual's abilities are much lower than can be measured by a test, and in the opposite direction, a ceiling effect can occur when individual's abilities are much higher than can be measured by a test. In the instance in which many individuals are subject to a floor effect, the arbitrary score of zero on a test does not map to a single location on the underlying construct, such that meaningful interpretation or differentiation of such individuals becomes either very difficult or even impossible. Wang, Zhang, McArdle, and Salthouse (2009) investigated how ceiling effects impact longitudinal research, reporting incorrect model selection, biased parameter estimation, and misinterpretation of parameters. Additionally, 18% or more individuals reaching the ceiling at one time point resulted in problems with the data analysis. Both CTT and IRT approaches result in arbitrary scales of measurement, and understanding how these two approaches function can help us understand their benefits and shortcomings.

The current study aims to demonstrate how differences in the two metrics of CTT and IRT can lead to different inferences even though the same true ability level underlies the construct both measure. Both CTT and IRT attempt to measure the true ability, or true score, of an individual based on either their overall score on a test or their responses to items on a test. Understanding these two theories and how they define true scores and standard errors of measurement helps highlight why scores from these two theories differ in their estimations of individual ability.

Classical Test Theory

In CTT, total test score provides an estimate of a person's ability. Difficulty of items can be calculated by taking the number of individuals who got an item right and dividing it by the total number of individuals. This tells us the proportion of individuals who got an item right and, in a sense, how difficult the item was. Additionally, one can compute a discrimination coefficient by correlating the scores on one item with the scores on all other items. A higher correlation tells us the item gives good information on differentiating people by ability. However, statistics resulting from CTT are dependent on the group of individuals from which the statistics were obtained (Lawson, 1991), and additionally may differ depending on groups of differing abilities (Hambleton, Swaminathan, Cook, Eignor, & Gifford, 1978). To illustrate this issue McKinley and Mills (1989) suggest the scenario in which examinees are assigned scores relating to the number of items they get correct, and the items, in turn, are given a difficulty statistic based on the proportion of individuals that get the item correct. If the two groups differ in ability level and take the same test the difficulty parameters estimated for each group will differ, and likewise, an individual may differ in observed ability if given two different

tests. Scores from a CTT approach are therefore arbitrary if different samples and groups can affect resulting scores in such a way that the metric depends on the individuals within a sample. This limitation of CTT has long been known, as Wright and Stone (1979) made reference to the instability of estimates in their foreword stating, "the so-called measurements that we now make in educational testing are no damn good!" (p. xi).

True Scores in Classical Test Theory

In CTT, scores on a measurement are typically the summed number of items an individual gets correct, an inherently ordinal scale variable. Conceptually, an individual's true score, τ, can be thought of as the score they should make on a specific measurement given their ability. However, this true score is not observable due to error in measurement, ε, which exists due to an innumerable number of causes such as how an individual is feeling to distractions influencing an individual's attention on a measurement. These errors are generally assumed to be unbiased with an expected value of zero (Lord, 1980). Due to these uncontrollable random errors, the closest approximation to an individual's true score on a specific measurement is their observed number-correct score, *x*. Therefore, we can consider an individual *i*'s observed score on a specific measurement, *m*, to be a combination of both their true score and random errors:

$$x_{mi} = \tau_{mi} + \varepsilon_{mi}, \quad (1)$$

of which we can only truly measure their observed score. By reducing errors in measurement we hope to obtain a better estimate of true scores.

This concept, discussed by Lord (1980), can be taken further when we consider what would happen if an individual could take a specific measurement an infinite number of times, with the individual's true score never changing, and each measurement being

unaffected by all other measurements. This is essentially treating the errors as uncorrelated from measurement to measurement. In this situation the average of all of the observed scores will equal the true score due to the random errors balancing and cancelling one another out. In this case, one can think of the observed scores x_{mi} as being caused by a random process and define a random variable X_{mi} upon the set of possible observed scores x_{mi}. In this infinite number of measurements on m we would expect different observed scores to occur at different frequencies, and that these frequencies would belong to a propensity distribution, F_{mi} (x_{mi}) of the random variable X_{mi}, with $F_{mi}(x_{mi}) = P(X_{mi} \leq x_{mi})$. Following this, the expected value of the observed scores is the same as the true score, or,

$$\tau_{mi} = \mathcal{E}[X_{mi}], \qquad (2)$$

where X_{mi} is the random variable with values of x_{mi}, $\mathcal{E}$ is the expectation, or mean, of the random variable with respect to the propensity distribution $F_{mi}(x_{mi})$, and τ_{mi} is the constant true score of the individual (Lord, Novick, & Birnbaum, 1968).

This relationship between true score and observed scores allows us to make some important observations about true scores. True score scale is determined by the observed score scale, and additionally though we cannot measure true scores they are theoretically related to observed scores based on the definition of true scores given by Equation 2 (Lord, Novick, & Birnbaum, 1968). Logically, it should follow that the expected value of the error of measurement can be defined by the equation,

$$\mathcal{E}[E_{mi}] = \mathcal{E}[X_{mi} - \tau_{mi}] = \tau_{mi} - \tau_{mi} = 0, \qquad (3)$$

showing that the errors are unbiased. Based on this and the previous arguments it becomes apparent that,

$$\tau_{mi} = \mathcal{E}[X_{mi} - E_{mi}] = \mathcal{E}[X_{mi}], \quad (4)$$

showing that true score and error are uncorrelated with one another, as the regression of E on τ would have a zero slope and therefore a zero correlation coefficient. From this fact it follows that errors between each of the infinite number of measurements will be uncorrelated with one another and have an expected value of zero. Despite each measurement having an error associated with it we should feel comfortable using x_{mi} as an estimate of true score, τ_{mi}.

Standard Errors of Measurement in Classical Test Theory

Of interest to this study is the fact that measurement error in CTT may differ for each individual. Though two individuals may share a common number-correct score, that does not necessarily mean their true score and error of measurement will be the same. Since the error of measurement can be defined by,

$$e_{mi} = x_{mi} - \tau_{mi}, \quad (5)$$

we can see that if an individual has a specific τ_{mi} that e_{mi} and x_{mi} will only differ by a constant, which results in the same standard deviation for each. This standard deviation is the standard error of measurement at τ_{mi}, expressed as $\sigma_{e_{mi}|\tau_{mi}}$ (Lord, Novick, & Birnbaum, 1968). From this, we can define thevariance error of measurement, s^2 of CTT as $\sigma^2_{e_{mi}|\tau_{mi}}$ averaged across all N individuals,

$$s^2 = \frac{1}{N}\sum^{N} \sigma^2_{e_{mi}|\tau_{mi}}. \quad (6)$$

That is, the variance error of measurement CTT and its square root, the standard error of measurement, are constant across the entire range of possible number-correct scores. This is a commonly known shortcoming since the assumption of a measure being equally

precise for individuals of varying ability levels does not hold up to scrutiny (Hambleton, Swaminathan, & Rogers, 1991).

Item Response Theory

IRT uses individual ability as well as item characteristics to predict responses, typically correct or incorrect, observed on a measure (De Ayala, 2009). Individual's ability and item difficulty are located on the same dimension, and regressing an individual's observed response on an item with these two locations forms the backbone of IRT estimation. These ability and item difficulties are still on an arbitrary metric, however, they are on the same continuous metric. In IRT, individuals have a singular latent ability level, but items may have multiple parameters relating to different item characteristics.

Similar to how individuals are believed to possess a true score for some measure, IRT invokes the idea of individuals having a true ability level, θ. In IRT, when considering dichotomous items, depending on their level of ability an individual will have a certain probability of getting the item correct. It is important to note that individuals with the same ability level will have the same probability of a correct response on a particular item. As θ increases, so does $P(\theta)$, that is, individuals with higher ability levels are more likely to correctly respond to an item. So, if P is the probability an individual gets an item correct, we might define Q as the probability an individual gets an item incorrect, and that $Q = 1 - P$.

The most simple of IRT models, the one-parameter logistic model, is defined by the equation,

$$P(X_{ki} = 1|\theta_i, \delta_k) = \exp(\theta_i - \delta_k) / (1 + \exp(\theta_i - \delta_k)). \quad (7)$$

In this equation, θ_i is the ability of person *i* and δ_k is the difficulty of item *k*, both of which are located on the same dimension. $P(X_{ki} = 1|\theta, \delta)$ represents the probability of getting an item correct given an ability level and item difficulty. The two-parameter model introduces the discrimination parameter, α_k:

$$P(X_{ki} = 1|\theta_i, \alpha_k, \delta_k) = \exp(\alpha_k(\theta_i - \delta_k)) / (1 + \exp(\alpha_k(\theta_i - \delta_k))). \quad (8)$$

This parameter relates how well an item differentiates individuals according to their location on the underlying dimension (De Ayala, 2009). The one-parameter model may actually involve a discrimination parameter, but it is constant across all items. In the case of the Rasch model, item discrimination of the one-parameter model is fixed at 1, effectively fixing the unit of measurement at the logit (De Ayala). Finally, the three-parameter model introduces the chance or guessing parameter, χ_k:

$$P(X_{ki} = 1|\theta_i, \alpha_k, \delta_k, \chi_k) = \chi_k + (1 - \chi_k)\frac{\exp(\alpha_k(\theta_i - \delta_k)}{1+\exp(\alpha_k(\theta_i - \delta_k)}. \quad (9)$$

This model takes into account the probability of responding correctly to an item based on chance.

Unlike CTT where item difficulty on a test depends on the individuals taking the test, IRT estimates should be unbiased across samples and tests (Embretson & Reise, 2000). McKinley and Mills (1989) suggest IRT analyses should result in sample free measurements, resulting in the same estimates regardless of the test they take. Additionally, item difficulty and discrimination parameters should remain stable across groups of individuals (Lawson, 1991). Tinsley and Dawis (1977) discussed the robustness and generality of the Rasch model, stating that while an individual's estimates are less biased by the difficulty of the test used, the precision of measurement still depends on how appropriate items are for an individual's ability level, and therefore

tailoring a test's difficulty to an individual's ability will maximize the model's usefulness over CTT.

In relation to arbitrary metrics, Becker and Forsyth (1992) investigated how a Thurstone (1925) scaling method and both one-parameter and three-parameter IRT models compare to one another. Thurstone scaling, an early attempt to measure attitudes based on individual's responses scaled based on total scores and the total scores of other individuals, and the Rasch model function very similarly to one another (Andrich, 1978). Becker and Forsyth's findings suggested that the two scaling methods result in similar results. However, they did suggest that scaling method has an impact on interpretations as well as expectations of growth for achievement tests. Andrich further discussed the differences between the two approaches, suggesting that the Rasch model allows for individual level parameters to be eliminated from the model resulting in a "sample-free" model, whereas Thurstone's approach results in parameters more directly tied to a specific sample. Grimm, Kuhl, and Zhang (2013) also suggested that the measurement model used can have an effect on estimates of growth trajectory, noting that IRT scaling imparts several advantages to studying change and ultimately recommend fitting IRT growth models when possible.

IRT scaling has been suggested as an approach to help alleviate problems with spurious interactions and increased Type I and Type II errors due to ordinal scaling in analyses (Embretson, 1996). Embretson used ANOVA to demonstrate that total scores from CTT may result in higher Type I and II error rates for interaction terms, while IRT scaled scores do so to a lesser degree. In an extension of Embretson's study, Kang and Waller (2005), demonstrated that for data sets that are well characterized by IRT models,

IRT scaling may control for spurious interactions and Type I errors in moderated multiple regression. Similarly, Morse, Johanson, and Griffeth (2012) found that the graded response models, too, was less likely to result in spurious interaction effects than number correct scores.

True Scores in Item Response Theory

Both CTT and IRT involve scoring items on a test, so as done in CTT we might define *x* as the number-correct score for a measurement, as in Lord (1980). Where CTT has a random variable X_{mi} with a distribution of scores x_{mi}, IRT instead has a distribution of scores x_{mi} for a given ability level θ_i. We may denote this frequency distribution as $G(x_{\text{mi}}|\theta_i)$, and if all *j* items on a measure *m* have the same item response functions it would be defined as the binomial distribution,

$$G(x_{mi}|\theta_i) = \binom{j}{x_{mi}} P^{x_{mi}} Q^{j-x_{mi}}, \quad (10)$$

and the generating function of $(Q + P)^j$ expresses $G(x_{mi}|\theta_i)$ for the number of possible values of x_{mi} for k = 0, 1, …, j. The items are not restricted to simply having the same item response functions, however, and the generalized binomial distribution can be generated from,

$$G(x_{mi}|\theta_i) = \prod_{k=1}^{j} (Q_k + P_k), \quad (11)$$

where each item, *k*, can have a different item response function. Lord states $G(x_{\text{mi}}|\theta_i)$ does not have a simple form, and defines the expected value of the number-correct score for a given ability level as,

$$\mathcal{E}[x_{mi}|\theta_i] = \sum_{k=1}^{j} P_k(\theta_i), \quad (12)$$

since the score of a measure is the sum of the score of the items, and in the case of dichotomous items the average item score is the same as the probability of getting an item correct. It follows that the expected value of the number correct score for an individual at a given ability level is,

$$\mathcal{E}[X_{mi}] = \sum_{k=1}^{j} P_k(\theta_i) \ or \ \tau_{mi} = \sum_{k=1}^{j} P_k(\theta_i), \quad (13)$$

logically relating the true score τ_{mi} with IRT's ability θ_i. It should be obvious that both τ_{mi} and θ_i measure the same thing, albeit on different scales of measurement.

