Regression-Based Norming for Psychological Tests and Questionnaires

Proefschrift ter verkrijging van de graad van doctor
aan Tilburg University
op gezag van de rector magnificus,
prof. dr. E. H. L. Aarts,
in het openbaar te verdedigen ten overstaan van een
door het college voor promoties aangewezen commissie
in de aula van de Universiteit

op woensdag 12 april 2017 om 16:00 uur

door

Hannah Elisabeth Maria Oosterhuis
geboren op 17 mei 1989 te Eindhoven

Promotores:

Prof. dr. K. Sijtsma

Prof. dr. L. A. van der Ark

Overige leden van de Promotiecommissie:

Prof. dr. J. K. L. Denollet

Prof. dr. G. J. P. van Breukelen

Prof. dr. M. Ph. Born

Dr. W. van der Elst

Dr. W. H. M. Emons

Table of Contents

Chapter 1

Introduction

Everyday psychological test and questionnaires are used to make important decisions in the lives of individuals, such as whether or not to hospitalize a mental patient, admit a prospective student to an educational program, or hire an applicant for a job. Norms are required to allow for a meaningful interpretation of an individual's raw test score. For example, knowing that a person has answered 37 out of 50 questions on a test correctly is uninformative, unless we are informed about the relative position of this person in a group to which we want to compare the person. If we know that a score of 37 is equal to or higher than the raw scores of 90% of individuals in the norm sample, we can infer that the individual has obtained a relatively high score on the test. However, if only 10% of individuals in the norm sample scored equal to or lower than 37, we would infer that a score of 37 is relatively low.

Usually, norms are estimated based on the raw scores of a group of people, the norm sample, who take the test during the test's construction phase. In addition to the mean and the standard deviation, test constructors might provide percentile ranks, stanines, standard scores or normalized standard scores for each possible raw score in the norm sample. Alternatively, test constructors can provide the raw scores that are associated with specific norm statistics, such as specific percentile ranks or standard scores. It is the test constructor's task to provide accurate norms, because without such norms test results cannot be meaningfully interpreted, rendering the test practically useless.

It is well known that a norm sample should be both representative of the population and large enough to ensure that the norms have small standard errors (e.g., Kline, 2000). Representativeness refers to the composition of the sample, which should accurately reflect the subgroup structure of the population that is relevant for the abilities or traits measured by the test. For example, covariates such as age group, urbanization grade, educational level, socioeconomic status, and religious affiliation each provide a different subgroup structure. Norms are estimated for each of the subgroups separately. However, not all covariates need be relevant for specific measurements. For example, religious

affiliation may not be relevant for the measurement of children's intelligence and as a result, it may be ignored when drawing a sample during the test construction phase.

When constructing a test, sample size refers to both the total norm sample and the size of the different subgroups that were formed based on relevant covariates. Each subgroup should be large enough to ensure the estimation of precise norms. The larger the number of subgroups, the larger the total norm sample. From a practical angle, collecting large and representative samples for each norm subgroup is difficult, time consuming, and costly, which can easily be resolved by limiting the number of subgroups. However, a problem arises when the covariates are continuous. For example, if age is a covariate, subgroups may be based on arbitrary age categories. To limit the number of subgroups in the total norm sample, age categories need to contain a wide range of ages. However, an age difference of just a few days might result in an entirely different interpretation of test scores of individuals who are close to the boundaries of these age categories. To prevent this ambiguity, test constructors can select a larger number of age categories, but this larger number requires a larger norm sample.

Given the overwhelming importance of norms in practical test use, it is surprising that so little research has been done after the question how to adequately balance the precision of norms and the size of the norm sample. This proposal provides an attempt to help solve the problem. Zachary and Gorsuch (1985; also see Bechger, Maris, & Hemker, 2009; Van Breukelen & Vlaeyen, 2005) proposed a more efficient norming procedure, called continuous norming or regression-based norming. They used covariates, such as age and gender, as independent variables in a linear regression equation to predict the raw test score. They used the corresponding empirical distribution of standardized residuals to estimate the norms. The Commissie Testaangelegenheden Nederland (COTAN; Evers et al., 2009) have provided Dutch test constructors sample-size guidelines for regression-based norming, but noted that research on this topic is badly needed. Hence, a systematic investigation of the precision of regression-based norming is required.

Typically, standard errors (SEs) and confidence intervals (CIs) based on these standard errors quantify the estimation precision of statistics. However, for most norm statistics, SEs are currently not available. Methods and software to derive SEs of norms from the test data should be developed and made available, because they allow test

constructors to investigate and demonstrate the precision of norms. SEs and CIs can also be used to determine whether estimated norms are precise enough for the intended use of the test. For example, if a test is used to make important decisions about individual test takers, the test's norms should be estimated with higher precision than for tests that are used for less important decisions (Evers et al., 2009). Furthermore, SEs and CIs allow for the comparison of the precision of different estimation methods under different circumstances.

The purported efficiency of the regression-based method may easily persuade test constructors and test publishers to abandon traditional norming, because the required subgroup sample size and the total sample size are smaller for regression-based norming. However, a useful application of regression-based norming requires that the assumptions of the linear regression model are consistent with the data. The assumptions include normality, linearity, and homoscedasticity (e.g., Fox, 1997), and several authors have argued that violations may lead to seriously biased norms (Semel, Wiig & Secord, 2004; Tellegen & Laros, 2011; Van der Elst et al., 2010). Hence, test constructors must investigate whether the regression-based norming procedure is robust against violations of the assumptions. Also see Semel et al. (2004) and Tellegen and Laros (2011), who suggested regression-based norming is not robust but without seriously investigating this claim. Results may also suggest the conditions under which regression-based norming produces unbiased results.

This proposal reports results of simulation studies with respect to the use of the linear regression model to estimate regression-based norms for psychological tests and questionnaires. In particular, the following research questions were addressed:

1. Given a particular sample size, does regression-based norming produce more precise estimates than traditional norming? (Chapter 2)
2. Can SEs be derived for the test-score standard deviation, percentiles rank scores, the boundaries of the stanines, and *Z*-scores? (Chapter 3)
3. What are the effects of violations of the assumptions of the linear regression model on the bias and the precision of regression-based norm estimates and the corresponding CIs? (Chapter 4)
4. In practice, which procedure should one follow when estimating regression-based norms? (Chapter 5)

Answers to these questions provide information about the correct application of regression-based norming, while considering the limitations of the method. Knowledge of the strengths and weaknesses of the regression-based method can help test constructors choose between different methods to estimate norms. Furthermore, the correct application of regression-based norming can reduce bias in interpreting test scores that would otherwise result from the arbitrary categorization of continuous covariates. In addition, instead of determining whether a norm sample is large enough, SEs and CIs for regression-based norms allow for a statistical basis that can be used to judge the precision of norm estimates. These SEs and CIs can also be used to compare norm estimation methods under different circumstances. Finally, by considering sampling error, decisions based on the comparison of test scores to regression-based norms are less likely to be erroneous.

Overview of the thesis

In Chapter 2, we compared the precision of norm estimates based on the traditional method and the regression-based method. The traditional method consisted of the division of the total sample into eight subgroups using age and gender as covariates, whereas the regression-based method consisted of a linear regression model in which raw test scores were predicted using age and gender. Using simulated data, we compared the sampling distribution of percentile estimates based on the traditional method and the regression-based method. The two norm estimation methods were compared for different test lengths (i.e., 10, 50 or 100 items), numbers of answer categories (i.e., 2 or 5), sample sizes (i.e.., ranging from 100 – 10,000), and strength of covariate effects (i.e., small, medium, or large).

Chapter 3 is dedicated to the derivation of SEs and CIs for the standard deviation, percentile rank scores, stanine boundaries, and *Z*-scores under the mild assumption that the raw test scores follow a multinomial distribution. The general framework used to derive the SEs consisted of two steps. The first step was to write the norm statistic as a function of the frequencies of the raw scores. In the second step, the delta method was used to approximate the variance of the norm statistic. An SPSS macro and *R* script are provided to guarantee the procedure to obtain SEs and CIs is easily accessible for researchers.

In Chapter 4, we investigated the effect of violations of the assumptions of the linear regression model on regression-based norm estimates. Using simulated data, we compared the bias and the precision of regression-based percentile estimates for violations of

linearity, independence between covariates and the residual, and homoscedasticity of the residual variances to percentile estimates based on a regression model without violations. The bias and the precision of the percentile estimates was investigated for different conditions of violation strength (i.e., weak, medium, or strong) and sample size (i.e., ranging from 100 – 5,000).

Chapter 5 uses example data to describe a procedure to obtain unbiased, precise regression-based norms. Topics discussed in this procedure include determining which covariates to include in the regression model, how violations of the model assumptions can be detected, and how sampling error of the norm statistics can be quantified and presented.

Chapter 2

Sample Size Requirements for Traditional and Regression-Based Norms

Abstract

Test norms enable determining the position of an individual test taker in the group. The most frequently used approach to obtain test norms is traditional norming. Regression-based norming may be more efficient than traditional norming and is rapidly growing in popularity, but little is known about its technical properties. A simulation study was conducted to compare the sample-size requirements for traditional and regression-based norming by examining the 95% interpercentile ranges for percentile estimates as a function of sample size, norming method, size of covariate effects on the test score, test length, and number of answer categories in an item. Provided the assumptions of the linear regression model hold in the data, for a subdivision of the total group into eight equal-size subgroups we found that regression-based norming requires samples 2.5 to 5.5 times smaller than traditional norming. Sample-size requirements are presented for each norming method, test length, and number of answer categories. We emphasize that additional research is needed to establish sample-size requirements when the assumptions of the linear regression model are violated.

This chapter has been published as Oosterhuis, H. E. M., Van der Ark, L. A., & Sijtsma, K. (2016). Sample size requirements for traditional and regression-based norms. *Assessment, 23,* 191-202.

2.1 Introduction

Tests are omnipresent in psychological research and in clinical, personality, health, medical, developmental, and personnel psychology practice. In research, tests provide measures of abilities, traits, and attitudes that are used as variables in regression models, factor models, structural equation models, and other statistical models used for testing hypotheses about behavior, and also in experiments as dependent variables. In practice, test scores may be used to diagnose patients for pathology treatment and couples for marriage counseling; to provide advice to people suffering from eating disorder, coronary patients coping with anxiety, and children suffering from developmental problems; and to predict job success for job applicants in industry, commercial organizations, education, and government. This study focuses on tests used in psychological practice for individual measurement, and searches for the smallest sample size allowing the precise determination of an individual's test score relative to the population to which (s)he belongs; this is the norming problem.

Norm scores are helpful for interpreting test performance. For example, an 8-year old boy was presented the Letter Digit Substitution Test (LDST; Jolles, Houx, Van Boxtel, & Ponds, 1995) and made 15 correct substitutions in 60 seconds, resulting in a test score of 15. The test score is not informative of his relative information processing ability unless one knows that 22% of his peers have a test score lower than 15; this information suggests that his ability is within normal limits (Van der Elst, Dekker, Hulst, & Jolles, 2012). Test-score distributions often differ between age groups, education-level groups, and so on. Test constructors regularly construct norm distributions for different subgroups. For example, compared to women men underreport depressive symptoms (Hunt, Auriemma, & Cashaw, 2003), which necessitates different norms for men and women. Norms are often presented as percentiles or are derived from standard scores (Kline, 2000, p. 59-63).

Two norming approaches are available (e.g., Bechger, Hemker, & Maris, 2009; Evers, Lucassen, Meijer, & Sijtsma, 2009). The most frequently used traditional norming approach entails estimating separate test-score distributions for different subgroups. Regression-based norming entails, for example, employing a regression model in which covariates are used to estimate a norm distribution. Compared to traditional norming, regression-based

norming is expected to require a smaller sample to obtain equally precise norms (Bechger et al., 2009).

The goals of this study were to investigate whether, given a particular sample size, regression-based norming produces more precise estimates than traditional norming, and for both methods to determine the minimally required sample sizes to obtain acceptable precision of the norm scores. The expected pay-off was to provide test constructors with reliable advice about minimum sample-size requirements for test-score norming and to suggest how to obtain more precise norms using regression-based norming rather than traditional norming.

This article is organized as follows. First, we explain traditional norming and regression-based norming. Next, we present the results of a simulation study that suggests the required minimum sample sizes to obtain precise norms for both norming approaches. Finally, we discuss practical implications and recommendations for future research.

2.2 Methods for Norming

Two methods for obtaining norms are available: traditional norming and regression-based norming. For both norming methods we discuss the selection of relevant covariates and their use in the norm estimation process. We also discuss which norm statistics are usually presented, and the advantages and disadvantages of both norming methods.

2.2.1 Traditional Norming

Traditional norming uses one or more covariates to define relevant subgroups and estimates the test-score distribution separately for each subgroup.

Selection and incorporation of covariates. Four strategies use the following criteria to select covariates: (1) statistical significance, (2) effect-size assessment, (3) statistical significance and effect-size assessment, and (4) stratification variables.

Statistical significance. Covariates can be tested for statistical significance. For example, covariates correlating significantly with the test score are selected for dividing the sample into subgroups (Grande, Romppel, Glaesmer, Petrowski, & Herrmann-Lingen, 2010). Similarly, significance tests based on analysis of variance (ANOVA), regression analysis, or Pearson's chi-squared test can be used to select covariates (Aardoom, Dingemans, Slof Op 't Land, & Van Furth, 2012; Mond , Hay, Rodgers, & Owen, 2006; Pedraza, Lucas, Smith, Petersen, Graff-Radford, & Ivnik, 2010).

Effect-size assessment. Crawford, Henry, Crombie, and Taylor (2001) used effect size to select covariates for determining the subgroups for the Hospital Anxiety and Depression Scale (HADS; Zigmond & Snaith, 1983). The authors found that males had a higher mean test score than females, and they also found modest positive correlations between the test score and age, level of education and social class. However, the authors ignored the modest correlations and only used gender to define subgroups. Furthermore, to define relevant subgroups, Crawford, Cayley, Lovibond, Wilson, and Hartley (2011) used only those covariates that correlated at least .20 with the test score, regardless of statistical significance.

Statistical significance and effect-size assessment. The information from significance testing and effect size can be combined to select covariates. For example, Glaesmer et al. (2012) used ANOVA to determine whether age and gender influenced test scores on the revised version of the Life Orientation Test Revised (LOT-R; Scheier, Carver, & Bridges, 1994). They only selected covariates that were statistically significant (ANOVA) and had at least a medium effect size (Cohen, 1992; Cohen's $d > .50$).

Stratification variables. In some studies, the stratification variables that were used to establish representativeness of the normative sample were also used as covariates for norming. For example, Krishnan, Sokka, Häkkinen, Hubert, and Hannonen (2004) used age and gender to select participants in the normative sample and subsequently used these stratification variables to define norm subgroups.

Estimation of norm statistics. Norm statistics are used to characterize the distribution of the test performance in each norm group. Test performance can be distinguished by the raw score, which is the sum of the item scores, and the test score, which is a transformation of the raw score meant to enhance the interpretation of test performance. Sometimes, test score and raw score coincide, for example, when the number-correct score on an educational test is reported together with the pass-fail score, which serves to interpret the raw score. Many transformations of raw scores to test scores exist, and these transformed test scores often serve as norm scores. Examples are standard scores and normalized standard scores (Kline, 2000, pp. 59-63), T-scores and stanines. The most frequently used transformation is the percentile score, defined as the percentage of individuals in the norm group who have the same raw score as a particular individual or a

lower raw score. For example, Crawford et al. (2001) presented gender-corrected percentiles for the HADS corresponding to each of the raw scores test takers can acquire. Also refer to the Wechsler Individual Achievement Test third edition (WIAT III; Wechsler, 2009), the Wide Range Achievement Test third edition (WRAT III; Wilkinson, 1993), and the Bender Visual-Motor Gestalt Test second edition (BVMG II; Brannigan & Decker, 2003).

Advantages and disadvantages. Traditional norming is simple. Norm statistics can be computed directly from the distribution of the test scores in each of the norm groups. The greatest disadvantage of traditional norming is that continuous covariates, such as age, have to be divided arbitrarily into mutually exclusive and exhaustive categories, which define separate norm groups. As a result of the arbitrariness, different choices of age categories can change the interpretation of an individual's test performance, depending on the norm group to which the individual is assigned (Parmenter, Testa, Schretlen, Weinstock-Guttman, & Benedict, 2009). A straightforward correction of the bias is to define more categories, but this also introduces smaller category sample sizes thus producing norms that have lower precision.

2.2.2 Regression-Based Norming

Selection and incorporation of covariates. Zachary and Gorsuch (1985) proposed linear regression to circumvent having to categorize continuous covariates; hence, the name regression-based norming. The model regresses the test score on one or more relevant covariates. Four strategies are used to select covariates: (1) stepwise regression, (2) simultaneous regression, (3) correlational analysis, and (4) theory-based selection.

Stepwise regression. Stepwise regression analysis is the most frequently used approach to select covariates for regression-based norming. For neuropsychological tests, covariates often include age, gender, and education (Parmenter et al., 2010). First, all covariates that are expected to predict the test score are simultaneously included in the regression model. Second, of all predictors having insignificant regression coefficients ($p > \alpha$; p is the probability of exceedance, α is the significance level), the predictor having the greatest p-value ($p > \alpha$) is deleted from the model. Third, the model including the remaining predictors is re-estimated. Fourth, in the new model the predictor having the highest p-value greater than α is deleted from the model. The procedure is repeated until all remaining covariates have regression weights significantly different from zero ($p < \alpha$).

Stepwise regression has several drawbacks. First, the overall significance level cannot be controlled because in each step multiple comparisons have to be performed for identifying the covariates to be deleted. Second, covariates such as age, gender and SES may not be the best predictors of the test score but they may be selected by a complex procedure such as stepwise regression that easily capitalizes on chance and thus likely produces results that are not replicable (Derksen & Keselman, 1992; Leigh, 1988).

Van der Elst, Hoogenhout, Dixon, De Groot, and Jolles (2011) used stepwise regression to estimate regression-based norms for the Dutch Memory Compensation Questionnaire (MCQ). The authors performed several regression analyses using the MCQ scale scores as dependent variables, and age, squared age (Parmenter et al., 2010; Van Breukelen & Vlaeyen, 2005; Van der Elst, Dekker, et al., 2012; Van der Elst, Ouwehand et al., 2012), gender, and education as predictors. All predictors having $p > .01$ were subsequently deleted from the model. Other authors employing stepwise linear regression include Heaton, Avitable, Grant, and Matthews (1999), Van Breukelen and Vlaeyen, (2005), Van der Elst, Dekker, et al. (2012), Van der Elst, Ouwehand, et al. (2012), Llinàs-Reglà, Vilalta-Franch, López-Pousa, Calvó-Perxas, & Olmo (2013), Roelofs et al. (2013a), Roelofs et al. (2013b), Vlahou et al. (2013), and Goretti et al. (2014).

***Simultaneous regression*.** Another possibility is to start with the regression model that contains all covariates, simultaneously test the regression coefficients for significance, and retain only those for which $p < \alpha$ (e.g., Conti, Bonazzi, Laiacona, Masina, & Coralli, 2014; Shi et al., 2014; Van der Elst et al., 2013; Yang et al., 2012). Unlike stepwise regression, simultaneous regression is done only once and thus suffers less from chance capitalization. For both approaches, the effect of chance capitalization is smaller as the sample is larger.

Correlational analysis. *C*orrelational analysis entails the selection of all covariates that have a significant correlation with the test score into the regression model (e.g., Cavaco et al., 2013a, 2013b; Kessels, Montagne, Hendriks, Perrett, & de Haan, 2014; Van den Berg et al., 2009). Compared to regression analysis, the method ignores the correlation between covariates and may be expected to explain less variance in the test score.

***Theory-based selection*.** Finally, we mention the possibility of choosing predictors on the basis of substantive theories about the attribute the test measures and previous research (e.g., Berrigan et al., 2014; Parmenter et al., 2010; Smerbeck et al., 2011, Smerbeck

et al., 2012). The absence of well-articulated theories or well-informed expectations from previous research renders the approach problematic.

Estimation of norm statistics. Van Breukelen and Vlaeyen (2005; also, Van der Elst et al., 2011) proposed a five-step procedure to estimate regression-based norm statistics: (a) Including covariates into the regression model. Let $X_1, \ldots, X_K$ represent the K covariates of interest. Continuous covariates can be added directly to the model and categorical covariates are replaced by dummy variables (Hardy, 1993). (b) Computing the predicted test scores. Let Y_+ be the observed test score, and let $\hat{Y}_+$ be the predicted test score. Let β_0 be the intercept and let $\beta_1, \ldots, \beta_K$ be the regression coefficients; then the regression equation equals

$$\hat{Y}_+ = \beta_0 + \beta_1 X_1 + \ldots + \beta_K X_K. \quad (1)$$

(c) Computing the residuals. Residuals are defined as $E = Y_+ - \hat{Y}_+$. (d) Standardizing the residuals. Index i enumerates the observations in the sample. Residuals are standardized by dividing them by their standard error,

$$S_E = \sqrt{\frac{\sum_{i=1}^{N} E_i^2}{N-k-1}}. \quad (2)$$

(e) Using the distribution of the standardized residuals to estimate norm statistics. The cumulative empirical distribution of the standardized residuals is used to estimate the norm statistics.

Advantages and disadvantages. We do not reiterate the method-specific disadvantages mentioned but rather mention two method-transcending advantages regression-based norming has relative to traditional norming. First, continuous covariates do not have to be categorized; thus, one avoids arbitrary decisions. Second, the method uses the entire norming sample to estimate the regression model and the norm statistics; thus, it is more efficient. A drawback of regression-based norming is that failure of the assumptions (i.e., normally distributed errors, homoscedasticity of the error variances, and linearity) in the data may bias the norms (Van der Elst et al., 2011). Alternatively, nonlinear regression models, having less stringent assumptions, may be used to obtain regression-based norms (e.g., Semel, Wiig, & Secord, 2004; Tellegen & Laros, 2011).

2.2.3 Norm Estimation Precision

Norms such as percentiles are influenced by sampling fluctuation. The required precision for norm estimates depends on the importance of the decisions made on the basis of the test score (Evers et al., 2009, p. 22). As a rule, more important decisions require norms having higher precision. Evers et al. (2009) proposed practical sample-size guidelines for norm groups that provide guidance to Dutch test constructors for choosing a sample size but have an insufficient statistical basis. The American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) provided guidelines for test construction (AERA, APA, & NCME, 1999) but without sample size recommendations.

The purpose of the current study was: Given a certain sample size, to determine the precision of an estimated percentile score for either traditional norming or regression-based norming. We used a simulation study, which allowed us to obtain the sampling distribution of the percentile estimates, and to control for the characteristics of the tests for which the data were simulated. The factors we used in the simulation design were derived from a literature review.

2.3 Method

Literature Review

Test constructors have to make decisions about the number of items in the test, the number of answer categories per item and how they are scored, the size of the normative sample and the covariates to be collected. For the simulation study, we reviewed the literature for 65 tests the Dutch Committee on Tests and Testing (COTAN) assessed between 2008 and 2012 so as to derive realistic approximations to the number of items, et cetera. We used freely accessible test reviews from the COTAN database (Egberink, Janssen, & Vermeulen, 2014). We assumed frequency distributions of the test characteristics of interest (number of items, number of item scores, sample size, and type of covariates) are representative for tests used in other Western countries and thus did not pursue test reviews from other test databases (e.g., the Buros Center for Testing).

The test review showed that across tests the number of items ranged from 14 to 681 (mean = 131, $Q_1 = 44.5$, $Q_2 = 89.5$ (median), $Q_3 = 159.25$). Tests containing at least 100 items consisted of several subtests each measuring a unique psychological attribute. Items

in tests had two score categories (46.9% of the tests), 3 or 4 ordered scores (26.2%), 5 ordered scores (23.3%), or more than 5 ordered scores (3.6%). The normative sample size varied greatly across tests, ranging from 122 to 96,582 participants in the complete sample. Sixty-eight percent of the normative samples contained between 500 and 2,500 participants. The covariates that were most often used to define norms were age (36.2% of the tests), gender (33.3%), and education level/ job position (30.4%). Approximately 40% of the tests were targeted at elementary school children between 4 and 12 years of age.

Population Model

The population model used to simulate respondents' test scores contains a dichotomous covariate (denoted X_1) representing gender and a continuous covariate (denoted X_2) representing age that are independent of each other. Both covariates were related to the attribute the test measures; the attribute was represented by a latent variable denoted θ. Latent variable θ determined test score Y_+; see Figure 2.1. Let N denote the size of the total normative sample. We simulated item scores and test scores as follows.

