Assessment
2021, Vol. 28(2) 503–517

DOI: 10.1177/1073191120913607
journals.sagepub.com/home/asm


Precision and Sample Size Requirements for Regression-Based Norming Methods for Change Scores

Zhengguo Gu[1], Wilco H. M. Emons[1], and Klaas Sijtsma[1]

Abstract

To interpret a person's change score, one typically transforms the change score into, for example, a percentile, so that one knows a person's location in a distribution of change scores. Transformed scores are referred to as norms and the construction of norms is referred to as norming. Two often-used norming methods for change scores are the regression-based change approach and the *T* Scores for Change method. In this article, we discuss the similarities and differences between these norming methods, and use a simulation study to systematically examine the precision of the two methods and to establish the minimum sample size requirements for satisfactory precision.

Keywords

change assessment, norming, regression-based change approach, regression-based norming, *T* Scores for Change method

This article concerns the practical use of change scores (e.g., posttest score minus pretest score) obtained from psychological tests or questionnaires for drawing inferences about change at the level of the *individual*. For example, if a patient is in treatment, the clinician may want to evaluate how this patient responds to the treatment in terms of change of mental health, distress, quality of life, or general functioning. Important questions include the following: Has the patient improved at all? Is the patient's improvement on track? Is the observed change practically important? The answers to these questions may be used to tailor future treatment of the patient.

Change scores often are based on counts of correct answers or sums of scores on rating scales observed at two measurement occasions. Without a frame of reference, scores are not directly usable for practical measurement (Angoff, 1984). For example, knowing that John had 18 out of 28 arithmetic items correct or that Mary scored 37 scale points out of 60 on an introversion scale means little if anything as long as their test scores cannot be related to a distribution of test scores or a performance standard with a well-established meaning. The same goes for change scores. Without an interpretative context, it is hard to say whether observed change of an individual is small or large, consistent with natural recovery, or lagging behind compared with the change of other patients undergoing the same treatment.

To provide a frame of reference, one needs transformed scores, known as norms (Allen & Yen, 2002). A well-known type of transformed scores is norm-referenced norms (Allen & Yen, 2002), locating an individual amid a norm group. Another type of transformed scores, criterion-referenced norms (Allen & Yen, 2002), refers to diagnostic cutoffs that patients have to pass to be admitted to a course or a therapy, or to qualify levels of severity (e.g., mild vs. strong depression). In change assessment, criterion-referenced norms include the minimal clinical important difference, which is the minimum change a patient must show to qualify change as having practical impact. In this article, we focus on norm-referenced norms and study two often-used regression-based norming methods for change scores. Study of criterion-referenced norms deserves full attention in a separate article.

The construction of norms is known as norming. Norming methods for change scores have received surprisingly little attention so far. In this study, we consider the simplest change score possible, which is the difference between the test score obtained after a treatment and the test score obtained prior to the treatment, known as posttest and pretest scores, respectively. Norming change scores can be challenging for at least two reasons. The first challenge is that, pretest scores, posttest scores, and consequently change scores contain measurement errors and therefore may be unreliable (Bereiter, 1963; Cronbach &

[1]Tilburg University, Tilburg, Netherlands

Corresponding Author:
Zhengguo Gu, Department of Methodology and Statistics, TSB, Tilburg University, PO Box 90153, Tilburg 5000LE, Netherlands.
Email: z.gu@tilburguniversity.edu

Furby, 1970; Linn & Slinde, 1977; O'Connor, 1972). In practice, interindividual change typically has small variance (Gu et al., 2018), which may also cause low change-score reliability. The second challenge, in particular in the context of mental health care, is the heterogeneity of a group of patients. Often one cannot simply compare the change of one patient to the change of all other patients. To monitor a patient's change, ideally the patient should be compared to patients of the same age, the same gender, the same comorbidity, or other relevant background variables. This means that researchers may need to take such information into account when constructing norms. Ideally, data are collected within all relevant subpopulations based on the relevant background variables, but because this approach may require a huge sample, it may be practically infeasible. A solution to the problem is continuous norming (Gorsuch, 1983), which results in norms for subgroups using statistical control, based on a relatively small sample.

The goals of this study were to gain more insight in the precision of norms for change scores and to derive sample size requirements for deriving reliable norms for change scores. The structure of this article is the following. First, we present a general, regression-based framework for norming change scores, which includes two popular norming methods for change scores as special cases. The two methods are the regression-based change approach (Van der Elst et al., 2008) and the *T* Scores for Change method (McSweeny et al., 1993). Second, using a simulation, we study the precision of norms developed by means of the regression-based change approach and the *T* Scores for Change method. Finally, we discuss the results and provide recommendations for minimum sample size needed to produce sufficiently precise norms.

Norming Methods for Change Scores

First, to provide an overview of the field and also as a precursor for the two methods we study in this research, we briefly discuss two approaches to quantifying individual change in the pretest–posttest design, which are the classical change-score approach (i.e., using the classical change score to quantify individual change) and the residual change-score approach (i.e., using the residual change score resulting from a regression model). Second, we present a general framework of norming change scores, which includes the regression-based change approach and the *T* Scores for Change method as special cases.

Two Approaches to Quantifying Individual Change

The classical change-score approach and the residual change-score approach (Willett, 1988) are widely adopted methods. Caruso (2004) discussed other methods for quantifying change that were used less frequently. For the classical change-score approach, we use the observable classical change score D, which is defined as the difference between the observable posttest score X_{post} and the observable pretest score X_{pre}; that is, $D = X_{\text{post}} - X_{\text{pre}}$. Notice that D, X_{pre}, and X_{post} are random variables, whose realizations are denoted by d, x_{pre}, and x_{post}, respectively. In classical test theory (Lord & Novick, 1968), an observable test score X is assumed to be the sum of a true score T and a random measurement error E (i.e., $X = T + E$). A person's true score is defined as the expectation across hypothetical independent administrations of the test, so that $T = \mathcal{E}(X)$ and $\mathcal{E}(E) = 0$. Let Δ denote the person's true change score, defined as $\Delta = T_{\text{post}} - T_{\text{pre}}$. Then, we can write

$$\begin{aligned} D &= X_{\text{post}} - X_{\text{pre}} \\ &= \left(T_{\text{post}} + E_{\text{post}}\right) - \left(T_{\text{pre}} + E_{\text{pre}}\right) \\ &= \left(T_{\text{post}} - T_{\text{pre}}\right) + \left(E_{\text{post}} - E_{\text{pre}}\right) \\ &= \Delta + \left(E_{\text{post}} - E_{\text{pre}}\right). \end{aligned}$$

Because $\Delta = \mathcal{E}(D)$, the classical change score D is an unbiased estimate of the true change Δ (Rogosa et al., 1982), but this has not withheld several researchers from questioning D as a useful measure of individual change (e.g., Bereiter, 1963; Cronbach & Furby, 1970; Linn & Slinde, 1977; O'Connor, 1972). Others have supported the use of D (e.g., Overall & Woodward, 1975; Rogosa et al., 1982; Williams & Zimmerman, 1977, 1996, Zimmerman & Williams, 1982a, 1982b). Gu et al. (2018) suggested that many negative beliefs about D, such as its alleged low reliability, are based on, for example, inappropriate assumptions, thus mitigating the criticism of classical change scores.

The residual change-score approach, also known as residual gains or base-free measurement of change, intends to correct for the correlation between pretest score and change score (Cronbach & Furby, 1970; Manning & Dubois, 1962; Willett, 1988) by means of the residualized posttest score. Let $\tilde{X}_{\text{post}}$ be the predicted posttest score obtained by regressing X_{post} on X_{pre}. Then, the residualized posttest score, denoted by R_D, is defined as $R_D = X_{\text{post}} - \tilde{X}_{\text{post}}$. The residualized posttest score for a person shows how much more or less he or she has changed compared with the predicted average change of others with the same observed pretest score. For example, a positive residualized posttest score means that the person's individual change is larger than the expected change of patients with the same observed pretest scores. A residualized posttest score of zero means that the person's change equals the average change given the same observed pretest scores, but it does *not* mean that the person has not changed at all. Residualized posttest scores have been used in studies that search for predictors of interindividual differences in change (e.g., Castro-Schilo & Grimm,

2018). These predictors may be of theoretical interest, for example, to know whether recovery after brain injury can be explained by age. Variables that have been shown to predict interindividual change may also help the interpretation of individual scores. However, for individual assessment one should take into account that residualized scores have a very specific interpretation, which, as we will argue, coincides with a normative interpretation of change scores.