Standard Errors of Measurement in Item Response Theory

Lord (1980) defines the variance of the number-correct score for a given ability level as,

$$\sigma^2[x_{mi}|\theta_i] = \sum_{k=1}^{j} P_k(\theta_i)Q_k(\theta_i), \quad (14)$$

based on the variance of a binomial distribution, although the actual variance will be less than the binomial variance if item parameters are not equal across all items. Based on Equation 14, in IRT, any given number-correct score will have a corresponding ability level coinciding with it. Similarly, for a specific τ_{mi} there is a corresponding specific ability level θ_i, so,

$$\sigma_{e_{mi}|\tau_{mi}} = \sigma_{e_{mi}|\theta_i}. \quad (15)$$

Equation 15 can then be adapted into,

$$\sigma^2{}_{e_{mi}|\theta_i} = \sum_{k=1}^{j} P_k(\theta_i)Q_k(\theta_i), \quad (16)$$

and it follows that $\sigma^2{}_{e_{mi}|\theta_i}$ approaches zero as $P_k(\theta_i) \rightarrow 0$ or $P_k(\theta_i) \rightarrow 1$ supporting the idea that the standard error of measurement for individuals of differing abilities are not the same across ability levels. Each ability level in IRT scaling will have a specific standard error of measurement common across all individuals at that same ability level, but standard errors of measurement between ability levels will differ depending on how extreme the ability level is. This makes sense when one considers that tests generally differentiate the average test takers from one another better than individuals at the high or low end of scales (Thorndike, 1982). Based on these definitions of standard errors of measurement CTT's estimates of ability will involve equal intervals between number-correct scores and the ability levels used in IRT will instead result in non-equivalent intervals between ability estimates. This explains why IRT estimates of ability follow an ogive pattern where individuals of high or low ability are much more dispersed on the scale than individuals around the average ability level.

Comparisons of Classical Test Theory and Item Response Theory

Due to the relationships between CTT and IRT, one can generalize concepts and adapt calculations and indices between the two (Bechger, Maris, Verstralen, & Béguin, 2003). Many studies have investigated the similarities and differences between estimates from these two theories. Overall, these results tend to support the idea that both CTT and IRT perform very similar to one another.

Lawson (1991) reported remarkable similarities between person and item parameter estimates using CTT sum scores and IRT estimates, noting very strong correlations between the two. However, despite the similarities, Lawson suggests that differences do occur towards the extremes of the ability distributions. Additionally,

Lawson suggests that IRT may be more useful in obtaining reliable scores in individualized testing, while CTT tends to focus on yielding reliable information for groups of examinees instead. Lawson suggests that IRT estimates might not be worth the effort of calculating due to the similarities between the two scales.

MacDonald and Paunonen (2002) used Monte Carlo techniques to compare item and person statistics based on both CTT and IRT. In all conditions, including different test lengths, item difficulties, and item discriminations, both CTT and IRT were very similar as well as accurate. Both CTT and IRT estimates were invariant across samples, interestingly with CTT estimates appearing more invariant across samples. In regards to item discrimination, IRT outperformed CTT estimates, and the CTT estimates were only accurate under a few conditions. MacDonald and Paunonen suggest using IRT parameters for item selection for their accuracy in both item difficulty and item discrimination parameters.

Sharkness and DeAngelo (2010) built and tested two scales using both CTT and IRT estimates to inform development. Both CTT and IRT estimates provided similar results about which items best measured the desired information, as well as which items should be removed due to either poor correlations with other items or very large discrimination parameters. Sharkness and DeAngelo suggest that the IRT estimates, however, provide evidence of which individuals the scale has more precision for, due to the fact the standard error of measurement in CTT applies to all individuals, and IRT has different standard errors of measurement for individuals.

Güler, Uyanık, and Teker (2014) examined the similarities between CTT and IRT parameters for one-, two-, and three-parameter models. Results from CTT and IRT

difficulty and discrimination parameters were correlated with one another, resulting in high correlations between both one- and two-parameter IRT estimates and CTT estimates. Correlations between the three-parameter IRT estimates and CTT estimates resulted in lower correlations between item parameters, and the three-parameter model was found to fit the data better overall. Ultimately, Güler, Uyanık, and Teker suggest there may not be much difference between using one- or two-parameter IRT models and CTT scoring, but when one suspects there is chance or guessing involved the three-parameter model should be chosen.

Scaling Impact on Measuring Change

Embretson (2007) performed a study to investigate and attempt to alleviate the aforementioned issues scaling introduces to growth data. For the first part of the investigation, three groups were generated according to a latent variable model of relationships over time. True trait scores were generated for five time points, with the true trait means increasing at a steady rate across groups, a strong relationship between adjacent time points, and constant variances at each time point. The three groups were generated with individual ability level means of -1 for the low ability level group, 0 for the moderate ability level group, and 1 for the high ability group at each group's initial time point, and variances of 1.0 at each time point. Each group increased by 0.5 points from time point to time point, passing through middle time points of 0, 1, and 2, and finally reaching averages of 1, 2, and 3 at the final time point. The data were generated this way to ensure a strong linear relationship within the data regardless of an individual's group membership. Three 30 item exams were generated according to three different means, each with a standard deviation in difficulty of 1. Each of these exams

corresponded to one of the three group's average ability level at the middle time points, 0 for the low difficulty test, 1 for the moderate difficulty test, and 2 for the high difficulty test. Embretson chose these values due to the fact that exams should be appropriate for the entire range of time points, and in this case the exams would each be appropriate for a different group at their middle time point. Since the exams and individuals were generated on the same scale according to a one-parameter item response model, individuals were simulated to take each exam by using the equation for a 1PL model with the individual abilities and item difficulties to determine the probability an individual would get each item correct. These probabilities were compared to a uniform random number, and if the individual probability was higher than the random number the individual was treated as getting the item correct. After obtaining item response vectors, the vectors were summed across for each individual at each exam to calculate an individual's number correct score as the CTT measure for the analysis.

Using both the true trait scores and the CTT measures for each individual, Embretson performed two separate Group by Time repeated measures analysis of variance to investigate trend effects in the generated data. Effect sizes for different effects were reported due to the fact that the large sample size of n = 1,000 for each group lead to even very small effects being significant. Embretson reported that while the CTT measures estimated the Group effect with relative accuracy compared to the true trait Group effect, the CTT measures were not as accurate for the within-groups effects. CTT measures underestimated the Time effect compared to the true trait Time effect. These differences were largest when a test did not match the ability level of the individuals within it, for instance when a low ability level group took a high difficulty exam. While

the true trait score Time by Group interaction was essentially 0 due to each group following the same trend of development, the CTT measures overestimated the interaction and this suggested different interpretations of group trends across time. For example, if one examined the groups with a low difficulty test the individuals in the low ability group appeared to increase in ability with a linear trend, a correct interpretation of what is truly happening with the true trait scores, however, the high ability individuals appear to increase with a quadratic trend, slowing in increase of ability at the later time points. Therefore, using these CTT measures leads to an invalid interpretation of the data.

In addition to investigating the trends with a repeated measures analysis of variance, Embretson also investigated gain scores at different time intervals to simulate conditions in which only two observations are available for analysis. Using the same data as previously, simple gain scores between the intervals of time points 1 and 3, time points 2 and 4, and time points 3 and 5 were calculated for each group across each test with the true trait scores as well as the CTT measures. The true trait scores exhibited a constant gain score of 1.0 regardless of the group or test. The CTT measures, however, displayed different patterns of gain scores depending on the difficulty of the test. The Easy and Hard tests displayed marked differences in gain scores depending on the group taking the exam, and although gain scores were more similar for the three groups when considering the moderate exam, the low and high ability groups still showed less of a gain than the moderate group. Once again, interpreting these CTT measure gain scores would lead to invalid inferences about the true trend of development for each group.

In order to attempt to alleviate the problems observed with the CTT measures, Embretson considered IRT scaling as an alternative to CTT scaling. IRT scaling stretches

out extremes of distributions, and as such small differences in individual ability will be scaled differently compared to CTT measures. The item difficulties and item response vectors generated previously were used in a Rasch IRT model to estimate ability levels according to the *expected a posterior* (EAP) method using a prior distribution with a mean of 0 and standard deviation of 1. As with the CTT measures, a Group by Time repeated measures analysis of variance was employed to investigate the trend of the IRT parameters. Though the IRT parameters did not estimate the Time by Group effects as zero, they were considerably better than those of the CTT measures, coming much closer to the true trait estimates than the CTT measures did. Additionally, the IRT parameters were used to generate gain scores between the same time periods as previously used for the CTT measures, resulting once again in improved effect sizes. The interpretation of the gain scores still depended on the difficulty of the test for which the gain scores were computed. Embretson suggests that although IRT parameters do help alleviate some of the problems in scaling compared to CTT measures, the mismatch between group ability levels and test difficulties remains a problem.

The Current Study

This study aims to replicate and extend Embretson's (2007) findings. While Embretson's study utilized one dataset with a linear trend of growth, this study will perform 1,000 replications for each different curvilinear trend of development: linear, quadratic, and cubic. By increasing the number of replications performed in the study it is possible to examine what factors influence analyses on longitudinal test scores, such as bias in estimates introduced by the CTT and IRT scoring processes compared to the true generated scores and the Type I and Type II error rates in identifying trends of change.

Additionally, various studies of both psychological development and skill development in young children and adolescents have shown that in addition to linear trends of development, quadratic and cubic trends may be observed in longitudinal research (Eisenberg, Carlo, Murphy, & Court, 2014; Ryoo et al., 2014). Therefore, the introduction of curvilinear developmental trends allow us to extend our knowledge to better understand how the different scaling of scores by CTT sum scores and IRT estimates affect our perception of developmental trends in situations beyond just a linear trend of development. Although IRT estimates may not always be as accepted as a scoring process in many testing situations, this study aims to show that IRT estimates can be used to better understand and recognize the developmental trends of individuals across time.

The research hypotheses for this study are as follows:

(1) IRT estimates will better represent the true linear developmental trend compared to CTT sum scores across the replications, as demonstrated in Embretson (2007).

(2) In quadratic and cubic trends of development, IRT estimates will also exhibit a more accurate depiction of the true trends of development than CTT sum scores, like in the linear condition.

(3) IRT estimates will lead to an increased number of correct identifications of trend of development compared to CTT sum scores, that is, the Type I and Type II error rates will be smaller for IRT estimates than CTT sum scores.

(4) IRT visual trends and estimates of trend will be closer to the true trend of development in the population than CTT sum scores.

In order to test these hypotheses an adequate number of replications must be used for analyzing Type I and Type II errors and bias, for this reason a simulation study was conducted.

Method

True Scores: The Criterion

To investigate different trends of development, three different latent variable models were created. A linear model, quadratic model, and cubic model of development were considered in this study. Following Embretson's (2007) design, each of the different models feature three groups differing only in average ability level, exhibiting strong relationships between adjacent time periods, constant variances across time, and an increase in ability of 2 points by the final time period. Each model consists of three groups: individuals with high average ability level (High), individuals with a moderate average ability level (Moderate), and individuals with a low average ability level (Low).

The intercept for each model depends on the group an individual is in: the Low group had an intercept at -1, the Moderate group had an intercept of 0, and the High group had an intercept of 1. The proportion of variance predicted from each successive time period to the prior time period was 0.81; this is represented in each model by a random intercept with a variance of 0.81. For each of the different models, the applicable linear, quadratic and cubic slopes had variances fixed at 0 to ensure the slope affected all individuals equally. Finally, all time periods had a time-specific change to each individual represented by a variance of 0.19 leading to a total variance of 1.0 at each time period based on the 0.81 variance of the intercepts and the 0.19 variance at the individual time periods. The variance of each slope was fixed at 0, meaning that this study only

utilized a random intercept and all individuals should increase at the same rate aside from their time specific influences and group intercept. True trait ability levels were generated according to the following models using the simsem: SIMulated Structural Equation Modeling package (Pornprasertmanit, Miller, & Schoemann, 2012) in R (R Core Team, 2015).

Linear Development Model

Figure 1 presents the linear model across five time periods. The mean structure (μ) loads on the intercept (I) with a value of -1 for the Low group, 0 for the Moderate group, and 1 for the High group, giving each of the groups their respective means at the first time period (θ t1). Additionally, the mean structure loads on the linear slope (L) at a value of 0.5, and paired with the variance of the linear slope being zero this causes the linear slope to affect all time periods (θ t1 through θ t5) the same. The variance in individual scores within a group comes from the intercept's variance of 0.81 and the time-specific variances of 0.19. Due to this structure, individuals would be expected to have highly correlated scores from time period to time period and all grow in ability in a predictable fashion.

Since the linear slope loads on each time period linearly, individuals would be expected to have an ability measure around their group's intercept at time period 1, and increase by around 0.5 for each following time period. Thus, the expected group means across time periods are as follows:

$$\text{Low Group } E(\bar{\theta}_{low}) = [\ -1, -.5, 0, .5, 1]$$

$$\text{Moderate Group } E(\bar{\theta}_{mod}) = [\ 0, .5, 1, 1.5, 2]$$

$$\text{High Group } E(\bar{\theta}_{high}) = [\ 1, 1.5, 2, 2.5, 3]$$

with a standard deviation of *SD* = 1 for each group at each time period.