First, each of the N simulated participants received scores for X_1 and X_2. Scores on X_1 (males = 0, females = 1) were randomly sampled from a Bernoulli distribution with probability $p = .5$. Scores on X_2 were randomly sampled from the uniform distribution on the interval [4, 12].

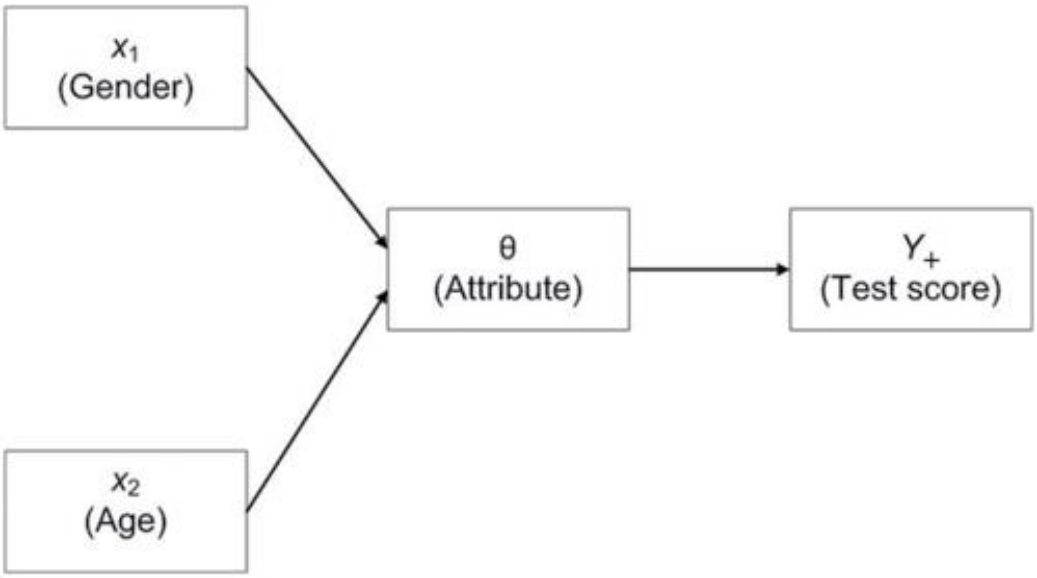

Figure 2.1. Population model for simulating test score (Y_+) based on latent variable (θ) and covariates (X_1, X_2).

Second, for each participant a θ score was randomly drawn from a normal distribution with mean $E(\theta|X_1, X_2)$, and unit variance, so that

$$E(\theta|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \quad (3)$$

thus assuming θ depends on covariates X_1 and X_2. The regression parameters β_0, β_1, and β_2 were chosen such that the squared multiple correlation (R^2) between θ and the covariates was either equal to 0, .065, .13 or .26. These values correspond to an absent, small, medium or large effect of covariates on θ, respectively (Cohen, 1992; $.02 \leq R^2 < .13$ is small, $.13 \leq R^2 < .26$ is medium, and $R^2 \geq .26$ is large). The covariates were uncorrelated and explained an equal portion of the variance of θ. As a result of the dummy coding, we have $E(\theta|X_1 = 0) < E(\theta|X_1 = 1)$ if $R^2 > 0$.

Third, for each of the participants an item-score vector was generated using the graded response model (GRM; Samejima, 1969). The simulated item scores are discrete; hence, the resulting test scores are also discrete and have a known score range based on the number of items and the number of item scores. Let the test consist of *J* items indexed *j*. Item scores are denoted Y_j, and items are scored $y = 0, \ldots, m$. Let α_j denote the discrimination parameter of item *j*, and let λ_{jy} denote the location parameter of score y of item *j*. The GRM is defined as

$$P(Y_j \geq y|\theta) = \frac{\exp[\alpha_j(\theta - \lambda_{jy})]}{1 + \exp[\alpha_j(\theta - \lambda_{jy})]}.$$

It may be noted that $P(Y_j \geq y|\theta) = 1$ for $y < 1$, and $P(Y_j \geq y|\theta) = 0$ for $y > m$. It follows that $P(Y_j = y|\theta) = P(Y_j \geq y|\theta) - P(Y_j \geq y + 1|\theta)$.

Table 2.1 shows the values for item parameters α_j and λ_{jy}. The range and the mean of the values were based on parameter estimates obtained from real psychological test data (Embretson & Reise, 2000, pp. 73, 101). Tests contained multiples of 10 items, and the item parameters were repeated so that the parameters of items 1, 11, and 21 were equal; the parameters of items 2, 12, and 22 were equal, et cetera. The item-score vectors were generated by means of random draws from a multinomial distribution with probabilities $P(Y_j = 0|\theta), \ldots, P(Y_j = m|\theta)$, for, $j = 1, \ldots, J$. The test score was obtained by means of $Y_+ = \sum Y_j$.

Table 2.1. *Graded Response Model Parameters for Dichotomous and Polytomous Items.*

		Dichotomous	Polytomous			
Item	α	λ	λ_1	λ_2	λ_3	λ_4
1	0.85	-2.25	-3.50	-1.10	-0.15	1.60
2	0.95	-1.75	-3.30	-1.00	-0.05	1.70
3	1.05	-1.25	-3.10	-0.90	0.05	1.80
4	1.15	-0.75	-2.90	-0.80	0.15	1.90
5	1.25	-0.25	-2.70	-0.70	0.25	2.00
6	1.35	0.25	-2.50	-0.60	0.35	2.10
7	1.45	0.75	-2.30	-0.50	0.45	2.20
8	1.55	1.25	-2.10	-0.40	0.55	2.30
9	1.65	1.75	-1.90	-0.30	0.65	2.40
10	1.75	2.25	-1.70	-0.20	0.75	2.50
Mean	1.30	0.00	-2.60	-0.65	0.30	2.05

Independent Variables

The five independent variables based on the literature review were the following:

1. *Test length* (*J*). The number of items was 10, 50, or 100.
2. *Number of item scores* (*m* + 1). The number of item scores was 2 (dichotomous items) or 5 (polytomous items).
3. *Sample size* (*N*). The 15 values for *N* were equal to 100, 500, 1,000, 1,500, 2,000, 2,500, 3,000, 3,500, 4,000, 5,000, 6,000, 7,000, 8,000, 9,000, and 10,000. The number of levels is relatively large so as to provide sufficient precision for determining sample-size recommendations.
4. *Covariate effects*. Covariates X_1 and X_2 had a multiple correlation with latent variable θ equal to 0 (no effect), .065 (small effect), .13 (medium effect), and .26 (large effect).
5. *Norming method*. Percentiles were estimated by means of the traditional norming method and the regression-based norming method.

Table 2.2 shows coefficient alpha (e.g., Cronbach, 1951) for each combination of test length, number of item scores, and size of covariate effect.

Table 2.2. *Summary of Simulated Test Scores (N = 1,000).*

Population Model			Test scores		
#Items	Item scores	R^2	Mean	SD	Coeff. Alpha
10	2	.00	4.8	2.1	.666
	2	.065	4.8	2.1	.667
	2	.13	4.7	2.1	.665
	2	.26	4.8	2.1	.666
	5	.00	21.1	7.0	.816
	5	.065	21.1	7.2	.816
	5	.13	21.5	7.0	.815
	5	.26	21.2	7.0	.815
50	2	.00	23.7	9.2	.911
	2	.065	23.3	9.1	.911
	2	.13	24.1	9.1	.911
	2	.26	23.7	9.1	.911
	5	.00	105.0	33.2	.957
	5	.065	106.9	31.3	.957
	5	.13	104.1	33.1	.957
	5	.26	107.4	32.5	.957
100	2	.00	48.5	17.9	.953
	2	.065	47.5	17.7	.953
	2	.13	46.6	17.5	.953
	2	.26	47.2	18.0	.954
	5	.00	209.0	63.7	.978
	5	.065	210.3	63.3	.978
	5	.13	213.5	66.9	.978
	5	.26	208.8	63.5	.978

Dependent Variables

The dependent variable was the precision of the estimates of the 50th, 75th, 90th, 95th, and 99th percentiles. Percentile values of 50, 75, 90, 95 and 99 are commonly presented as norms (Bride, 2007; Glaesmer et al., 2012; Krishnan et al., 2004; and Wizniter et al., 1992)

or cut-off scores in testing practice (Crawford & Henry, 2003; Crawford et al., 2001; Lee, Loring, & Martin, 1992; Mond et al. 2006; Murphy & Barkley, 1996; Posserud, Lundervold, & Gillberg, 2006; Van den Berg et al., 2009; Van Roy, Grøholt, Heyerdahl, & Clench-Aas, 2006; Wozencraft & Wagner, 1991). Based on the assumption that the sampling variance of the 1st, 5th, 10th, and 25th percentile is the same as that of the 99th, 95th, 90th, and 75th percentiles, respectively, we did not include the low percentiles in the study. The assumption is only valid if the distribution of test scores and residuals is symmetrical. Indeed we found that the scores in the norm groups and the residuals were approximately normally distributed for both norming methods.

Precision was operationalized as the 95% interpercentile range (IPR). IPR is the difference between the 97.5th percentile and the 2.5th percentile of an estimate's sampling distribution, here a percentile's sampling distribution. If percentile scores are estimated with higher precision, the IPR is smaller. We constructed the IPR of a particular percentile on the basis of 1,000 random samples.

Use of Y_+ would cause IPRs for tests with a larger number of items or with a larger number of item scores to be larger due to the larger range of X_1 and render results for different tests incomparable. Thus, for each of the simulated total normative samples, we used the corresponding mean and standard deviation to transform test score Y_+ into *Z*-scores. As a result, remaining differences between IPRs were due to a difference in precision rather than scale differences. For each of the conditions, Table 2 presents the mean and the standard deviation of test scores in a total normative sample of size *N* = 1,000.

To estimate the percentiles using the traditional norming approach, covariates $X_1$and X_2 were used to divide the total normative sample into eight separate norm groups. Scores on X_2 were divided into four age categories: $4 \leq X_2 < 6$ (first category), $6 \leq X_2 < 8$ (second category), et cetera. Given that scores 0 and 1 on X_1 had equal probabilities and scores on X_2 were draws from a uniform distribution the eight groups had the same size as the norm group for which norms were estimated. Hence, it sufficed to report results only for one group; we arbitrarily chose $X_1 = 0$ and X_2= $6 \leq X_2 < 8$ (second category).

To estimate percentiles based on the regression-based norming approach, $X_1$and X_2 served as independent variables in the linear regression model (Equation 1). The

standardized test score ($Z_{Y_+} = (Y_+ - \bar{Y}_+)/S_{Y_+}$) rather than Y_+ served as the dependent variable. We did not divide the residuals by their standard error (Equation 2). Using the standardized test score as the dependent variable and not standardizing the residuals has the advantage that the IPRs for both the regression-based approach and the traditional approach are expressed in the same metric

Analyses

First, for the 50th, 75th, 90th, 95th, and 99th percentiles we used an ANOVA to investigate the main effects and the two-way interaction effects on IPR that included sample size. Eta-squared (η^2) was used to interpret the effect sizes: $\eta^2 > .14$ (large effect), $\eta^2 > .06$ (medium), and $\eta^2 > .01$ (small) (Cohen, 1992). Let SS_{effect} be the sum of squares corresponding to a particular main or interaction effect that is of interest, and let SS_{total} be the total sum of squares, then η^2 for the effect equals

$$\eta^2 = \frac{SS_{\text{effect}}}{SS_{\text{total}}}$$

Each design cell contained one observation, which was the IPR based on 1,000 simulated samples.

Second, for each of the five percentiles we graphically displayed the IPR as a function of sample size. Separate curves were provided for each test characteristic that had a statistically significant ($p < .05$) effect that is at least small ($\eta^2 > .01$). Researchers can use the curves to determine the required sample size for their norming research given the desired precision of the percentile scores and the characteristics of the test.

Third, for each percentile we computed the ratio of the IPRs for traditional norming and regression-based norming, as a function of sample size. For given sample size, the ratio shows the precision of traditional norming relative to regression-based norming. For example, if for a given sample size the ratio equals 4, then the precision of regression-based norming is 4 times better than that of traditional norming.

2.4 Results

Analyses of Variance

For the 50th, 75th, 90th, 95th, and 99th percentile, Table 2.3 shows the effect size (η^2) corresponding to the main effects and the interaction effects on the IPRs.

Table 2.3. *Effect Sizes (η^2) Based on ANOVAs Performed on IPR of Percentiles.*

	Percentile				
	50	75	90	95	99
Main effects					
N	**.492****	**.460****	**.472****	**.509****	**.557****
Norming Method	**.253****	**.293****	**.271****	**.304****	**.303****
Effect of Covariates	.006**	.003**	.005**	.001**	.000
Answer Categories	.000	.001**	.008**	.007**	**.013****
Test Length	.000	.001**	.004**	.001**	.001
Interactions					
*N**Norming Method	**.205****	**.194****	**.186****	**.144****	**.091****
*N** Effect of Covariates	.003	.002	.002	.002	.002
*N**Answer Categories	.000	.000	.001	.002**	.001*
*N**Test Length	.003*	.000	.001	.000	.001
Complete Model	.963**	.954**	.950**	.970**	.969**

Note. ANOVAs = analyses of variance; IPR = interpercentile range. Effect sizes >.01 are in boldface.*$p < .05$. **$p < .01$.

Interaction effects. For each of the five percentiles, the interaction effect between sample size *N* and norming method on IPR was large ($\eta^2 > .14$). Thus, for traditional norming and regression-based norming the relationship between *N* and IPR is different. Alternatively, one could say that for a particular sample size the methods produce different IPRs. As the estimated percentile increases, the proportion of variance explained by the interaction decreases suggesting that for the different methods the difference between the IPRs depends less on *N* as the percentile is more extreme. The significance of the interaction term prohibits the interpretation of the main effects of sample size and norming method. All other interaction effects were negligible ($\eta^2 < .01$; Table 3); hence, they were ignored.

Main effects. For each of the five percentiles, sample size *N* and norming methods had large main effects ($\eta^2 > .14$). For the 99th percentile, the main effect of number of answer categories was small ($\eta^2 > .01$) but all other main effects were negligible ($\eta^2 < .01$).

The Relation Between Sample Size and IPR

For the 50th, 75th, 90th, 95th, and 99th percentile, figures 2.2 to 2.6 show the relationship between sample size *N* (horizontal axis) and IPR (vertical axis). The figures show two main results. First, for fixed *N*, regression-based norming produces a smaller IPR than traditional norming. Hence, regression-based norming is more efficient than traditional norming. The explanation is that regression-based norming estimates norms based on the entire sample, whereas traditional norming estimates norms in each separate subgroup. Second, for small sample sizes, the effect of increasing the sample size on IPR is large but this effect decreases rapidly as sample size increases. Similarly, for continuous variables, the standard error is inversely related to the square root of *N* (e.g., Mood, Graybill, & Boes, 1974, section VI-5); for our discrete data, figures 2.2 to 2.6 show a similar relationship.

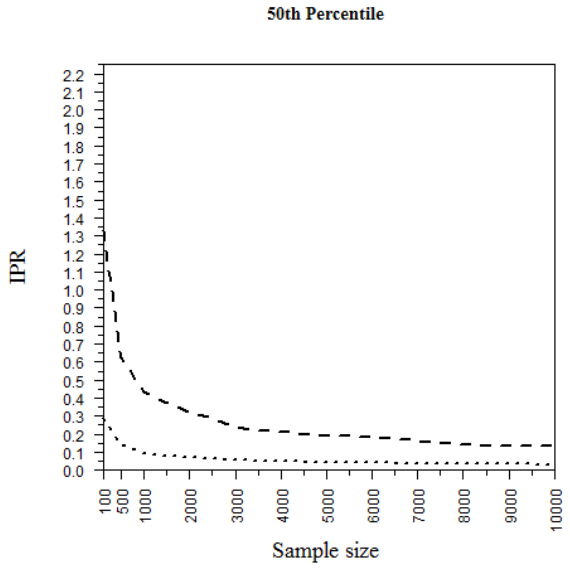

Figure 2.2. Interpercentile range for the 50th percentile estimate: traditional norming (dashed) and regression-based norming (dotted).

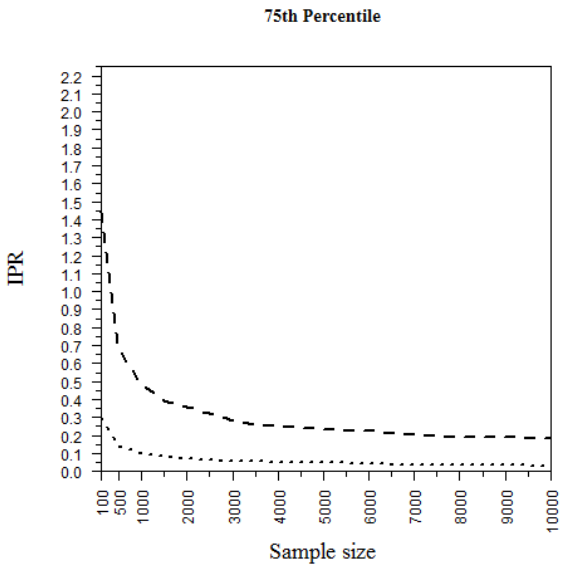

Figure 2.3. Interpercentile range for the 75th percentile estimate for traditional norming (dashed) and regression-based norming (dotted).

For the 50th, 75th, 90th and 95th percentile (see figures 1.2 to 1.5), no effects other than norming method were included, resulting in two curves. For the 99th percentile (Figure 1.6), for both norming procedures a fixed *N* produces a smaller IPR for polytomous-item tests than for dichotomous-item tests. Also, IPR increased as percentiles were more extreme. Hence, extreme percentiles require a larger sample size to obtain a required precision.

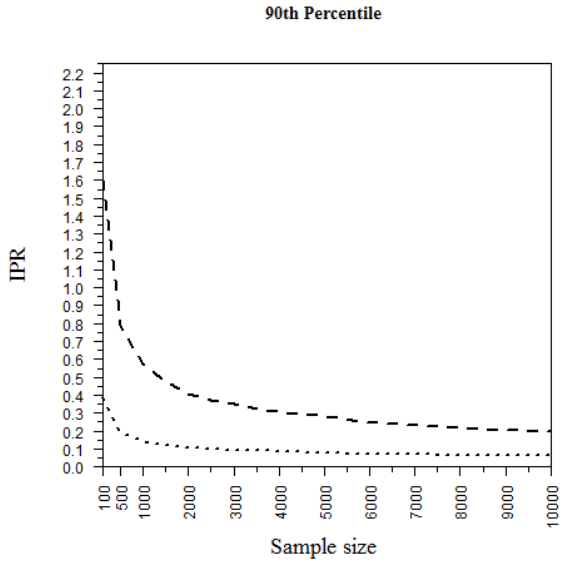

Figure 2.4. Interpercentile range for the 90th percentile estimate: traditional norming (dashed) and regression-based norming (dotted).

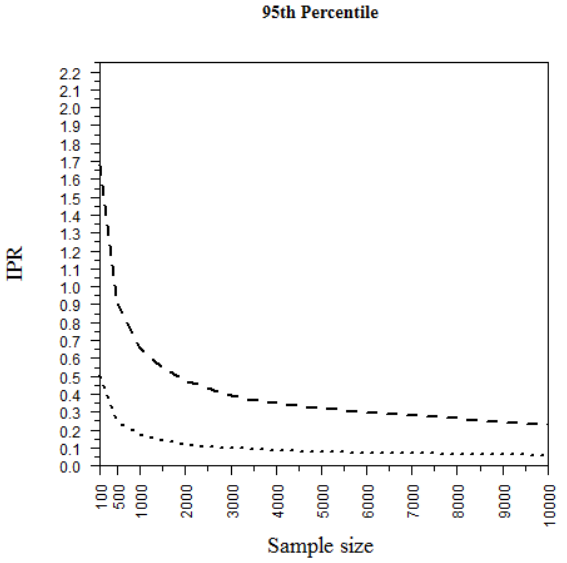

Figure 2.5. Interpercentile range for the 95th percentile estimate: traditional norming (dashed) and regression-based norming (dotted).

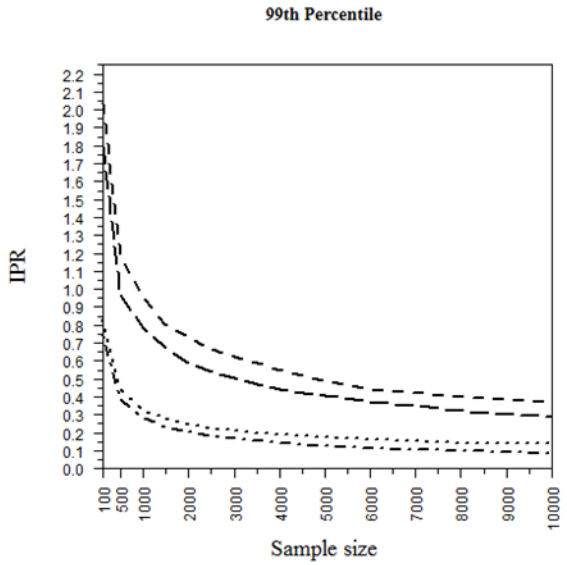

Figure 2.6. Interpercentile range for the 99th percentile estimate: traditional norming with dichotomous items (dashed), traditional norming with polytomous items (long dashed), regression-based norming with dichotomous items (dotted), and regression-based norming with polytomous items (dotted-dashed).

IPR Ratio of Traditional Norming Versus Regression-Based Norming

Table 2.4 shows a summary of the ratios of the IPR of traditional norming and regression-based norming. For each percentile and each *N*, estimation precision is higher for regression-based norming, which is indicated by a ratio larger than 1. The absolute difference between the two methods' estimation precision is largest for small *N* and decreases as *N* increases, and eventually levels off. However, the IPR ratio between the two methods did not depend on *N*. The same relationship between sample size and the standard errors of percentiles has been described for continuous data (Mood et al., 1974, section VI-5). The IPR ratio ranged from 2.4 to 5.6. The smallest ratio (i.e., 2.4) was found for the 99th percentile when the test consisted of polytomous items, and the largest ratio (i.e., 5.6) was found for the 75th percentile.

Table 2.4. *Summary of Ratio between IPR of Traditional and Continuous Norming for Given N.*

	IPR Ratio		
Percentile	Min.	Max.	Mean (SD)
50	3.99	4.74	4.36 (0.26)
75	4.60	5.59	5.02 (0.30)
90	3.01	4.15	3.62 (0.36)
95	3.33	4.08	3.90 (0.19)
99 dichotomous	2.44	3.01	2.82 (0.16)
99 polytomous	2.41	3.36	3.01 (0.27)

2.5 Discussion

We studied the precision of percentile estimates expressed by IPRs to derive sample-size requirements for traditional and regression-based norming. For both norming approaches, precision of the percentile estimates was also examined as a function of size of covariate effects on the test score, number of item scores, and test length.

From the results, test constructors can determine the sample size required to obtain percentile estimates with a particular degree of precision. Suppose a dichotomous 50-item test is used for important decisions for which the 75th percentile is crucial. In this case

precise estimation is required. The test constructor therefore selects a maximum IPR of .1 standard deviations. In our study, for a 50-item dichotomous test .1 standard deviation corresponds to approximately 1 score unit. Hence, most percentile estimates differ by at most 1 score unit. If traditional norming is used, one needs $N > 10{,}000$ to obtain the required precision. However, for regression-based norming $N = 1{,}000$ suffices.

Another example concerns a polytomous 100-item test intended for less important decisions using the 95th percentile. The test constructor selects a maximum IPR of half a standard deviation. For a 100-polytomous item test, this value corresponds to an IPR of approximately 32 score units. For traditional norming, $N = 1{,}500$ is required, and for regression-based norming, $100 < N < 500$ is sufficient.

The finding that regression-based norming requires smaller samples than traditional norming is consistent with the sample size guidelines Evers et al. (2009, pp. 22-23) presented. For regression-based norming with eight norm groups, the authors recommended sample sizes one third the sample sizes for traditional norming. We found that as the percentiles were further away from the median, the difference between the two norming methods was smaller.

For both norming approaches, we also found that IPR grew larger as the estimated percentiles lay further away from the mean. In general, estimating the tails of a distribution requires larger samples. Thus, in order to choose a sample size test constructors first need to decide which percentiles are important for the use of the test, because more extreme percentiles require larger samples. For continuous data, the required sample size to estimate a percentile with a certain precision can be obtained analytically (e.g., Mood et al., 1974, section VI-5).