Regression-Based Norming for Change Scores: A General Framework

Regression-based norming for test scores (Van Breukelen & Vlaeyen, 2005; Zachary & Gorsuch, 1985) generates the reference distribution of test scores by means of regression analysis. Let the random vector $\mathbf{W} = [W_1, W_2, \ldots, W_K]$ contain K relevant covariates, whose realization is denoted by $\mathbf{w} = [w_1, w_2, \ldots, w_K]$. Let k index covariates. Regression-based norming assumes that the average test score can be predicted by relevant covariates, so that

$$\mathcal{E}(X|\mathbf{W} = \mathbf{w}) = \beta_0 + \beta_1 w_1 + \ldots + \beta_K w_K, \quad (1)$$

where β_0 is the intercept and $\beta_1, \ldots, \beta_K$ are regression coefficients for covariates. It may be noted that Equation (1) may also include quadratic terms such as w_k^2 and include interactions between covariates. An observable test score, denoted by X, deviates from $\mathcal{E}(X|\mathbf{W} = \mathbf{w})$ by residual $\epsilon = X - \mathcal{E}(X|\mathbf{W} = \mathbf{w})$. Computing ϵ for the persons belonging to the subpopulation satisfying $\mathbf{W} = \mathbf{w}$, we obtain a distribution of ϵs, reflecting the relative position of persons within the same subpopulation (i.e., $\mathbf{W} = \mathbf{w}$) on X. When deriving norm statistics in practice, one first estimates $\beta_0, \beta_1, \ldots, \beta_K$ in Equation (1), estimates denoted by $b_0, b_1, ..., b_K$, and then computes the residuals for all persons in the sample. Let x_i denote the test score for person i ($i = 1, \ldots, N$), let $w_{i1}, ..., w_{iK}$ denote person i's scores on the K covariates, and let e_i denote the residual for person person i. Then,

$$e_i = x_i - (b_0 + b_1 w_{i1} + \ldots + b_K w_{iK}).$$

The distribution of $e_i s$ ($i = 1, \ldots, N$) is used to compute norm statistics (e.g., percentiles).

Regression-based norming for test scores can readily be extended to norming change scores. Replacing X in Equation (1) with an observable change score D, one obtains the population model

$$\mathcal{E}(D|\mathbf{W} = \mathbf{w}) = \beta_0 + \beta_1 w_1 + \ldots + \beta_K w_K. \quad (2)$$

An observable change score, denoted by D, deviates from $\mathcal{E}(D|\mathbf{W} = \mathbf{w})$ by residual $\varepsilon = D - \mathcal{E}(D|\mathbf{W} = \mathbf{w})$. The distribution of εs reflects the relative position of persons among other persons within the same subpopulation defined by $\mathbf{w}$. Specifically, $\mathcal{E}(D|\mathbf{W} = \mathbf{w})$ is the average change in the subpopulation satisfying $\mathbf{W} = \mathbf{w}$. If $\varepsilon > 0$, then the person's change score, D, is larger than the average change (i.e., $\mathcal{E}(D|\mathbf{W} = \mathbf{w})$) in the corresponding subpopulation. In practice, when computing norm statistics, one first estimates the parameters in Equation (2), then computes the residuals for all the persons in the sample, and finally uses the distribution of the residuals to generate norm statistics.

One may notice that Equation (2) is a general framework for norming change scores, and that we have not discussed which covariates $\mathbf{W}$ should be included. In practice, covariates should be selected based on substantive arguments based on theory and domain knowledge, and so on. In addition, Oosterhuis et al. (2016) suggested that the selection might also be supported by statistical procedures, such as stepwise regression. Interestingly, depending on whether the pretest score is used as a covariate, the literature has identified two regression-based norming methods based on the general framework (Equation 2) that are frequently used in practice. They are the regression-based change approach (Van der Elst et al., 2008) and the *T* Scores for Change method (McSweeny et al., 1993). In the remainder of this section, we present the two norming methods and discuss their similarities and differences.

The regression-based change approach (Van der Elst et al., 2008) assumes that change scores can be predicted by means of a few relevant covariates, and that the pretest score is not included as one of the covariates in Equation (2). The model used is

$$D = \beta_0 + \beta_1 w_1 + \ldots + \beta_K w_K + \varepsilon. \quad (3)$$

Because Equation (3) directly models the change score, the regression-based change approach follows the classical change-score approach to quantifying individual change. The *T* Scores for Change method (McSweeny et al., 1993) requires that the pretest score is included as a covariate in Equation (2), so that

$$D = \beta'_0 + \beta'_1 w_1 + \ldots + \beta'_0 w_K + \beta'_{X_{\text{pre}}} X_{\text{pre}} + \varepsilon'. \quad (4)$$

Adding X_{pre} to both sides of Equation (4), we obtain

$$X_{\text{post}} = \beta'_0 + \beta'_1 w_1 + \ldots + \beta'_0 w_K + \underbrace{(\beta'_{X_{\text{pre}}} + 1)}_{\beta''_{X_{\text{pre}}}} X_{\text{pre}} + \varepsilon' \quad (5)$$

$$\Rightarrow X_{\text{post}} = \beta'_0 + \beta'_1 w_1 + \ldots + \beta'_0 w_K + \beta''_{X_{\text{pre}}} X_{\text{pre}} + \varepsilon', \quad (6)$$

which is the model McSweeny et al. (1993) proposed. Equation (6) shows that the *T* Scores for Change method follows the residual change-score approach to quantifying change, which at first glance appears to be different from the regression-based change approach. However, as we have shown from Equations (4) to (6), the *T* Scores for Change method and the regression-based change approach

together define a general framework for norming change scores (Equation 2), and the only difference between the two methods resides in the inclusion of X_{pre} in Equation (6).

Here, we notice a special case of including X_{pre} as a covariate when the model also includes a categorical variable, such as gender. Suppose that one of the covariates in the model of the regression-based change approach (i.e., Equation 3) is a categorical variable, denoted by g, such that

$$D = \beta_0 + \beta_g g + \beta_1 w_1 + \ldots + \beta_K w_K + \varepsilon. \quad (7)$$

Equation (7) often is referred to as CHANGE, which is a method using the classical change score as the dependent variable to analyze the pretest–posttest control-group design (Van Breukelen, 2013). On the other hand, suppose that the *T* Scores for Change method (Equation 6) includes a categorical variable, so that Equation (6) becomes

$$X_{\text{post}} = \beta'_0 + \beta'_g g + \beta'_1 w_1 + \ldots + \beta'_0 w_K + \beta''_{X_{\text{pre}}} X_{\text{pre}} + \varepsilon'. \quad (8)$$

Then Equation (8) becomes the analysis of covariance (ANCOVA) model. CHANGE and ANCOVA may cause contradictory results with respect to group-mean differences in nonrandomized studies when groups differ on average at pretest, where ANCOVA indicates a mean effect, whereas CHANGE does not, which is known as Lord's paradox (Lord, 1967; Van der Elst et al., 2008). Van Breukelen (2013) formally examined this issue and showed that applying ANCOVA and CHANGE to the same data could result in completely different conclusions, because ANCOVA assumes absence of such a group effect at pretest and CHANGE assumes presence of a group effect (Van Breukelen, 2013). The implication of Van Breukelen's research for our study is that, when using the *T* Scores for Change method, one may find that β'_g in Equation (8) is significant, but when using the regression-based change approach, one may find that β_g in Equation (7) is not significant.

Before concluding the section, we remind the reader that regression-based norming requires the assumptions associated with regression analysis to hold for the application of interest. For detailed discussions on this topic, we refer the reader to Oosterhuis (2017).

Deriving Norm Statistics

The two norming methods produce norm statistics in the same manner. To describe the procedure, we use the regression-based change approach as an example. The steps are the following.

Step 1: Compute for each person $i\,(i = 1,\ldots,N)$ the predicted change score, denoted by $\tilde{d}_i$, by means of

$$\tilde{d}_i = b_0 + b_1 w_{i1} + \ldots + b_K w_{iK}, \quad (9)$$

where b_0 denotes the sample intercept and $b_1,...,b_K$ denote the sample regression coefficients.

Step 2: Compute residual e_i, for person i, as

$$e_i = d_i - \tilde{d}_i, \quad (10)$$

where d_i is the observed change score for person i.

Step 3: One may use the distribution of e_is to gauge norm statistics, such as percentiles. Sometimes, researchers transform e_is into standardized scores, when, for example, using the *T* Scores for Change method:

Step 3*: Compute the standardized e_i,

$$z_{e_i} = \frac{e_i}{SD_e}, \quad (11)$$

where

$$SD_e = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{N - (K+1)}} \quad (12)$$

A few remarks are in order. First, when the *T* Scores for Change method is used, in Step 1, one computes the predicted posttest score $\tilde{x}_{\text{post}}$ instead of $\tilde{d}$ based on Equation (6), and then computes the residuals. Second, the *T*-scores for Change method owes its name from the fact that standardized residuals are rescaled to scores with $M = 50$ and $SD = 10$, which produces *T* scores (Allen & Yen, 2002): $T_i = 50 + 10 \times z_{e_i}\,(i = 1,\ldots,N)$. However, residuals may also be transformed to other well-known scales, such as Wechsler scores (i.e., IQ scores) by using the formula $IQ_i = 100 + 15 \times z_{e_i}\,(i = 1,\ldots,N)$. Which scale is convenient depends on the application envisaged. Third, in practice, when a new patient arrives and is measured, the practitioner uses the fitted regression model and the distribution of residuals from the norm samples to obtain the normed scores for the new patient using tables or dedicated software (e.g., De Vroege et al., 2018).

In this article, we used a simulation study to examine the two norming methods under various conditions and investigate the precision of estimated norms and the minimum sample size needed to obtain norms of high precision.