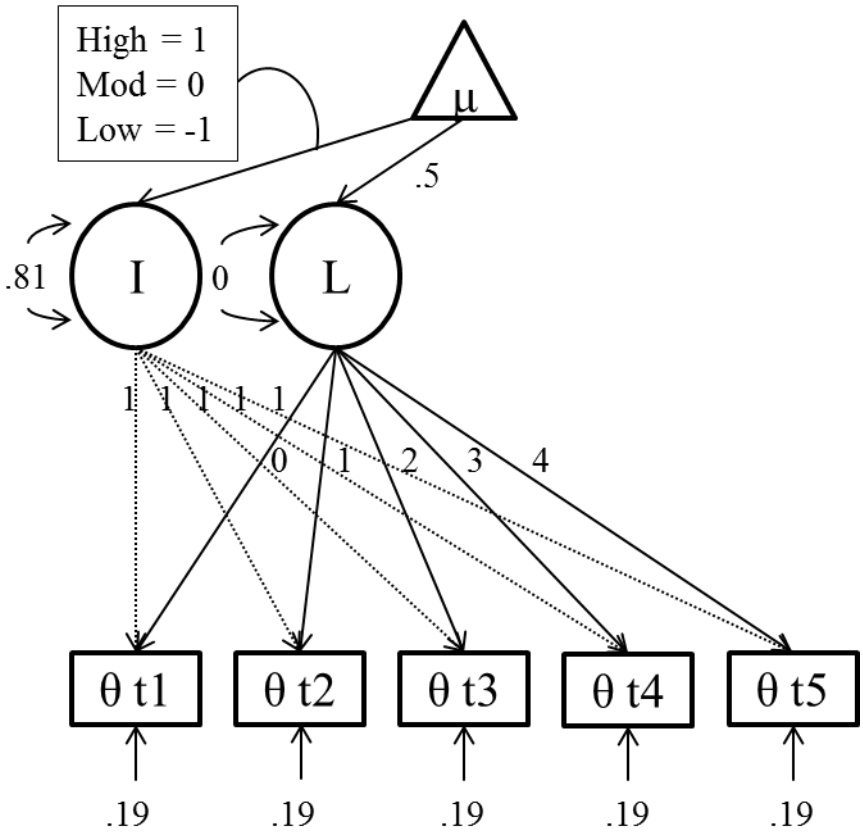

FIGURE 1. True trait linear model for simulation study.

Quadratic Development Model

Figure 2 presents the quadratic model across five time periods. As seen in Figure 2, a quadratic slope (Q) has been added to the previous model and the mean structure has changed accordingly. The quadratic slope loads onto each of the time periods increasing quadratically, allowing for the quadratic trend to be created. Rather than increasing by a set amount at each time period, the quadratic model instead increases rapidly at the beginning and flattens out towards the final two time periods, but still ends with each group having their average ability increased 2 points between the first and final time

period. The linear slope having a mean of 1.04 and the quadratic slope having a mean of -0.135 allow for this trend.

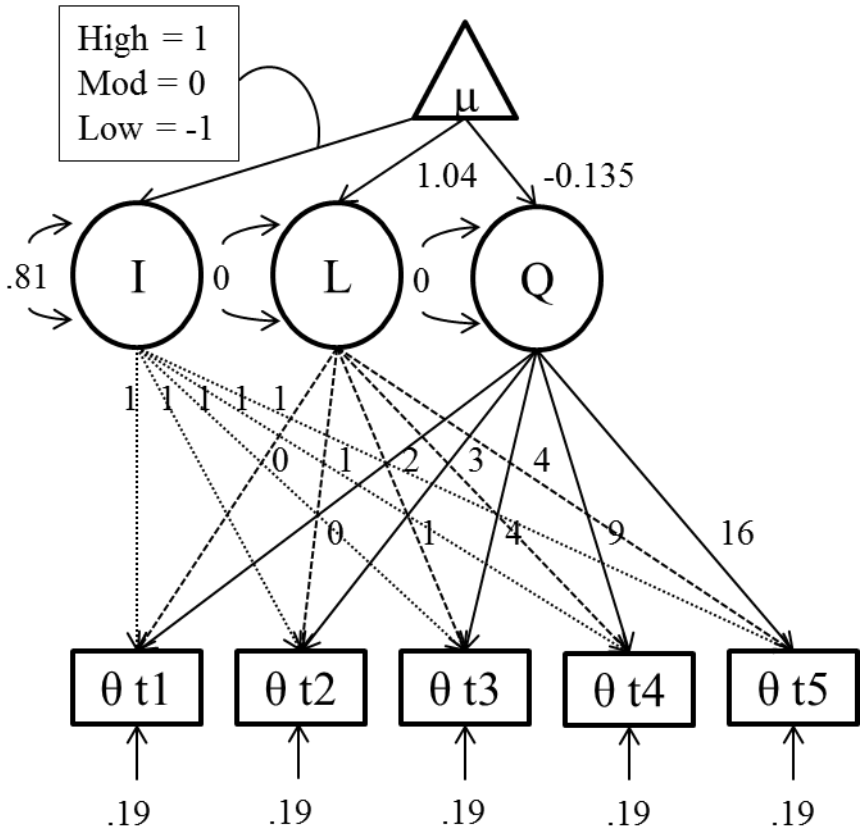

FIGURE 2. True trait quadratic model for simulation study.

The resulting expected means for the groups at each time period are as follows:

$$\text{Low Group } E(\bar{\theta}_{low}) = [\ -1, -.095, 0.540, .905, 1]$$

$$\text{Moderate Group } E(\bar{\theta}_{mod}) = [\ 0, .905, 1.540, 1.905, 2]$$

$$\text{High Group } E(\bar{\theta}_{high}) = [\ 1, 1.905, 2.540, 2.905, 3]$$

with a standard deviation of *SD* = 1 for each group at each time period.

Cubic Development Model

Figure 3 presents the cubic model across five time periods. Finally, in the cubic model presented in Figure 3 a cubic slope (C) is added to the model to create a cubic

trend of development. This cubic slope loads onto the time periods with a cubic function in order to achieve the desired effect. A mean of 1.567 for the linear slope, -0.8 for the quadratic slope, and .133 for the new cubic slope lead to a cubic trend of ability which increases between the first two time periods, remains flat until the fourth time period, and then increases again until the final time period.

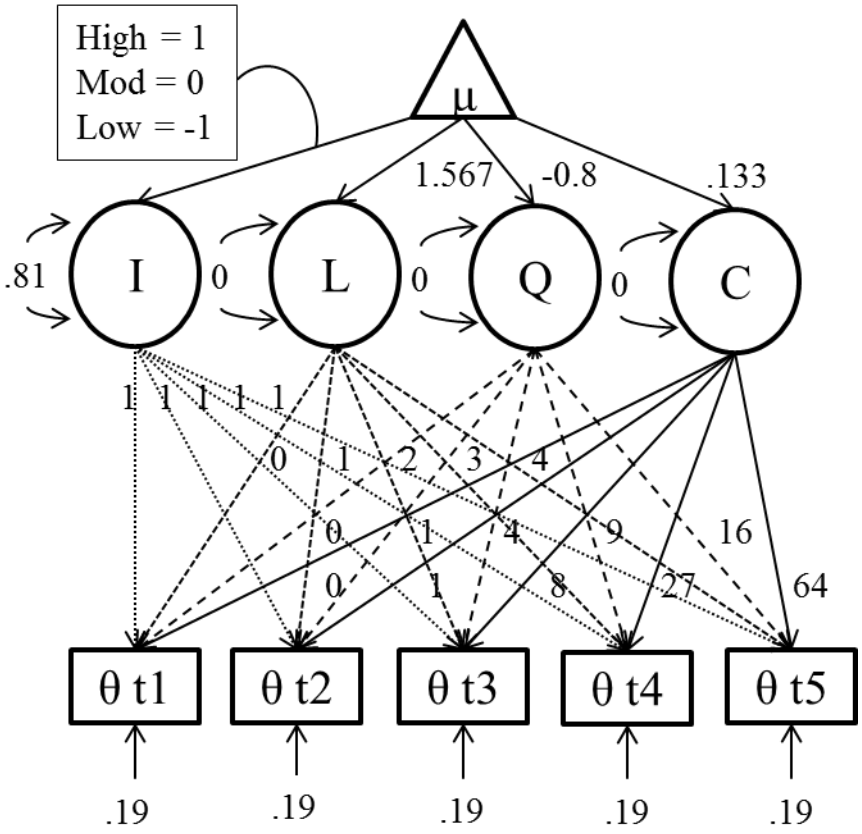

FIGURE 3. True trait cubic model for simulation study.

As with the linear and quadratic models, each group's average ability increased 2 points between the first and final time period. The expected means for each group at each time period are as follows:

Low Group $E(\bar{\theta}_{low})$ = [-1, -0.1, 0, 0.1, 1]

Moderate Group $E(\bar{\theta}_{mod})$ = [0, 0.9, 1, 1.1, 2]

High Group $E(\bar{\theta}_{high})$ = [1, 1.9, 2, 2.1, 3]

with a standard deviation of $SD = 1$ for each group at each time period.

Simulated Tests

In order to inspect the effects of a mismatch between individual ability and test difficulty, tests were generated with items of varying difficulties. At each time period individuals would be simulated to have taken three tests: a test with low difficulty (Easy), a test with medium difficulty (Medium), and a test with a higher difficulty (Hard). Each test contained a random sample of 30 items generated based on a one-parameter logistic item response model. In the one-parameter model, items only differ based on their difficulty parameter. At each time period a new batch of items was to be generated for a test, and items on a test would have an average difficulty relating to the test which they were part of, a standard deviation of 1, and a normal distribution of item difficulties.

Generated item difficulties correspond with the same scale as the individuals were generated on. The Easy test had a mean difficulty of 0, the Medium test had a mean difficulty of 1, and the Hard test had a mean difficulty of 2. These mean difficulties were chosen by Embretson (2007) to align with the middle time period, Time 3, for each of the respective groups of the same ability level. This practice is common in studies concerning fixed content tests, because the tests should be appropriate for the entire time period which leads to selecting a test that corresponds to the middle time period (Embretson, 2007). In the linear model, the Low ability group has an average ability of 0 at time period 3, corresponding with the mean difficulty of 0 on the Easy test. Similarly, the

Moderate group has an average ability of 1 at time period 3, and the High ability group has an average ability of 2 at time period 3, corresponding to the Medium test and the Hard Test mean difficulties respectively.

Since the previous study did not involve any trend beyond linear, consideration of these choices were put into the models used to generate the person ability data. Similar to the linear model, the cubic model average abilities of each group correspond to the mean difficulties of each test at time period 3. In the quadratic model, having the corresponding average group abilities and test difficulties match at time period 3 would not make sense due to the pattern of development, instead the quadratic model has average abilities of each group correspond with the test mean difficulties at time period 4.

Response Vectors and Scoring the Simulated Tests

Once individual abilities and item difficulties were generated, individuals could be simulated to take the three exams at each time period. In IRT the item difficulties and person abilities are on the same scale, making direct comparisons between them possible. To generate item response data, the one-parameter model in Equation 7 was applied to calculate the probability of each simulated individual answering the simulated items correctly. For each of the four different models 30 items were generated for each test difficulty, leading to 90 probabilities being calculated for each person at each of the five time periods. After the probabilities of getting an item correct were calculated, they were compared to a uniform random distribution of probabilities ranging from 0 to 1.0. If the calculated probability was higher than the uniform randomly generated probability the person was assumed to answer the item correctly. Each item across the three different tests and five different time periods were scored for each of the four models, and then

compiled into a vector of responses in the format of 0 for wrong responses and 1 for correct responses.

For the CTT scoring the response vectors were simply summed across to find the total number of correct responses for an individual. This led to each individual having a summed score at each of five time points for the three exams in each of the four models, analogous to the true score criterion abilities for each individual. Though CTT total scores will typically be transformed into a standardized score such as a T-score based on a norm group, for this study they were left in their unstandardized form because they display the same characteristics as the standardized scores (Embretson, 2007, p. 78). The sum scores ranged from 0 to 30 due to the length of the tests.

IRT person ability estimates were acquired using the ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses package (Rizopoulos, 2006). Item estimation was based on a modified 3 parameter logistic IRT model with item discrimination parameters fixed at 1 and the guessing parameters fixed at 0, essentially a Rasch model. Marginal Maximum Likelihood was used for estimation. Factor scores were generated using Empirical Bayes estimates. Item response vectors were scored for each test across the three groups of individuals at each time period creating five IRT estimates of ability for each individual, analogous to the true score criterion abilities and the CTT sum scores for each individual. Person ability levels were estimated based on the known values of item difficulty, meaning calibration was not performed as in Embretson (2007). IRT ability scores were scaled to have a mean of 0 and standard deviation of 1.

The Main Simulation Process

For the main simulation, 1,000 replications of three different models were generated: a linear, quadratic, and cubic model. Within each model, each group was generated with a sample size of n = 1,000, for a total sample size of N = 3,000 for each replication. Each model generated ability level estimates according to their specifications for the five different time periods, resulting in each individual having 5 true abilities, one for each time point. These data sets were used as the criterion true scores for the purpose of analyzing the data. After the criterion true scores were generated, Easy, Medium, and Hard tests were generated for each of the five time periods in each dataset. Each individual was simulated to take the exam using the one-parameter model to generate item responses. After these item responses were generated for each individual at each time, a CTT sum score was computed and an IRT score was estimated.