For both norming methods, the estimation of the 99th percentile was more precise for polytomous than dichotomous items. The explanation may be that in the highest score range polytomous items provide more score diversity than dichotomous items, resulting in narrower IPRs for the norm estimates relative to the total scale length. It should be noted that little score diversity prohibits distinguishing between individuals in the higher score range even if estimation precision is high. For example, the 90th percentile for a 10-item dichotomous might be estimated with high precision to be equal to a score of 10. However, for the 99th percentile one might estimate the same value of 10.due to the scale having only

eleven values in total, the two highest being 9 and 10. Thus, one cannot distinguish individuals located in the top 10% and the top 1%. If precise estimation of extreme percentiles is important, we recommend a larger number of items, if possible polytomous items. Regression-based norming uses the relationship between covariates and the test score to adjust the discrete test scores, which results in a non-discrete distribution of residuals enabling distinguishing different extreme scores. If dichotomous items must be used, regression-based norming enables high precision and also enables distinguishing different high-scoring individuals.

The covariates influenced the mean test score of the norm groups but not the distribution shape; hence, the value of the multiple correlation between covariates and test score did not affect the precision of norm estimation. We notice that in real-data research one usually does not know the model that generated the data, and in simulation research one has to choose a plausible candidate. Using the much-used nonlinear GRM for data generation allowed us to study the effect of number of items and number of response categories on precision. Our aim was comparing traditional and regression-based norming. Hence, we checked two conditions. First, the nonlinear GRM produced test scores that are nonlinearly related to the GRM's latent variable and, second, the linear regression assumptions of homoscedasticity, linearity, and normality are satisfied in the generated data. We found that the relation between test score and latent variable was approximately linear and that model violations were negligible. Hence, we concluded that the corresponding percentiles are unbiased. The results were based on plots of the raw scores as a function of latent variable θ, plots of the standardized residuals as a function of standardized predicted values, qq-plots, and histograms of both the test scores and the standardized residuals (e.g., Tabachnick & Fidell, 2012, pp. 85-86, 97), and can be obtained upon request from the first author.

Further research may investigate the effect of failure of the assumptions of the regression model, which are heteroscedasticity, non-linearity, and non-normality, on norm estimation precision using the regression-based norming method. Other topics are model misspecification and the effect of unequal sample sizes in covariate groups on norm estimation precision.

Chapter 3

Standard Errors and Confidence Intervals of Norm Statistics for Psychological and Educational Tests

Abstract

Norm statistics allow for the interpretation of scores on psychological and educational tests, by relating the test score of an individual test taker to the test scores of individuals belonging to the same gender, age, or education groups, et cetera. Given the uncertainty due to sampling error, one would expect researchers to report standard errors for norm statistics. In practice, standard errors are seldom reported; they are either unavailable or derived under strong distributional assumptions that may not be realistic for test scores. We derived standard errors for four norm statistics (standard deviation, percentile ranks, stanine boundaries and Z-scores) under the mild assumption that the test scores are multinomially distributed. A simulation study showed that the standard errors were unbiased and that corresponding Wald-based confidence intervals had good coverage. Finally, we discuss the possibilities for applying the standard errors in practical test use in education and psychology. The procedure is provided via the R (R Core Team, 2015) function check.norms, which is available in the mokken package (Van der Ark, 2012).

This chapter has been accepted for publication as Oosterhuis, H. E. M., Van der Ark, L. A., & Sijtsma, K. (in press). Standard errors and confidence intervals of norm statistics for psychological and educational tests. *Psychometrika.*

3.1 Introduction

Norm statistics allow for the interpretation of scores on educational and psychological tests, by relating the test score of an individual test taker to the test scores of a group of individuals having, for example, the same gender, age, or education level. Examples of norm statistics frequently used in practice are percentile ranks, linear standard scores such as *Z*-scores, and normalized standard scores such as stanines (Mertler, 2007, Module 6). A norm statistic obtained from a normative sample should be viewed as a point estimate of the norm in the population (Crawford, Garthwaite, & Slick, 2009), which means the norm estimate should be accompanied by an indication of estimation precision.

The publication manual of the American Psychological Association (2010, p. 34) also requires that when point estimates are provided the authors "always include an associated measure of variability (precision), with an indication of the specific measure used (e.g., the standard error)". In addition, on the same page the publication manual strongly recommends reporting confidence intervals (CIs). CIs, of which Wald-based CIs are the most commonly used, are directly related to the standard errors (SEs). Let $\hat{\theta}$ be the point estimate with $SE_{\hat{\theta}}$, let α be the two-tailed *p*-value and $z_{\alpha/2}$ the corresponding *Z*-score, then the limits of the 100(1-α)% Wald-based CI are

$$\hat{\theta} \pm z_{\alpha/2} \cdot SE_{\hat{\boldsymbol{\theta}}}. \quad (1)$$

Hence, CIs for norm scores serve to remind us that norm estimates based on normative samples are fallible and quantify the degree of fallibility that is caused by estimation imprecision (Crawford, Cayley, Lovibond, Wilson, & Hartley, 2011). However, norm constructors commonly fail to provide SEs or CIs to quantify estimation precision of the norms (e.g., Aardoom, Dingemans, Slof Op't Land, & Van Furth, 2012; Cavaco et al., 2013; Glaesmer et al., 2012; Goretti et al., 2014; Grande, Romppel, Glaesmer, Petrowski, & Herrmann-Lingen, 2010; mer, Montagne, Hendriks, Perrett, & De Haan, 2014; Mond, Hay, Rodgers, & Owen, 2006; Palomo et al., 2011; Sartorio et al., 2013; Shi et al., 2014), because for many norm statistics the SEs are unknown, difficult to derive, or if available not included in popular software. For example, Crawford et al. (2011) described a procedure for obtaining CIs for percentile norms, but their cannot be used for estimating CIs for other

norm statistics such as stanine boundaries. For other norm statistics, the SEs are known only under strong assumptions. For example, the SE of the standard deviation assumes data to be normally distributed, but test scores often are discrete and their distribution skewed because the test is too difficult or the attribute measured is rare (test-score distribution skewed to the right), or the test is too easy or the attribute is highly prevalent (test-score distribution skewed to the left).

We derived SEs for the test-score standard deviation, percentile rank scores, the boundaries of the stanines, and *Z*- scores. The SEs were derived under the mild assumption that raw test scores follow a multinomial distribution. Let a trial have three or more possible outcomes, and let *N* represent the number of independent identical trials. The probabilities of obtaining a certain outcome follow a multinomial distribution (Agresti, 2013, p. 6). This model can easily be extended to the raw scores obtained by means of psychological and educational tests. For example, let the number of items answered correctly be the raw score on such a test. The administration of the test to a norm sample of *N* respondents then corresponds to *N* trials, and each value of the raw score corresponds to a trial outcome. The probabilities of obtaining each raw score then follow a multinomial distribution. Although the method is based on a discrete distribution, the test scores need not be integers. Hence, the method can also be applied to continuous measures such as reaction time or blood pressure. For large *N*, the multivariate normal central limit theorem (see, e.g., Rao, 1973, p. 128) ascertains that the multinomial distribution is close to the multivariate normal distribution if the model parameters are not near the boundaries of the parameter space (i.e., probabilities are not close to 0 or 1). Hence, if *N* is large enough, the method described in this article can also be used for data that are (approximately) normal.

This article is structured as follows. First, we provide an example that illustrates the use of SEs and CIs for norm statistics in practice. Second, we explain the general framework for deriving SEs for norm statistics. Third, using this general framework we derived the SEs for the test-score standard deviation, the *Z*- scores, the percentile rank scores, and the stanine boundaries. Fourth, using a simulation study we investigated the bias and the precision of the derived SE estimates and the coverage of the corresponding Wald-based

CIs. Finally, we briefly discuss the results and provide computer code to obtain the SEs for norm statistics.

3.2 An Illustration of Using Norms With and Without Standard Errors

Because estimation precision, quantified by SEs or CIs, usually is not taken into account when test constructors present norm statistics, norm statistics are often presented and interpreted as if they were parameters. Ignoring imprecision of norm statistics may produce the wrong conclusions about test performance and can have serious consequences for individual test takers if decisions are based on their test performance. As an example, Figure 3.1 shows the scores on Social Skills measured by means of the Preschool and Kindergarten Behavior Scales (PKBS; Merrell, 1994) for three five-year old boys named Oliver, Jack, and Harry. The horizontal lines at raw scores 59 and 75 correspond to the 5th and 20th percentile ranks of the score distribution, respectively. These percentile ranks were estimated in a norm sample, and according to the PKBS test manual (Merrell, 1994), a score below the 5th percentile rank (i.e., raw score 59) indicates a significant deficit of social skills, a score between the 5th and 20th percentile rank (i.e., raw scores 59 and 76, respectively) indicates a moderate deficit, and a score above the 20th percentile rank (i.e., raw score 76) indicates average social skills. In what follows, we distinguish the influence of random measurement error on test performance, typical of classical test theory, and the influence of sampling error on norm statistics.

In Figure 3.1a, influence of measurement error and sampling error were not taken into account for the individual scores and the norm score, respectively, which means all values were treated as population values. Oliver, Jack, and Harry had raw scores equal to 56, 82, and 61, respectively. Based on Figure 1a, we conclude that Oliver has a significant social skills deficit, Jack has average social skills, and Harry has a moderate deficit.

Next, we assess influence of random measurement error on test performance, and follow the classical test model, which assumes that an observable test score (X) can be decomposed into an unobservable true score (T) and unobservable random measurement error (E), so that $X = T + E$. In Figure 3.1b, the norm value was treated as a population value, but random measurement error related to individual test scores was taken into account using a 68% CI, or a score band. The score band contained the likely values of the true score, and was based on the standard error of measurement (SEM), which indicates the

magnitude of measurement error associated with a particular test score and in the classical model is equal for all test scores. Hence, the width of the score band indicates the degree to which an individual's test score is expected to vary across replications. Several methods for estimating the SEM are available (Brennan & Lee, 1999; Lee, Brennan, & Kolen, 2000). If the score band contained the norm value, the individual's true score was not significantly different from the norm value based on a 68% confidence level. Oliver's score band (i.e., [53.2; 58.8]), was located completely below the 5th percentile rank (i.e., raw score 59), meaning we conclude that Oliver has a significant deficit. Jack's score band (i.e., [79.2; 84. 8]) was located above the 20th percentile rank (i.e., raw score 76), meaning we conclude that Jack's skills are average. However, Harry's score band (i.e., [58.2; 63.8]) contained the 5th percentile rank (i.e., raw score 59), meaning we are uncertain whether Harry's deficit is significant or moderate.

In Figure 3.1c, in addition to random measurement error for the individual test scores, sampling error was taken into account for the percentile rank values. The horizontal dotted lines corresponded to the boundaries of the 95% CIs for the percentile ranks. Using the heuristic rule that an overlap of 25% or less between the CIs of two statistics suggests a significant difference between statistics (Van Belle, 2003, Section 2.6), we conclude that Jack's true score differed significantly from the 20th percentile rank, indicating that he has average social skills. On the other hand, we conclude Oliver's and Harry's true scores did not differ significantly from the 5th percentile rank, which means that for both boys we are not sure whether they have a significant or a moderate deficit.

This example shows that the interpretation of the boys' test results changed depending on whether measurement error and sampling error was taken into account for the individual test results and for the norm values, respectively. Especially for individuals that score close to norm values, taking into account the sampling error of the norm values can have important consequences for the interpretation of their test scores.

Crawford and Howell (1998) have argued that treating norm statistics as population values is justifiable if the sample is large enough. However, there are no clear size requirements for a norm sample. For example, Evers, Lucassen, Meijer, and Sijtsma (2009) have provided Dutch test constructors with practical guidelines for the size of norm groups, but these guidelines lack a sufficient statistical basis. The American Educational

Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) have provided guidelines for test construction (AERA, APA, & NCME, 1999), without sample size recommendations.

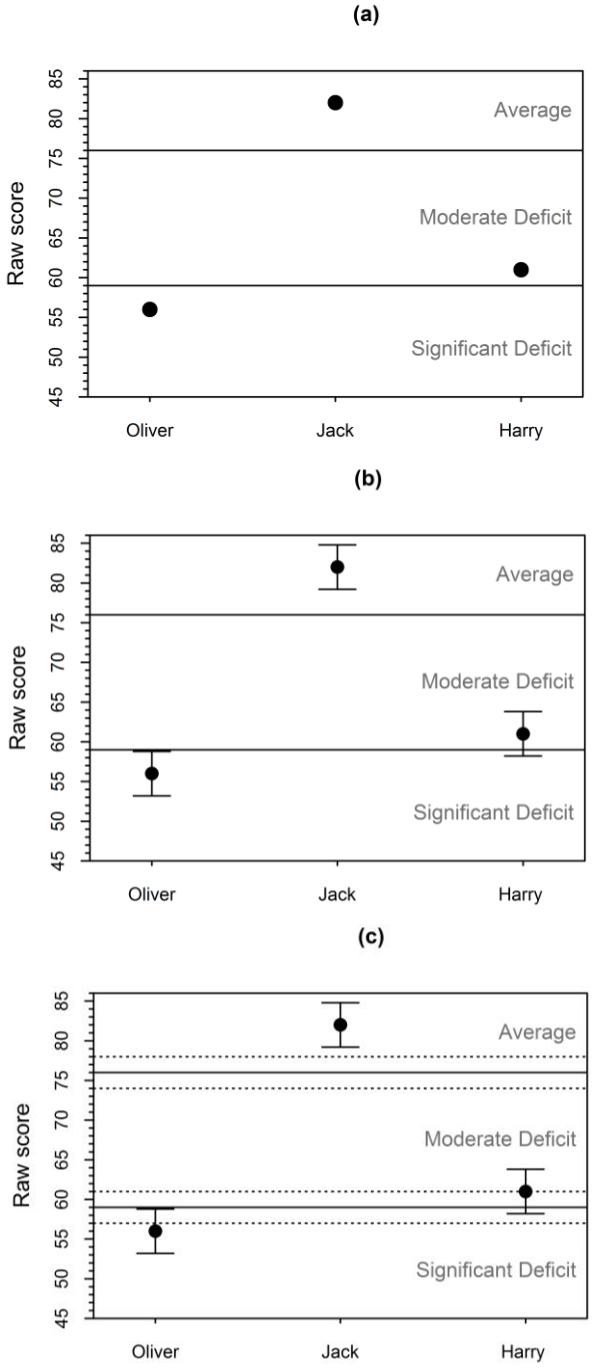

Figure 3.1. Comparison of raw scores to percentile norms when (a) no measurement or norm sampling error, (b) only measurement error, or (c) both measurement and norm sampling error are taken into account.

In addition to sample size, precision of a norm estimate also depends on the statistic's location in the norm sample distribution. Norms that are based on extreme test scores are expected to show more error variation, because they are less likely to occur in a sample. For example, for equal norm-sample size, Oosterhuis, Van der Ark, and Sijtsma (2016b) found that raw scores associated with percentile ranks further away from the median were estimated with less precision than percentile ranks closer to the median. Hence, a norm sample might be large enough for precise estimation of the median or the mean, but not for extreme percentile scores. SEs and CIs can be used to quantify the sampling error associated with each norm statistic and might be used to determine whether the sample was large enough to obtain precise estimates for all norm values of interest.

3.3 A General Framework for Deriving SEs under a Multinomial Distribution

3.3.1 A Two-Step Procedure

We used a general framework consisting of two steps to compute the SEs of the norm statistics (e.g., Bergsma, Croon, & Hagenaars, 2009; Kuijpers, Van der Ark, & Croon, 2013a). The first step is to write the norm statistic as a function of the frequencies of the raw scores. Suppose the raw scores, obtained from administrating a test to a sample of N respondents, are collected in an $N \times 1$ vector $\mathbf{x}$. It may be noted that the unique realizations of the raw score need not be integers and need not include all possible realizations. The k unique realizations of the raw score $r_{(1)}, \ldots, r_{(k)}$ $(r_{(i)} < r_{(i+1)})$, for $i = 1, \ldots, k$ are contained in a $k \times 1$ vector $\mathbf{r}$, and the expected and observed frequencies of the realizations $r_{(1)}, \ldots, r_{(k)}$ are collected in a $k \times 1$ vector $\mathbf{m}$ and a $k \times 1$ vector $\widehat{\mathbf{m}}$, respectively. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_l)'$ denote a vector containing l —possibly mutually dependent—norms, such as the $l = k$ Z-scores or the $l = 8$ boundaries of the stanines. The first step encompasses showing that $\widehat{\boldsymbol{\theta}} = \mathbf{g}(\widehat{\mathbf{m}})$, where $\mathbf{g}(\cdot)$ is a vector function. For a single norm, such as the mean or the standard deviation, vector $\widehat{\boldsymbol{\theta}}$ reduces to scalar $\hat{\theta}$, and $\widehat{\boldsymbol{\theta}} = \mathbf{g}(\widehat{\mathbf{m}})$ reduces to $\hat{\theta} = g(\widehat{\mathbf{m}})$. The technique to obtain $\mathbf{g}(\widehat{\mathbf{m}})$ is called the *generalized exp-log notation* and is described later.

In the second step, the delta method (e.g., Agresti, 2013, §16.1.4; Kendall & Stuart, 1977, pp. 246-247), is used to approximate the variance of the norm statistic. If $\widehat{\mathbf{m}}$ is a

consistent estimator, it converges to its true value $\mathbf{m}$, and the central limit theorem can be applied to obtain asymptotical normality,

$$(\hat{\mathbf{m}} - \mathbf{m}) \xrightarrow{D} N(0, \mathbf{V}_{\hat{\mathbf{m}}}), \tag{2}$$

where $\mathbf{V}_{\hat{\mathbf{m}}}$ is the covariance matrix of $\hat{\mathbf{m}}$. Let $\mathbf{D}(\mathbf{y})$ be a diagonal matrix with the elements of vector $\mathbf{y}$ on the diagonal, and let $\mathbf{Y}'$ be the transpose of $\mathbf{Y}$. Under a multinomial distribution, let $\hat{\mathbf{V}}_{\hat{\mathbf{m}}}$ be the sample estimate of the variance of $\mathbf{g}(\hat{\mathbf{m}})$,

$$\hat{\mathbf{V}}_{\hat{\mathbf{m}}} = \mathbf{D}(\hat{\mathbf{m}}) - \hat{\mathbf{m}}N^{-1}\hat{\mathbf{m}}' \tag{3}$$

(e.g., Agresti, 2013, p. 6). Using the first two terms of the Taylor series, $\mathbf{g}(\hat{\mathbf{m}})$ can be approximated by

$$\mathbf{g}(\hat{\mathbf{m}}) \approx \mathbf{g}(\mathbf{m}) + \mathbf{G}'(\hat{\mathbf{m}} - \mathbf{m}), \tag{4}$$

where $\mathbf{G}$ is the matrix of first partial derivatives, or the Jacobian, of $\mathbf{g}(\hat{\mathbf{m}})$ with respect to $\hat{\mathbf{m}}$, evaluated at $\mathbf{m}$. Equation 4 implies that the variance of $\mathbf{g}(\hat{\mathbf{m}})$ is approximately

$$\mathbf{V}_{\mathbf{g}(\hat{\mathbf{m}})} \approx \mathbf{G}(\hat{\mathbf{V}}_{\hat{\mathbf{m}}})\mathbf{G}'. \tag{5}$$

Therefore, the delta method implies that

$$\big(\mathbf{g}(\hat{\mathbf{m}}) - \mathbf{g}(\mathbf{m})\big) \xrightarrow{D} N(0, \mathbf{G}\mathbf{V}_{\hat{\mathbf{m}}}\mathbf{G}'). \tag{6}$$

Based on Equation 6, the sample estimate of the asymptotic variance of $\mathbf{g}(\hat{\mathbf{m}})$ is

$$\begin{aligned} \hat{\mathbf{V}}_{\mathbf{g}(\hat{\mathbf{m}})} &= \mathbf{G}\hat{\mathbf{V}}_{\hat{\mathbf{m}}}\mathbf{G}' \\ &= \mathbf{G}(\mathbf{D}(\hat{\mathbf{m}}) - \hat{\mathbf{m}}N^{-1}\hat{\mathbf{m}}')\mathbf{G}' \\ &= \mathbf{G}\mathbf{D}(\hat{\mathbf{m}})\mathbf{G}' - \mathbf{G}\hat{\mathbf{m}}N^{-1}\hat{\mathbf{m}}'\mathbf{G}'. \end{aligned} \tag{7}$$

By taking the square root of the diagonal elements of $\hat{\mathbf{V}}_{\mathbf{g}(\hat{\mathbf{m}})}$, the sample estimate of the asymptotic SEs for $\mathbf{g}(\hat{\mathbf{m}})$ are obtained. In some cases, Equation 7 can be further reduced. Let the constant $t > 0$, if $\mathbf{g}(\hat{\mathbf{m}}) = \mathbf{g}(t\hat{\mathbf{m}})$, then vector function $\mathbf{g}(\hat{\mathbf{m}})$ is homogeneous of order 0. If $\mathbf{g}(\hat{\mathbf{m}})$ is homogeneous of order 0, then $\mathbf{G}\hat{\mathbf{m}}N^{-1}\hat{\mathbf{m}}'\mathbf{G}' = \mathbf{0}$, and Equation 7 reduces to

$$\hat{\mathbf{V}}_{\mathbf{g}(\hat{\mathbf{m}})} = \mathbf{G}\mathbf{D}(\hat{\mathbf{m}})\mathbf{G}' \tag{8}$$

(e.g., Bergsma, 1997, Appendix D). For example, if $\mathbf{g}(\hat{\mathbf{m}})$ is homogeneous of order 0, then it does not matter whether absolute frequencies ($\hat{\mathbf{m}}$) or relative frequencies (e.g., proportions $\hat{\mathbf{P}} = \hat{\mathbf{m}}/N$) are used as the argument of $\mathbf{g}$, because for $t = 1/N$, $t\hat{\mathbf{m}} = \hat{\mathbf{P}}$.

3.3.2 Generalized exp-log notation

In general, obtaining function $\mathbf{g}(\widehat{\mathbf{m}})$ and Jacobian $\mathbf{G}$ requires tedious derivations. The generalized exp-log notation (Bergsma, 1997; Grizzle, Starmer, & Koch, 1969; Kritzer, 1977) is a method that alleviates this problem by rewriting $\mathbf{g}(\widehat{\mathbf{m}})$ using a series of q appropriate design matrices $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_q$ (to be explained below), such that

$$\mathbf{g}(\widehat{\mathbf{m}}) = \mathbf{A}_q \cdot \exp(\mathbf{A}_{q-1} \cdot \log(\mathbf{A}_{q-2} \cdot \exp(\ldots \log(\mathbf{A}_3 \cdot \exp(\mathbf{A}_2 \cdot \log(\mathbf{A}_1 \cdot \widehat{\mathbf{m}})))))). \quad (9)$$

The reason for using such complex notation to write $\mathbf{g}(\widehat{\mathbf{m}})$ is to make it relatively easy to derive the Jacobian using the chain rule (e.g., Larson & Edwards, 2013, pp. 129-135). The Jacobian for Equation 9 can be derived using the following steps (Bergsma, 1997, pp. 66-68; Kritzer, 1977; Kuijpers et al., 2013a). First, a series of $q + 1$ functions $\mathbf{g}_0, \mathbf{g}_1, \ldots, \mathbf{g}_q$ is defined, such that

$$\mathbf{g}_0 = \widehat{\mathbf{m}}, \quad (10)$$

and, for $i = 1, \ldots, q$,

$$\mathbf{g}_i = \log(\mathbf{A}_i \cdot \mathbf{g}_{i-1}), \text{ if } i \text{ is an odd number}, \quad (11)$$

and

$$\mathbf{g}_i = \exp(\mathbf{A}_i \cdot \mathbf{g}_{i-1}), \text{ if } i \text{ is an even number}. \quad (12)$$

The last function in this series is

$$\mathbf{g}(\widehat{\mathbf{m}}) = \mathbf{g}_q = \mathbf{A}_q \cdot \mathbf{g}_{q-1}. \quad (13)$$

Second, let $\mathbf{G}_0, \mathbf{G}_1, \ldots, \mathbf{G}_q$ denote the matrices of the partial derivatives of $\mathbf{g}_0, \mathbf{g}_1, \ldots, \mathbf{g}_q$, with respect to $\widehat{\mathbf{m}}$, respectively. Let $\mathbf{I}_p$ denote an identity matrix of order p. Standard calculus shows that

$$\mathbf{G}_0 = \frac{\partial \mathbf{g}_0}{\partial \widehat{\mathbf{m}}'} = \mathbf{I}_k . \quad (14)$$

Furthermore, let $\mathbf{Y}^{-1}$ be the inverse of $\mathbf{Y}$. For $i = 1, \ldots, q - 1$,

$$\mathbf{G}_i = \frac{\partial \mathbf{g}_i}{\partial \widehat{\mathbf{m}}'} = \mathbf{D}^{-1}(\mathbf{A}_i \cdot \mathbf{g}_{i-1}) \cdot \mathbf{A}_i \cdot \mathbf{G}_{i-1}, \text{ if } i \text{ is an odd number}, \quad (15)$$

and

$$\mathbf{G}_i = \mathbf{D}\big(\exp(\mathbf{A}_i \cdot \mathbf{g}_{i-1})\big) \cdot \mathbf{A}_i \cdot \mathbf{G}_{i-1}, \text{ if } i \textit{ is} \text{ an even number}. \quad (16)$$

Finally, the last function in the series is

$$\mathbf{G} = \mathbf{G}_q = \mathbf{A}_q \cdot \mathbf{G}_{q-1}. \quad (17)$$

3.3.3 SEs for Norm Statistics

To illustrate the derivation of SEs for norm statistics, the case of the sample mean is discussed extensively. For other norm statistics, the derivation is only outlined, and details can be found in appendices A to D.