Method

Data Generation

Population Model. We assumed that pretest score, X_{pre}, was determined by latent variable θ_{pre} representing the attribute scale of interest at pretest. To identify the scale, we assumed that θ_{pre} followed a standard normal distribution, $\theta_{\text{pre}} \sim N(0,1)$. We assumed that the change on the attribute scale, denoted by θ_D, was partly predicted by (1) θ_{pre}, (2) a dichotomous covariate, for example, gender, denoted by X_{gender}, and (3) a continuous covariate, age, denoted by X_{age}, and assumed that θ_{pre}, X_{gender}, and X_{age} were independent

of one another. We also assumed that the unexplained part of θ_D was subsumed under random residual E, so that

$$\theta_D = \beta_0 + \beta_1 X_{\text{gender}} + \beta_2 X_{\text{age}} + \beta_{\theta_{\text{pre}}} \theta_{\text{pre}} + E. \quad (13)$$

Thus, we assumed that the variance of θ_D, denoted by $\sigma^2_{\theta_D}$, was partly explained by the variances of X_{gender}, X_{age}, and θ_{pre}, and partly explained by the variance of E. Including E in Equation (13) is important, because it is exactly the residuals that we intended to norm. We further assumed that, if X_{gender}, X_{age}, and θ_{pre} exert no effect on θ_D (i.e., $\beta_1 = \beta_2 = \beta_{\theta_{\text{pre}}} = 0$), then

$$\theta_D \sim N(.75,.14) \text{ and } N(.75,1.14). \quad (14)$$

Gu et al. (2018) showed that the choice of $\theta_D \sim N(.75,.14)$ corresponds to 75% of the group showing a minimal important difference (MID; Norman et al., 2003; Schünemann & Guyatt, 2005; see supplementary Appendix A, available online). MID refers to the minimal change that clinicians consider important. Gu et al. (2018) showed that the larger the $\sigma^2_{\theta_D}$, the higher the change-score reliability. Therefore, manipulating the value of $\sigma^2_{\theta_D}$ enabled us to examine how change-score reliability influenced the norms. In cases where X_{gender}, X_{age}, and θ_{pre} exerted effect on θ_D, we also assumed $\sigma^2_{\theta_D} = .14$ and 1.14, respectively. In the remainder of this section, we use $\sigma^2_{\theta_D} = .14$ to show how to obtain β_1, β_2, and $\beta_{\theta_{\text{pre}}}$ in situations where X_{gender}, X_{age}, and θ_{pre} exerted effect on θ_D. Appendix B (supplementary table available online) presents the parameter values used in the simulation study.

We further assumed that posttest score, X_{post}, was determined by latent variable θ_{post} representing the attribute scale of interest at posttest, and was computed by taking the sum of θ_{pre} and θ_D (i.e., $\theta_{\text{post}} = \theta_{\text{pre}} + \theta_D$). Alternatively, one may first simulate θ_{pre} and θ_{post}, and then define $\theta_D = \theta_{\text{post}} - \theta_{\text{pre}}$, but, unlike our approach, this approach does not allow directly manipulating $\sigma^2_{\theta_D}$.

Correlation $\rho_{\theta_{\text{pre}} D}$ is of interest in psychological and educational research (e.g., Bryk & Raudenbush, 1987; Gu et al., 2018; Hertzog et al., 2008; Linn & Slinde, 1977; Raykov, 1993; Rogosa et al., 1982; Werker & Lalonde, 1988). We considered $\rho_{\theta_{\text{pre}}\theta_D} = 0$, .1 (small effect size, choice explained in Appendix C; Cohen, 1992), and −.1. One may notice that

$$\rho_{\theta_{\text{pre}}\theta_D} = \frac{\text{COV}(\theta_D, \theta_{\text{pre}})}{\sqrt{\sigma^2_{\theta_D} \times \sigma^2_{\theta_{\text{pre}}}}} = \frac{\beta_{\theta_{\text{pre}}}}{\sigma_{\theta_D}}, \quad (15)$$

where $\sigma^2_{\theta_{\text{pre}}} = 1$. Thus, when $\sigma_{\theta_D} = \sqrt{.14}$, $\rho_{\theta_{\text{pre}} D} = 0$, .1, and −.1 correspond to $\beta_{\theta_{\text{pre}}} = 0$, .037, and −.037, respectively.

Consistent with Oosterhuis et al. (2016), we assumed that covariate X_{gender} followed a Bernoulli distribution with probability .5, and covariate X_{age} followed a uniform distribution on the interval [4, 12]. We chose β_1 and β_2 such that variance of θ_D explained by X_{gender}, X_{age}, and θ_{pre}, denoted by R^2, corresponded to small effect size (i.e., $R^2 = .065$), medium effect size (i.e., $R^2 = .13$), and large effect size (i.e., $R^2 = .26$; Cohen, 1992), respectively. The two covariates in Equation (13) explained equal proportions of the variance of θ_D. We refer to Appendix C (supplementary material available online) where we show how to obtain β_1 and β_2 in Equation (13), and why correlation $\rho_{\theta_{\text{pre}}\theta_D}$ is restricted to have small effect size.

Test Characteristics and Item Parameters. We considered tests containing 10, 20, and 40 items (Jabrayilov et al., 2016; Kruyen et al., 2013). Polytomous items were simulated using the graded response model (GRM; Samejima, 1969), which is a common choice in simulation research, because the GRM enables the easy manipulation of a test with desirable features. Because the GRM assumes that latent variable θ is a nonlinear function of test score X, and the regression models we use for norming X are linear functions of the covariates, one might object to the use of the GRM. However, it is well-known that for most tests θ and X correlate high (Macdonald & Paunonen, 2002). For example, in an empirical study, Fan (1998) found that the correlation between θ and X could be higher than .9.

Suppose item j has $M+1$ ordered scores, so that realization x has values $x = 0,\ldots,M$. Consistent with Likert-type scales, we chose $M = 4$. Let α_j be the slope parameter and let $\beta_{jx} (x = 1,\ldots,M)$ be the threshold parameter; then, the GRM models the probability of obtaining a score $X_j \geq x$ as

$$P\left(X_j \geq x|\theta\right) = \frac{\exp\left[\alpha_j\left(\theta - \beta_{jx}\right)\right]}{1 + \exp\left[\alpha_j\left(\theta - \beta_{jx}\right)\right]}, x = 1,\ldots,M. \quad (16)$$

Slope parameter α_j was sampled from a uniform distribution $U(1.5, 2.5)$, and an average threshold $\bar{\beta}_j$ was sampled from $U(0,1.25)$ (Emons et al., 2007; Jabrayilov et al., 2016). Individual β_{jk}s were chosen as: $\beta_{j1} = \bar{\beta}_j - .75$, $\beta_{j2} = \bar{\beta}_j - .25$, $\beta_{j3} = \bar{\beta}_j + .25$, and $\beta_{j4} = \bar{\beta}_j + .75$. For $M = 1$, dichotomous item scores were generated. Parameter α_j was sampled from a uniform distribution $U(1.5, 2.5)$, and parameter β_j was sampled from $U(0,1.25)$.

Simulation Design

The completely crossed design had 3×2×15×3×3×2=1,620 cells and included the following factors:

1. Test length: 10, 20, and 40 items.
2. Number of item scores: 2 and 5.
3. Sample size: 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, and 1500.
4. Effect size of covariates: R^2 = .065, .13, and .26.

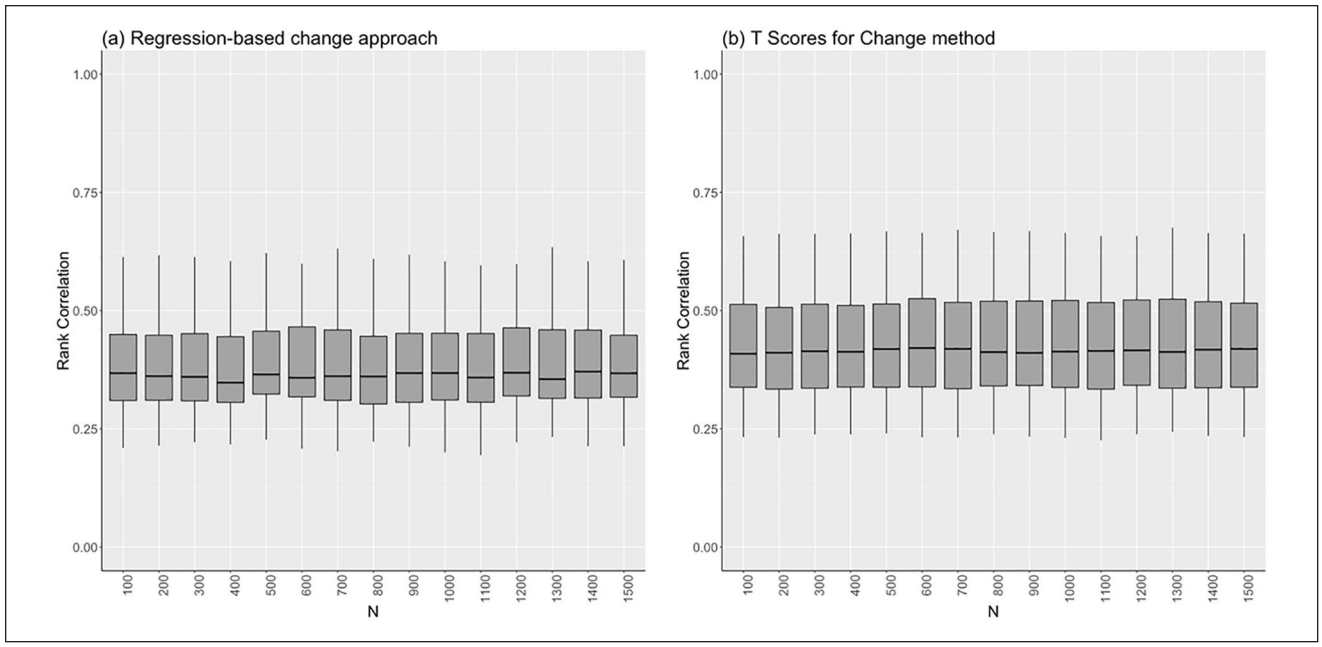

Figure 1. Boxplots of $r_{E,D}$ (panel a) and $r_{E,T}$ (panel b), when $N = 100, 200, ..., 1500$.