Analysis Plan

In order to investigate the development trends for each test, individual performances were averaged across group and then average group performance was again averaged across replication. To investigate Type I and Type II errors, for the True Score data as well as the six combinations of score scaling type (CTT or IRT) and test difficulty (Easy, Medium, Hard) a profile analysis for each replication was run using PROC GLM to generate output using SAS software (SAS Institute Inc., 2015).

For each condition, contrasts of one polynomial higher than generated were used to determine Type I error rates for classifying the true trend as showing a more complex trend than generated. A Type I error rate of 5% was considered acceptable. For Type II errors, contrasts of the polynomial matching the condition were performed and the

proportion of replications that correctly detected the linear condition were recorded. Type II error rates below 20% were considered acceptable. Finally, to investigate how closely both CTT and IRT scaling estimates matched the true trend, for each condition semipartial $\hat{\eta}^2$ statistics were calculated for each effect in the profile analysis, and these effect sizes were compared to those of the true trend for that condition. To investigate what factors influence the differences between effect size estimates in each condition, 2 x 3 ANOVAs were performed using scaling type and test difficulty as factors for each effect.

Results

Visual Trends of Development

The resulting average group scores were plotted across time for each condition and test difficulty combination to be analyzed visually. The axes on each plot correspond to the minimum and maximum range of scores or ability estimates obtained in each test in order to focus more closely on the differences in trends of development. In each plot, panel A contains the true score trend from which the individuals were generated. In each condition the true score trend follows the simulated trend perfectly. As generated, the three groups do not differ in their rate of development across the five time periods. The groups display noticeable differences in ability. In each plot, panel B contains the scaled scores for the Easy test, panel C contains the Medium test, and panel D the Hard test.

Linear condition

Figure 4 contains the trends observed for CTT scaling and Figure 5 contains the trends observed for IRT scaling. For CTT scaling in Figure 4 panel B, the Low ability group shows a linear trend of development. The Moderate and High ability groups, however, are not parallel to the Low ability group's trend and instead show a more quadratic trend. This same pattern appears for IRT scaling in Figure 5 panel B, however, the Moderate ability group appears more parallel to the Low ability group, and the High ability group appears slightly more linear. In both Figure 4 and 5, panel C shows a linear trend for the Moderate ability group, while both the High and Low ability groups bow outwards. For CTT scaling, this bow is slightly more pronounced than for IRT scaling.

Finally, panel D in both Figures 4 and 5 appear to mirror the trends seen in panel B, with the high ability group appearing more linear and the Low ability group showing more of a quadratic trend. Also of note is the fact that in both Figures 4 and 5, in panel B the difference in average ability at time point 1 appears larger than differences at time point 5. The High and Moderate ability groups appear much closer together at the final time point. The reverse pattern is seen in panel D, where the Low and Moderate ability groups appear closer together at time point 1. For CTT scaling, these differences are more apparent. Essentially, when test difficulty and group ability match the resulting trend appears much more similar to the true trend as seen in panel A.

Quadratic condition

Figure 6 contains the trends observed for CTT scaling and Figure 7 contains the trends observed for IRT scaling. As in the Linear condition, for both CTT and IRT scaling, when test difficulty and group ability match the trends tend to match the true

trends seen in panel A. When the mismatch between difficulty and ability is high, the trends seen appear more linear than they should. This is more apparent in panel D, in which the Low ability group appears very flat. IRT scaling resulted in slightly more parallel trends across groups than CTT scaling. Once again, in both Figure 6 and Figure 7 in panel B and D, the differences at time point 1 and 5 appear different. CTT scaling again results in differences that appear slightly larger compared to the IRT scaling.

Cubic condition

Figure 8 contains the trends observed for CTT scaling and Figure 9 contains the trends observed for IRT scaling. As in the previous condition, mismatch between difficulty and group ability result in slightly more linear trends of development. Group differences at time point 1 and 5 again appear different than the true differences between groups. The trends seen with IRT scaling appear slightly more equidistant and parallel than those for CTT scaling.

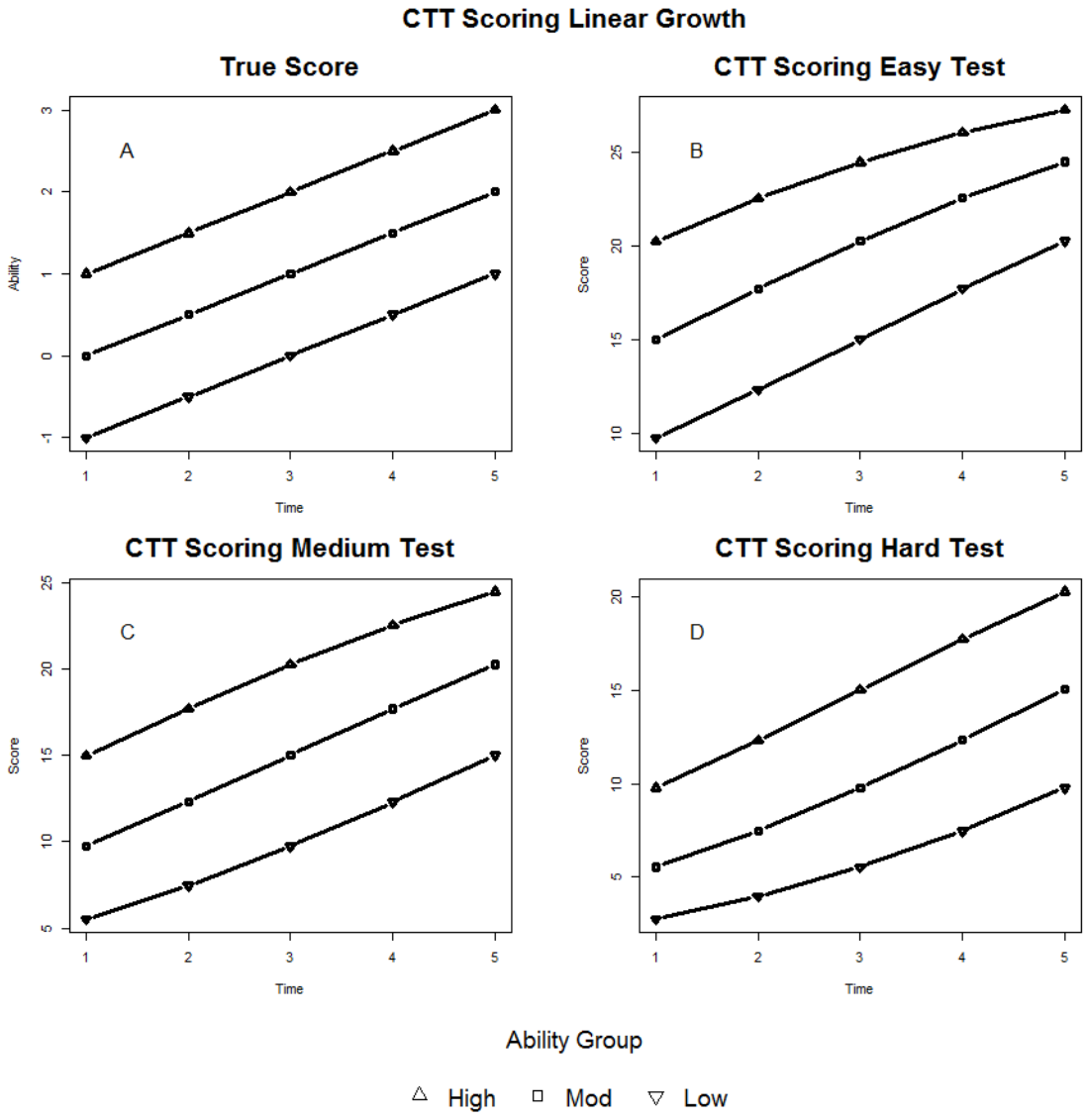

FIGURE 4. Trends of development for CTT scaling in the Linear condition.

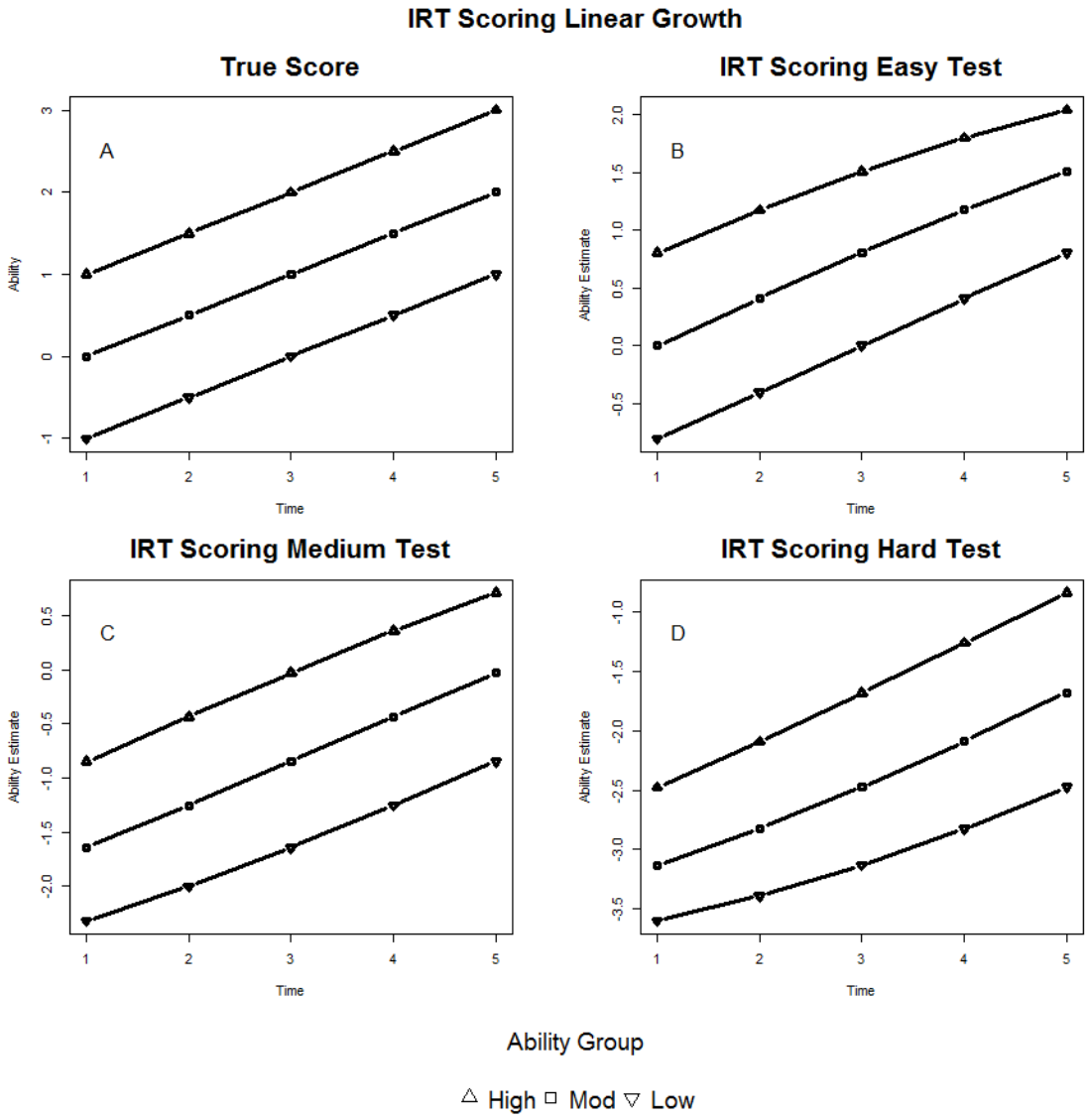

FIGURE 5. Trends of development for IRT scaling in the Linear condition.