Mean test score. Sample mean $\bar{X} = \sum_i X_i / N$ (i indexes persons) can be cast in the generalized exp-log notation (Equation 5), using three design matrices as follows. Let $\mathbf{1}_p$ be a vector of ones of length p, let $\mathbf{A}_1$ be a $2 \times k$ matrix, defined as $\mathbf{A}_1 = [\mathbf{r} \quad \mathbf{1}_k]'$, let $\mathbf{A}_2$ be a 1×2 vector, defined as $\mathbf{A}_2 = [1 \quad -1]$, let $\mathbf{A}_3$ be a 1 x 1 matrix containing the scalar 1, that is, $\mathbf{A}_3 = [1]$. It follows that $\mathbf{g}_0 = \widehat{\mathbf{m}}$ (Equation 10); next that

$$\mathbf{g}_1 = \log(\mathbf{A}_1 \cdot \widehat{\mathbf{m}}) = \log([\mathbf{r} \quad \mathbf{1}_k]' \cdot \widehat{\mathbf{m}}) = \log\left(\begin{bmatrix}\mathbf{r}' \cdot \widehat{\mathbf{m}} \\ \mathbf{1}_k' \cdot \widehat{\mathbf{m}}\end{bmatrix}\right) = \begin{bmatrix}\log(\sum X_i) \\ \log(N)\end{bmatrix} \tag{18}$$

(Equation 11); then

$$\mathbf{g}_2 = \exp(\mathbf{A}_2 \cdot \mathbf{g}_1) = \exp\left([1 \quad -1] \cdot \begin{bmatrix}\log(\sum X_i) \\ \log(N)\end{bmatrix}\right) = \left[\frac{\sum X_i}{N}\right], \tag{19}$$

(Equation 12); and finally that

$$\mathbf{g}(\widehat{\mathbf{m}}) = \mathbf{g}_3 = \mathbf{A}_3 \cdot \mathbf{g}_2 = [1] \cdot \left[\frac{\sum X_i}{N}\right] = \frac{\sum X_i}{N} \tag{20}$$

(Equation 13), which equals the sample mean. It may be noted that $\mathbf{g}(\widehat{\mathbf{m}})$ is homogenous of order 0.

Next, the Jacobian matrices $\mathbf{G}_0$, $\mathbf{G}_1$, $\mathbf{G}_2$, and $\mathbf{G}_3$ can be derived for $\mathbf{g}_0$, $\mathbf{g}_1$, $\mathbf{g}_2$, and $\mathbf{g}_3$, respectively. First, it follows that $\mathbf{G}_0 = \mathbf{I}_k$ (Equation 14). Second, let $\mathbf{y}/p$ denote the element-wise division of $\mathbf{y}$ by p. Then $\mathbf{G}_1$ is the $2 \times k$ matrix,

$$\mathbf{G}_1 = \frac{\partial \mathbf{g}_1}{\partial \widehat{\mathbf{m}}'} = \mathbf{D}^{-1}[\mathbf{A}_1 \cdot \mathbf{g}_0] \cdot \mathbf{A}_1 \cdot \mathbf{G}_0 = \mathbf{D}^{-1}\begin{bmatrix}\sum X_i \\ N\end{bmatrix} \cdot \begin{bmatrix}\mathbf{r}' \\ \mathbf{1}_k'\end{bmatrix} \cdot \mathbf{I} = \begin{bmatrix}\mathbf{r}'/\sum X_i \\ \mathbf{1}_k'/N\end{bmatrix} \tag{21}$$

(Equation 15). Third, $\mathbf{G}_2$ is the $1 \times k$ vector,

$$\mathbf{G}_2 = \frac{\partial \mathbf{g}_2}{\partial \widehat{\mathbf{m}}'} = \mathbf{D}[\exp(\mathbf{A}_2 \cdot \mathbf{g}_1)] \cdot \mathbf{A}_2 \cdot \mathbf{G}_1 = \left[\frac{\sum X_i}{N}\right] \cdot [1 \quad -1] \cdot \begin{bmatrix}\mathbf{r}'/\sum X_i \\ \mathbf{1}_k'/N\end{bmatrix} \tag{22}$$

$$= (\mathbf{r}' - \bar{X} \cdot \mathbf{1}_k')/N$$

(Equation 16). Finally, $\mathbf{G}_3$ is the $1 \times k$ vector,

$$\mathbf{G} = \mathbf{G}_3 = \mathbf{A}_3 \cdot \mathbf{G}_2 = [1] \cdot \left((\mathbf{r}' - \bar{X} \cdot \mathbf{1}_k')/N\right) = (\mathbf{r}' - \bar{X} \cdot \mathbf{1}_k')/N \tag{23}$$

(Equation 17).

Because $\mathbf{g}(\widehat{\mathbf{m}})$ (Equation 20) is homogenous of order 0, the variance of $\mathbf{g}(\widehat{\mathbf{m}})$ (Equation 8) is approximated by

$$V_{\bar{X}} \approx \mathbf{GD}(\hat{\mathbf{m}})\mathbf{G}' = [(\mathbf{r}' - \bar{X} \cdot \mathbf{1}_{\boldsymbol{k}}')\mathbf{D}(\hat{\mathbf{m}})(\mathbf{r} - \bar{X} \cdot \mathbf{1}_{\boldsymbol{k}})]/N^2 \quad (24)$$
$$= \sum_{j=1}^{k} \hat{m}_j \left[(r_j - \bar{X})/N\right]^2.$$

Note that Equation 24 equals the variance of a population of size *N* divided by *N.* The approximated SE for the sample mean is then given by

$$S_{\bar{X}} = \sqrt{V_{\bar{X}}}. \quad (25)$$

Note that Equation 25 equals the standard deviation of a population of size *N* divided by the square root of *N.*

Standard deviation. The unbiased sample estimator of the standard deviation of the test scores equals

$$s_X = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2}. \quad (26)$$

Let *i* and *j* index realizations **r**, let δ_{ij} be Kronecker's delta ($\delta_{ij} = 1$ if $i = j$, and $\delta_{ij} = 0$ if $i \neq j$), let $d_i = r_i - \bar{X}$, let $e = \frac{1}{N-1}$, and let $SS = \sum(X_i - \bar{X})^2$. In Appendix A we show that the standard error of s_X can be approximated by

$$S_{s_X} \approx 0.5 s_X \sqrt{\sum_{\boldsymbol{i}}\sum_{\boldsymbol{j}}\left(\frac{d_i^2}{SS} - e\right)\left(\frac{d_j^2}{SS} - e\right)\left(\delta_{ij}\hat{m}_i - \frac{\hat{m}_i\hat{m}_j}{N}\right)}. \quad (27)$$

For large *N*, Equation 27 reduces to

$$S_{s_X} \approx 0.5 s_X \sqrt{\sum_{\boldsymbol{i}} \hat{m}_i \left(\frac{d_i^2}{SS} - \frac{1}{N}\right)^2}. \quad (28)$$

Ahn and Fessler (2003) also derived a SE estimate for the sample standard deviation under the assumption that the data are normally distributed,

$$\dot{S}_{S_X} = \frac{s_X}{\sqrt{2(N-1)}}. \quad (29)$$

Percentile rank scores. Percentile rank score PR_x is the percentage of individuals in the normative sample that have raw score x or lower. For example, if $X = 20$ is associated with $PR_{20} = 90$, 90% of individuals in the normative sample have a raw score of 20 or lower. If researchers use percentile ranks as norm statistics, they either associate the raw scores with particular percentile ranks (e.g., 75 or 95) or they present the percentile ranks for each possible raw score. Percentile ranks can also be used as cut-off scores. For example, a job selection test may use the raw score corresponding to the 95th percentile as the cut-off score. Then, individuals are only selected if they belong to the highest 5% of the

raw-score distribution compared to individuals having, for example, the same age, sex, or education. Let x be a particular raw score on the test of interest, and let $P(\cdot)$ denote a sample proportion, then PR_x is estimated as

$$\begin{aligned} PR_x &= \{P(X < x) + 1/2 \cdot P(X = x)\} \cdot 100 \\ &= 50 \cdot P(X \leq x) + 50 \cdot P(X < x). \end{aligned} \tag{30}$$

In Appendix B, we show that the sample estimate of the asymptotic standard error of PR_x equals

$$S_{PR_x} \approx \frac{50}{N} \sqrt{\sum_i \hat{m}_i \big(\gamma_{xi} - P(X < x) - P(X \leq x)\big)^2}, \tag{31}$$

where $\gamma_{xi} = 0$ if $x < i$, $\gamma_{xi} = 1$ if $x = i$, and $\gamma_{xi} = 2$ if $x > i$.

Stanines. Stanines are norm scores 1, 2, …, 9 that correspond to ranges of the normal distribution. The first stanine corresponds to raw scores smaller than 1.75 standard deviations below the mean, and the ninth stanine corresponds to raw scores larger than 1.75 standard deviations above the mean. The remaining seven stanines (stanines 2, 3, …, 8) have a width of half a standard deviation and divide the range between 1.75 standard deviations below the mean and 1.75 standard deviations above the mean in segments 0.5 standard deviations wide. The raw-score mean is located in the middle of the fifth stanine. When stanines are used as norm statistics, the raw scores that correspond to the eight boundaries of the nine stanines are presented. Let **f** be an 8 x 1 vector containing values starting from -1.75 and ending at 1.75 with increments of .50. The elements of **f** are denoted f_b ($b = 1, \ldots, 8$) and correspond to the eight boundaries of the stanines. On the scale of the raw scores, the bth boundary of the stanines is given by

$$St_b = f_b \cdot s_X + \bar{X}. \tag{32}$$

Let $d_i^* = {d_i}^2/SS = (r_i - \bar{X})^2/[\sum X_i^2 - (\sum X_i)^2/N]$. In Appendix C we show that the standard error of St_b can be approximated by

$$S_{St_b} \approx \sqrt{\sum_{j=1}^{k} \sum_{i=1}^{k} \left[\delta_{ij} \hat{m}_j - \frac{\hat{m}_i \hat{m}_j}{N}\right] \cdot \left[\frac{f_b s_X}{2} \cdot (d_i^* - e) + \frac{d_i}{N}\right] \cdot \left[\frac{f_b s_X}{2} \cdot \left(d_j^* - e\right) + \frac{d_j}{N}\right]}. \tag{33}$$

For large N, Equation 33 reduces to

$$S_{St_b} \approx \sqrt{\sum_{i=1}^{k} \hat{m}_i \cdot \left[\frac{f_b\, s_X}{2} \cdot \left(d_i^* - \frac{1}{N}\right) + \frac{d_i}{N}\right]^2}. \tag{34}$$

***Z*-scores.** *Z*-scores are expressed in standard deviations relative to the group mean. For instance, if the mean raw score equals 15 and the standard deviation equals 2, a raw score equal to 18 corresponds to a *Z*-score equal to (18-15)/2 = 1.5. When Z-scores are used as norm statistics, either a selection of the raw scores or all raw scores are presented with the corresponding *Z*-scores. The *Z*-score corresponding to X_h is computed as

$$Z_h = (X_h - \bar{X})/s_X = X_h/s_X - \bar{X}/s_X. \quad (35)$$

In Appendix D we show that the standard error of Z_h can be approximated by

$$S_{Z_h} \approx \sqrt{\sum_{j=1}^{k}\sum_{i=1}^{k} t\left[-\bar{X}\cdot\left(\frac{r_i}{\sum X}-\frac{1}{N}-u_i\right)-r_h u_i\right]\cdot\left[-\bar{X}\left(\frac{r_j}{\sum X}-\frac{1}{N}-u_j\right)-r_h u_j\right]}, \quad (36)$$

where $t = \frac{1}{s_X^2}\left(\delta_{ij}\hat{m}_j - \frac{\hat{m}_i\hat{m}_j}{N}\right)$, and $u_i = .5(d_i^* - e)$.

For large *N*, Equation 36 reduces to

$$S_{z_h} \approx \sqrt{\sum_i \frac{\hat{m}_i}{s_X^2}\cdot\left(-\bar{X}\cdot\left(\frac{r_i}{\sum X_i}-\frac{1}{N}+u_i\right)-r_h u_i\right)^2}. \quad (37)$$

3.4 Simulation Study

We performed a simulation study to investigate the bias and the precision of the SE estimates of the standard deviation, the percentile rank scores, the stanine boundaries, and the *Z*-scores (equations 23, 26, 29, and 32, respectively), as well as the coverage probability of Wald CIs based on the SE estimates of these norm statistics for different sample and test characteristics.

Simulation Model

We simulated data using the 2-parameter logistic model (2PLM; Birnbaum, 1968; Van der Linden & Hambleton, 1997). The 2PLM describes response probabilities to each item *j* (with scores *x* = 0, 1) by means of a logistic function with a slope parameter α_j and a location parameter β_j. The probability of a score of $X_j = 1$ on item *j* equals

$$P(X_j = 1|\theta) = \frac{\exp[\alpha_j(\theta-\beta_j)]}{1+\exp[\alpha_j(\theta-\beta_j)]}. \quad (38)$$

Slope parameter α_j = 0.85, 0.95, 1.05,..., 1.75, where *j* = 1, 2, ..., 10. For tests consisting of more than 10 items, the first ten slope parameters $\alpha_1, \ldots, \alpha_{10}$ had the same slope values as slope parameters $\alpha_{11}, \ldots, \alpha_{20}$ and $\alpha_{21}, \ldots, \alpha_{30}$; etcetera. Location parameter β_j = -2.25, -2.15, -2.05, ..., 2.25, where *j* = 1, 2, ..., 10. For tests consisting of more than 10

items, the first ten difficulty parameters $\beta_1, \ldots, \beta_{10}$ had the same difficulty values as location parameters $\beta_{11}, \ldots, \beta_{20}$ and $\beta_{21}, \ldots, \beta_{30}$; etcetera.

First, θ-values were randomly drawn from a standard normal distribution. Second, each θ-value was inserted in Equation 38, and the resulting probabilities were used to generate item-score vectors by means of random draws from a multinomial distribution. The raw score for each θ-value was then computed as $X = \sum X_j$ and could range from 0 to J, the total number of items in the test. The simulated item scores were discrete; hence, the resulting raw test scores were also discrete. Given that item difficulty was symmetrically dispersed around 0, and that θ was sampled from a normal distribution, the raw scores were approximately normally distributed, especially for a larger number of items. The simulation study consisted of Q = 10,000 replications, which guaranteed stability of the results.

Dependent Variables

Bias of the SEs. To estimate the bias of the SEs, we first computed the standard deviation of the norm estimates across the Q replications. Let $\bar{\theta} = \frac{1}{Q}\sum_{q=1}^{Q} \hat{\theta}_q$, then

$$s_{\hat{\theta}} = \sqrt{\frac{1}{Q-1}\sum_{q=1}^{Q}\left(\hat{\theta}_q - \bar{\theta}\right)^2}. \quad (39)$$

Standard deviation $s_{\hat{\theta}}$ estimates the variance of the norm statistic across replications, and serves as a gold standard for the SE. Let $\hat{S}_{\hat{\theta}_q}$ denote the estimated SE of the qth norm statistic. Then, the bias of the SE estimator equals

$$bias.se = \frac{1}{Q}\sum_{q=1}^{Q}\left(\hat{S}_{\hat{\theta}_q} - s_{\hat{\theta}}\right). \quad (40)$$

RMSE of the SEs. The root mean square error (RMSE) was used as a measure of the precision of the SE estimator of a statistic. To estimate the RMSE of the SEs we first computed $s_{\hat{\theta}}$ (Equation 39) The RMSE is then computed as follows.

$$RMSE_{\hat{S}_{\hat{\theta}}} = \sqrt{\frac{1}{Q}\sum_{q=1}^{Q}\left(\hat{S}_{\hat{\theta}_q} - s_{\hat{\theta}}\right)^2}. \quad (41)$$

The bias and the RMSE of the standard deviation and the stanine boundaries could not always be readily compared across design cells because the bias and RMSE of these norm statistics are scale-dependent. To alleviate this problem we divided the bias and the RMSE

of the SEs of these norm statistics by the average standard deviation of the simulated scores; that is, we divided by

$$\bar{s}_Y = \frac{1}{Q}\sum_{q=1}^{Q} s_{Y_q}, \quad (42)$$

where s_{Y_q} is the standard deviation of the simulated raw scores in the qth replication. Hence, bias and RMSE are expressed on a scale with mean 0 and standard deviation 1. Dividing the bias and RMSE by $\bar{s}_Y$ guaranteed a fair comparison of the bias and the RMSE of the standard deviation and the stanine boundaries for different sample and test characteristics. For the bias and the RMSE of the percentile rank scores, expressed on a scale from 0 to 100, and *Z*-scores expressed on a scale with mean 0 and variance 1, such a correction was not necessary.

Coverage of the 95% CI. To investigate the coverage of the 95% CI, in each replication we first constructed a Wald-based CI (Equation 1) for $\hat{\theta}_q$. The 95% coverage was defined as the proportion of replications in which CI contained $\bar{\theta}$. Given that the lowest and highest scores might not be observed in all replicated samples, only replications in which a particular raw score was observed were used to compute the coverage.

Independent Variables

In the simulation study, the test and norm sample characteristics were varied as follows.

Number of items (*J*). The number of items was equal to either *J* = 10, 30, or 50. These values were chosen to cover the range of test lengths often found in practice (Oosterhuis et al., 2016b).

Sample size (*N*). The sample size was equal to either *N* = 500, 1,000, 1,500, 2,000, or 2,500. These values were chosen to cover sample sizes often found in practice (Oosterhuis et al., 2016b).

3.5 Results

For tests with 10, 30, and 50 items, $\bar{s}_Y$ (Equation 42) was equal to 2.076, 5.530, and 8.966, respectively. These values were used to adjust the bias and the RMSE of the SEs of the standard deviation and the stanine boundaries.

Standard Deviation

For none of the combinations of sample size and test length, the SE of the standard deviation (Equation 27) showed bias, whereas the SE of the standard deviation derived by Ahn & Fessler (2003; Equation 29), $\dot{S}_{S_X}$ (Equation 29), was positively biased (Table 1). The bias of $\dot{S}_{S_X}$ decreased as sample size increased. Precision of our SE was higher than the precision of $\dot{S}_{S_X}$ (Table 1). For both SEs, precision increased as sample size increased. The precision of both SEs did not depend on test length. The coverage of the 95% CIs that were based on our SE were close to .95 for all combinations of sample size and test length, whereas the coverage of the CIs based on $\dot{S}_{S_X}$was too high (Table 3.1).

Table 3.1. *Standardized Bias, Standardized RMSE, and Coverage Probability of 95% CIs of the Estimated SEs of the Standard Deviation.*

		Bias		RMSE		Coverage	
Items	*N*	S_{S_X}	$\dot{S}_{S_X}$*	S_{S_X}	$\dot{S}_{S_X}$*	S_{S_X}	$\dot{S}_{S_X}$*
	500	-0.0001	0.0037	0.0010	0.0038	.949	.973
	1,000	-0.0001	0.0026	0.0005	0.0026	.951	.974
10	1,500	-0.0001	0.0021	0.0003	0.0021	.947	.971
	2,000	0.0000	0.0019	0.0002	0.0019	.950	.97
	2,500	0.0000	0.0017	0.0002	0.0017	.951	.976
	500	-0.0004	-0.0033	0.0012	0.0034	.946	.970
	1,000	-0.0002	0.0024	0.0006	0.0024	.949	.972
30	1,500	0.0000	0.0021	0.0004	0.0021	.944	.971
	2,000	-0.0001	0.0016	0.0003	0.0017	.948	.971
	2,500	0.0000	0.0016	0.0002	0.0016	.949	.973
	500	-0.0002	0.0035	0.0011	0.0036	.951	.975
	1,000	-0.0001	0.0024	0.0006	0.0025	.952	.973
50	1,500	-0.0001	0.0020	0.0004	0.0020	.952	.975
	2,000	0.0001	0.0019	0.0003	0.0019	.950	.973
	2,500	-0.0001	0.0014	0.0003	0.0015	.95	.972

Note. *SE based on Ahn & Fessler (2003, Equation 29).

Percentiles

For a test consisting of 10 items, for none of the sample sizes, the SEs of the percentile rank scores (Equation 31) showed bias. On the other hand, for tests consisting of ≥ 30 items, the lowest and highest raw scores showed small positive bias (Figure 3.2). For these extreme raw scores, bias decreased as sample size increased. For N = 500, the RMSE

of the SE estimates increased as test length increased (Figure 3.2). For other sample sizes, test length did not influence estimation precision. The RMSE of the SEs first increased as scores became more extreme, but decreased for the most extreme scores. For all conditions and percentile rank scores, the coverage of the 95% CIs was generally close to .95, except when the raw scores were closer to the extremes of the scale (Figure 3.3). For these extreme raw scores, the coverage was lower than .95. Furthermore, a sharp increase of the coverage was visible for the most extreme raw scores.

The increase of the RMSE and decrease of the coverage for scores further away from the mean can be explained by the decrease of the observed scores further away from the mean, whereas the decrease of the RMSE and increase of the coverage for the most extreme scores was caused by the proximity of the corresponding percentile ranks to 0 and 100, respectively. Considering that the extreme raw scores had a low prevalence in the samples, in absolute terms they did not change much from sample to sample. This resulted in a lower RMSE and a higher coverage.

We further investigated which percentile rank scores were associated with the extreme raw scores that corresponded to low coverage. We found that coverage of the CIs was close to .95 when $N > 500$ and the population percentile rank scores were ≥ 1% or ≤ 99%. For a test consisting of 10 items (Figure 3.3), percentile ranks ≥1% and ≤ 99% corresponded to raw scores 2 and 10, respectively. For tests containing 30 or 50 items (Figure 3.3), the raw scores for percentile ranks ≥1% and ≤ 99% were equal to 4 and 28, or 7 and 45, respectively. For $N = 500$, the coverage of the CIs was close to .95 when the population percentile rank score was ≥ 2.5% and ≤ 97.5. These percentile rank scores corresponded to raw scores 2 and 9, 6 and 26, and 9 and 43 for tests containing 10, 30, and 50 items, respectively.