5. Correlation between θ_{pre} and θ_D: $\rho_{\theta_{pre}\theta_D} = 0$, .1, and −.1.
6. Variance of θ change: $\sigma^2_{\theta_D} = .14$ and 1.14.

For each cell, we generated data using the following steps.

Step 1: Item parameters were sampled.

Step 2: A sample of person parameters at pretest were randomly drawn from the θ_{pre} distribution ($\theta_{pre} \sim N(0,1)$), and a sample of person-change parameters were randomly drawn based on Equation (13). Based on these samples, the sample of person parameters at posttest was obtained using $\theta_{post} = \theta_{pre} + \theta_D$. Item-score data sets of pretest and posttest were generated given θ_{pre} and θ_{post}.

Step 3: For each design cell, we repeated Step 2 1,000 times, and as a result, 1,000 item-score data sets of pretest and posttest administrations were generated.

Step 4: For each data set, we computed standardized residuals based on the regression-based change approach and the *T* Scores for Change method.

Dependent Variables and Data Analysis

Rank Correlation. Using Kendall's tau (Kendall, 1938), we computed three different rank correlations. They are (a) Rank correlation (denoted by $r_{E,T}$) between *E* (Equation 13) and the standard residuals produced by the *T* Scores for Change method; (b) Rank correlation (denoted by $r_{E,D}$) between *E* (Equation 13) and the standard residuals produced by the regression-based change approach; and (c) Rank correlation (denoted by $r_{D,T}$) between the standard residuals produced by the regression-based change approach and the *T* Scores for Change method. It may be noted that $r_{E,T}$ almost equal to 1 means that the relative position in the sample norm-distribution produced by the *T* Scores for Change method preserves the relative position in the *E* distribution in Equation (13). Rank correlations $r_{E,D}$ and $r_{D,T}$ are interpreted similarly.

Precision. We considered the precision of the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th percentiles of the standardized residuals generated by the regression-based change approach and the *T* Scores for Change method. Precision expressed by the standard deviation of the sampling distribution is not suitable for percentile estimates that typically are not normally distributed. Alternatively, we used the 95% interpercentile range (IPR), which is a distribution-free measure defined as the difference between the 97.5th percentile and the 2.5th percentile of the distribution of standardized residuals, based on 1,000 data sets in each design cell. The higher the IPR, the lower the precision.

Results

Rank Correlations

Figure 1 presents boxplots of $r_{E,D}$ and $r_{E,T}$ against sample size *N*. Because rank correlations remained approximately the same as sample size *N* increased, we singled out sample size $N = 1,000$ and presented in Table 1 the median values of rank correlations $r_{E,T}$, $r_{E,D}$, and $r_{D,T}$, estimated change-score reliability (denoted by $r_{DD'}$) using coefficient

Table 1. The Median Values of Estimated Rank Correlations, Estimated Change-Score Reliability ($r_{DD'}$) Using Coefficient α (Cronbach, 1951), Sample Variances of Pretest and Posttest ($S^2_{X_{pre}}$ and $S^2_{X_{post}}$), Sample Correlation Between the Pretest and the Posttest ($r_{X_{pre}X_{post}}$), and Sample Correlation Between the Pretest and Change ($r_{X_{pre}D}$), When $N = 1,000$.

	Rank correlation							
	$r_{E,T}$	$r_{E,D}$	$r_{D,T}$	$r_{DD'}$ *	$S^2_{X_{pre}}$	$S^2_{X_{post}}$	$r_{X_{pre}X_{post}}$	$r_{X_{pre}D}$
Test length								
10 Items	.35	.32	.77	.68	55.17	55.21	.63	−.41
20 Items	.41	.37	.82	.80	228.32	181.05	.67	−.32
40 Items	.44	.39	.80	.88	909.76	667.98	.69	−.36
Number of item scores								
2	.37	.33	.78	.78	31.84	36.51	.63	−.40
5	.45	.39	.81	.84	426.62	474.97	.68	−.32
Effect size of covariates								
$R^2 = .065$	.48	.46	.83	.81	125.14	146.47	.67	−.30
$R^2 = .13$	.44	.39	.78	.81	121.63	130.34	.64	−.38
$R^2 = .26$	.37	.31	.71	.83	118.73	115.34	.61	−.50
Population correlation between θ_{pre} *and* θ_D								
$\rho_{\theta_{pre}\theta_D} = 0$	.42	.37	.79	.82	124.95	129.43	.63	−.37
$\rho_{\theta_{pre}\theta_D} = -.1$	.43	.37	.77	.82	124.95	124.90	.61	−.39
$\rho_{\theta_{pre}\theta_D} = .1$	.41	.37	.82	.82	124.95	136.81	.65	−.32
Variance of θ_D								
$\sigma^2_{\theta_D} = 0.14$	.34	.32	.85	.70	124.95	131.52	.79	−.25
$\sigma^2_{\theta_D} = 1.14$	.52	.45	.73	.89	124.95	125.77	.51	−.47

Note. *$r_{DD'}$ was computed based on Lord and Novick (1968, p. 76) using coefficient α. Let $r_{X_{pre}X'_{pre}}$ denote the estimated reliability at pretest using α. Let $r_{X_{post}X'_{post}}$ denote the estimated reliability at posttest using α. Then, $r_{DD'} = \frac{r_{X_{pre}X'_{pre}} S^2_{X_{pre}} + r_{X_{post}X'_{post}} S^2_{X_{post}} - 2r_{X_{pre}X_{post}} s_{X_{pre}} s_{X_{post}}}{S^2_{X_{pre}} + S^2_{X_{post}} - 2r_{X_{pre}X_{post}} s_{X_{pre}} s_{X_{post}}}$.

As an aside, the reader may notice that, for the last 5 rows, $S^2_{X_{pre}}$s are exactly the same, which is due to the same seed used (please see the R script in the supplementary material, available online).

α (Cronbach, 1951), sample variances of pretest and posttest scores (denoted by $S^2_{X_{pre}}$ and $S^2_{X_{post}}$), sample correlation between pretest and posttest scores (denoted by $r_{X_{pre}X_{post}}$), and sample correlation between pretest scores and change scores (denoted by $r_{X_{pre}D}$). To interpret the results, we use the first row in Table 1 (i.e., test length equal to 10 items) as an example. Recall that 1,000 item-score data sets were generated. First, for each cell, we computed the mean rank correlation $r_{E,T}$ across the 1,000 item-score data sets. Then, the median of the rank correlation $r_{E,T}$ was determined across all the mean rank correlations for the other four design factors, including the number of item scores, the effect size of covariates, the population correlation between θ_{pre} and θ_D, and the variance of θ_D, and was found to equal .35.

Based on Table 1, we made the following observations. First, the median of $r_{E,T}$ was slightly higher than the median of $r_{E,D}$. In general, both $r_{E,T}$ and $r_{E,D}$ were lower than .5, suggesting that, given the current simulation setup, the relative position in the sample norm-distribution produced by the two norming methods largely differed from the relative position in the E distribution in Equation (13). The low $r_{E,T}$ and $r_{E,D}$ might be attributed to the measurement errors randomly generated in the data generating procedure. Second, the median of $r_{D,T}$ in general was higher than .7. This suggests that the regression-based change approach and the T Scores for Change method generated fairly comparable norm statistics. Third, increasing test length and number of item scores caused higher rank correlations. Finally, larger variance of θ_D and smaller effect sizes of covariates were positively associated with higher rank correlation.