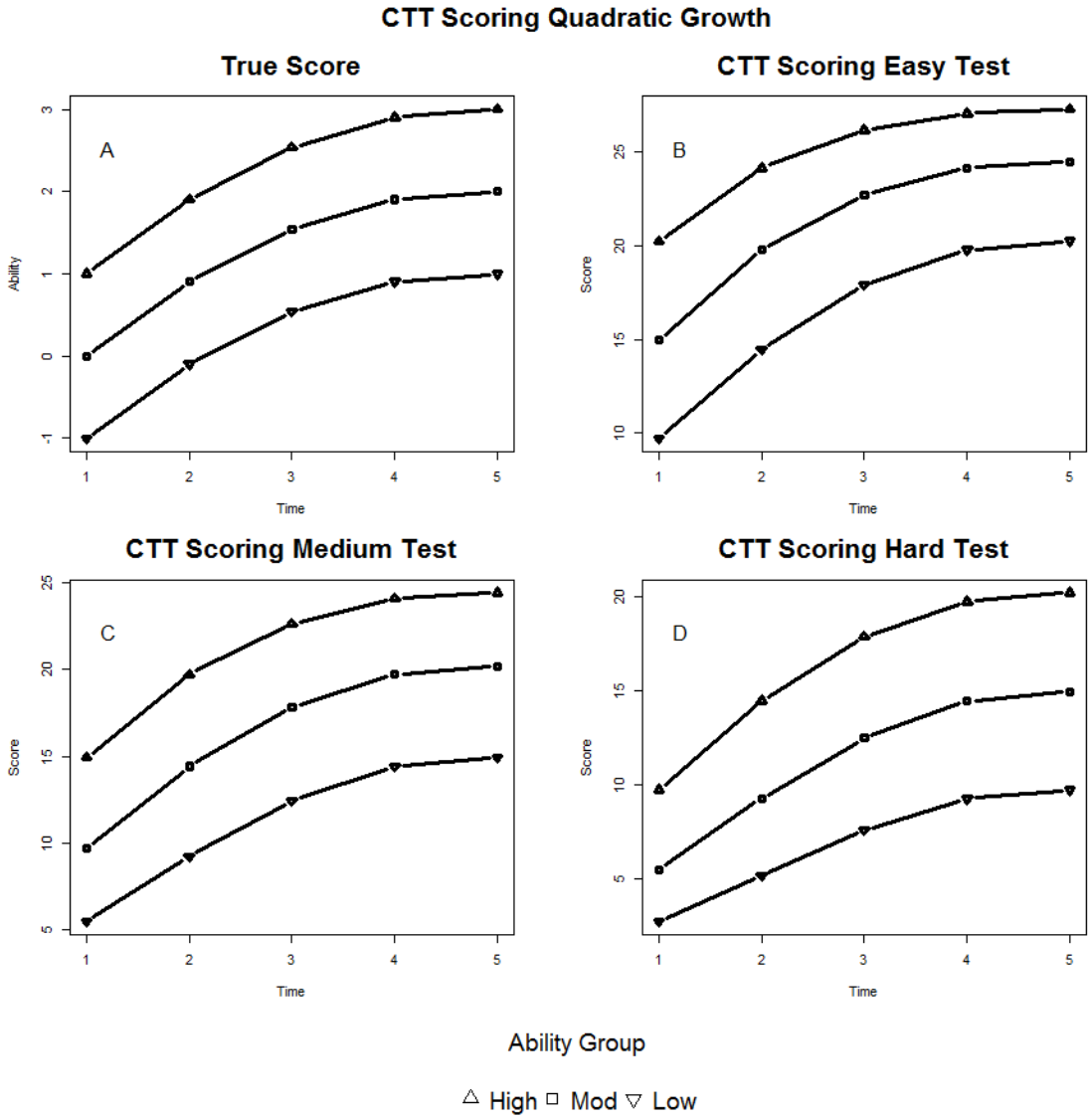

FIGURE 6. Trends of development for CTT scaling in the Quadratic condition.

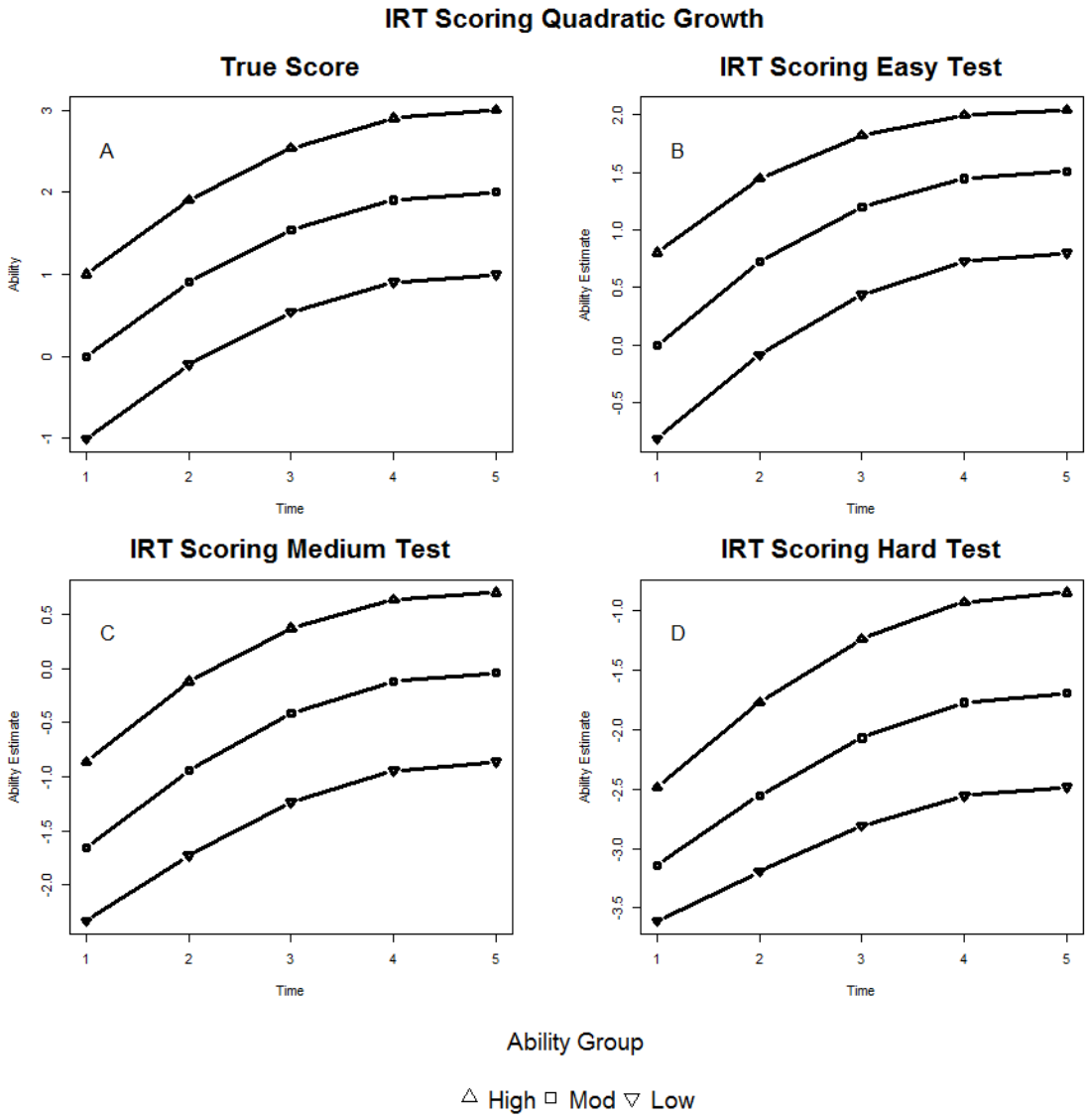

FIGURE 7. Trends of development for IRT scaling in the Quadratic condition.

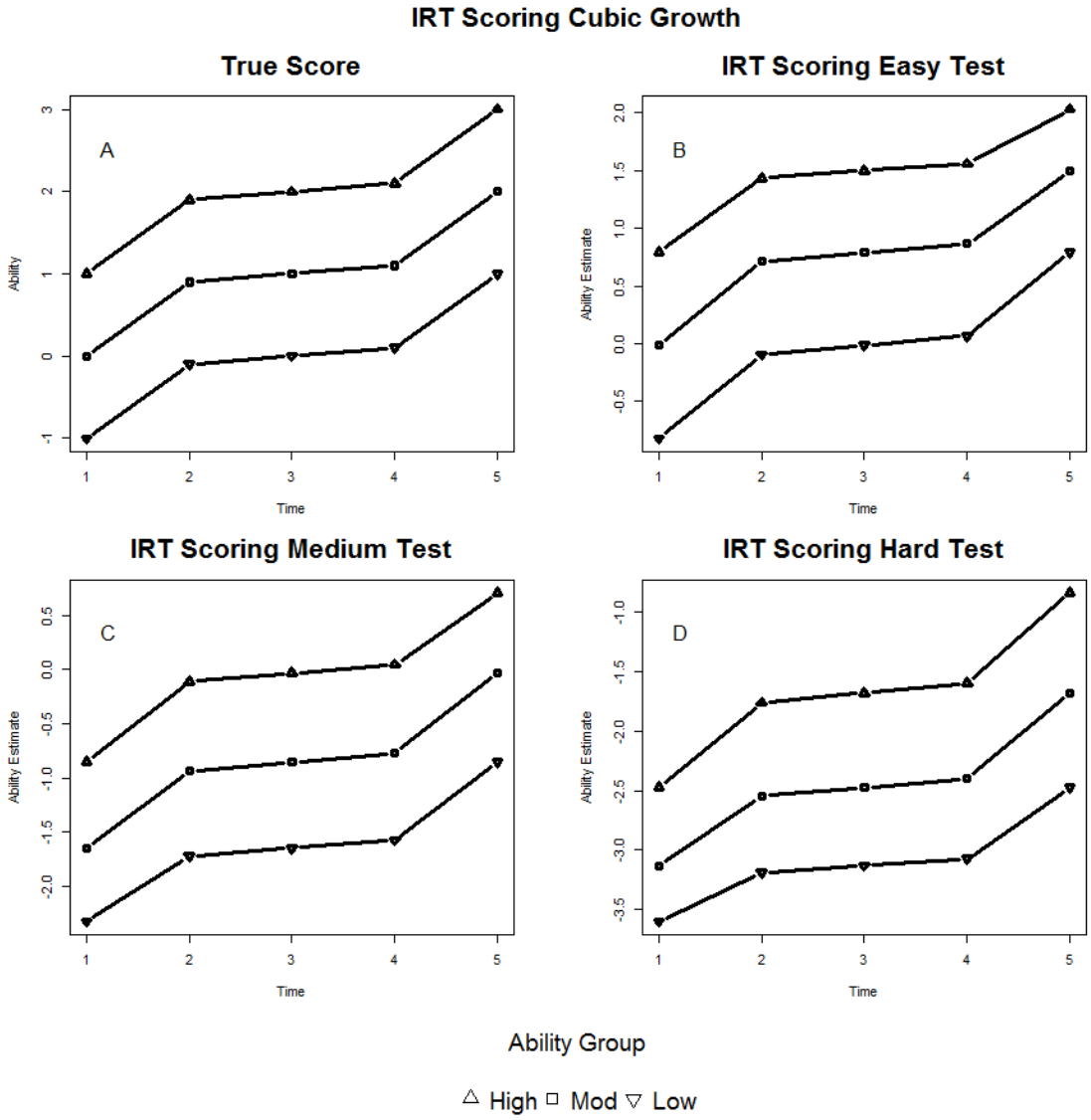

FIGURE 8. Trends of development for CTT scaling in the Cubic condition.

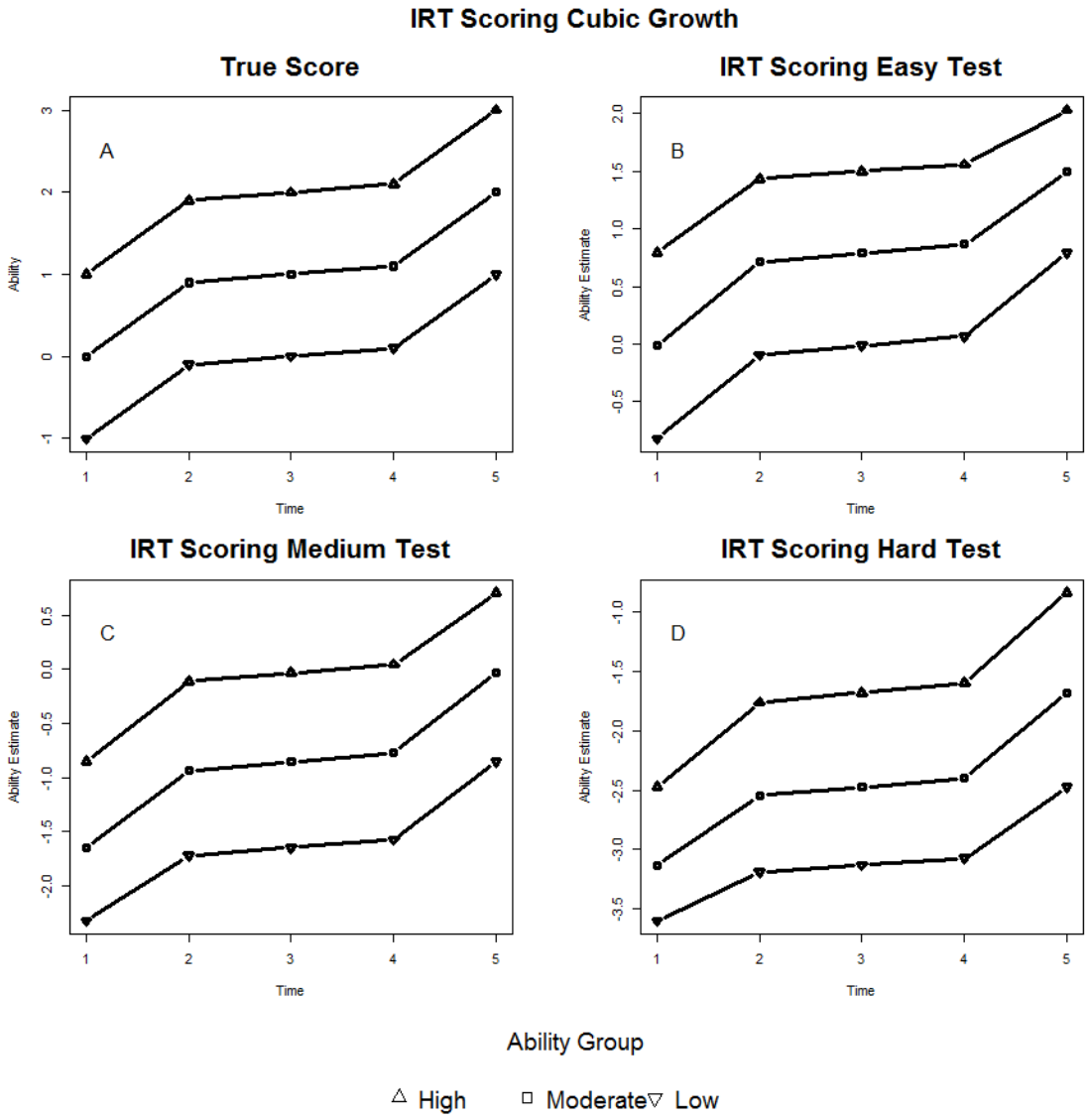

FIGURE 9. Trends of development for IRT scaling in the Cubic condition.