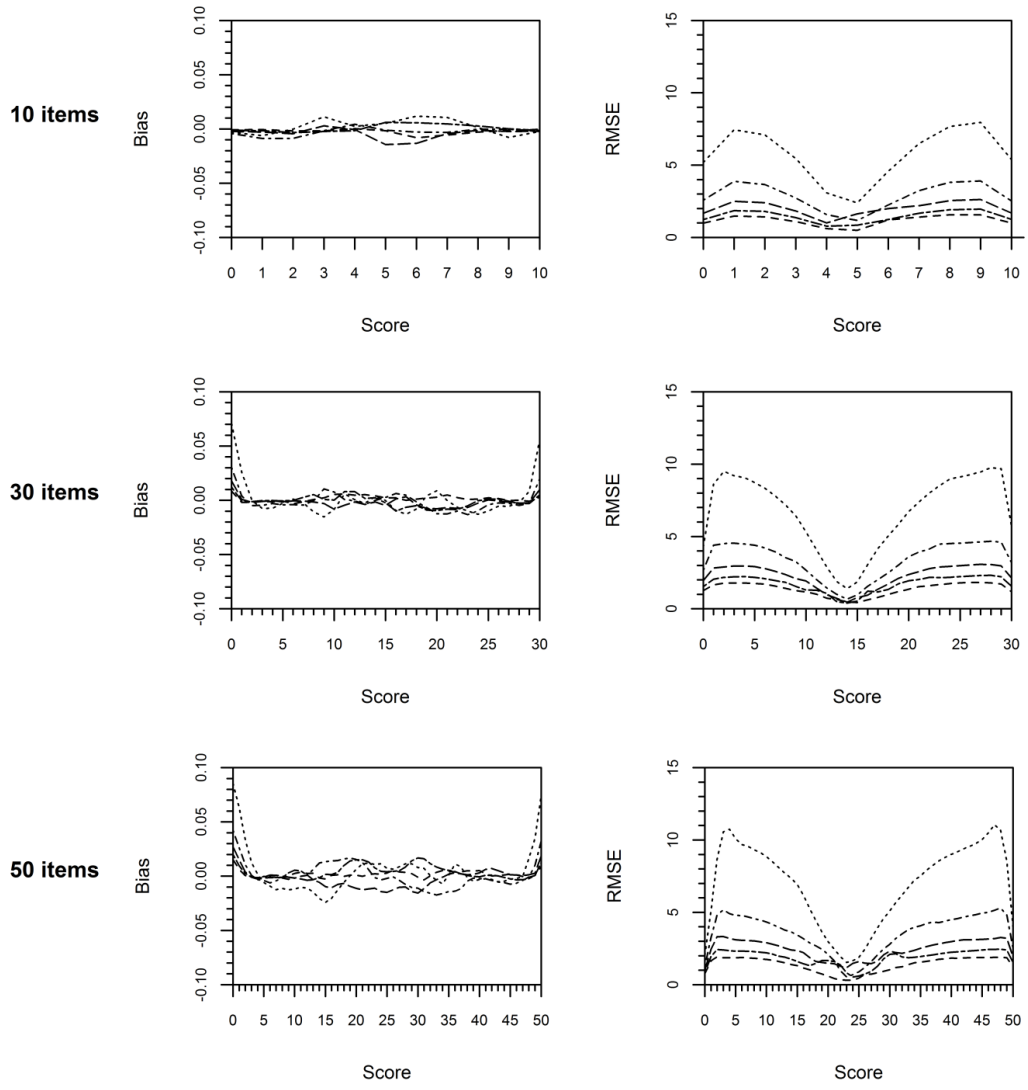

Figure 3.2. Bias and RMSE of percentile rank SEs for *N* = 500 (dotted), 1,000 (dot-dashed), 1,500 (long dashed), 2,000 (two-dashed), or 2,500 (dashed).

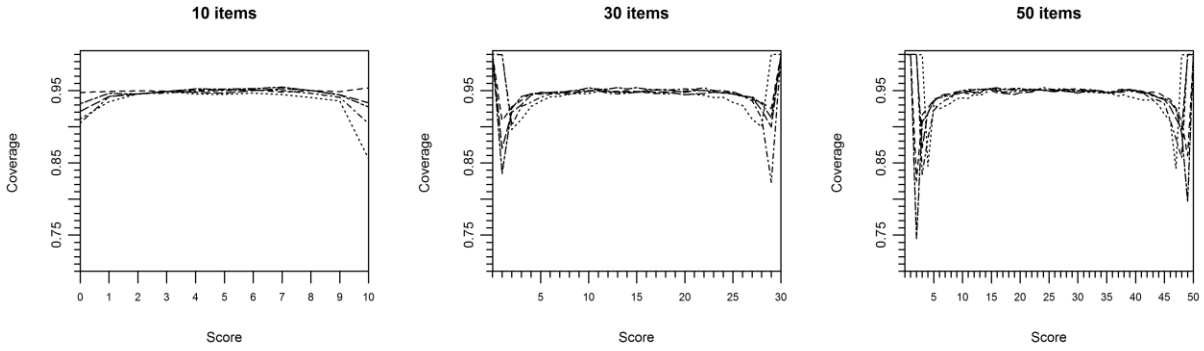

Figure 3.3. Coverage of 95% CIs for percentile ranks for *N* = 500 (dotted), 1,000 (dot-dashed), 1,500 (long dashed), 2,000 (two-dashed), or 2,500 (dashed).

Stanine Boundaries

For none of the combinations of sample size and test length, the SEs of the stanine boundaries (Equation 33) showed bias (Figure 3.4). Precision of the SE estimates increased as sample size increased (Figure 3.4). Precision was lower for the lowest and highest stanine boundaries, which can be explained by the smaller number of observations located further away from the mean. For all conditions and stanine boundaries, the coverage of the 95% CIs was close to .95 (Figure 3.5).

Z-Scores

For none of the combinations of sample size and test length, the SEs of the *Z*-score estimates (Equation 36) showed bias (Figure 3.6). Precision of the SE estimates increased as sample size increased (Figure 3.6). For all conditions and *Z*-scores, the coverage of the 95% CIs was close to .95 (Figure 3.5).

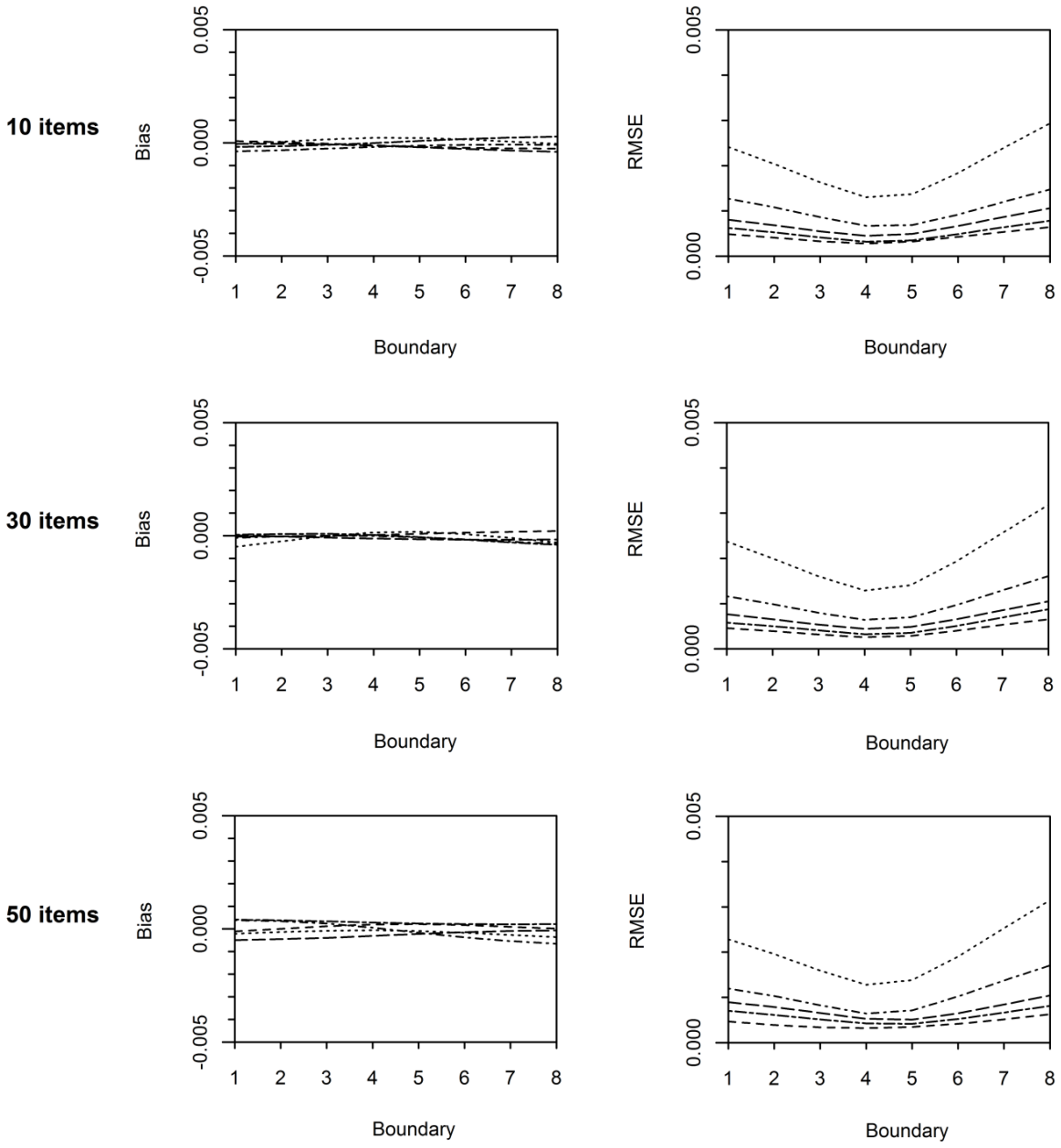

Figure 3.4. Bias and RMSE of stanine boundary SEs for *N* = 500 (dotted), 1,000 (dot-dashed), 1,500 (long dashed), 2,000 (two-dashed), or 2,500 (dashed).

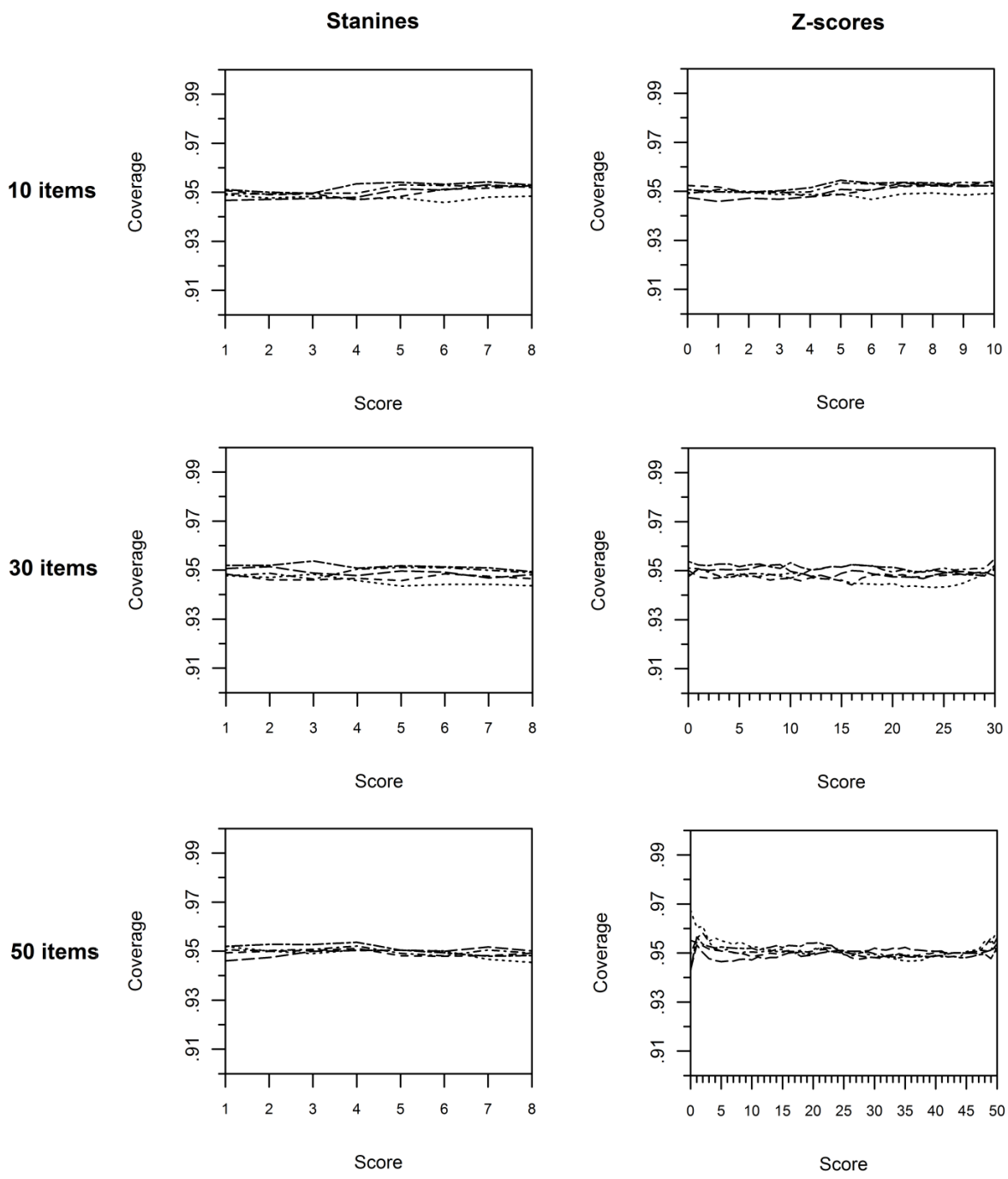

Figure 3.5. Coverage of 95% CIs for stanine boundaries and Z-scores for *N* = 500 (dotted), 1,000 (dot-dashed), 1,500 (long dashed), 2,000 (two-dashed), or 2,500 (dashed).

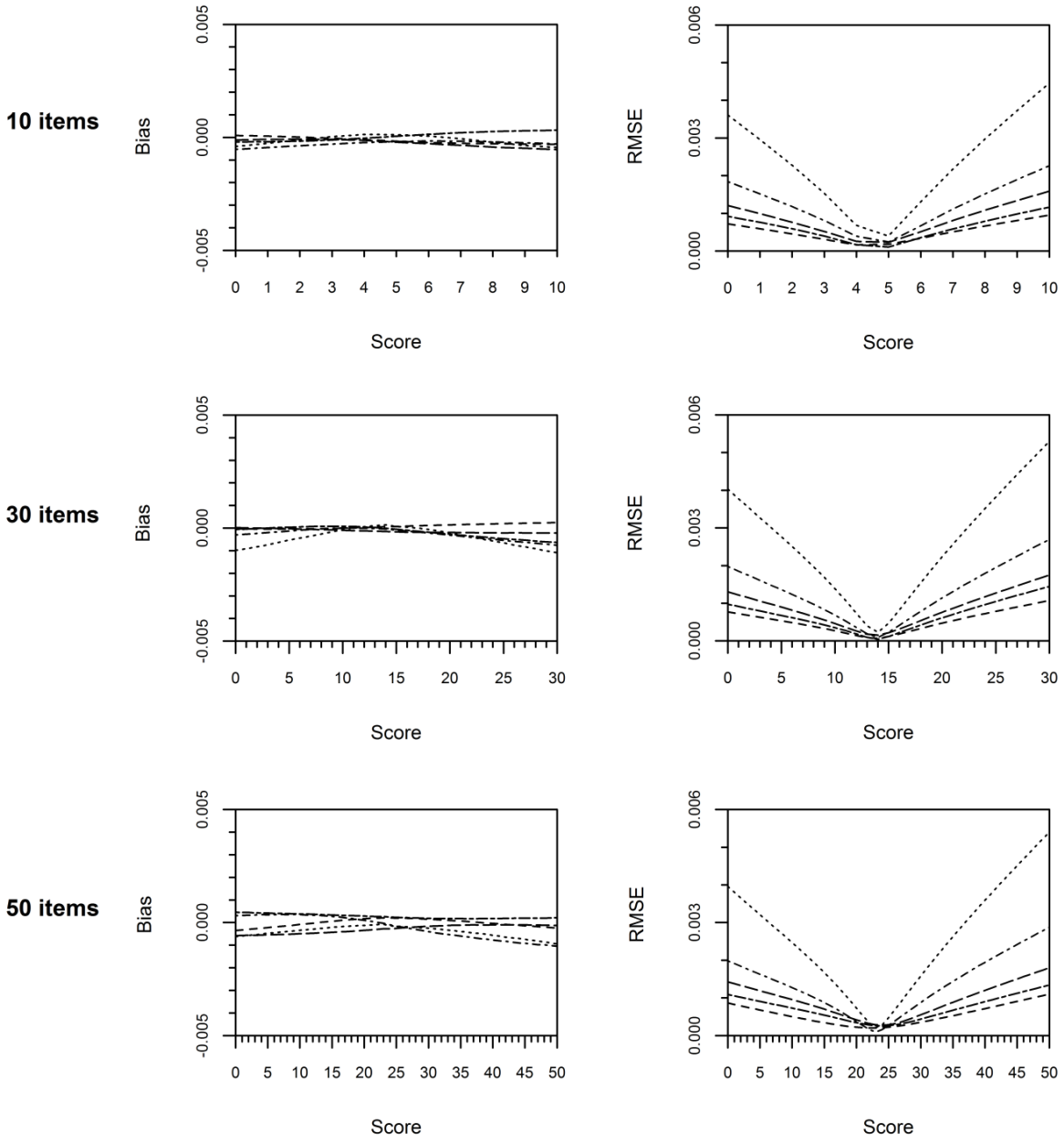

Figure 3.6. Bias and RMSE of *Z*-score SEs for *N* = 500 (dotted), 1,000 (dot-dashed), 1,500 (long dashed), 2,000 (two-dashed), or 2,500 (dashed).

3.6 Discussion

We derived SEs for the standard deviation, percentile rank scores, stanine boundaries, and *Z*-scores, assuming that the data were multinomially distributed. This is a reasonable assumption, which does not restrict the applicability of the SEs. The SEs were derived using the delta method in conjunction with the generalized exp-log notation. A simulation study suggested that for all practical purposes the SEs of the standard deviation, the stanine boundaries, and the Z-scores were unbiased, whereas the SE of the standard deviation Ahn and Fessler (2003, Equation 29) proposed was found to be biased. We also found that Wald-based CIs based on the SEs of the standard deviation, the stanines and the Z-scores had good coverage.

SEs of the percentile ranks were unbiased, unless the statistic was close to 0 or 100. We found that small samples (i.e., $N < 1{,}000$) resulted in unbiased SE estimates and CIs with good coverage if the percentile ranks were ≥ 2.5% and ≤ 97.5%. SEs or CIs for more extreme percentile ranks (e.g., the 99-th percentile rank) were unbiased and had good coverage if a larger sample (i.e., $N \geq 1{,}000$) was used to obtain unbiased estimates. It is well-known that the Wald CI may have poor coverage for small samples and for parameters close to the boundary of the parameter space (e.g., Agresti & Min, 2001). Score and profile likelihood CIs for contingency tables (Lang, 2008; also see Agresti, 2012, section 2.3.3) are range preserving, and in the future, they may provide an alternative for norm statistics close to the boundary of the parameter space. Currently these CIs are not ready to be used in the application at hand as models for norm statistics have not yet been developed. If the intended use of the test requires the distinction between extreme percentile ranks (i.e., < 2.5% and > 97.5%), the poor coverage of CIs for extreme percentile ranks can be alleviated by increasing the size of the normative sample. This will result in narrower CIs, so that the boundaries of the CI are less likely to cross the parameter boundaries. Alternatively, if a larger normative sample is not feasible, the methods described by Crawford et al. (2011) can be used to derive SEs. However, these methods can only be used to calculate SEs for percentile ranks, which means test constructors still need to employ a different method to estimate SEs for other norm statistics, such as stanine boundaries. Given that the Wald-

based CIs performed very well in most cases and given the computational complexity in score and profile likelihood CIs, we recommend using Wald CIs.

The same simulation study also suggested that precision of the SE estimates of the percentiles ranks, the stanine boundaries, and the Z-scores increased as the number of observations for the corresponding raw scores increased. Hence, the norm estimates that are of interest for the intended use of the test should be based on a sufficient number of observations. For example, if extreme percentile ranks or Z-scores are of interest, a larger sample is required, whereas the SE of the standard deviation can be precisely estimated in relatively small norm samples.

The procedure to derive SEs assumes that a simple sampling design was used. However, normative samples are often based on stratified sampling, which means the SEs should be adjusted to represent the sampling design. For a range of different sampling designs, Lehtonen and Pahkinen (2004, pp. 61-64) provide an overview of methods to adjust SEs of point estimates. These methods are based on the design effect of the type of sampling that was used to obtain the normative sample, and can be readily applied to the SEs derived by our procedure.

The delta method only uses the first partial derivatives to approximate the function of a random variable, which means this method assumes that the function is linear in the expected range of the parameter. If the function is highly non-linear and the data show considerable variation, the delta method might not result in a close approximation of the variance of the function (Cooch & White, 2015, Appendix B). For percentile rank scores, the delta method can be readily used to approximate the SEs of percentile rank scores, because these statistics are linear combinations of the observed raw score frequencies. On the other hand, the standard deviation, the stanine boundaries and the Z-scores are non-linear transformations. Incorporating higher-order Taylor expansions in the derivation of the corresponding SE might results in closer approximations of the SEs. However, such higher-order Taylor expansions very quickly result in lengthy and complicated derivations, which according to the results of our simulation study are unnecessary. Furthermore, the delta method is based on the central limit theorem and therefore cannot be applied to small samples. However, using small samples is not advocated in establishing norms for psychological and educational tests.

The methodology proposed in this article can also be applied to other statistics that can be written as a function of the observed frequencies of the unique realizations of the scores. For example, SEs have been derived for Cronbach's alpha (Kuijpers, Van der Ark, & Croon, 2013b;) and Coefficient *H* (Kuijpers et al., 2013a; Van der Ark, Croon, & Sijtsma, 2008). Further research is necessary to determine whether the methodology in this article can be applied to other functions, such as coefficients, association measures, or goodness-of-fit statistics.

We recommend test constructors to provide CIs for norm values that are presented in norm tables. Individual test results can then be compared to the CIs of the norms, instead of the point estimates. As a result, sampling error of norm values is taken into account and, especially for test scores close to the norm values, the comparison of test scores to the norms is less likely to result in erroneous decisions. Applied researchers can easily access the procedure for obtaining SEs and CIs for the mean, standard deviation, percentile ranks scores, stanine boundaries, and *Z*-scores via the *R* (*R* Core Team, 2015) function *check.norms*, which is available in the *mokken* package (Van der Ark, 2012).

3.7 Appendix A

The sample estimate of the standard deviation, s_X, can be written using the generalized exp-log notation (Equation 5) as

$$s_X = \mathbf{A}_5 \cdot \exp(\mathbf{A}_4 \cdot \log(\mathbf{A}_3 \cdot \exp(\mathbf{A}_2 \cdot \log(\mathbf{A}_1 \cdot \widehat{\mathbf{m}})))). \quad (43)$$

Let $\mathbf{y}^{(2)}$ be the vector containing the squares of the elements in $\mathbf{y}$. Then, the $4 \times k$ matrix $\mathbf{A}_1$ equals

$$\mathbf{A}_1 = [\mathbf{r} \quad \mathbf{r}^{(2)} \quad \mathbf{1}_k \quad \mathbf{1}_k]', \quad (44)$$

the 4×4 matrix $\mathbf{A}_2$ equals

$$\mathbf{A}_2 = \begin{bmatrix} 2 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \quad (45)$$

the 2×4 matrix $\mathbf{A}_3$ equals

$$\mathbf{A}_3 = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \quad (46)$$

the 1×2 vector $\mathbf{A}_4$ equals

$$\mathbf{A}_4 = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \end{bmatrix}, \quad (47)$$

and $\mathbf{A}_5$ equals the scalar 1. It follows that $\mathbf{g}_0 = \widehat{\mathbf{m}}$, the 4×1 vector $\mathbf{g}_1$ equals

$$\mathbf{g}_1 = \log(\mathbf{A}_1 \cdot \mathbf{g}_0) = \log([\mathbf{r} \quad \mathbf{r}^{(2)} \quad \mathbf{1}_k \quad \mathbf{1}_k]' \cdot \widehat{\mathbf{m}}) = \begin{bmatrix} \log(\sum X_i) \\ \log(\sum X_i^2) \\ \log(N) \\ \log(N) \end{bmatrix}, \tag{48}$$

the 4×1 vector $\mathbf{g}_2$ equals

$$\mathbf{g}_2 = \exp(\mathbf{A}_2 \cdot \mathbf{g}_1) = \exp\left(\begin{bmatrix} 2 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} \log(\sum X_i) \\ \log(\sum X_i^2) \\ \log(N) \\ \log(N) \end{bmatrix}\right) = \begin{bmatrix} \frac{(\sum X_i)^2}{N} \\ \sum X_i^2 \\ N \\ 1 \end{bmatrix}, \tag{49}$$

the 2×1 vector $\mathbf{g}_3$ equals

$$\mathbf{g}_3 = \log(\mathbf{A}_3 \cdot \mathbf{g}_2) = \log\left(\begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} \frac{(\sum X_i)^2}{N} \\ \sum X_i^2 \\ N \\ 1 \end{bmatrix}\right) \tag{50}$$

$$= \begin{bmatrix} \log(SS) \\ \log(N-1) \end{bmatrix},$$

where $SS = \sum X_i^2 - \frac{(\sum X_i)^2}{N}$, and

$$\mathbf{g}_4 = \exp(\mathbf{A}_4 \cdot \mathbf{g}_3) = \exp\left(\begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \cdot \begin{bmatrix} \log(SS) \\ \log(N-1) \end{bmatrix}\right) = \sqrt{\frac{SS}{N-1}}, \tag{51}$$

because $\mathbf{A}_5 = 1$, $\mathbf{g}(\widehat{\mathbf{m}}) = \mathbf{g}_5 = \mathbf{g}_4 = s_X$.