Relationship Between IPR and Sample Size

For each percentile and for the T Scores for Change method, Figure 2 shows the box plots of IPR against sample size across all 1,620 cells. We chose not to plot the percentiles generated by the regression-based change approach because the results were similar to Figure 2. The figure shows that, as sample size grew, the norms were more precise, which reflects the well-known, inverse relation between sample size and sampling variance for all statistics based on a sample of independent observations. Specifically, estimation precision sharply increased as sample size increased from 100 to 500. As sample size reached 1,500, the increase of

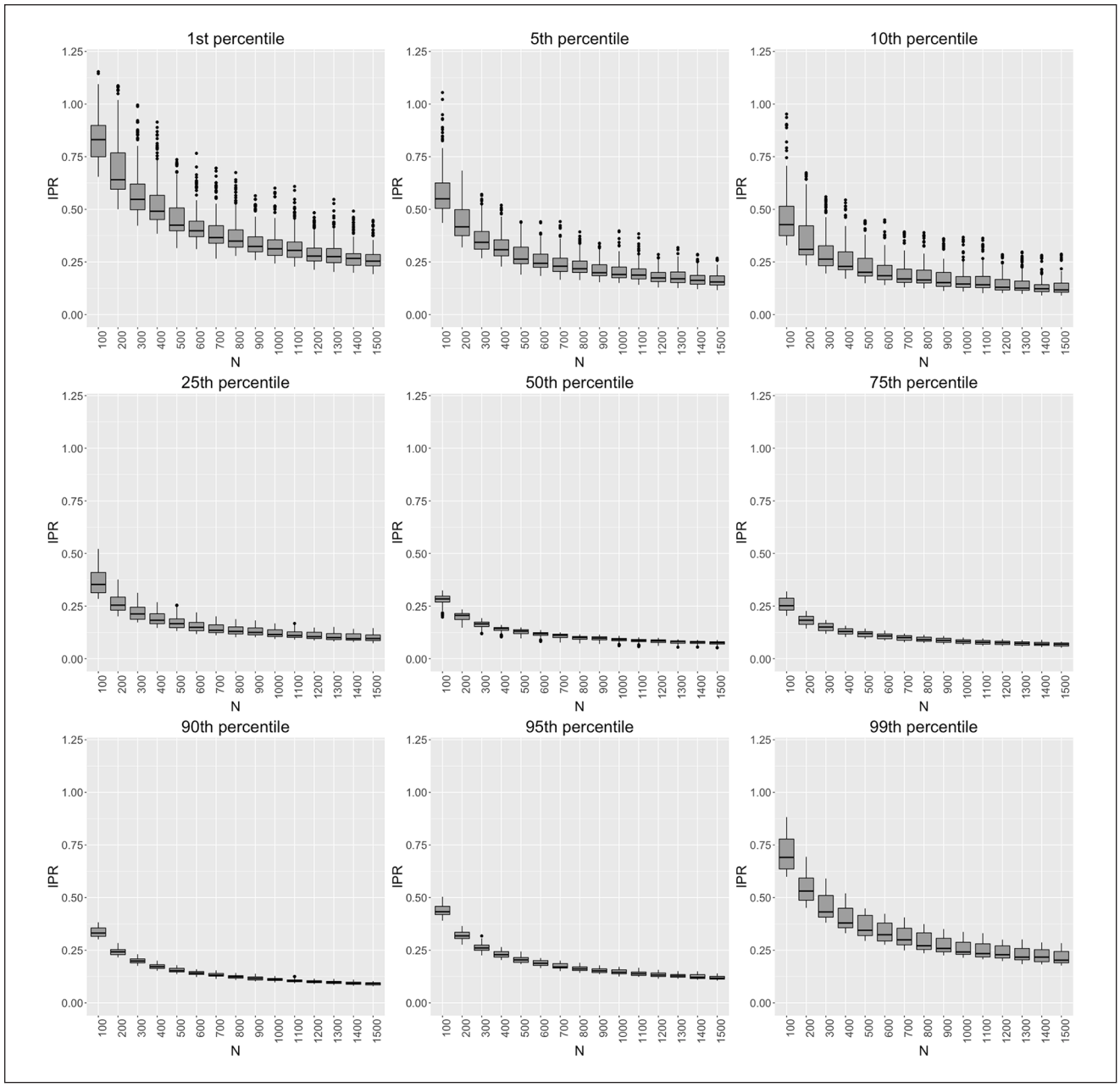

Figure 2. Relationship between sample size (*N*) and interpercentile range (IPR) for the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th percentiles generated by the *T* Scores for Change method.

precision leveled off. The figure suggests that, when designing a study for normative data, a sample size between 500 and 1,500 may be desirable, but based on this study suggestions for exact sample sizes do not seem to be realistic.

Relationship Between IPR and the Other Five Design Factors

Figures (3) through (7) present the relationship between IPR generated by means of the *T* Scores for Change method and the other five design factors, which were test length, number of item scores, effect size of covariates, correlation between θ_{pre} and θ_D, and variance of θ change. In general, the figures suggest that the five design factors did not noticeably influence the precision by which percentiles of the change-score distribution were estimated, especially for the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles. An interesting result is the absence of an effect of test length on the IPRs, and thus the precision by which norms were obtained. This may be surprising because shorter tests

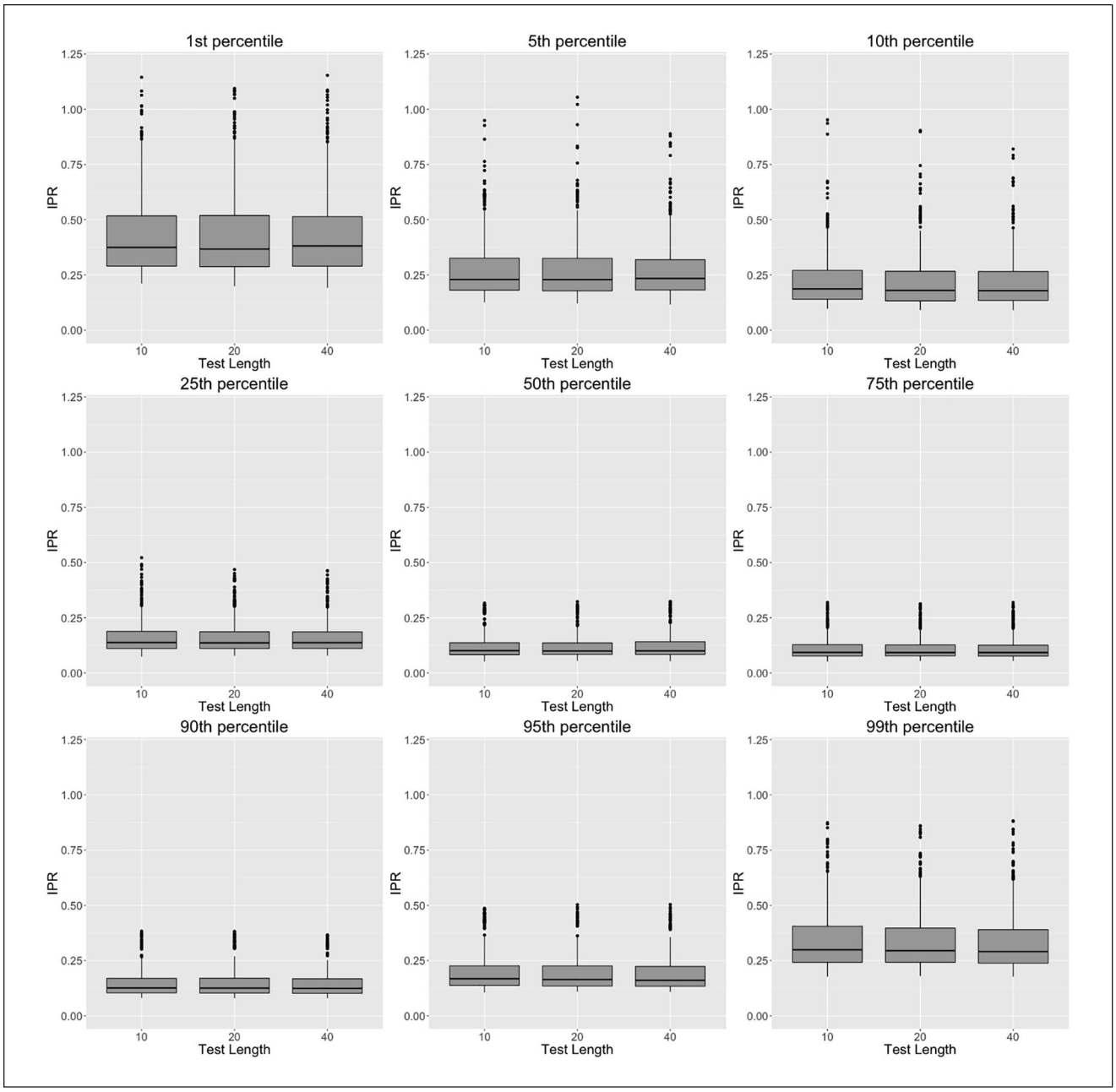

Figure 3. Relationship between test length and interpercentile range (IPR) for the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th percentiles generated by the *T* Scores for Change method.

generate less reliable test scores, and therefore one would expect lower precision for shorter tests. However, the IPR reflect variability in the estimated norms across samples and this variability is explained by both random sampling of persons and random measurement errors. Because we see high impact of sample-size changes on the IPRs and not for test length, results suggest that the cross-sample variability in the percentiles arising from random person sampling outweighs the variability from measurement errors. This result is consistent with a study by Sijtsma and Emons (2011) who found that the power of an independent samples *t* test only changed a little when test length was manipulated but greatly as sample size was varied.

Although the simulation was a fully crossed factorial design, we chose not to use analysis of variance (ANOVA) to analyze IPR results. The reasons are that, first, the normality assumption of ANOVA was severely violated, and second, although the log-transformed IPR to some extent reduced the problem of nonnormality, ANOVA results showed that almost all the design factors were significant.