Type I and II Errors

Linear condition

In total there were five effects considered for the analyses, the effects of linear and quadratic Time, the effect of Group, and the two interactions between Group and linear and quadratic Time. The percentage of replications in which these effects were found significant can be found in Table 1.

Table 1

Percent of Significant Effects for the Linear Condition with Quadratic and Linear Contrasts

		Easy Test		Medium Test		Hard Test	
Effect	True	CTT	IRT	CTT	IRT	CTT	IRT
Time	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
$Time^2$	0.0	98.9	84.6	53.0	8.9	99.3	99.0
Group	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
Time*Group	0.0	24.2	36.0	95.7	76.5	100.0	100.0
$Time^2$*Group	0.0	82.5	40.4	99.8	54.8	83.9	21.3

Note. Effects generated with a non-zero value appear in boldface.

For Type I error rates, the effect of quadratic Time as well as the two interaction effects should be considered, as they were not generated to appear in the data. For the effect of quadratic Time, IRT scaling resulted in lower Type I error rates than CTT for all difficulties. The difference in errors was largest for the Medium test, and smallest for the Hard test. In the Easy test, CTT resulted in a lower Type I error rate for the interaction between linear Time and Group. CTT and IRT performed the same in the Hard test for the interaction between linear Time and Group. IRT outperformed CTT for every other interaction, and to a considerably higher degree for the interactions between quadratic Time and Group. In no tests did either CTT or IRT result in acceptable Type I error rates for any of the three effects.

For Type II error rates, the effect of linear Time as well as the Group effect should be considered. For all difficulties and scaling types, these two effects were found to be significant 100% of the time, with CTT and IRT demonstrating the same Type II error rates for detecting the trends.

Quadratic condition

In total there were seven effects considered for the analyses. In this condition the effect of cubic Time and the interaction between Group and cubic Time were introduced. The percentage of replications with significant effects can be found in Table 2.

Table 2

Percent of Significant Effects for the Quadratic Condition with Linear, Quadratic, and Cubic Contrasts

		Easy Test		Medium Test		Hard Test	
Effect	True	CTT	IRT	CTT	IRT	CTT	IRT
Time	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
$Time^2$	**100.0**	**100.0**	**100.0**	**99.4**	**97.6**	**15.7**	**17.9**
$Time^3$	0.0	63.8	3.3	1.5	2.5	73.0	66.9
Group	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
Time*Group	0.0	19.6	0.5	87.7	55.0	100.0	99.8
$Time^2$*Group	0.0	14.4	0.1	70.0	11.5	57.2	13.9
$Time^3$*Group	0.0	23.3	0.1	38.3	3.4	9.1	0.8

Note. Effects generated with a non-zero value appear in boldface.

For Type I error rates the effect of cubic Time as well as the three interaction effects should be considered, as they were not generated to appear in the data. For the effect of cubic Time, IRT outperformed CTT in the Easy and Hard test, while CTT marginally outperformed IRT in the Medium test. Both CTT and IRT had acceptable Type I error rates in the Medium test, and IRT had an acceptable rate for the Easy Test. For all three of the interaction effects IRT resulted in smaller Type I error rates. The

difference between CTT and IRT error rates was most pronounced for the Easy test, and least pronounced in the Hard test. IRT had acceptable Type I error rates for all of the interactions in the Easy test, and only the interaction between cubic Time and Group for the Medium and Hard tests. CTT did not have acceptable Type I error rates for any of the interactions.

For Type II error rates the effects of linear and quadratic Time and the effect of Group should be considered. Only for the Medium and Hard conditions non-significant effects were observed. for the effect of quadratic Time. In the Medium test CTT marginally outperformed IRT with higher rates of significance, while in the Hard test the opposite was true. In all conditions except for the Hard test, Type II error rate was more than acceptable. The Hard test resulted in severely increased Type II error rates at detecting the true effect for both scaling types.

Cubic condition

In total there were nine effects considered for the analyses. In this condition the effect of quartic Time and the interaction between Group and quartic Time were introduced. The percentage of replications with significant effects can be found in Table 3.

For the Type I errors the effect of quartic Time as well as the four interaction effects should be considered, as they were not generated to appear in the data. For the quartic effect of Time, CTT and IRT performed the same in the Medium test, and IRT marginally outperformed CTT in the Easy and Hard test. Type I error rates were at acceptable levels across all three tests for the effect of quartic Time. The four interactions behaved similarly to the interactions from the Quadratic condition, with IRT

outperforming CTT in all cases, and the rates increasing along with test difficulty. Only the interaction between linear Time and Group in the Medium Test and Hard test, and the interaction between quadratic Time and Group did not show acceptable Type I error rates.

Table 3

Percent of Significant Effects for the Cubic Condition with Linear, Quadratic, Cubic, and Quartic Contrasts

		Easy Test		Medium Test		Hard Test	
Effect	True	CTT	IRT	CTT	IRT	CTT	IRT
Time	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
Time^2	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
Time^3	**93.8**	**99.2**	**96.7**	**72.6**	**65.6**	**14.5**	**15.4**
Time^4	0.0	5.8	1.4	0.2	0.2	4.8	3.9
Group	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**
Time*Group	0.0	4.4	0.2	21.8	8.4	94.2	75.6
Time^2*Group	0.0	0.0	0.1	2.1	0.7	13.1	4.1
Time^3*Group	0.0	0.2	0.1	1.1	0.1	1.7	0.8
Time^4*Group	0.0	0.8	0.2	1.0	0.1	0.7	0.4

Note. Effects generated with a non-zero value appear in boldface.

For Type II error rates the effects of linear, quadratic, and cubic Time and the effect of Group should be considered. For all difficulties and scaling types lower numbers of significant effects were found for the effect of cubic Time. In both the Easy and Medium test, CTT detected more significant effects than IRT, and this difference was much larger for the Medium test. On the Hard test, IRT marginally outperformed CTT. Type II error rates were acceptable for both scaling types for the Easy Test, high for the Medium test, and as in the Quadratic condition, there were severe Type II error rates when detecting the true effects in the Hard test.

Effect Sizes

Average effect sizes for the effects from the profile analyses run for each test difficulty and scaling type were calculated and tabulated below for each of the three

conditions. Each condition uses the effects from the profile analysis with the polynomial contrast matching the trend the condition was generated to follow.

Linear condition

The three effects considered for these analyses were the effect of linear Time, the effect of Group, and the interaction between linear Time and Group. The average effect sizes for linear development in can be found in Table 4.

Table 4

Average Effect Sizes for the Linear Condition with Linear Contrasts

		Easy Test		Medium Test		Hard Test	
Effect	True	CTT	IRT	CTT	IRT	CTT	IRT
Time	.2310	.2024	.2060	.2092	.2102	.2025	.2035
Group	.1024	.1309	.1182	.0956	.0897	.0604	.0617
Time*Group	.0000	.0057	.0026	.0007	.0005	.0057	.0049

Note. Effect sizes reported are the semipartial $\hat{\eta}^2$.

For the effect of linear Time, both CTT and IRT underestimated the true effect size for all difficulties. IRT estimates were marginally closer to the true effect size than CTT. For the effect of Group, both CTT and IRT overestimated the effect in the Easy test, and underestimated the effect in the Medium and Hard test. For the Easy and Hard Tests, IRT was closer to the true effect size, while CTT was closer for the Medium test. For the interaction effect, CTT and IRT both overestimated the effect, which was generated to be zero. For each difficulty, IRT were closer to the true value.

Table 5 contains the effect sizes for the two additional effects obtained when the linear data were run with quadratic contrasts. The True Score effect sizes were zero, as generated, and both CTT and IRT scaling estimates were close to or at zero. IRT estimates were marginally closer to the true value.

Table 5

Average Effect Sizes for the Linear Condition with Quadratic Contrasts

		Easy Test		Medium Test		Hard Test	
Effect	True	CTT	IRT	CTT	IRT	CTT	IRT
$Time^2$	.0000	.0006	.0003	.0000	.0000	.0006	.0006
$Time^2$*Group	.0000	.0003	.0002	.0005	.0002	.0004	.0001

Note. Effect sizes reported are the semipartial $\hat{\eta}^2$.

Quadratic condition

In addition to the three effects in the linear condition, the effects of quadratic Time as well as the interaction between quadratic Time and Group were considered for these analyses. The average effect sizes for quadratic development can be found in Table 6.

Table 6

Average Effect Sizes for the Quadratic Condition with Quadratic and Linear Contrasts

		Easy Test		Medium Test		Hard Test	
Effect	True	CTT	IRT	CTT	IRT	CTT	IRT
Time	.0786	.0841	.0782	.0729	.0706	.0585	.0585
$Time^2$	.0231	.0281	.0249	.0216	.0204	.0150	.0150
Group	.0680	.0901	.0773	.0601	.0568	.0357	.0372
Time*Group	.0000	.0014	.0004	.0005	.0004	.0028	.0021
$Time^2$*Group	.0000	.0027	.0000	.0039	.0002	.0013	.0009

Note. Effect sizes reported are the semipartial $\hat{\eta}^2$.

For the two effects of Time, CTT outperformed IRT in the Medium test, IRT outperformed CTT in the Easy test, and the two performed similarly for the Hard test. For the effect of Group, both CTT and IRT overestimated the effect for the Easy test, and underestimated for the other two tests. CTT estimates were closer to the true value for the Medium test, while IRT estimates were closer for the Easy and Hard test. For both

interactions, the effect sizes were overestimated by both IRT and CTT, but IRT estimates were closer to zero.

Table 7 contains the effect sizes for the two additional effects obtained when the quadratic data were run with cubic contrasts. The True Score effect sizes were zero, as generated, and both CTT and IRT scaling estimates were close to or at zero.

Table 7

Average Effect Sizes for the Quadratic Condition with Cubic Contrasts

		Easy Test		Medium Test		Hard Test	
Effect	True	CTT	IRT	CTT	IRT	CTT	IRT
$Time^3$	.0000	.0002	.0000	.0000	.0000	.0002	.0002
$Time^3$*Group	.0000	.0001	.0001	.0002	.0000	.0001	.0000

Note. Effect sizes reported are the semipartial $\hat{\eta}^2$.

Cubic condition

In addition to the five effects in the quadratic condition, the effects of cubic Time as well as the interaction between cubic Time and Group were considered for these analyses. The average effect sizes for cubic development can be found in Table 8.

For the effect of linear Time, CTT estimates were closer to the true value in the Easy and Medium Test, while IRT estimates were closer in the Hard test. For the effect of quadratic Time, the two scaling types performed similarly in the Easy test, and IRT estimates were closer for the Medium and Hard test. The effect of cubic Time was more closely estimated by IRT in all three difficulties. For the effect of Group, IRT estimates were closer for the Easy and Hard test, while CTT performed better for the Medium test. IRT and CTT scaling performed similarly for the interaction between cubic Time and Group for the Medium test. For all other interactions, IRT outperformed CTT with

estimates of effect size closer to zero. The effect sizes for the two additional effects obtained when the cubic data were run with quartic contrasts were essentially zero.

Table 8

Average Effect Sizes for Cubic Development in a Cubic Profile Analysis

		Easy Test		Medium Test		Hard Test	
Effect	True	CTT	IRT	CTT	IRT	CTT	IRT
Time	.0372	.0360	.0354	.0325	.0324	.0276	.0281
$Time^2$	.0240	.0218	.0218	.0209	.0211	.0191	.0194
$Time^3$	.0247	.0210	0.215	.0215	.0219	.0210	.0213
Group	.0654	.0767	.0707	.0544	.0532	.0349	.0371
Time*Group	.0000	.0005	.0002	.0003	.0002	.0014	.0011
$Time^2$*Group	.0000	.0004	.0002	.0001	.0000	.0007	.0006
$Time^3$*Group	.0000	.0006	.0003	.0000	.0000	.0006	.0005

Note. Effect sizes reported are the semipartial $\hat{\eta}^2$.

ANOVAs Between Scaling Type and Test Difficulty

Linear condition

Three 2 x 3 ANOVAs were performed using scaling type and test difficulty as factors for the semipartial η^2s of each effect. In each of the three ANOVAs the two main effects as well as their interaction were found to be significant. The ANOVA source table for the average effect size for the effect of Time can be found in Table 9.

Table 9

ANOVA Between Scaling Type and Test Difficulty for the Effect Sizes of Linear Time in the Linear Condition

Source	*df*	*SS*	*MS*	*F*	η^2
Scaling Type	1	.0052	.0052	234.95*	.0220
Test Difficulty	2	.0511	.0255	1165.34*	.2161
S x D	2	.0021	.0011	47.98*	.0089
Error	5994	.1313	.0000		

Note. * $p < .0001$

For the effect sizes for the Time effect, test difficulty appears to be the most important factor in the differences between effect sizes for the effect of Time, explaining around 21% of the variance in effect sizes between conditions. The interaction between scaling type and difficulty explains very little variance in effect sizes, and scaling type explains a small amount of variance in effect sizes.

The ANOVA source table for the average effect size for the effect of Group can be found in Table 10.