After some tedious algebra, it may be verified that $\mathbf{G}_0 = \mathbf{I}_k$,

$$\mathbf{G}_1 = \mathbf{D}^{-1}[\mathbf{A}_1 \cdot \mathbf{g}_0] \cdot \mathbf{A}_1 \cdot \mathbf{G}_0 \tag{52}$$

$$= \mathbf{D}^{-1}\begin{bmatrix} \sum_{i=1}^N X_i \\ \sum_{i=1}^N X_i^2 \\ N \\ N \end{bmatrix} \cdot [\mathbf{r} \quad \mathbf{r}^{(2)} \quad \mathbf{1}_k \quad \mathbf{1}_k]' \cdot \mathbf{I} = \begin{bmatrix} \frac{\mathbf{r}'}{\sum X_i} \\ \frac{\mathbf{r}^{(2)\prime}}{\sum X_i^2} \\ \frac{\mathbf{1}_k'}{N} \\ \frac{\mathbf{1}_k'}{N} \end{bmatrix},$$

$$\mathbf{G}_2 = \mathbf{D}[\exp(\mathbf{A}_2 \cdot \mathbf{g}_1)] \cdot \mathbf{A}_2 \cdot \mathbf{G}_1 \tag{53}$$

$$= \mathbf{D}\begin{bmatrix} \frac{(\sum X_i)^2}{N} \\ \sum X_i^2 \\ N \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 2 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} \frac{\mathbf{r}'}{\sum X_i} \\ \frac{\mathbf{r}^{(2)\prime}}{\sum X_i^2} \\ \frac{\mathbf{1}_k'}{N} \\ \frac{\mathbf{1}_k'}{N} \end{bmatrix} = \begin{bmatrix} 2\mathbf{r}'\bar{X} - \bar{X}^2 \\ \mathbf{r}^{(2)\prime} \\ \mathbf{1}_k' \\ \mathbf{0}_k' \end{bmatrix}.$$

Then, $\mathbf{G}_3$ is a $2 \times k$ matrix,

$$\mathbf{G}_3 = \mathbf{D}^{-1}[\mathbf{A}_3 \cdot \mathbf{g}_2] \cdot \mathbf{A}_3 \cdot \mathbf{G}_2 \tag{54}$$

$$= \mathbf{D}^{-1}\begin{bmatrix} SS \\ N-1 \end{bmatrix} \cdot \begin{bmatrix} -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} 2\mathbf{r}'\bar{X} - \bar{X}^2 \\ \mathbf{r}^{(2)\prime} \\ \mathbf{1}_k' \\ \mathbf{0}_k' \end{bmatrix} = \begin{bmatrix} \frac{\mathbf{d}^{(2)\prime}}{SS} \\ \mathbf{e}' \end{bmatrix},$$

where the $k \times 1$ vector $\mathbf{d} = \mathbf{r} - \bar{X}$ and $\mathbf{e} = \mathbf{1}_k/(N-1)$,

and $\mathbf{G}_4$ is the $1 \times k$, vector

$$\mathbf{G}_4 = \mathbf{D}[\exp(\mathbf{A}_4 \cdot \mathbf{g}_3)] \cdot \mathbf{A}_4 \cdot \mathbf{G}_3 \tag{55}$$

$$= \mathbf{D}[s_X] \cdot \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \cdot \begin{bmatrix} \frac{\mathbf{d}^{(2)}}{SS} \\ \mathbf{e} \end{bmatrix} = .5s_X\left(\frac{\mathbf{d}^2}{SS} - \mathbf{e}\right).$$

Finally, one can derive that $\mathbf{G} = \mathbf{G}_5 = \mathbf{G}_4$.

Because the sample estimate of the standard deviation is obtained by dividing the sum of squared deviation scores by $N-1$, $\mathbf{g}(\widehat{\mathbf{m}})$ (Equation 50) is not homogeneous of order 0. As a result, the asymptotic variance is approximated by $V_{S_X} \approx \mathbf{G}\mathbf{D}(\widehat{\mathbf{m}})\mathbf{G}' - \mathbf{G}\widehat{\mathbf{m}}N^{-1}\widehat{\mathbf{m}}'\mathbf{G}'$ (cf. Equation 2). Inserting the Jacobian $\mathbf{G}$ (Equation 54) in Equation 2 results in the sample estimate of the asymptotic variance of s_X,

$$V_{s_X} \approx 0.25s_X^2 \cdot \sum_{\boldsymbol{i}}\sum_{\boldsymbol{j}}\left(\frac{d_i^2}{SS} - e\right)\left(\frac{d_j^2}{SS} - e\right)\left(\delta_{ij}\widehat{m}_i - \frac{\widehat{m}_i\widehat{m}_j}{N}\right). \tag{56}$$

For large N,

$$V_{s_X} \approx 0.25s_X^2 \cdot \sum_{\boldsymbol{i}}\widehat{m}_i\left(\frac{d_i^2}{SS} - \frac{1}{N}\right)^2. \tag{57}$$

3.8 Appendix B

The sample estimates of the percentile ranks, collected in vector $\mathbf{PR} = (PR_{r_1} \quad \cdots \quad PR_{r_k})'$ (Equation 26), can be written as

$$\mathbf{PR} = \mathbf{A}_3 \cdot \exp(\mathbf{A}_2 \cdot \log(\mathbf{A}_1 \cdot \widehat{\mathbf{m}})). \tag{58}$$

Let $\mathbf{L}_p$ be a lower triangular matrix of ones. The $(k+1) \times k$ matrix $\mathbf{A}_1$ equals

$$\mathbf{A}_1 = \begin{bmatrix} \mathbf{L}_k \\ \mathbf{1}_k' \end{bmatrix}, \tag{59}$$

and the $k \times (k+1)$ matrix $\mathbf{A}_2$ equals

$$\mathbf{A}_2 = [\mathbf{I}_k \quad -\mathbf{1}_k]. \tag{60}$$

Let $\mathbf{A}_3$ be a lower bidiagonal $k \times k$ matrix in which each nonzero element equals 50; that is,

$$\mathbf{A}_3 = 50\left(\mathbf{I}_k + \begin{bmatrix} \mathbf{0}_{k-1}' & 0 \\ \mathbf{I}_{k-1} & \mathbf{0}_{k-1} \end{bmatrix}\right). \tag{61}$$

It follows that $\mathbf{g}_0 = \hat{\mathbf{m}}$. Let $\hat{\mathbf{m}}^* = (\sum_{i=1}^{1} \hat{\mathrm{m}}_i, \sum_{i=1}^{2} \hat{\mathrm{m}}_i, \ldots, \sum_{i=1}^{k} \hat{\mathrm{m}}_i)$ be a vector of cumulative frequencies; that is, $\hat{\mathbf{m}}^* = \mathbf{L}_k \cdot \hat{\mathbf{m}}$, and let $\mathbf{P}^* = \hat{\mathbf{m}}^*/N = \left(P(X \le r_1), \ldots, P(X \le r_k)\right)'$.

Function $\mathbf{g}_1$ is a $(k+1) \times 1$ vector,

$$\mathbf{g}_1 = \log(\mathbf{A}_1 \cdot \hat{\mathbf{m}}) = \log\left(\begin{bmatrix} \mathbf{L}_k \\ \mathbf{1}_k' \end{bmatrix} \cdot \hat{\mathbf{m}}\right) = \log\begin{bmatrix} \hat{\mathbf{m}}^* \\ N \end{bmatrix}, \tag{62}$$

$\mathbf{g}_2$ is a $k \times 1$ vector,

$$\mathbf{g}_2 = \exp(\mathbf{A}_2 \cdot \mathbf{g}_1) = \exp\left([\mathbf{I}_k \quad -\mathbf{1}_k] \cdot \log\begin{bmatrix} \hat{\mathbf{m}}^* \\ N \end{bmatrix}\right) = \mathbf{P}^*, \tag{63}$$

and $\mathbf{g}_3$ is a $k \times 1$ vector,

$$\mathbf{g}(\hat{\mathbf{m}}) = \mathbf{g}_3 = \mathbf{A}_3 \cdot \mathbf{g}_2 = 50\left(\mathbf{I}_k + \begin{bmatrix} \mathbf{0}_{k-1}' & 0 \\ \mathbf{I}_{k-1} & \mathbf{0}_{k-1} \end{bmatrix}\right) \cdot \mathbf{P}^* = \mathbf{PR}\,. \tag{64}$$

It follows that $\mathbf{G}_0 = \mathbf{I}$, the $(k+1) \times k$ matrix $\mathbf{G}_1$ equals

$$\mathbf{G}_1 = \mathbf{D}^{-1}[\mathbf{A}_1 \cdot \mathbf{g}_0] \cdot \mathbf{A}_1 \cdot \mathbf{G}_0 = \mathbf{D}^{-1}\left(\begin{bmatrix} \hat{\mathbf{m}}^* \\ N \end{bmatrix}\right) \cdot \begin{bmatrix} \mathbf{L}_k \\ \mathbf{1}_k' \end{bmatrix}, \tag{65}$$

the $k \times k$ matrix $\mathbf{G}_2$ equals

$$\mathbf{G}_2 = \mathbf{D}[\exp(\mathbf{A}_2 \cdot \mathbf{g}_1)] \cdot \mathbf{A}_2 \cdot \mathbf{G}_1 = \mathbf{D}(\mathbf{P}^*) \cdot [\mathbf{I}_k \quad -\mathbf{1}_k] \cdot \mathbf{D}^{-1}\left(\begin{bmatrix} \hat{\mathbf{m}}^* \\ N \end{bmatrix}\right) \cdot \begin{bmatrix} \mathbf{L}_k \\ \mathbf{1}_k' \end{bmatrix}, \tag{66}$$

and $k \times k$ matrix $\mathbf{G}_3$ equals

$$\begin{aligned} \mathbf{G} &= \mathbf{G}_3 = \mathbf{A}_3 \cdot \mathbf{G}_2 \\ &= 50\left(\mathbf{I}_k + \begin{bmatrix} \mathbf{0}_{k-1}' & 0 \\ \mathbf{I}_{k-1} & \mathbf{0}_{k-1} \end{bmatrix}\right) \cdot \mathbf{D}([\mathbf{P}^*]) \cdot [\mathbf{I}_k \quad -\mathbf{1}_k] \cdot \mathbf{D}^{-1}\left(\begin{bmatrix} \hat{\mathbf{m}}^* \\ N \end{bmatrix}\right) \cdot \begin{bmatrix} \mathbf{L}_k \\ \mathbf{1}_k' \end{bmatrix}. \end{aligned} \tag{67}$$

Some tedious algebra shows that the elements of $\mathbf{G}$ equal

$$G_{gi} = \frac{50}{N} \cdot \begin{cases} 2 - P_{g-1}^* - P_g^* & \text{if } g > i \\ 1 - P_{g-1}^* - P_g^* & \text{if } g = i, \\ 0 - P_{g-1}^* - P_g^* & \text{if } g < i \end{cases} \tag{68}$$

where $P_x^* = P(X \leq x)$. Because the percentile ranks are homogeneous of order 0, the asymptotic variance is approximated by $V_{\mathbf{PR}} \approx \mathbf{GD}(\widehat{\mathbf{m}})\mathbf{G}'$ (cf. Equation 3). The sample estimates of the elements of the asymptotic covariance matrix of **PR** are:

$$V_{\mathbf{PR}_{gh}} \approx \frac{2{,}500}{N^2} \sum_i \widehat{\mathrm{m}}_i (\gamma_{gi} - P_{g-1}^* - P_g^*)(\gamma_{hi} - P_{h-1}^* - P_h^*) \tag{69}$$

for $g, h = 1, \ldots, k$.

3.9 Appendix C

The sample estimates of the 8 boundaries of the stanines (Equation 28) can be written using the generalized exp-log notation (Equation 5) as follows

$$\mathbf{St}_{\boldsymbol{b}} = \mathbf{A}_5 \cdot \exp(\mathbf{A}_4 \cdot \log(\mathbf{A}_3 \cdot \exp(\mathbf{A}_2 \cdot \log(\mathbf{A}_1 \cdot \widehat{\mathbf{m}}))))). \tag{70}$$

Let $\mathbf{A}_1$ be the $4 \times k$ matrix,

$$\mathbf{A}_1 = [\mathbf{r} \quad \mathbf{r}^{(2)} \quad \mathbf{1}_k \quad \mathbf{1}_k]', \tag{71}$$

let $\mathbf{A}_2$ be the 5×4 matrix

$$\mathbf{A}_2 = \begin{bmatrix} 2 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \tag{72}$$

let $\mathbf{A}_3$ be the 3×5 matrix

$$\mathbf{A}_3 = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \tag{73}$$

let $\mathbf{A}_4$ be the 2×3 matrix,

$$\mathbf{A}_4 = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{74}$$

and let $\mathbf{A}_5$ be the 8×2 matrix

$$\mathbf{A}_5 = [\mathbf{f} \quad \mathbf{1}_8]. \tag{75}$$

It follows that $\mathbf{g}_0 = \widehat{\mathbf{m}}$, $\mathbf{g}_1$ is the 4×1 vector

$$\mathbf{g}_1 = \log(\mathbf{A}_1 \cdot \mathbf{g}_0) = \log([\mathbf{r} \quad \mathbf{r}^{(2)} \quad \mathbf{1}_k \quad \mathbf{1}_k]' \cdot \widehat{\mathbf{m}}) = \log\left(\begin{bmatrix} \sum X_i \\ \sum X_i^2 \\ N \\ N \end{bmatrix}\right), \tag{76}$$

$\mathbf{g}_2$ is the 5×1 vector

$$\mathbf{g}_2 = \exp(\mathbf{A}_2 \cdot \mathbf{g}_1) = \exp\left(\begin{bmatrix} 2 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \cdot \log\left(\begin{bmatrix} \sum X_i \\ \sum X_i^2 \\ N \\ N \end{bmatrix}\right)\right) = \begin{bmatrix} \frac{(\sum X_i)^2}{N} \\ \sum X_i^2 \\ N \\ \bar{X} \\ 1 \end{bmatrix}, \tag{77}$$

$\mathbf{g}_3$ is the 3×1 vector

$$\mathbf{g}_3 = \log(\mathbf{A}_3 \cdot \mathbf{g}_2) = \log\left(\begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} \frac{(\sum X_i)^2}{N} \\ \sum X_i^2 \\ N \\ \bar{X} \\ 1 \end{bmatrix}\right), \tag{78}$$

$$= \log\left(\begin{bmatrix} SS \\ N-1 \\ \bar{X} \end{bmatrix}\right),$$

$\mathbf{g}_4$ is the 2×1 vector

$$\mathbf{g}_4 = \exp(\mathbf{A}_4 \cdot \mathbf{g}_3) = \exp\left(\begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \log\left(\begin{bmatrix} SS \\ N-1 \\ \bar{X} \end{bmatrix}\right)\right) = \begin{bmatrix} S_X \\ \bar{X} \end{bmatrix}, \tag{79}$$

and $\mathbf{g}_5$ is the 8×1 vector

$$\mathbf{g}_5 = \mathbf{A}_5 \cdot \mathbf{g}_4 = [\mathbf{f} \quad \mathbf{1}_8] \cdot \begin{bmatrix} S_X \\ \bar{X} \end{bmatrix} = \mathbf{g}(\hat{\mathbf{m}}) = \mathbf{St_b}. \tag{80}$$

Next, $\mathbf{G}_0 = \mathbf{I}_k$. $\mathbf{G}_1$ is the $4 \times k$ matrix,

$$\mathbf{G}_1 = \mathbf{D}^{-1}[\mathbf{A}_1 \cdot \mathbf{g}_0] \cdot \mathbf{A}_1 \cdot \mathbf{G}_0 = \mathbf{D}^{-1} \begin{bmatrix} \sum X_i \\ \sum X_i^2 \\ N \\ N \end{bmatrix} \cdot [\mathbf{r} \quad \mathbf{r}^{(2)} \quad \mathbf{1}_k \quad \mathbf{1}_k]' = \begin{bmatrix} \frac{\mathbf{r}'}{\sum X_i} \\ \frac{\mathbf{r}^{(2)\prime}}{\sum X_i^2} \\ \frac{\mathbf{1}_k'}{N} \\ \frac{\mathbf{1}_k'}{N} \end{bmatrix}. \tag{81}$$

Then $\mathbf{G}_2$ is the $5 \times k$ matrix

$$\mathbf{G}_2 = \mathbf{D}[\exp(\mathbf{A}_2 \cdot \mathbf{g}_1)] \cdot \mathbf{A}_2 \cdot \mathbf{G}_1 \tag{82}$$

$$= \mathbf{D}\left(\begin{bmatrix} \frac{(\sum X_i)^2}{N} \\ \sum X_i^2 \\ N \\ \bar{X} \\ 1 \end{bmatrix}\right) \begin{bmatrix} 2 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \frac{\mathbf{r}'}{\sum X_i} \\ \frac{\mathbf{r}^{(2)\prime}}{\sum X_i^2} \\ \frac{\mathbf{1}_k'}{N} \\ \frac{\mathbf{1}_k'}{N} \end{bmatrix} = \begin{bmatrix} 2\mathbf{r}'\bar{X} - \bar{X}^2 \\ \mathbf{r}^{(2)\prime} \\ \mathbf{1}_k' \\ \frac{\mathbf{d}'}{N} \\ \mathbf{0}_k' \end{bmatrix},$$

$\mathbf{G}_3$ is the $3 \times k$ matrix

$$\mathbf{G}_3 = \mathbf{D}^{-1}[\mathbf{A}_3 \cdot \mathbf{g}_2] \cdot \mathbf{A}_3 \cdot \mathbf{G}_2 \tag{83}$$

$$= \mathbf{D}^{-1}\left(\begin{bmatrix} SS \\ N-1 \\ \bar{X} \end{bmatrix}\right) \cdot \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 2\mathbf{r}'\bar{X} - \bar{X}^2 \\ \mathbf{r}^{(2)\prime} \\ \mathbf{1}_k' \\ \frac{\mathbf{d}'}{N} \\ \mathbf{0}_k' \end{bmatrix} = \begin{bmatrix} \mathbf{d}^{*\prime} \\ \mathbf{e}' \\ \frac{\mathbf{d}'}{\sum X_i} \end{bmatrix},$$

where $\mathbf{d}^* = \frac{\mathbf{d}^{(2)}}{SS}$. Then $\mathbf{G}_4$ is the $2 \times k$ matrix

$$\mathbf{G}_4 = \mathbf{D}[\exp(\mathbf{A}_4 \cdot \mathbf{g}_3)] \cdot \mathbf{A}_4 \cdot \mathbf{G}_3 \tag{84}$$

$$= \begin{bmatrix} s_X & 0 \\ 0 & \bar{X} \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{d}^{*\prime} \\ \mathbf{e}' \\ \frac{\mathbf{d}'}{\sum X_i} \end{bmatrix} = \begin{bmatrix} \frac{s_X}{2} \cdot (\mathbf{d}^{*\prime} - \mathbf{e}') \\ \frac{\mathbf{d}'}{N} \end{bmatrix},$$

and $\mathbf{G}_5$ is the $8 \times k$ matrix

$$\mathbf{G}_5 = \mathbf{G} = \mathbf{A}_5 \cdot \mathbf{G}_4 = [\mathbf{f} \quad \mathbf{1}_8] \begin{bmatrix} \frac{s_X}{2} \cdot (\mathbf{d}^{*\prime} - \mathbf{e}') \\ \frac{\mathbf{d}'}{N} \end{bmatrix} = \frac{\mathbf{f} \cdot s_X}{2} \cdot (\mathbf{d}^{*\prime} - \mathbf{e}') + \frac{\mathbf{1}_8 \cdot \mathbf{d}'}{N}. \tag{85}$$

The stanine boundaries are not homogeneous of order 0, because $\mathbf{St_b}$ is obtained by using s_X. As a result, the asymptotic variance is approximated by $\mathbf{V_{St}} \approx \mathbf{G}\mathbf{V}_{\hat{\mathbf{m}}}\mathbf{G}'$ (cf. Equation 2). Some tedious algebra shows that the sample estimates of the elements of the asymptotic covariance matrix of $\mathbf{V_{St}}$ are

$$V_{\mathbf{St}_{gh}} \approx \sum_{j=1}^{k} \sum_{i=1}^{k} t \cdot \left[\frac{f_g \cdot s_X}{2} \cdot (d_i^* - e) + \frac{d_i}{N}\right] \cdot \left[\frac{f_h \cdot s_X}{2} \cdot \left(d_j^* - e\right) + \frac{d_j}{N}\right], \tag{86}$$

where $t = \delta_{ij}\hat{\mathrm{m}}_j - \frac{\hat{\mathrm{m}}_i \hat{\mathrm{m}}_j}{N}$. For large N, Equation 84 reduces to

$$V_{\mathbf{St}_{gh}} \approx \sum_{i=1}^{k} \hat{\mathrm{m}}_i \cdot \left[\frac{f_g \cdot s_X}{2} \cdot \left(d_i^* - \frac{1}{N}\right) + \frac{d_i}{N}\right] \cdot \left[\frac{f_h \cdot s_X}{2} \cdot \left(d_i^* - \frac{1}{N}\right) + \frac{d_i}{N}\right]. \tag{87}$$

3.10 Appendix D

The sample estimate of the k standardized scores corresponding to $\mathbf{r}$, collected in a $k \times 1$ vector $\mathbf{z}$ (Equation 31), can be written as

$$\mathbf{z} = \mathbf{A}_7 \cdot \exp(\mathbf{A}_6 \cdot \log(\mathbf{A}_5 \cdot \exp(\mathbf{A}_4 \cdot \log(\mathbf{A}_3 \cdot \exp(\mathbf{A}_2 \cdot \log(\mathbf{A}_1 \cdot \hat{\mathbf{m}})))))). \tag{88}$$

Let $\mathbf{A}_1$ be the $(k+2) \times k$ matrix

$$\mathbf{A}_1 = [\mathbf{I}_k \quad \mathbf{1}_k \quad \mathbf{1}_k]'. \tag{89}$$

Let $\oplus$ indicate the direct sum, for example: $\mathbf{X} \oplus \mathbf{Y} = \begin{bmatrix} \mathbf{X} & 0 \\ 0 & \mathbf{Y} \end{bmatrix}$, then $\mathbf{A}_2$ is the $(k+1) \times (k+2)$ matrix

$$\mathbf{A}_2 = \mathbf{I}_k \oplus [1 \quad -1], \tag{90}$$

$\mathbf{A}_3$ is the $(k+4) \times (k+1)$ matrix

$$\mathbf{A}_3 = [\mathbf{r} \quad \mathbf{r}^{(2)} \quad \mathbf{1}_k \quad \mathbf{1}_k]' \oplus \mathbf{r}, \tag{91}$$

$\mathbf{A}_4$ is the $(5+k) \times (4+k)$ matrix

$$\mathbf{A}_4 = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 2 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \oplus \mathbf{I}_k, \tag{92}$$

$\mathbf{A}_5$ is the $(k+3) \times (k+5)$ matrix

$$\mathbf{A}_5 = 1 \oplus [-1 \quad 1] \oplus [1 \quad -1] \oplus \mathbf{I}_k, \tag{93}$$