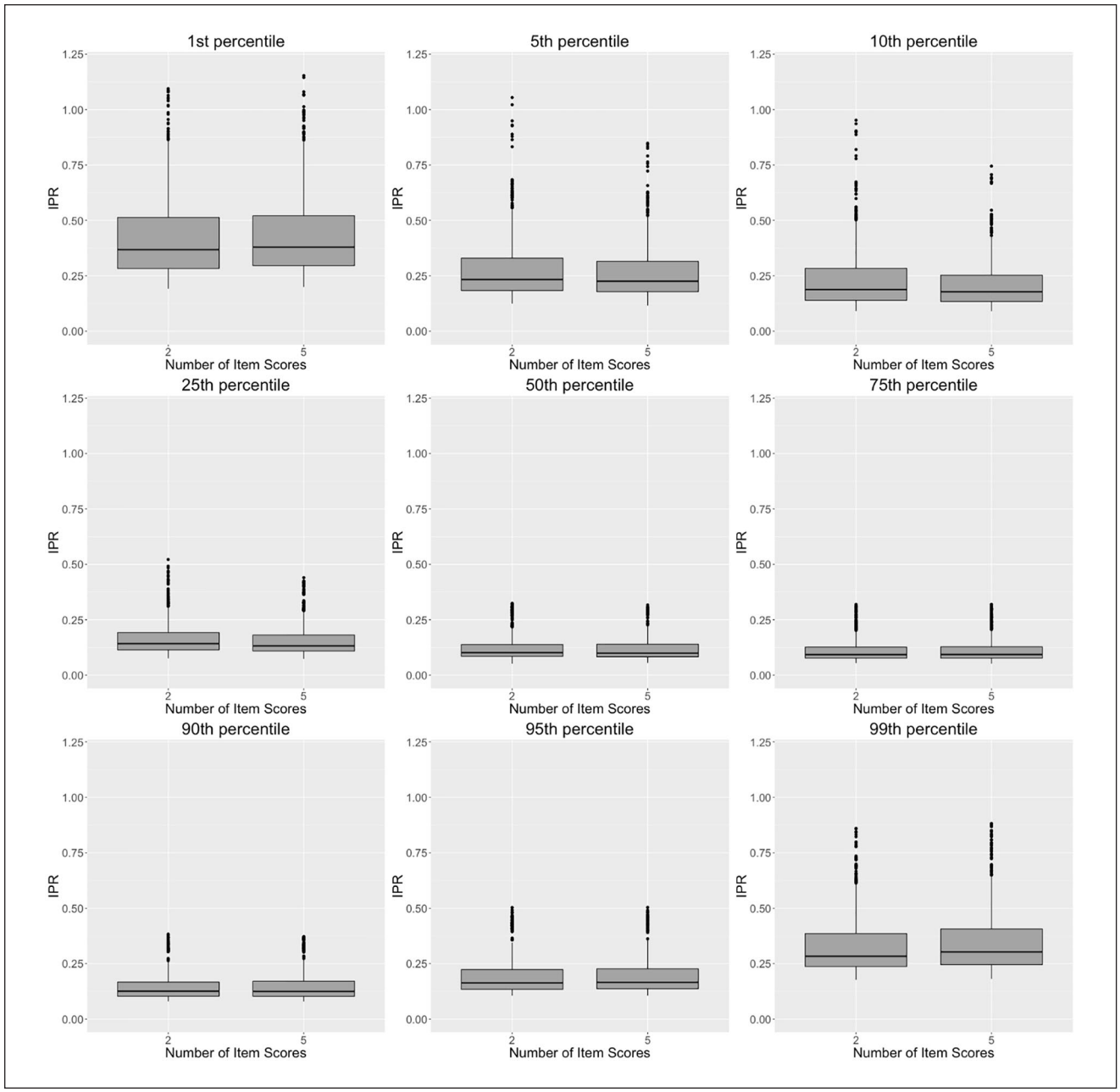

Figure 4. Relationship between number of item scores and interpercentile range (IPR) for the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th percentiles generated by the *T* Scores for Change method.

The reasons for significance were considerable sample size, hence large power, and outliers, of which there were quite a few; see Figures (3) through (7).

Discussion

Because of their simplicity and popularity, we limited attention to change scores obtained from pretest–posttest designs, but change scores are not limited to such designs. For example, to monitor cognitive decline in elderly people, norms developed for repeated administration of neuropsychological tests can offer diagnostic assistance. We showed that the regression-based change approach and the *T* Scores for Change method originated from the same general, regression-based framework for norming change scores. We advise test constructors to make critical decisions about the covariates, such as the pretest score, that they decide to include in the model.

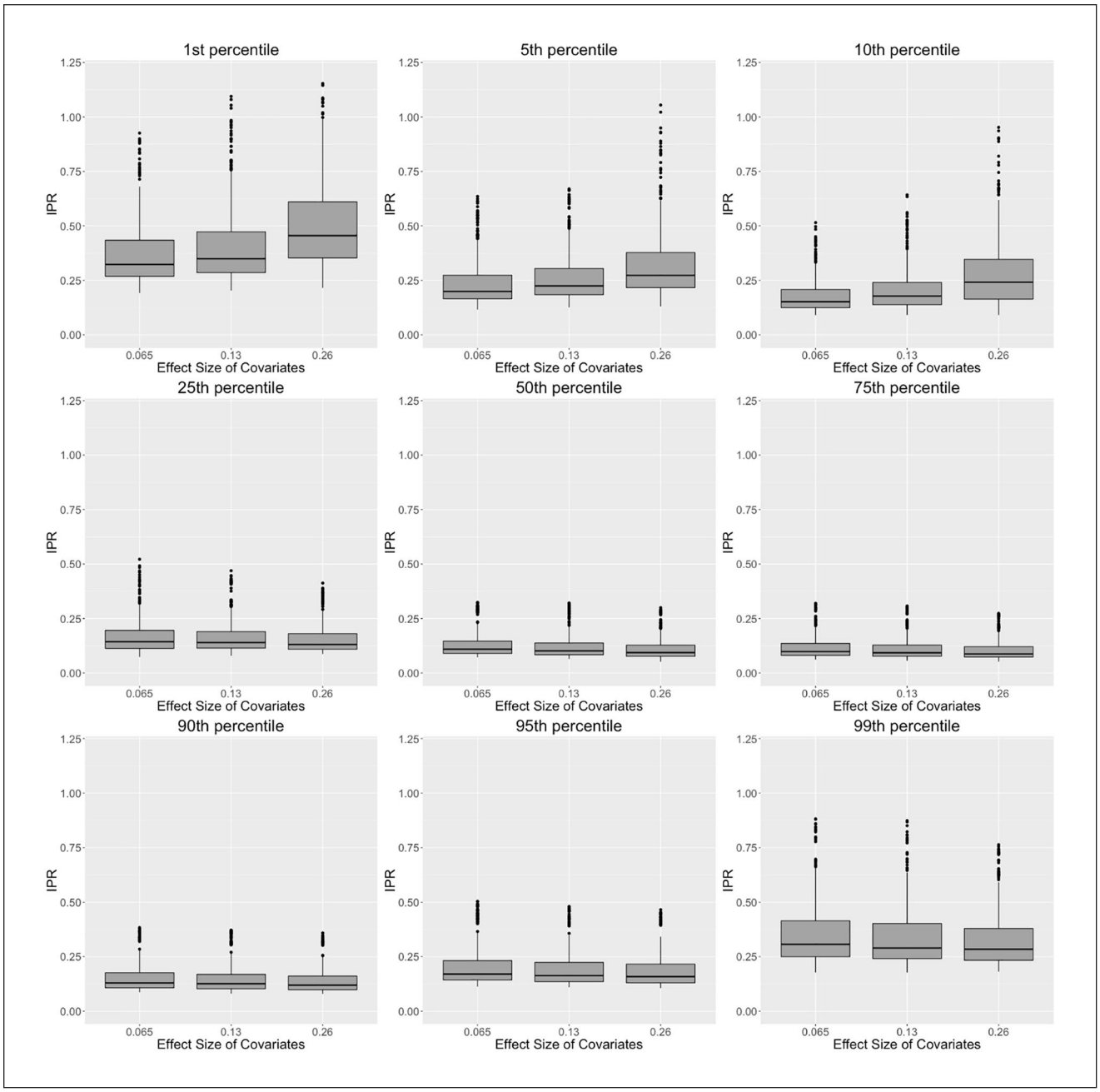

Figure 5. Relationship between effect size of covariates and interpercentile range (IPR) for the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th percentiles generated by the *T* Scores for Change method.

Norming change scores is a challenging task. Our simulation study showed that the relative position of persons in the sample norm-distribution produced by the two norming methods largely differed from the relative position of persons in the error distribution, as witnessed by the low rank correlations. This suggests that decision-making (e.g., whether a patient's health condition has improved compared with a normative sample) solely based on norm statistics may lead to biased conclusions. More studies are needed to understand the cause of low rank correlations to improve the norming methods.

Increasing sample size greatly improved the precision of norms, but the benefit of a larger sample size diminished quickly as the sample grew larger than 1,500 observations. Our simulation study suggested that a sample size of 500 is a reasonable minimum for norming change scores. Increasing the sample size to about 1,000 is still beneficial, but samples larger than 1,500 offer little improvement.