Table 10

ANOVA Between Scaling Type and Test Difficulty for the Effect Sizes of Group in the Linear Condition

Source	*df*	*SS*	*MS*	*F*	η^2
Scaling Type	1	.0498	.0498	1009.45*	.0113
Test Difficulty	2	4.0260	2.0130	40796.34*	.9108
S x D	2	.0489	.0244	495.29*	.0111
Error	5994	.2958	.0000		

Note. * $p < .0001$

For the effect sizes of the Group effect, test difficulty had a very large η^2, explaining a large portion of the differences between effect sizes for the effect of Group. The interaction between scaling type and test difficulty explains nearly as much variance as scaling type itself, but both only explained around 1% of the variance.

The ANOVA source table for the average effect size for the effect of the interaction between Time and Group can be found in Table 11. For the effect sizes of the interaction between the effect of Time and Group, test difficulty explained a larger amount of variance than scaling type or the interaction between scaling type and test difficulty.

Table 11

ANOVA Between Scaling Type and Test Difficulty for the Effect Sizes of the Interaction Between Linear Time and Group in the Linear Condition

Source	*df*	*SS*	*MS*	*F*	η^2
Scaling Type	1	.0030	.0030	1474.04	.0725
Test Difficulty	2	.0242	.0121	6018.97	.5845
S x D	2	.0022	.0011	558.53	.0531
Error	5994	.0120	.0000		

Note. * p < .0001

Both scaling type and the interaction between the two main factors explained more variance for the interaction between Time and Group than for the main effects of Time and Group, at around 7% and 5% respectively.

Quadratic condition

Five 2 x 3 ANOVAs were performed using scaling type and test difficulty as the factors for each effect. The main effects as well as the interactions between the main effects were found to be significant in all five ANOVAs. The ANOVA source table for the average effect size for the effect of Time can be found in Table 12, and for the quadratic effect of Time in Table 13.

Table 12

ANOVA Between Scaling Type and Test Difficulty for Effect Sizes of Linear Time in the Quadratic Condition

Source	*df*	*SS*	*MS*	*F*	η^2
Scaling Type	1	.0112	.0112	1279.03*	.0190
Test Difficulty	2	.5178	.2589	29583.90*	.8769
S x D	2	.0090	.0045	515.85*	.0152
Error	5994	.0525	.0000		

Note. * *p* < .0001

For the effect sizes of the linear and quadratic effect of Time, test difficulty had a very large η^2, explaining the majority of the difference between effect sizes across conditions. In the Quadratic condition, the interaction between scaling type and difficulty explained a similar amount of variance to scaling type at around 2%, something not seen in the ANOVAs for the Linear condition.

Table 13

ANOVA Between Scaling Type and Test Difficulty for Effect Sizes of Quadratic Time in the Quadratic Condition

Source	*df*	*SS*	*MS*	*F*	η^2
Scaling Type	1	.0033	.0033	1586.31*	.0218
Test Difficulty	2	.1328	.0664	32350.90*	.8789
S x D	2	.0027	.0013	654.71*	.0179
Error	5994	.0123	.0000		

Note. * $p < .0001$

The ANOVA source table for the average effect size for the effect of Group can be found in Table 14.

Table 14

ANOVA Between Scaling Type and Test Difficulty for the Effect Sizes of Group in the Quadratic Condition

Source	*df*	*SS*	*MS*	*F*	η^2
Scaling Type	1	.0358	.0358	1429.1*	.0145
Test Difficulty	2	2.2289	1.1145	44532.5*	.9032
S x D	2	.0530	.0265	1058.7*	.0215
Error	5994	.1500	.0000		

Note. * $p < .0001$

Differences in the effect sizes of the Group effect were largely explained by test difficulty as it explained nearly 90% of the variance between conditions. The interaction

between scaling type and test difficulty explained a larger portion of variance than scaling type alone, but still only explained around 2% of the variance.

The ANOVA source table for the average effect size for the effect of the interaction between Time and Group can be found in Table 15, and for the effect of the interaction between quadratic Time and Group in Table 16.

Table 15

ANOVA Between Scaling Type and Test Difficulty for the Effect Sizes of the Interaction between Linear Time and Group in the Quadratic Condition

Source	*df*	*SS*	*MS*	*F*	η^2
Scaling Type	1	.0006	.0006	2367.30*	.0870
Test Difficulty	2	.0046	.0023	9370.96*	.6667
S x D	2	.0002	.0001	439.89*	.0290
Error	5994	.0015	.0000		

Note. * $p < .0001$

As in the Linear condition, for the effect sizes of the interaction between Time and Group as well as the interaction between quadratic Time and Group, test difficulty explained a majority of the variance. Scaling type explained a larger amount of variance than for the main effects of Time and Group, now explaining roughly 8% of the variance in effect sizes for the interactions.

Table 16

ANOVA Between Scaling Type and Test Difficulty for the Effect Sizes of the Interaction Between Quadratic Time and Group in the Quadratic Condition

Source	*df*	*SS*	*MS*	*F*	η^2
Scaling Type	1	.000105	.000105	2577.00*	.0796
Test Difficulty	2	.000950	.000475	11717.40*	.7197
S x D	2	.000023	.000011	277.4*	.0174
Error	5994	.000243	.000000		

Note. * $p < .0001$

Cubic condition

Seven 2 x 3 ANOVAs were performed using scaling type and test difficulty as the factors for each effect. All main effects, except for one, as well as the interactions between the main effects were found to be significant in all seven ANOVAs. Only in the ANOVA for the effect sizes of linear Time was scaling type found to be non-significant. The ANOVA source table for the average effect size for the effect of linear Time can be found in Table 17, for the effect of quadratic Time in Table 18, and the effect of cubic Time in Table 19.

Table 17

ANOVA Between Scaling Type and Test Difficulty for Effect Sizes of Linear Time in the Cubic Condition

Source	*df*	*SS*	*MS*	*F*	η^2
Scaling Type	1	.000005	.000005	1.926	.0000
Test Difficulty	2	.062758	.031379	13325.763*	.8131
S x D	2	.000304	.000152	64.646*	.0039
Error	5994	.014114	.000002		

Note. * $p < .0001$

Test difficulty explained around 80% of the variance between effect sizes of the linear effect of Time, 50% of the variance for the quadratic effect of Time, and only 5% of the variance for the cubic effect of Time. The interaction between scaling type and test difficulty, explained almost no variance between the conditions. Scaling type similarly explained next to none of the variance between conditions for the linear and quadratic effects of Time, however, for the cubic effect of Time it explained around 3% of the variance, putting it close to explaining as much as test difficulty, something unseen in the Linear and Quadratic conditions.

Table 18

ANOVA Between Scaling Type and Test Difficulty for Effect Sizes of Quadratic Time in the Cubic Condition

Source	*df*	*SS*	*MS*	*F*	η^2
Scaling Type	1	.000066	.000066	57.740*	.0048
Test Difficulty	2	.006747	.003373	2972.529*	.4949
S x D	2	.000018	.000009	7.934*	.0013
Error	5994	.006802	.000001		

Note. * $p < .0001$

Table 19

ANOVA Between Scaling Type and Test Difficulty for Effect Sizes of Cubic Time in the Cubic Condition

Source	*df*	*SS*	*MS*	*F*	η^2
Scaling Type	1	.000265	.000265	239.25*	.0364
Test Difficulty	2	.000375	.000188	169.42*	.0515
S x D	2	.000007	.000003	3.15°	.0010
Error	5994	.006638	.000000		

Note. * $p < .0001$, ° $p < .05$

The ANOVA source table for the average effect size for the effect of Group can be found in Table 20.

Table 20

ANOVA Between Scaling Type and Test Difficulty for Effect Sizes of Group in the Cubic Condition

Source	*df*	*SS*	*MS*	*F*	η^2
Scaling Type	1	.0044	.0044	223.0*	.0027
Test Difficulty	2	1.4255	.7128	36503.9*	.8779
S x D	2	.0169	.0085	433.2*	.0104
Error	5994	.1170	.0000		

Note. * $p < .0001$

As in the Linear and Quadratic conditions, test difficulty explained a large proportion of the variance in effect sizes for the effect of Group. The interaction between

scaling type and test difficulty as well as scaling type explained very little of the variance between the conditions, around 1% for the interaction and almost none for scaling type.

The ANOVA source table for the average effect size for the effect of the interaction between Time and Group can be found in Table 21, for the interaction between quadratic Time and Group in Table 22, and for the interaction between cubic Time and Group in Table 23.

Table 21

ANOVA Between Scaling Type and Test Difficulty for the Effect Sizes of the Interaction between Linear Time and Group in the Cubic Condition

Source	*df*	*SS*	*MS*	*F*	η^2
Scaling Type	1	.00007	.00007	1012.79*	0.0417
Test Difficulty	2	.00117	.00059	8450.68*	0.6964
S x D	2	.00002	.00001	143.75*	0.0119
Error	5994	.00042	.00000		

Note. * $p < .0001$

For the interactions between Time and Group, test difficulty remained the factor explaining much of the variance between conditions, however, this amount dropped from the effect sizes of linear Time to cubic Time from around 70% to around 36%. As in the Linear and Quadratic conditions, scaling type was a bit more important for the interactions, explaining from 2% to 3% of the variance in effect sizes between conditions. The interaction between scaling type and test difficulty explained from 1% to 3% of the variance in effect sizes between conditions.

Table 22

ANOVA Between Scaling Type and Test Difficulty for the Effect Sizes of the Interaction between Quadratic Time and Group in the Cubic Condition

Source	*df*	*SS*	*MS*	*F*	η^2
Scaling Type	1	.00002	.00002	632.58*	0.0274
Test Difficulty	2	.00031	.00015	4680.41*	0.4247
S x D	2	.00002	.00001	153.00*	0.0274
Error	5994	.00020	.00000		

Note. * $p < .0001$

Table 23

ANOVA Between Scaling Type and Test Difficulty for the Effect Sizes of the Interaction between Cubic Time and Group in the Cubic Condition

Source	*df*	*SS*	*MS*	*F*	η^2
Scaling Type	1	.00002	.00002	651.93*	.0303
Test Difficulty	2	.00024	.00012	3498.01*	.3636
S x D	2	.00002	.00001	293.87*	.0303
Error	5994	.00020	.00000		

Note. * $p < .0001$

Consolidated Results

Type I and II errors

Table 24 displays the number of times each scaling type had the same rates or better rates than the other scaling type across test difficulty and consolidated into effects relating to Time, Group, and interactions between the two.

Table 24

Number of Times Each Scaling Type had the Same or Better Type I and II Error Rates

	Easy Test		Medium Test		Hard Test	
Effects	CTT	IRT	CTT	IRT	CTT	IRT
Time Related	6	8	8	6	5	8
Group	3	3	3	3	3	3
Interactions	1	8	1	9	0	9

For Time related effects, IRT had better Type I and II error rates for the Easy and Hard tests, and CTT had better rates for the Medium test. The two strategies performed similarly to one another overall for the Time related effects. Both scaling types performed the same when it came to error rates for the effect of Group. For the interaction effects, IRT nearly always had better error rates when compared to CTT.

Effect sizes

Table 25 displays the number of times each scaling type had the same effect size estimate or better estimates than the other scaling type across test difficulty and consolidated into effect sizes relating to Time, Group, and interactions between the two.

Table 25

Number of Times Each Scaling Type Was the Same or Closer to the True Effect Size

	Easy Test		Medium Test		Hard Test	
Effects	CTT	IRT	CTT	IRT	CTT	IRT
Time Related	2	5	3	3	2	6
Group	0	3	3	0	0	3
Interactions	0	6	1	6	0	6

For Time related effect sizes, IRT estimates were closer to the true effect size more often in the Easy and Hard tests. The two scaling types performed similarly in the Medium test. For the Group effect sizes, IRT estimates were closer to the true value in the Easy and Hard tests, while CTT was better in the Medium test. As with error rates, IRT nearly always had better estimates of effect size for the interaction effect sizes.

Discussion

Hypothesis 1

IRT estimates were hypothesized to better represent the true linear developmental trend compared to CTT sum scores across replications, as demonstrated in Embretson (2007). The results from this study replicated those found previously. The visual inspections of the trends between the CTT and IRT scaling methods mirrored those found in Embretson's research. Visual inspection of the linear trends of both CTT and IRT revealed that group trends with IRT scaling were slightly more parallel to one another. The axes in Figure 4 for panels B, C and D clearly demonstrate how CTT estimates may differ based on different tests, as discussed by McKinley and Mills (1989). While the IRT estimates are all on different arbitrary scales and the scores should not be directly compared to one another, the estimates clearly depend on the difficulty of the test, as demonstrated in Figure 5 where the visual trends in panels B, C, and D are clearly different from one another. This supports Tinsley and Dawis (1977) who suggested that even though IRT estimates may be less biased by test difficulty, the precision of measurement will still depend on how appropriate items are for the individual. However, if the two estimates were put on the same metric, one would find the trends to be the same, as the one-parameter IRT model will place individuals with the same number correct score on the same ability level (De Ayala, 2009). Slight floor effects for the Low ability group in the Hard test, and ceiling effects for the High ability group in the Easy test for CTT scaling support Wang, Zhang, McArdle, and Salthouse's (2009) claim that ceiling effects can impact longitudinal research by causing researchers to select incorrect

models and biasing parameter estimation, as CTT estimates were not as accurate as those from IRT.

The effect sizes for both scaling methods resulted in biased estimates. IRT scaling resulted in effect size estimates closer to the true value than CTT scaling for all effects besides the group effect in the Medium test. Of note is that CTT and IRT performed more similarly in the Medium test, where less of a mismatch between test difficulty and group ability existed. IRT performing better when there was a larger amount of mismatch between test difficulty and group ability supports the claim by Lawson (1991) that IRT tends to differ from CTT towards extremes of the ability distribution.