$\mathbf{A}_6$ is the $(1+k) \times (3+k)$ matrix

$$\mathbf{A}_6 = \begin{bmatrix} 1 & -\frac{1}{2} & \frac{1}{2} & \mathbf{0}_k' \\ \mathbf{0}_k & -\mathbf{1}_k \cdot \frac{1}{2} & \mathbf{1}_k \cdot \frac{1}{2} & \mathbf{I}_\mathrm{k} \end{bmatrix}, \tag{94}$$

and $\mathbf{A}_7$ is the $k \times (k+1)$ matrix

$$\mathbf{A}_7 = [-\mathbf{1}_k \quad \mathbf{I}_k]. \tag{95}$$

It follows that $\mathbf{g}_0 = \widehat{\mathbf{m}}$, the $(k+2) \times 1$ vector $\mathbf{g}_1$ equals

$$\mathbf{g}_1 = \log(\mathbf{A}_1 \cdot \widehat{\mathbf{m}}) = \log([\mathbf{I}_k \quad \mathbf{1}_k \quad \mathbf{1}_k]' \cdot \widehat{\mathbf{m}}) = \log([\widehat{\mathbf{m}}' \quad N \quad N]'), \tag{96}$$

the $(k+1) \times 1$ vector $\mathbf{g}_2$ equals

$$\mathbf{g}_2 = \exp(\mathbf{A}_2 \cdot \mathbf{g}_1) = \exp(\mathbf{I}_k \oplus [1 \quad -1] \cdot \log([\widehat{\mathbf{m}}' \quad N \quad N]')) = [\widehat{\mathbf{m}}' \quad 1]', \tag{97}$$

the $(k+4) \times 1$ vector $\mathbf{g}_3$ equals

$$\mathbf{g}_3 = \log(\mathbf{A}_3 \cdot \mathbf{g}_2) = \log\big(([\mathbf{r} \quad \mathbf{r}^{(2)} \quad \mathbf{1}_k \quad \mathbf{1}_k]' \oplus \mathbf{r}) \cdot [\widehat{\mathbf{m}}' \quad 1]'\big) = \log\begin{bmatrix} \sum X_i \\ \sum X_i^2 \\ N \\ N \\ \mathbf{r} \end{bmatrix}, \tag{98}$$

the $(k+5) \times 1$ vector $\mathbf{g}_4$ equals

$$\mathbf{g}_4 = \exp(\mathbf{A}_4 \cdot \mathbf{g}_3) \tag{99}$$

$$= \exp\left(\left(\begin{bmatrix} 1 & 0 & 0 & -1 \\ 2 & 0 & 0 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix} \oplus \mathbf{I}_k\right) \cdot \log\begin{bmatrix} \sum X_i \\ \sum X_i^2 \\ N \\ N \\ \mathbf{r} \end{bmatrix}\right) = \begin{bmatrix} \bar{X} \\ \frac{(\sum X_i)^2}{N} \\ \sum X_i^2 \\ N \\ 1 \\ \mathbf{r} \end{bmatrix},$$

the $(k+3) \times 1$ vector $\mathbf{g}_5$ equals

$$\mathbf{g}_5 = \log(\mathbf{A}_5 \cdot \mathbf{g}_4) \tag{100}$$

$$= \log\left([1 \oplus [-1 \quad 1] \oplus [1 \quad -1] \oplus \mathbf{I}_k] \cdot \begin{bmatrix} \bar{X} \\ \frac{(\sum X_i)^2}{N} \\ \sum X_i^2 \\ N \\ 1 \\ \mathbf{r} \end{bmatrix}\right) = \log\begin{bmatrix} \bar{X} \\ SS \\ N-1 \\ \mathbf{r} \end{bmatrix},$$

the $(k+1) \times 1$ vector $\mathbf{g}_6$ equals

$$\mathbf{g}_6 = \exp(\mathbf{A}_6 \cdot \mathbf{g}_5) \tag{101}$$

$$= \exp\left(\begin{bmatrix} 1 & -\frac{1}{2} & \frac{1}{2} & \mathbf{0}_k' \\ \mathbf{0}_k & -\mathbf{1}_k\frac{1}{2} & \mathbf{1}_k\frac{1}{2} & \mathbf{I}_k \end{bmatrix} \cdot \log\begin{bmatrix} \bar{X} \\ SS \\ N-1 \\ \mathbf{r} \end{bmatrix}\right) = \begin{bmatrix} \frac{\bar{X}}{s_X} \\ \frac{\mathbf{r}}{s_X} \end{bmatrix},$$

and the $k \times 1$ vector $\mathbf{g}_7$ equals

$$\mathbf{g_7} = \mathbf{g}(\hat{\mathbf{m}}) = \mathbf{A}_7 \cdot \mathbf{g}_6 = [-\mathbf{1}_k \quad \mathbf{I}_k] \cdot \begin{bmatrix} \frac{\bar{X}}{s_X} \\ \frac{\mathbf{r}}{s_X} \end{bmatrix} = \begin{bmatrix} \mathbf{d} \\ \frac{}{s_X} \end{bmatrix} = \mathbf{z}\,. \tag{102}$$

Next, $\mathbf{G}_0 = \mathbf{I}_k$, the $(k+2) \times k$ matrix $\mathbf{G}_1$ equals

$$\mathbf{G}_1 = \mathbf{D}^{-1}[\mathbf{A}_1 \cdot \mathbf{g}_0] \cdot \mathbf{A}_1 \cdot \mathbf{G}_0 \tag{103}$$

$$= \mathbf{D}^{-1}([\hat{\mathbf{m}}' \quad N \quad N]') \cdot [\mathbf{I}_k \quad \mathbf{1}_k \quad \mathbf{1}_k]' = \begin{bmatrix} \mathbf{D}^{-1}(\hat{\mathbf{m}}) \\ 1_k'/N \\ 1_k'/N \end{bmatrix},$$

the $(k+1) \times k$ matrix $\mathbf{G}_2$ equals

$$\mathbf{G}_2 = \mathbf{D}[\exp(\mathbf{A}_2 \cdot \mathbf{g}_1)] \cdot \mathbf{A}_2 \cdot \mathbf{G}_1 \tag{104}$$

$$= \mathbf{D}([\hat{\mathbf{m}} \quad 1]') \cdot \left[\mathbf{I}_k \oplus [1 \quad -1]\right] \cdot \begin{bmatrix} \mathbf{D}^{-1}(\hat{\mathbf{m}}) \\ 1_k'/N \\ 1_k'/N \end{bmatrix} = \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0}_k' \end{bmatrix},$$

the $(k+4) \times k$ matrix $\mathbf{G}_3$ equals

$$\mathbf{G}_3 = \mathbf{D}^{-1}\left(\begin{bmatrix}\sum X_i\\ \sum X_i^2\\ N\\ N\\ \mathbf{r}\end{bmatrix}\right)\cdot[[\mathbf{r} \quad \mathbf{r}^{(2)} \quad \mathbf{1}_k \quad \mathbf{1}_k]' \oplus \mathbf{r}]\cdot\begin{bmatrix}\mathbf{I}_k\\ \mathbf{0}_k'\end{bmatrix} = \begin{bmatrix}\mathbf{r}'/\sum X_i\\ \mathbf{r}^{(2)\prime}/\sum X_i^2\\ \mathbf{1}_k'/N\\ \mathbf{1}_k'/N\\ \mathbf{0}_{k\times k}\end{bmatrix}, \tag{105}$$

the $(k+5)\ \times\ k$ matrix $\mathbf{G}_4$ equals

$$\mathbf{G}_4 = \mathbf{D}[\exp(\mathbf{A}_4\cdot\mathbf{g}_3)]\cdot\mathbf{A}_4\cdot\mathbf{G}_3 \tag{106}$$

$$= \mathbf{D}\left(\begin{bmatrix}\bar{X}\\ \frac{(\sum X_i)^2}{N}\\ \sum X_i^2\\ N\\ 1\\ \mathbf{r}\end{bmatrix}\right)\cdot\left[\begin{bmatrix}1 & 0 & 0 & -1\\ 2 & 0 & 0 & -1\\ 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 1\\ 0 & 0 & 1 & -1\end{bmatrix}\oplus\mathbf{I}_k\right]\cdot\begin{bmatrix}\mathbf{r}'/\sum X_i\\ \mathbf{r}^{(2)\prime}/\sum X_i^2\\ \mathbf{1}_k'/N\\ \mathbf{1}_k'/N\\ \mathbf{0}_{k\times k}\end{bmatrix} = \begin{bmatrix}\frac{\mathbf{d}'}{N}\\ \mathbf{r}'2\bar{X}-\bar{X}^2\\ \mathbf{r}^{(2)\prime}\\ \mathbf{1}_k'\\ \mathbf{0}_{(k+1)\times k}\end{bmatrix},$$

the $(k+3)\ \times\ k$ matrix $\mathbf{G}_5$ equals

$$\mathbf{G}_5 = \mathbf{D}^{-1}[\mathbf{A}_5\cdot\mathbf{g}_4]\cdot\mathbf{A}_5\cdot\mathbf{G}_4 \tag{107}$$

$$= \mathbf{D}^{-1}\begin{bmatrix}\bar{X}\\ SS\\ N-1\\ \mathbf{r}\end{bmatrix}\cdot[1\oplus[-1 \quad 1]\oplus[1 \quad -1]\oplus\mathbf{I}_k]\cdot\begin{bmatrix}\frac{\mathbf{d}'}{N}\\ \mathbf{r}'2\bar{X}-\bar{X}^2\\ \mathbf{r}^{(2)\prime}\\ \mathbf{1}_k'\\ \mathbf{0}_{(k+1)\times k}\end{bmatrix}$$

$$= \begin{bmatrix}\frac{\mathbf{r}'}{\sum X_i}-\frac{1}{N}\\ \mathbf{d}^{*\prime}\\ \mathbf{e}'\\ \mathbf{0}_{k\times k}\end{bmatrix},$$

and the $(k+1)\times\ k$ matrix $\mathbf{G}_6$ equals

$$\mathbf{G}_6 = \mathbf{D}[\exp(\mathbf{A}_6\cdot\mathbf{g}_5)]\cdot\mathbf{A}_6\cdot\mathbf{G}_5 \tag{108}$$

$$= \mathbf{D}\left(\begin{bmatrix}\frac{\bar{X}}{s_X}\\ \frac{\mathbf{r}}{s_X}\end{bmatrix}\right)\cdot\begin{bmatrix}1 & -\frac{1}{2} & \frac{1}{2} & \mathbf{0}_k'\\ \mathbf{0}_k & -\mathbf{1}_k\frac{1}{2} & \mathbf{1}_k\frac{1}{2} & \mathbf{I}_k\end{bmatrix}\cdot\begin{bmatrix}\frac{\mathbf{r}'}{\sum X_i}-\frac{1}{N}\\ \mathbf{d}^{*\prime}\\ \mathbf{e}'\\ \mathbf{0}_{k\times k}\end{bmatrix} = \begin{bmatrix}\frac{\bar{X}}{s_X}\cdot\left[\frac{\mathbf{r}'}{\sum X}-\frac{\mathbf{1}_k'}{N}-.5\left(\mathbf{d}^{*\prime}-\mathbf{e}'\right)\right]\\ \frac{-.5\mathbf{r}}{s_X}\cdot\left(\mathbf{d}^{*\prime}-\mathbf{e}'\right)\end{bmatrix}.$$

Finally, the $k\ \times\ k$ matrix $\mathbf{G}_7$ equals

$$\mathbf{G} = \mathbf{A}_7\cdot\mathbf{G}_6 = [-\mathbf{1}_k \quad \mathbf{I}_k]\cdot\begin{bmatrix}\frac{\bar{X}}{s_X}\cdot\left[\frac{\mathbf{r}'}{\sum X}-\frac{\mathbf{1}_k'}{N}-.5\left(\mathbf{d}^{*\prime}-\mathbf{e}'\right)\right]\\ \frac{-.5\mathbf{r}}{s_X}\cdot\left(\mathbf{d}^{*\prime}-\mathbf{e}'\right)\end{bmatrix} \tag{109}$$

$$= \frac{\mathbf{1}_k}{s_X} \cdot \left\{ -\bar{X} \left[\frac{\mathbf{r}'}{\sum X} - \frac{\mathbf{1}_k'}{N} - .5 \left(\mathbf{d}^{*\prime} - \mathbf{e}' \right) \right] - .5\mathbf{r} \left(\mathbf{d}^{*\prime} - \mathbf{e}' \right) \right\},$$

with elements

$$G_{ij} = \frac{1}{s_X} \cdot \left\{ -\bar{X} \left[\frac{r_j}{\sum X} - \frac{1}{N} - .5\left(d_j^* - e \right) \right] - .5 r_i \left(d_j^* - e \right) \right\}. \quad (110)$$

The Z-scores are not homogeneous of order 0, because $\mathbf{z}$ is obtained by using s_X. As a result, the asymptotic variance is approximated by $\mathbf{V_z} \approx \mathbf{G V_{\hat{m}} G}'$ (cf. Equation 2). Some tedious algebra shows that the sample estimate of the asymptotic covariance matrix of $\mathbf{V_z}$ is

$$V_{Z_{gh}} \approx \sum_{j=1}^{k} \sum_{i=1}^{k} t \left[-\bar{X} \cdot \left(\frac{r_i}{\sum X} - \frac{1}{N} - u_i \right) - r_g u_i \right] \cdot \left[-\bar{X} \left(\frac{r_j}{\sum X} - \frac{1}{N} - u_j \right) - r_h u_j \right], \quad (111)$$

Where $u_i = .5[d_i^* - e]$. For large N,

$$V_{Z_{gh}} \approx \sum_i \frac{\hat{\mathrm{m}}_i}{s_X^2} \left[-\bar{X} \left(\frac{r_i}{\sum X} - \frac{1}{N} - u_i \right) - r_g u_i \right] \cdot \left[-\bar{X} \left(\frac{r_j}{\sum X} - \frac{1}{N} - u_j \right) - r_h u_j \right]. \quad (112)$$

Chapter 4

The Effect of Assumption Violations on Regression-Based Norms

Abstract

We performed a simulation study to investigate the bias and the precision of regression-based percentile estimates, and the coverage of corresponding confidence intervals (CIs) when the assumptions of linearity, independence between population error term and covariates, and homoscedasticity of the error variances were violated. The results showed that the strength of assumption violations, sample size, and value of the estimated percentiles influenced the bias and coverage of corresponding 95% CIs. Although precision of the estimates was not influenced by assumption violations, higher sample size and percentile ranks closer to 100 resulted in estimates that were more precise. We advise test constructors to investigate assumption violations when estimating regression-based norms, because assumption violations can cause substantial bias in both norm estimates and corresponding CIs.

4.1 Introduction

Tests are omnipresent in psychological research and in clinical, personality, health, medical, developmental, and personnel psychology practice. Researchers often present norm scores in an effort to help colleagues interpret raw test scores. Examples of norm scores are percentiles, stanines, and standardized scores (Kline, 2000, pp. 59-63). An increasingly popular method to estimate norm scores is regression-based norming. Regression-based norming (Zachary & Gorsuch, 1985) involves the use of an ordinary least squares (OLS) regression model to estimate a norm distribution. In psychological test data, OLS regression assumptions are often violated, possibly producing biased norm estimates (Van der Elst, Hoogenhout, Dixon, De Groot, & Jolles, 2011). This study focuses on tests used in psychological practice for individual measurement, and investigates the effects of violations of the OLS regression model assumptions on the bias and the precision of regression-based norm estimates as well as corresponding confidence intervals (CIs).

Traditionally, norm estimation involves determining norm distributions in separate groups defined, for example, by gender and age. Such group variables are covariates expected to influence test scores. Regression-based norming has two advantages over traditional norm estimation. The first advantage is that regression-based norming allows for the direct inclusion of both categorical (e.g., gender) and continuous (e.g., age) covariates into the model. For example, compared to women, men underreport depressive symptoms (Hunt, Auriemma, & Cashaw, 2003). As a result, gender should be used as a covariate to obtain different norms for men and women for tests assessing depression. Using the traditional norming method, continuous covariates (e.g., age) have to be divided arbitrarily into mutually exclusive and exhaustive categories (e.g., Bechger, Hemker, & Maris, 2009; Evers, Lucassen, Meijer, & Sijtsma, 2009). Because of the arbitrariness, different categorizations can change the interpretation of an individual's test performance, depending on the norm group to which the individual is assigned (Parmenter, Testa, Schretlen, Weinstock-Guttman, & Benedict, 2010). The second advantage is that regression-based norming is more efficient than the traditional method. Oosterhuis Van der Ark, and Sijtsma (2016b) found that, compared to the traditional norming method, regression-based norming requires a smaller sample to obtain equally precise norms. The explanation for greater efficiency is that regression-based norming uses the entire sample to estimate

norms. Traditional norming, on the other hand, requires a division of the total norming sample into smaller subgroups, which subsequently are used to estimate norms that are less precise than would have been realized had the whole sample been used.

This article is organized as follows. First, we describe the procedure to estimate regression-based norms. Second, we explain the assumptions of the OLS regression model and describe consequences of violating the assumptions in the context of OLS regression and norm estimation. Third, we present the results of simulation studies that suggests that assumption violations may strongly influence the bias and the precision of regression-based norms and the coverage of corresponding 95% CIs. Finally, we discuss practical implications and recommendations for future research.

4.2 Estimation of Regression-Based Norms

A five-step procedure can be used to estimate regression-based norms (Van Breukelen & Vlaeyen, 2005; Van der Elst et al., 2011): (1) Include K relevant covariates, denoted $X_1, \ldots, X_K$, into the regression model. Continuous covariates can be added directly to the model and categorical covariates are replaced by dummy variables (Hardy, 1993). (2) Compute the predicted test scores. Let Y_+ be the observed test score, and let $\hat{Y}_+$ be the predicted test score. Let B_0 be the intercept and let $B_1, \ldots, B_K$ be the regression coefficients; then the regression equation equals

$$\hat{Y}_+ = B_0 + B_1 X_1 + \cdots + B_K X_K. \quad (1)$$

(3) Compute the residuals. Residuals are defined as

$$e = Y_+ - \hat{Y}_+. \quad (2)$$

(4) Standardize the residuals. Index i enumerates the observations in the sample. Residuals are standardized by dividing them by their standard error,

$$SE_e = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{N-K-1}}, \quad (3)$$

where N is the total sample size. (5) Use the cumulative empirical distribution of the standardized residuals to estimate the norm statistics. For example, Van der Elst et al. (2011) converted the standardized residuals into percentiles based on the empirical cumulative distribution function of the standardized residuals.

4.3 Assumptions of the Linear Regression Model

The assumptions of the OLS regression model are commonly referred to as the Gauss-Markov assumptions and refer to the population regression equation (Berry, 1993, p. 12). Let β_0 be the population intercept, let $\beta_1, \ldots, \beta_K$ be the population regression coefficients corresponding to the covariates included in the model, and let ε be the population error term. This error term encompasses both the effects of excluded variables that influence Y_+ as well as an intrinsically random component due to the nature of human behavior. The population model for the test score of individual *i* is then

$$Y_{+_i} = \beta_0 + \beta_1 X_{1_i} + \cdots + \beta_K X_{K_i} + \varepsilon_i. \quad (4)$$

If the Gauss-Markov assumptions are met, OLS estimators are the best linear unbiased estimators (BLUE; Berry, 1993, pp. 18-19). This means that the estimators are both unbiased and efficient, which are desirable properties. If an estimator is unbiased, the expected value of the estimator equals the corresponding population value. Furthermore, an estimator is efficient if it provides the most precise estimates of a specific set of linear unbiased estimators.

Regression-based norming uses the OLS estimators to compute $\hat{Y}_+$ and bases its norm estimates on the corresponding distribution of residuals. As a result, a violation of the Gauss-Markov assumptions might also have consequences for regression-based norm estimates. More specifically, if the OLS estimators are biased due to an assumption violation, because $\hat{Y}_+$ is biased, the residuals are also expected to be biased. Next, we describe the Gauss-Markov assumptions in more detail.

1) Linearity and additivity. Linearity means that Y_+ is a linear function of the covariates that are included in the model. A non-additive association occurs when the association between a covariate and Y_+ varies as a function of one or more of the other covariates. If the regression parameters of the covariates are non-linear or non-additive, they can sometimes be specified correctly using a transformation of the covariates. For example, Van der Elst et al. (2011) added both Age and Age squared to the regression model to account for a possible curvilinear effect of age (also see Van der Elst, Dekker, Hurks, & Jolles, 2012; Van der Elst, Ouwehand, Van der Werf, Kuyper, Lee, & Jolles, 2012). Furthermore, a multiplicative term (i.e., an interaction term) between two or more

covariates can be added to the model to account for non-additive associations. If linearity and additivity are obtained through such transformations, the model is said to be intrinsically linear and additive (Berry, 1993, p. 60). If the relationship between Y_+ is non-linear or non-additive and this relationship is not correctly represented by the regression model, $\hat{Y}_+$ is biased and norm estimates that are based on $\hat{Y}_+$ are also expected to be biased (Berry, 1993, pp. 28-29) .

2) $\boldsymbol{E(\varepsilon_i) = 0}$. The mean of the error term in Equation 3 should be equal to zero. A non-zero mean can occur, for example, if Y_+ is measured with error equal for all observations, one of the excluded variables is correlated with at least one of the covariates in the model, or when the data distribution is truncated (Berry, 1993, pp. 41-45). This assumption has the least important consequences for the regression model estimates, because its violation only results in a change of the intercept, which absorbs the expected value of the error term (Berry & Feldman, 1985, p. 73). For estimating norms it is of no consequence whether the non-zero mean of the error term is included in the intercept or in the distribution of residuals, because the relative position of individuals in the norm sample remains the same. Hence, we did not investigate a violation of this assumption.

3) $\boldsymbol{Cov(X_i, \varepsilon_i) = 0}$**.** The error term has to be independent of the covariates in the model. Henceforth, we refer to this assumption as the independence assumption. Considering that the error term consists of random error and effects of excluded covariates, a violation of this assumption might occur if an excluded relevant covariate correlates with the included covariate(s). Reciprocal causation, if present, causes this assumption to be violated (Berry, 1993, pp. 27-29). A feedback loop between Y_+ and one or more of the covariates produces reciprocal causation. In this case, the error term and the covariates influence Y_+, which in turn influences the value of the covariate(s). As a result, the error term and the covariate(s) must correlate. Warlaumont, Richards, Gilkerson, and Oller (2014) described an example of a feedback loop, and found that adults were more likely to respond to a child's vocalizations if they were speech-related. In turn, a child was more likely to speak instead of using other vocalizations if the child's previous speech-related vocalizations had received an immediate response from an adult. If covariates in the regression model are correlated with the error term, because of a feedback loop or

because of the omission of an important covariate from the regression model, $\hat{Y}_+$ is biased and norm estimates that are based on $\hat{Y}_+$ are also expected to be biased (Berry, 1993, pp. 41-45).

4) $\boldsymbol{Cov(\varepsilon_i, \varepsilon_j) = 0}$**.** Each observation in the sample must be independent of each of the other observations. This implies that the individual error terms are also independent of each other. This assumption is often violated for longitudinal data and for data collected across spatial units (Neter, Kutner, Nachtsheim, & Wasserman, 1996, pp. 104-105). The closer they are in time, the stronger the association between data collected at different time points. Dependence due to proximity in time can occur in cross-sectional, longitudinal, and panel data. Data collected across spatial units cause data measured closer geographically to have a higher association than data measured further apart. We did not investigate this assumption, because the nature of data collection can predict a violation of this assumption beforehand and models are available that accommodate correlated errors (Neter et al., 1996, p. 125).

5) Homoscedasticity. Homoscedasticity means that the variance of the error distribution is constant for all levels of the covariates (Fox, 1997, pp. 272-274). A violation of this assumption is referred to as heteroscedasticity. Heteroscedasticity can have several causes. One cause is model misspecification, which happens when, for example, an important variable or interaction effect is omitted from the model, or when the model has an incorrect functional form, for example, linear when log-linear is appropriate. Heteroscedasticity is also likely to occur when a high value on a covariate is a necessary but not a sufficient condition for a high value of Y_+. For example, age has a positive effect on income, but age is not a sufficient condition for a high income, because education level and other covariates also influence income level. Hence, at a low age, differences between individuals with respect to income are small. However, as age increases, highly educated individuals' income increases faster than lowly educated individuals' income. As a result, as age increases income differences between individuals become larger, producing heteroscedasticy. Other causes of heteroscedasticity are heterogeneous data (i.e., data having large differences in values on the covariates or Y_+), data that contain error, have a large range, have skewed distributions, or suffer from outliers. Heteroscedasticity is

expected to lead to biased norm estimates, because the error variances conditional on *X* are variable and therefore unequal to the unconditional error variance of the regression model (Berry, 1993, pp. 62-66). As a result, using the unconditional residuals (Equation 2) to estimate norms leads to an underestimation of the true norms conditional on *X* if the conditional error variance is larger than the unconditional error variance. On the other hand, if the conditional error variance is smaller than the unconditional error variance, the norms are overestimated.

6) Normality. Normality means that the errors in the model are normally distributed. This assumption also implies that the distribution of Y_+ given the covariate values is normally distributed. The normality assumption is often justifiable in the social and behavioral sciences, because the error term is assumed to consist of random effects of excluded variables and random measurement error (Neter et al., pp. 29-30). If the normality assumption is violated, the OLS estimators are still the best linear unbiased estimators (BLUE; Berry, 1993, pp. 18-22), but statistical hypothesis testing based on a normal distribution might not be valid. However, due to the central limit theorem, if the sample is larger than 50 observations (Casson & Farmer, 2014), the regression model usually is robust against a violation of the normality assumption (Neter et al., pp. 29-30). Norm samples are typically much larger than 50 observations, which means non-normality of the residuals has little to no consequences for hypothesis tests of the regression model estimates. Furthermore, norm statistics such as percentiles are often distribution-free. This means that regression-based norms can be estimated regardless of the shape of the empirical distribution of the residuals, as long as the distribution shape is constant conditional on the covariate variables. Hence, we did not investigate a violation of the normality assumption.