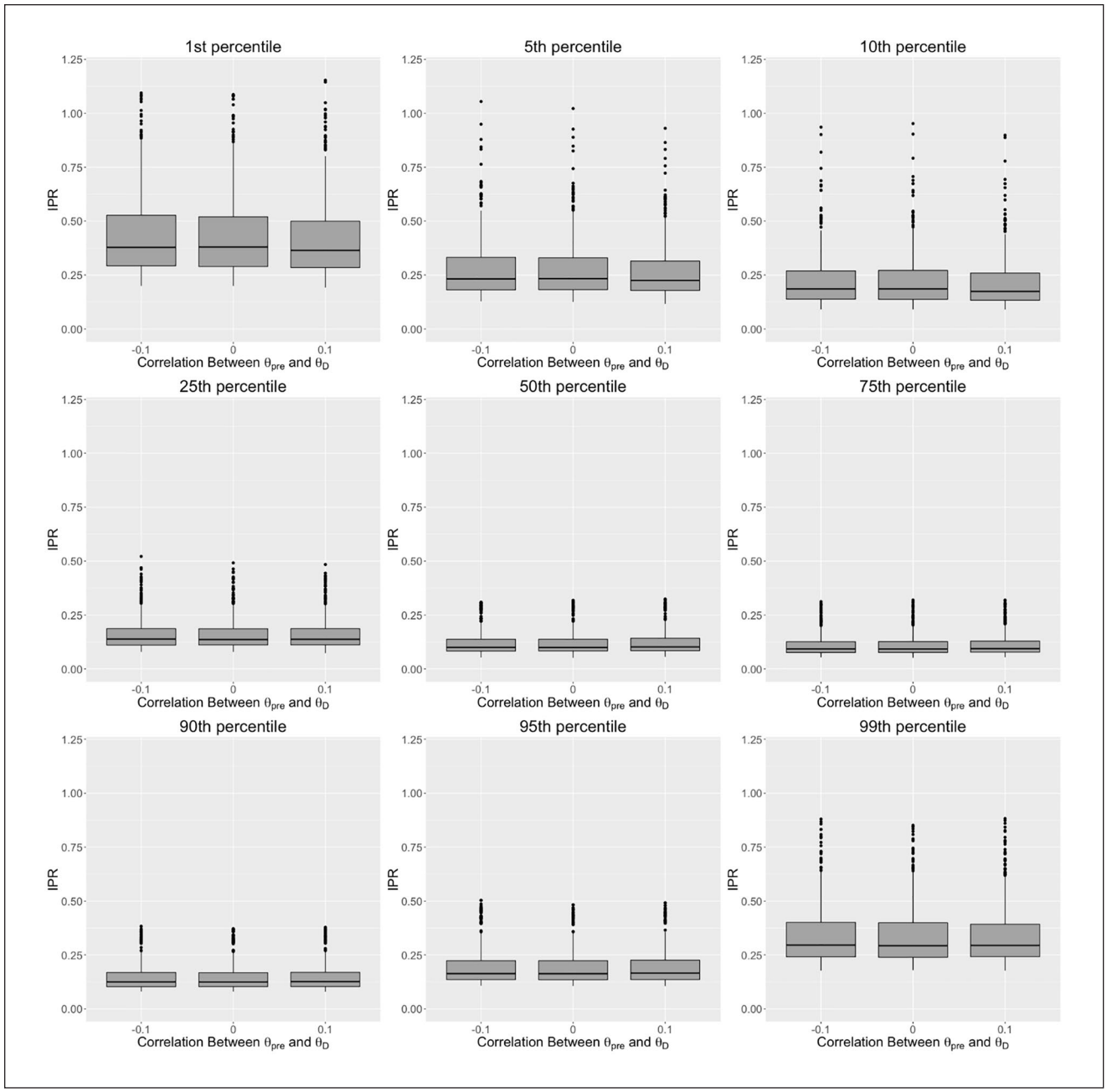

Figure 6. Relationship between correlation between θ_{pre} and θ_D and interpercentile range (IPR) for the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th percentiles generated by the *T* Scores for Change method.

In the simulation study, we assumed that θ_{pre}, X_{gender}, and X_{age} were independent from each other, but that θ_D was dependent on X_{gender} and X_{age}. It is common that change due to treatment is dependent on gender and age. For example, treatment effectiveness of substance use disorders is associated with gender (Polak et al., 2015). Older people are less responsive to medication and psychotherapy for anxiety disorders than younger people (Wetherell et al., 2013). θ_{pre} may also depend on X_{gender} and X_{age}. The presence of correlation between pretest score and covariates such as gender and age does not affect comparisons between individuals with the same covariate values, but it does complicate comparisons between individuals differing in gender or age. Specifically, the two norming methods discussed in this manuscript give different regression weights for the covariates if they correlate with the pretest score. That in turn can result in different orderings of

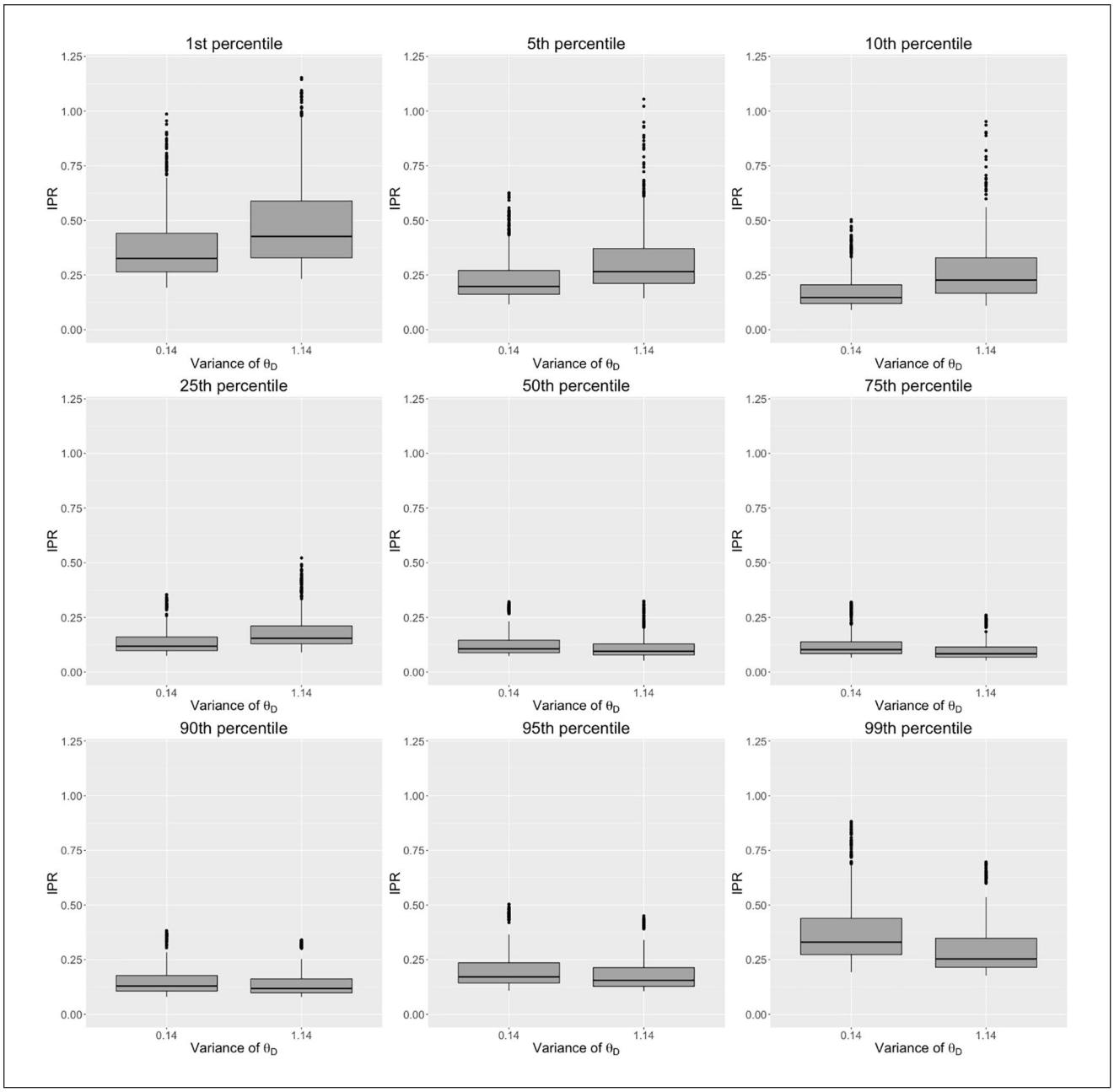

Figure 7. Relationship between variance of θ change and interpercentile range (IPR) for the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th percentiles generated by the *T* Scores for Change method.

individuals with respect to their standardized residual, even if the individuals have the same pretest score and also have the same posttest score. However, the two methods can also give different orderings of individuals of the same gender and age if these individuals differ in pretest score. This is because the two methods give a different regression weight to the pretest score when adjusting the posttest score.

The two regression-based norming methods for change scores assume a linear relationship between observed change scores and covariates. In the simulation study, we used a linear model to model the relation between θ_D and covariates X_{gender} and X_{age}, and we generated the item-score data by means of the GRM, which posits a nonlinear relation between item scores and latent variables. Several authors have noticed that the resulting test score correlates high in the nineties with the latent variable and, as a result, the use of the GRM in generating item-score data sets is common practice in psychometrics and has also been used

in studying regression-based norming methods (Oosterhuis et al., 2016). However, we believe that it may be interesting to examine the potential influence of the nonlinearity caused by the GRM in simulation studies on norming methods.