Hypothesis 2

To test the second hypothesis in this study, Embretson's (2007) design was extended to include both quadratic and cubic trends of development. In the investigation of whether IRT estimates will also exhibit more accurate depictions of true trends of development than CTT sum scores for quadratic and cubic trends, the results indicate that IRT scaling continues to outperform CTT scaling regardless of the trend of development. Especially in conditions in which test difficulty and ability level deviated from one another by larger degrees, in the Easy and Hard conditions, IRT tended to outperform CTT by a higher degree. Again, for both the quadratic and cubic trends, slight floor effects for the Low ability group in the Hard test, and ceiling effects for the High ability group in the Easy test for CTT scaling supported Wang, Zhang, McArdle, and Salthouse (2009), as the estimates from CTT scaling were not as accurate as those from IRT scaling.

Based on the consolidated effect size results from the linear, quadratic, and cubic trends in Table 25, for the Easy and Hard tests IRT estimates of effect sizes were more

accurate than CTT estimates. CTT and IRT performed more similarly in the Medium test, with CTT estimates of the Group effect being better than those from IRT. Once again, Lawson's (1991) statement that IRT estimates function differently than CTT scores when mismatch between difficulty and ability level exist is supported. IRT estimates were nearly always better for the interaction effects, again in agreement with previous literature that IRT scaling can help with spurious interactions (Embretson, 1996; Kang & Waller, 2005; Morse et al., 2012).

Hypothesis 3

IRT estimates were hypothesized to lead to an increased number of correct identifications of trend of development compared to CTT sum scores. Type I error rates were smaller for IRT estimates of trend across a majority of the profile analyses run, as suggested in the literature (Embretson, 1996; Morse et al., 2012). However, CTT sum scores did have lower Type I error rates for the higher order trends of Time with the Medium test. This result seems reasonable based on the literature that spurious interactions are more exacerbated when difficulty and ability are mismatched (Kang & Waller, 2005).

For the Easy and Hard tests, IRT outperformed CTT in regards to Type I error rates for the interactions between Time and Group, especially for the higher level interaction effects which should have all been zero, that is, non-significant. This finding follows with previous research that IRT scaling can alleviate issues with spurious interactions (Embretson, 1996; Kang & Waller, 2005; Morse et al., 2012). Both scaling types, however, did tend to have increased Type I error rates, only tending to fall below the acceptable rate in the Cubic condition for the higher level interactions, which may be

due to the sample size and high parameterization of the model. While the interactions between the different trends of Time and test difficulty were often found to be statistically significant, the actual effect sizes for the interactions were small enough to suggest the interactions were not practically meaningful.

Due to the large sample sized used in the study, the analyses related to this hypothesis were overpowered. As such, results from the Type II error rates are less conclusive. Both CTT and IRT scaling correctly identified the linear trend of Time and the Group effect across all test difficulties for the Linear condition, showing no difference from one another. For both the Cubic and Quadratic condition, IRT and CTT correctly identified lower order trends of Time and the Group effect across all test difficulties. In the Quadratic condition, the true quadratic trend of Time was detected correctly more often for CTT than IRT in the Medium test, while IRT detected the trend correctly more often in the Hard test, both had acceptable Type II error rates in the Medium test, while the rates were severely high in the Hard test. In the Cubic condition a similar pattern was seen, but the true cubic trend of Time was detected more often when using CTT scaling for the Easy and Medium tests, and more often when using IRT once again in the Hard test. For this condition error rates were acceptable for the Easy difficulty test, while the Medium test displayed higher rates, and the Hard test had higher rates still. These results seem to indicate that depending on test difficulty and how complex of a trend is present in the data, different Type II error rates may be observed. The results do not seem to support the literature that IRT scaling will result in fewer Type II error rates (Embretson, 1996), as CTT outperformed IRT in two conditions, while IRT only outperformed CTT in two.

Hypothesis 4

IRT estimates were hypothesized to result in more visually apparent trends of development than CTT sum scores. While the development plots appear quite similar, IRT estimates do seem to show trends closer to those of the true trend. Again, this is more apparent with a greater mismatch between test difficulty and ability level. The differences seen between trends in the two scaling types support Schulz and Nicewander's (1997) statement that measurement scale may change how growth functions are seen across time. These visual inspections of trend seem to be of great use in determining the actual trend of development seen in data, as even though for example in the Linear condition quadratic trends were suggested by analyses, investigations of the visual trends may lead one to reject that idea as the trends still appeared to be quite linear. These findings support the fact that researchers and practitioners alike should always use every resource available to them when making decisions based on data. Ultimately, based on the trends seen for CTT and IRT, though IRT does appear marginally better the two scaling types agree with one another to a very high degree as discussed in the literature (Güler et al., 2014; Lawson, 1991; Macdonald & Paunonen, 2002; Sharkness & DeAngelo, 2010).

Implications

The common theme seen across the visual trends of development, Type I and II errors, as well as the investigations into the effect sizes of trends of Time, Group, and the interactions between them, is that test difficulty has the largest impact on the differences seen across conditions. This finding supports long standing beliefs that tests should be developed to best fit the group of individuals which the test is intended to target.

Although interactions between test difficulty and scaling type tended to be significant, the actual effect sizes of the interactions were quite small. Similarly, the effect sizes of the scaling type were small as well. This tells us that while there is something going on between scaling type and test difficulty it does not explain much of the differences seen in effect sizes and the two different scaling types perform quite similarly to one another.

Overall, despite the effect sizes of scaling type being quite small, IRT estimates tended to outperform CTT estimates when test difficulty and ability level were not aligned. Additionally, IRT estimates of the effect sizes of interactions tended to be smaller than those from CTT. While CTT did outperform IRT for lower order trends of Time as well as Group differences on the Medium test, the two scaling types performed more similarly in those cases.

The results from this study have two main implications. The first implication is that the metric of a scale does have an impact, IRT and CTT had differences, and if one wishes to have the most accurate estimates IRT tends to perform just as well as or better than CTT depending on test difficulty and ability level of test takers. The second implication is that both IRT and CTT scaling result in quite similar results, so similar decisions will ultimately be made after investigating visual trends, significance tests, and effect size.

Limitations and Future Research

In the current study, rather large differences between groups were utilized. If differences between groups were more similar, or one group with a normal distribution of scores was used, the same patterns might not have been observed. Group effects in the trend analyses may not have been detected with perfect accuracy, which would allow for

better investigation of differences between IRT and CTT. Future research should investigate how smaller differences between groups affect the outcomes of scaling type on trend development.

The current study utilized rather large samples, making it difficult to investigate the effects of IRT and CTT on Type II errors. Future research should investigate smaller numbers of participants, as IRT estimation does not perform as well without large enough samples, and this may very well impact its performance. The large sample size resulted in the profile analyses and ANOVAs being overpowered, which is why many of the Type II error rates were nearly zero and nearly all effects in the 2 x 3 ANOVAs performed on effect sizes were significant. Type I error rates may have been too conservative, as the 5% error rate was not observed in the analyses using the true scores. This is likely due to the strong relationships between time points with relatively little error.

Test lengths of 30 were utilized in the current study, which may mean that results from the study do not generalize to shorter or longer scales. Future research should investigate whether using more or fewer items impacts which scaling type results in more accurate estimates. IRT scaling may perform worse at smaller test lengths, especially if calibration is required. Finally, different numbers of time points may be of interest to future researchers, as the time spans and number and structure of time points vary greatly in longitudinal studies (Card & Little, 2007).

Finally, the current research utilized a Rasch model, meaning results may not generalize to other IRT models. Future research may also wish to investigate whether using a two- or three-parameter IRT model to score responses has an impact on the accuracy of IRT. In the current study both CTT and IRT performed rather similarly, as

suggested by previous research (Lawson, 1991; Macdonald & Paunonen, 2002; Sharkness & DeAngelo, 2010). Based on the results in Güler, Uyanık, and Teker (2014), two-parameter models will also correlate highly with to CTT scores, but three-parameter models may result in different results, provided issues in model fit are addressed.

References

Andrich, D. (1978). Relationships Between the Thurstone and Rasch Approaches to Item Scaling. *Applied Psychological Measurement*, *2*, 451–462.

Bechger, T. M., Maris, G., Verstralen, H. H. F. M., & Béguin, A. A. (2003). Using Classical Test Theory in Combination with Item Response Theory. *Applied Psychological Measurement*, *27*, 319–334.

Becker, D., & Forsyth, R. (1992). An empirical investigation of Thurstone and IRT methods of scaling achievement tests. *Journal of Educational Measurement*, *29*, 341–354.

Blanton, H., & Jaccard, J. (2006a). Arbitrary metrics in psychology. *American Psychologist*, *61*, 27–41.

Blanton, H., & Jaccard, J. (2006b). Arbitrary metrics redux. *American Psychologist*, *61*, 62–71.

Card, N. A., & Little, T. D. (2007). Longitudinal modeling of developmental processes. *International Journal of Behavioral Development*, *31*, 297–302.

Davison, M. L., & Sharma, A. R. (1990). Parametric statistics and levels of measurement: Factorial designs and multiple regression. *Psychological Bulletin*, *107*, 394–400.

De Ayala, R. J. (2009). *Theory and practice of item response theory*. New York, NY: Guilford.

Eisenberg, N., Carlo, G., Murphy, B., & Court, P. Van. (2014). Prosocial Development in Late Adolescence : A Longitudinal Study Prosocial Development A Longitudinal Study in Late Adolescence: *Child Development*, *66*, 1179–1197.

Embretson, S. (1996). Item Response Theory Models and Spurious Interaction Effects in Factorial ANOVA Designs. *Applied Psychological Measurement*, *20*, 201–212.

Embretson, S. (2006). The continued search for nonarbitrary metrics in psychology. *The American Psychologist*, *61*, 50–5; discussion 62–71.

Embretson, S. (2007). Impact of measurement scale in modeling development processes and ecological factors. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studites* (pp. 63–87). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Grimm, K. J., Kuhl, A. P., & Zhang, Z. (2013). Measurement Models, Estimation, and the Study of Change. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*, 504–517.

Güler, N., Uyanık, G. K., & Teker, G. T. (2014). Comparison of classical test theory and item response theory in terms of item parameters. *European Journal of Research on Education*, *2*, 1–6.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. (1978). Developments in Latent Trait Theory: Models, Technical Issues, and Applications. *Review of Educational Research*, *48*, 467–510.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. SAGE Publications.

Kang, S.-M., & Waller, N. G. (2005). Moderated Multiple Regression, Spurious Interaction Effects, and IRT. *Applied Psychological Measurement*, *29*, 87–105.

Lawson, S. (1991). One parameter latent trait measurement: Do the results justify the effort? In B. Thompson (Ed.), *Advances in Educational Research: Substantive Findings, Methodological Developments* (Vol. 1, pp. 159–168). Greenwich, CT: JAI.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*.

Macdonald, P., & Paunonen, S. V. (2002). A Monte Carlo Comparison of Item and Person Statistics Based on Item Response Theory versus Classical Test Theory. *Educational and Psychological Measurement*, *62*, 921–943.

McKinley, R. L., & Mills, C. N. (1989). Item response theory: Advances in achievement and attitude measurement. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 1, pp. 71–135). Greenwich, CT: JAI Press.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.

Morse, B. J., Johanson, G. a., & Griffeth, R. W. (2012). Using the Graded Response Model to Control Spurious Interactions in Moderated Multiple Regression. *Applied Psychological Measurement*, *36*, 122–146.

Pornprasertmanit, S., Miller, P., & Schoemann, A. (2012). simsem: SIMulated Structural Equation Modeling. Retrieved from http://cran.r-project.org/package=simsem

R Core Team. (2015). R: A language and environment for statistical computing. Vienna, Austria. Retrieved from http://www.r-project.org/

Rizopoulos, D. (2006). ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses. *Journal of Statistical Software, 17*(5), 1–25. Retrieved from http://www.jstatsoft.org/v17/i05/

Ryoo, J. H., Molfese, V. J., Heaton, R., Zhou, X., Brown, E. T., Prokasky, A., & Davis, E. (2014). Early Mathematics Skills From Prekindergarten to First Grade: Score Changes and Ability Group Differences in Kentucky, Nebraska, and Shanghai Samples. *Journal of Advanced Academics, 25*, 162–188.

SAS Institute Inc. (2015). The SAS System for Windows. Release 9.3. Cary, NC: SAS Institute Inc.

Schulz, E. M., & Nicewander, W. A. (1997). Grade equivalent and IRT representations of growth. *Journal of Educational Measurement, 34*, 315–331.

Sharkness, J., & DeAngelo, L. (2010). Measuring Student Involvement: A Comparison of Classical Test Theory and Item Response Theory in the Construction of Scales from Student Surveys. *Research in Higher Education, 52*, 480–507.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science (New York, N.Y.), 103*(2684), 677–680.

Thorndike, R. L. (1982). *Applied Psychometrics*. Boston, MA: Houghton Mifflin.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 16*, 433–451.

Tinsley, H. E. A., & Dawis, R. V. (1977). Test-Free Person Measurement with the Rasch Simple Logistic Model. *Applied Psychological Measurement, 1*, 483–487.

Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2009). Investigating Ceiling Effects in Longitudinal Data Analysis. *Multivariate Behavioral Research, 43,* 476–496.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: Mesa Press.