In recent years, new methods have been applied to norming test scores. For example, the Box-Cox Power Exponential model, which is based on the Generalized Additive Model for Location, Scale, and Shape (GAMLSS; Rigby & Stasinopoulos, 2005), has been used to norm IQ scores (Voncken et al., 2017). GAMLSS-based models allow for the flexible specification of a raw test score distribution, including the mean, variance, skewness, and kurtosis and therefore may be better suited in practice where the empirical item-score data set does not show desirable features such as a normal distribution of residuals. Thus, for future research, it might be interesting to investigate how new methods, such as the GAMLSS-based models, can be used to norm change scores.

Declaration of Conflicting Interests

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

Funding

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

ORCID iD

Zhengguo Gu https://orcid.org/0000-0002-5953-6673

Supplemental Material

Supplemental material for this article is available online.

References

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.

Angoff, W. H. W. (1984). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Princeton, NJ: Educational Testing Service.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in measuring change* (pp. 3-20). Madison, WI: University of Wisconsin Press.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, *101*(1), 147-158. https://doi.org/10.1037/0033-2909.101.1.147

Caruso, J. C. (2004). A comparison of the reliabilities of four types of difference scores for five cognitive assessment batteries. *European Journal of Psychological Assessment*, *20*(3), 166-171. https://doi.org/10.1027/1015-5759.20.3.166

Castro-Schilo, L., & Grimm, K. J. (2018). Using residualized change versus difference scores for longitudinal research. *Journal of Social and Personal Relationships*, *35*(1), 32-58. https://doi.org/10.1177/0265407517718387

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155-159. https://doi.org/10.1037/0033-2909.112.1.155

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. https://doi.org/10.1007/BF02310555

Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin*, *74*(1), 68-80. https://doi.org/10.1037/h0029382

De Vroege, L., Emons, W. H. M., Sijtsma, K., & Van der Feltz-Cornelis, C. M. (2018). Psychometric properties of the Bermond-Vorst Alexithymia Questionnaire (BVAQ) in the general population and a clinical population. *Frontiers in Psychiatry*, *9*, Article 111. https://doi.org/10.3389/fpsyt.2018.00111

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods*, *12*(1), 105-120. https://doi.org/10.1037/1082-989X.12.1.105

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, *58*(3), 357-381. https://doi.org/10.1177/0013164498058003001

Gorsuch, R. L. (1983, August). *The theory of continuous norming* [Paper presentation]. 91st Annual Convention of the American Psychological Association. Anaheim, CA.

Gu, Z., Emons, W. H. M., & Sijtsma, K. (2018). Review of issues about classical change scores: A multilevel modeling perspective on some enduring beliefs. *Psychometrika*, *83*(3), 674-695. https://doi.org/10.1007/s11336-018-9611-3

Hertzog, C., von Oertzen, T., Ghisletta, P., & Lindenberger, U. (2008). Evaluating the power of latent growth curve models to detect individual differences in change. *Structural Equation Modeling*, *15*(4), 541-563. https://doi.org/10.1080/10705510802338983

Jabrayilov, R., Emons, W. H. M., & Sijtsma, K. (2016). Comparison of classical test theory and item response theory in individual change assessment. *Applied Psychological Measurement*, *40*(8), 559-572. https://doi.org/http://doi.org/10.1177/0146621616664046

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1/2), 81-93. https://doi.org/10.2307/2332226

Kruyen, P. M., Emons, W. H. M., & Sijtsma, K. (2013). On the shortcomings of shortened tests: A literature review. *International Journal of Testing*, *13*(3), 223-248. https://doi.org/10.1080/15305058.2012.703734

Linn, R. L., & Slinde, J. A. (1977). The determination of the significance of change between pre- and posttesting periods. *Review of Educational Research*, *47*(1), 121-150. https://doi.org/10.3102/00346543047001121

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, *68*(5), 304-305. https://doi.org/10.1037/h0025105

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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*(6), 921-943. https://doi.org/10.1177/0013164402238082

Manning, W. H., & Dubois, P. H. (1962). Correlational methods in research on human learning. *Perceptual and Motor Skills*, *15*(2), 287-321. https://doi.org/10.2466/pms.1962.15.2.287

McSweeny, A. J., Naugle, R. I., Chelune, G. J., & Lüders, H. (1993). "T Scores for Change": An illustration of a regression approach to depicting change in clinical neuropsychology. *Clinical Neuropsychologist*, *7*(3), 300-312. https://doi.org/10.1080/13854049308401901

Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care*, *41*(5), 582-592. https://doi.org/10.1097/01.MLR.0000062554.74615.4C

O'Connor, E. F. (1972). Extending classical test theory to the measurement of change. *Review of Educational Research*, *42*(1), 73-97. https://doi.org/10.3102/00346543042001073

Oosterhuis, H. E. M. (2017). *Regression-based norming for psychological tests and questionnaires* [Doctoral dissertation]. Tilburg University, the Netherlands.

Oosterhuis, H. E. M., Van der Ark, L. A., & Sijtsma, K. (2016). Sample size requirements for traditional and regression-based norms. *Assessment*, *23*(2), 191-202. https://doi.org/10.1177/1073191115580638

Overall, J. E., & Woodward, J. A. (1975). Unreliability of difference scores: A paradox for measurement of change. *Psychological Bulletin*, *82*(1), 85-86. https://doi.org/10.1037/h0076158

Polak, K., Haug, N. A., Drachenberg, H. E., & Svikis, D. S. (2015). Gender considerations in addiction: Implications for treatment. *Current Treatment Options in Psychiatry*, *2*(3), 326-338. https://doi.org/10.1007/s40501-015-0054-5

Raykov, T. (1993). A structural equation model for measuring residualized change and discerning patterns of growth or decline. *Applied Psychological Measurement*, *17*(1), 53-71. https://doi.org/10.1177/014662169301700110

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(3), 507-554.

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, *92*(3), 726-748. https://doi.org/10.1037/0033-2909.92.3.726

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, Monograph, No. 17.

Schünemann, H. J., & Guyatt, G. H. (2005). Commentary: Goodbye M(C)ID! Hello MID, where do you come from? *Health Services Research*, *40*(2), 593-597. https://doi.org/DOI 10.1111/j.1475-6773.2005.0k375.x

Sijtsma, K., & Emons, W. H. M. (2011). Advice on total-score reliability issues in psychosomatic measurement. *Journal of Psychosomatic Research*, *70*(6), 565-572. https://doi.org/10.1016/J.JPSYCHORES.2010.11.002

Van Breukelen, G. J. P. (2013). ANCOVA versus CHANGE from baseline in nonrandomized studies: The difference. *Multivariate Behavioral Research*, *48*(6), 895-922. https://doi.org/10.1080/00273171.2013.831743

Van Breukelen, G. J. P., & Vlaeyen, J. W. S. (2005). Norming clinical questionnaires with multiple regression: The Pain Cognition List. *Psychological Assessment*, *17*(3), 336-344. https://doi.org/10.1037/1040-3590.17.3.336

Van der Elst, W., Van Boxtel, M. P. J. J., Van Breukelen, G. J. P. P., & Jolles, J. (2008). Detecting the significance of changes in performance on the Stroop Color-Word Test, Rey's Verbal Learning Test, and the Letter Digit Substitution Test: The regression-based change approach. *Journal of the International Neuropsychological Society*, *14*(1), 71-80. https://doi.org/10.1017/S1355617708080028

Voncken, L., Albers, C. J., & Timmerman, M. E. (2017). Model selection in continuous test norming with GAMLSS. *Assessment*, *26*(7), 1329-1346. https://doi.org/10.1177/1073191117715113

Werker, J. F., & Lalonde, C. E. (1988). Cross-language speech perception: Initial capabilities and developmental change. *Developmental Psychology*, *24*(5), 672-683. https://doi.org/10.1037/0012-1649.24.5.672

Wetherell, J. L., Petkus, A. J., Thorp, S. R., Stein, M. B., Chavira, D. A., Campbell-Sills, L., Craske, M. G., Sherbourne, C., Bystritsky, A., Sullivan, G., & Roy-Byrne, P. (2013). Age differences in treatment response to a collaborative care intervention for anxiety disorders. *British Journal of Psychiatry: The Journal of Mental Science*, *203*(1), 65-72. https://doi.org/10.1192/bjp.bp.112.118547

Willett, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education*, *15*(1), 345-422. https://doi.org/10.3102/0091732X015001345

Williams, R. H., & Zimmerman, D. W. (1977). The reliability of difference scores when errors are correlated. *Educational and Psychological Measurement*, *37*(3), 679-689.

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement*, *20*(1), 59-69. https://doi.org/10.1177/014662169602000106

Zachary, R. A., & Gorsuch, R. L. (1985). Continuous norming: Implications for the WAIS-R. *Journal of Clinical Psychology*, *41*(1), 86-94. https://doi.org/10.1002/1097-4679(198501)41:1<86::AID-JCLP2270410115>3.0.CO;2-W

Zimmerman, D. W., & Williams, R. H. (1982a). Gain scores in research can be highly reliable. *Journal of Educational Measurement*, *19*(2), 149-154. https://doi.org/10.1111/j.1745-3984.1982.tb00124.x

Zimmerman, D. W., & Williams, R. H. (1982b). On the high predictive potential of change and growth measures. *Educational and Psychological Measurement*, *42*(4), 961-968. https://doi.org/10.1177/001316448204200403