We used a simulation study to determine the effects of violations of linearity, independence, and homoscedasticity of the error variances on the bias and precision of regression-based percentile estimates and corresponding 95% CIs. Using a simulation study allowed us to obtain the sampling distribution of the percentile estimates and to manipulate the characteristics of the regression model used for simulating the data.

4.4 Preliminaries

We chose to investigate percentiles, because they are commonly presented as norms (Bride, 2007; Glaesmer et al., 2012; Krishnan, Sokka, Häkkinen, Hubert, & Hannonen, 2004; Wizniter et al., 1992) or as cutoff scores in testing practice (Crawford & Henry, 2003; Crawford, Henry, Crombie, & Taylor, 2001; Lee, Loring, & Martin, 1992; Mond, Hay, Rodgers, & Owen, 2006; Murphy & Barkley, 1996). To estimate the percentiles, we used the Harrell-Davis (Harrell & Davis, 1982) quantile estimator. The H-D estimator provides an exact bootstrap estimate based on a weighted linear combination of order statistics and gives the greatest weight to traditional nonparametric quantile estimators (Harrell & Davis, 1982). A great advantage of this estimator is that it is distribution-free and more efficient in small samples than the traditional quantile estimators that use only one or two order statistics. Furthermore, the H-D estimator and the traditional estimators are asymptotically equivalent. Exact bootstrap estimates of the SEs corresponding to the percentiles can be obtained using a jackknife variance estimator for quantiles (Hutson & Ernst, 2000), and can be used to construct CIs. The procedures to obtain quantiles estimates and the corresponding SEs are available in the *R* package *Hmisc* (Harrell Jr., 2015). We also investigated traditional quantile estimators provided in *SPSS* and *R* (available from the *quantile* function in *R* (*R* Core team, 2013) and CIs based on the Marritz-Jarett SE estimator for quantiles (Wilcox, 2012, p 68.), but the results from these estimators were highly similar to the results presented here; thus, we did not consider these estimators in this study.

Oosterhuis et al. (2016b) found that when using regression-based norming, for a given sample size, percentiles further away from the median had a lower precision than percentiles closer to the median. Hence, in the current study we estimated the 75th, 90th, and 99th percentile rank to investigate whether the consequences of assumption violations depended on percentile value. We did not study the 1st, 10th, and 25th percentile rank, because the distributions we investigated were normal and therefore symmetrical, which means our results can be readily generalized to the 1st, 10th, and 25th percentile ranks.

The seriousness of the problems caused by assumption violations also depends on sample size. For example, instead of eliminating bias, increasing sample size can actually increase bias resulting from heteroscedasticity (Hayes, 2007; Long & Ervin, 2000).

Furthermore, SEs usually decrease as sample size increases, which also results in narrower CIs. If the estimates are biased, CIs are also biased and a larger sample size might result in lower coverage, because the CIs are located in the wrong place of the distribution.

The effects of assumption violations on the estimators and the predictions of the regression model also depend on the strength of the violation (Berry & Feldman, 1985). In this study, the strength of a linearity violation depended on the degree of non-linearity of the covariate effect in the population regression model. The strength of the violation of independence depended on the correlation between the covariate(s) in the model and excluded covariates that had an effect on Y_+. The stronger the correlation between covariates, the larger the influence of one covariate on the effect (i.e., partial regression slope) of another covariate in the model. Hence, omitting a covariate from the regression model can bias the effect of covariates that were included in the regression model, and bias is expected to be greater for covariates that have a higher correlation. The strength of a violation of homoscedasticity was evident from the differences between the conditional error variances. Larger differences resulted in a stronger violation of homoscedasticity.

4.5 General Method

We used four different population models to simulate raw rest scores, denoted Y_+, and covariates X_1 and X_2. Prior to discussing these population models, we first discuss relevant details of the simulation study. We simulated Q = 10,000 replicated samples, which each contained N observations. In all population models, the scale for Y_+ was fixed by choosing $\bar{Y}_+ = 0$ and $s^2_{Y_+} = 1$. The sample linear regression model estimated from the simulated data was equal to

$$Y_+ = B_0 + B_1 X_1 + e, \quad (5)$$

where $e \sim N(0, s_e^2)$. This model assumed that the population effect of X_1 on Y_+ was linear, that the population model only contained X_1 (i.e., X_2 is omitted from the model), and that the population error variance, σ_ε^2, was homoscedastic.

Regression-Based Norms

To obtain regression-based norms, we first estimated the sample regression model (Equation 5) based on the simulated scores. Second, for each individual *i* in the norm sample, we computed e_i (Equation 2), which resulted in the empirical distribution of

unstandardized residuals. Finally, using the H-D quantile estimator, we estimated the e that corresponded to the 75th, 90th, and 99th percentile ranks (PRs) in the empirical distribution of standardized residuals, $e(75)$, $e(90)$, and $e(99)$, respectively. The resulting values were expressed on the scale of e (i.e., zero mean and variance s_e^2).

To illustrate these three steps, Table 4.1 provides the scores on X_1 and Y_+ for the first 8 individuals in a simulated norm sample (N = 100). First, we estimated the sample regression model (Equation 5),

$$\hat{Y}_+ = 0.0316 + 0.3526 \cdot X_1. \quad (6)$$

Second, we implemented the scores on X_1 into Equation 6, to obtain the predicted raw test scores for the 8 individuals (Table 4.1). For example, individual 1 had $X_1 = 0.78$, which means $\hat{Y}_+ = 0.0316 + 0.3526 \cdot 0.78 = 0.3066$. Third, we obtained residuals e_i by implementing Y_+ and $\hat{Y}_+$ into Equation 2. For example, for individual 1, $e_i = Y_+ - \hat{Y}_+ = 1.61 - 0.31 = 1.30$, which shows that individual 1 scored 1.3 points higher than expected based on his X_1-score. Table 1 shows the 8 e_i values. Finally, we used the HD-estimator to estimate $e(75)$, $e(90)$, and $e(99)$ based on the sample distribution of unstandardized residuals. In this example, $e(75)$ = 0.669, $e(90)$ = 1.108, and $e(99)$ = 2.187. Thus, 75%, 90% and 99% of the individuals in the simulated norm sample had an unstandardized residual of at most 0.669, 1.108, and 2.187, respectively.

Table 4.1. Example to Compute Regression-Based Norms.

i	X_1	Y_+	$\hat{Y}_+$	e_i
1	0.78	1.61	0.31	1.30
2	0.09	0.71	0.06	0.65
3	0.77	0.10	0.30	-0.21
4	-0.42	-0.12	-0.17	-0.01
5	-0.29	-0.01	-0.07	0.06
6	-0.59	0.53	-0.18	0.70
7	-1.56	0.36	-0.52	0.88
8	.032	1.15	0.14	1.00

Sample Distribution of $Y_+|X_1$

We transformed the $e(\text{PR})$-values to the scale of the sample distribution of $Y_+|X_1$ as follows,

$$T_e(\text{PR}) = e(\text{PR}) + \hat{Y}_+\big|X_1. \quad (7)$$

We computed $\hat{Y}_+\big|X_1$ by implementing X_1-values into Equation 5:

$$\hat{Y}_+\big|X_1 = B_0 + B_1 \cdot X_1, \quad (8)$$

We used $X_1 = 0$, $X_1 = 1$, and $X_1 = 2$, because these X_1-values covered the range in which most positive scores on the standard normal covariate X_1 were located.

Based on the estimated regression model in Equation 6, for $X_1 = 0$, 1, and 2,

$$\hat{Y}_+\big|(X_1 = 0) = 0.03158 + 0.35261 \cdot 0 = 0.0316, \quad (9)$$

$$\hat{Y}_+\big|(X_1 = 1) = 0.03158 + 0.35261 \cdot 1 = 0.3842, \quad (10)$$

$$\hat{Y}_+\big|(X_1 = 2) = 0.03158 + 0.35261 \cdot 2 = 0.7368, \quad (11)$$

respectively. Hence, for $X_1 = 0$, implementing these values into Equation 7 for $e(75)$, $e(90)$, and $e(99)$ resulted in,

$$T_e(75)|(X_1 = 0) = e(75) + \hat{Y}_+\big|(X_1 = 0) \quad (12)$$
$$= 0.669 + 0.0316 = 0.7005,$$

$$T_e(90)|(X_1 = 0) = e(90) + \hat{Y}_+\big|(X_1 = 0) \quad (13)$$
$$= 1.108 + 0.0316 = 1.1396,$$

and

$$T_e(99)|(X_1 = 0) = e(99) + \hat{Y}_+\big|(X_1 = 0) \quad (14)$$
$$= 2.187 + 0.0316 = 2.2182.$$

Similarly, the values for $T_e(75)$, $T_e(90)$, and $T_e(99)$ could be computed for $X_1 = 1$ and $X_1 = 2$ by replacing $\hat{Y}_+\big|(X_1 = 0)$ with $\hat{Y}_+\big|(X_1 = 1)$ (Equation 10) and $\hat{Y}_+\big|(X_1 = 2)$ (Equation 11), respectively.

Population Distribution of $Y_+|X_1$

The population distribution of $Y_+|X_1$ was normal with mean $E(Y_+|X_1)$ and variance $\sigma^2_{Y_+|X_1}$. The values for $E(Y_+|X_1)$ and $\sigma^2_{Y_+|X_1}$ are specified in the sections describing the population models. Let θ be a population PR equal to either 75, 90, or 99 and let $\hat{\theta}$ be the PR in the population distribution of $Y_+|X_1$ that corresponded to $T_e(\text{PR})|X_1$. For $X_1 = 0$, $X_1 = 1$,

and $X_1 = 2$, we used the *pnorm* function in *R* (*R* Core team, 2016) to find $\hat{\theta}$ in a normal distribution with mean $E(Y_+|X_1)$ and variance $\sigma^2_{Y_+|X_1}$. If the sample distribution of $Y_+|X_1$ was equal to the population distribution of $Y_+|X_1$, $\hat{\theta} = \theta = 75$, 90, and 99 for $T_e(75)$, $T_e(90)$, and $T_e(99)$, respectively.

For example, let the population distribution for $X_1 = 0$ be normal with $E(Y_+|X_1) = 0$ and $\sigma^2_{Y_+|X_1} = .82$. In this distribution, the values of $T_e(75) = .7005$, $T_e(90) = 1.1396$, and $T_e(99) = 2.2182$ corresponded to the 43.5th, 62.6th, and 93.5th PR. Hence, the residuals that corresponded to the 75th, 90th, and 99th percentile rank based on the sample regression model, actually corresponded to the 43.5th, 62.6th, and 93.5th percentile rank in the distribution of residuals based on the population regression model.

Dependent Variables

Bias of norm estimates. The first dependent variable we estimated was the bias of the PR estimates, which was based on *Q* replications and computed as

$$bias = \frac{1}{Q}\sum_{q=1}^{Q}(\hat{\theta}_q - \theta). \quad (15)$$

Precision. The second dependent variable was the precision of the norms, which was quantified by the standard deviation (SD) of the PR estimates. The SD was computed as

$$SD_{\hat{\theta}} = \sqrt{\frac{1}{Q-1}\sum_{q=1}^{Q}(\hat{\theta}_q - \bar{\theta})^2}, \quad (16)$$

where $\bar{\theta}$ was the average PR estimate over *Q* replications, $\frac{1}{Q}\sum \hat{\theta}_q$.

Coverage probability of 95% CIs. The third dependent variable was the coverage of the 95% CI. First, in each replication we constructed a Wald-based CI for $\hat{\theta}_q$. Let α be the two-tailed *p*-value and $z_{\alpha/2}$ the corresponding *Z*-score, then the limits of the 100(1-α)% Wald-based CI are

$$\hat{\theta}_q \pm z_{\alpha/2} \cdot SE_{\hat{\theta}}. \quad (17)$$

The 95% coverage was defined as the proportion of replications in which the CI contained θ.

Independent Variables

The three independent variables that were included in the simulation study were the following:

(1) *Sample size (N).* Sample sizes were 100, 500, 1,000, 2,500, or 5,000, based on a literature review by Oosterhuis et al. (2016b).

(2) *Percentile values.* Estimated percentiles were the 75th, 90th, and 99th.

(3) *Violation strength*. The violation of the assumptions was either weak, medium, or strong. The labels "weak", "medium", and "strong" were arbitrary and only served to indicate the relative strength of the violation compared to the other conditions. Details for each type of assumption violation are provided in the sections describing the corresponding population models.

Analyses

For each population model, we performed three analyses of variance (ANOVAs) for each of the dependent variables, which resulted in twelve ANOVAs. We used the ANOVAs to investigate the main effects, and the two-way and three-way interaction effects of the independent variables on the bias of the percentile estimates, the SD of the PR estimates, and the coverage of the 95% Wald-based CIs. Eta-squared (η^2) was used to interpret the effect sizes: $\eta^2 > .14$ (large effect), $\eta^2 > .06$ (medium), and $\eta^2 > .01$ (small) (Cohen, 1992). Let SS_{effect} be the sum of squares corresponding to a particular main or interaction effect, and let SS_{total} be the total sum of squares, then η^2 for the effect equals

$$\eta^2 = \frac{SS_{effect}}{SS_{total}}. \quad (18)$$

We only considered statistically significant effects that were at least small ($\eta^2 > .01$). A significant main effect of an independent variable was only investigated if the independent variable was not included in a statistically significant interaction term. Otherwise, we only investigated the significant interaction effect. Each design cell contained one observation, which was the dependent variable based on Q simulated samples.

4.6 No Assumption Violations

Method

Population regression model. The population model without assumption violations contained one standard normal continuous covariate, denoted X_1. Each of the N observations randomly received scores for X_1, and for each observation a Y_+ score could be computed so that

$$Y_+ = \beta_0 + \beta_1 X_1 + \varepsilon, \quad (19)$$

where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. The degree of error in the model was fixed using the squared multiple correlation between the covariates and the raw test score, denoted R^2. Let $\sigma_{\hat{Y}_+}^2$ and σ_ε^2 be the variance of $\hat{Y}_+$ and ε, respectively. Then

$$R^2 = \frac{\sigma_{\hat{Y}_+}^2}{\sigma_{Y_+}^2} = 1 - \frac{\sigma_\varepsilon^2}{\sigma_{Y_+}^2}. \quad (20)$$

Hence, R^2 is the proportion explained variance. We chose R^2 =.18, which indicates a medium effect of X_1 on Y_+ (R^2 > .02 represents a small effect, R^2 > .13 medium, and R^2 > .26 large; Cohen, 1992). As a result,

$$\sigma_\varepsilon^2 = (1 - R^2)\sigma_{Y_+}^2 = 1 - R^2 = .82. \quad (21)$$

The Appendix shows that, to guarantee that $R^2 = .18$, $\bar{Y}_+ = 0$ and $s_{Y_+}^2 = 1$, the regression parameters in Equation 19 had to be chosen to be equal to $\beta_0 = 0$, and $\beta_1 = \sqrt{.18}$.

Population distribution of $Y_+|X_1$. Based on the population model without assumption violations (Equation 19),

$$E(Y_+|X_1) = \beta_0 + \beta_1 X_1, \quad (22)$$

and $\sigma_{Y_+|X_1}^2$ was equal to the unconditional variance of the residuals (Equation 21) $\sigma_{Y_+|X_1}^2 = \sigma_\varepsilon^2 = .82$.

Results and Discussion

Table 4.2 shows the effect sizes (η^2) corresponding to the ANOVAs that were performed.

Intercept and slope. Percentile value and *N* did not significantly influence B_0 and B_1 ($F(1,3) = .876$, n.s., $\eta^2 = .226$, and $F(1,3) = .097$, n.s., $\eta^2 = .031$, respectively), and both estimators were unbiased (bias equal to 0.0004 and -0.0001 for B_0 and B_1, respectively). This was expected, because the population regression model and sample regression model were equal.

Table 4.2. For No Assumption Violations, Effect Sizes (η^2) Based on ANOVAs.

	Bias			RMSE			Coverage		
	$X_1 = 0$	$X_1 = 1$	$X_1 = 2$	$X_1 = 0$	$X_1 = 1$	$X_1 = 2$	$X_1 = 0$	$X_1 = 1$	$X_1 = 2$
Main effects									
N	.205†	.273†	.311*	.317*	.293*	.272*	.151	.247†	.091†
PV	.029	.004	.004	.254*	.282*	.307**	.219†	.003	.653**
Interactions									
*N**PV	.028	.006	.002	.064	.073	.080	.080	.082	.023
Total	.262	.283	.317	.635	.648	.659	.450	.332	.767

Note. ANOVAs = analyses of variance. PV = percentile value. †p <.10. *p <.05. **p < .01.

Bias. Figure 4.1 shows the effect of *N* on the bias of the percentile estimates; the estimates were unbiased, except for *N* = 100. In this case, we found that bias increased as X_1 increased, although the amount of bias was small regardless. This was expected, because the sample regression model was unbiased.

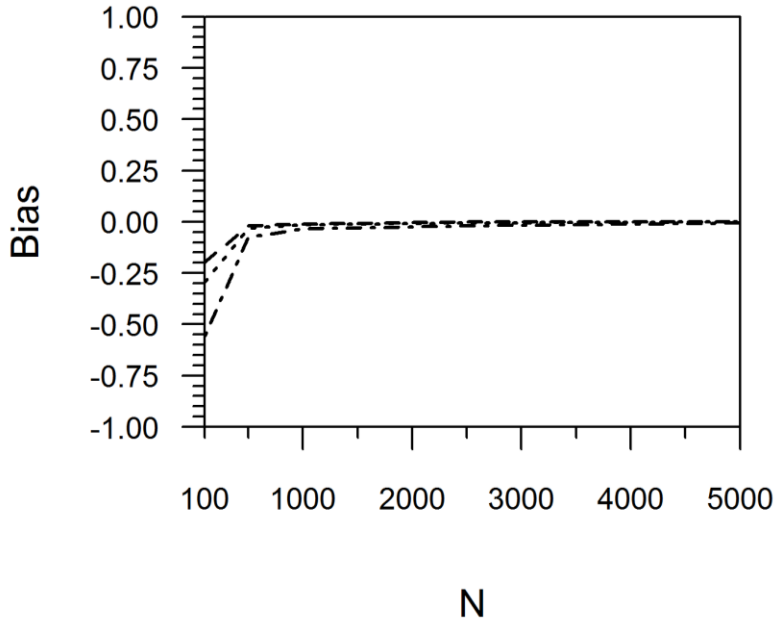

Figure 4.1. Bias of percentile estimates for $X_1 = 0$ (dashed), $X_1 = 1$ (dotted), and $X_1 = 2$ (dot-dash) when no assumption was violated.

Precision. The percentile estimates showed substantial variation, and this variation decreased as sample size increased (Figure 4.2). As X_1 increased, the standard deviation of the estimates increased, and the effect of sample size on the precision of the percentile estimates was also stronger for higher X_1-values. This suggests that PR values that were estimated for more extreme values of X_1, which were less likely to occur in the sample, required a larger sample size to obtain a certain precision level than PR values estimated for more common values of X_1. This effect was strongest for the 75th percentile. We also found that the precision of the PR estimates was greater for higher PR values (i.e., closer to 100), and that the effect of *N* on the SD of the estimates was stronger for lower percentile values (i.e., closer to 50). The explanation for these results is that, compared to the center of a normal distribution, a unit change in PR requires a larger difference in the corresponding value in the tails of a normal distribution. For example, in a standard normal distribution, the 50th and 51st percentile rank correspond to a value of 0 and 0.025, respectively, whereas the 98th and 99th percentile rank correspond to values of 2.054 and 2.326, respectively.

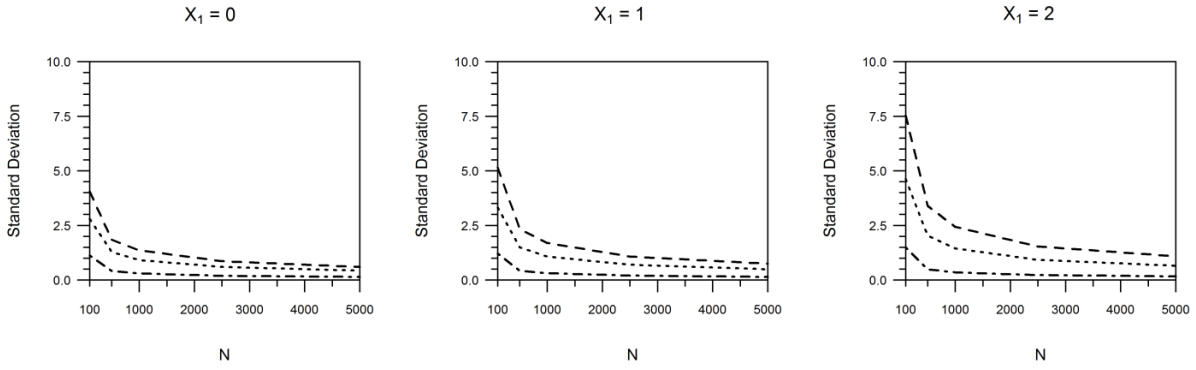

Figure 4.2. Standard deviation of the 75th (dashed), 90th (dotted), and 99th (dot-dash) percentile estimates when no assumption was violated.

Coverage. Figure 4.3 shows that the coverage of the 95% CIs for the PR estimates decreased as X_1 increased. For all X_1-values, coverage was lower for *N* = 100 compared to the other sample sizes. Furthermore, for X_1 = 0, coverage was close to .95, except for the 99th percentile when *N* = 100, in which case coverage was lower than .95. For X_1 = 1 and X_1 = 2, coverage was closer to .95 for higher PR values. Hence, coverage increased and was closer to .95 for percentile estimates that were based on a larger number of observations.

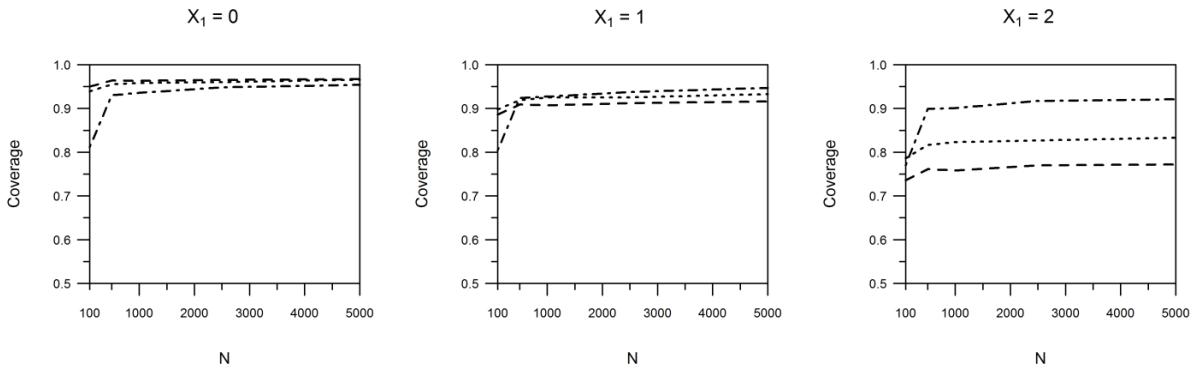

Figure 4.3. Coverage of 95% CIs for the 75th (dashed), 90th (dotted), and 99th (dot-dash) percentile estimates when no assumption was violated.

4.7 Linearity Violation

Method

Population regression model. The population model with a violation of linearity contained one standard normal continuous covariate, denoted X_1, and a quadratic term, denoted X_1^2. Each of the N observations randomly received scores for X_1, and for each observation a Y_+ score was computed so that

$$Y_+ = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon. \quad (23)$$

The Appendix shows that the values for β_0, β_1, and β_2 depended on f, which was an integer associated with the strength of the linearity assumption violation. Let β_2 depend on the value of β_1 so that $\beta_2 = f \cdot \beta_1$. We chose $R^2 = .18$, which represents a medium effect of the covariate(s) on Y_+ (Cohen, 1992). As a result, $\beta_0 = -\beta_2$ and $\beta_1 = \sqrt{.18/(1 + 2f^2)}$ (see Appendix).

Population distribution of $Y_+|X_1$. For a violation of linearity (Equation 23),

$$E(Y_+|X_1) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2, \quad (24)$$

and $\sigma^2_{Y_+|X_1}$ was equal to the unconditional variance of the residuals (Equation 21), $\sigma^2_{Y_+|X_1} = \sigma^2_\varepsilon = .82$.

Violation strength. For violations of the linearity assumption, the value of *f* was equal to .10, .25, or .40 for a weak, medium, and strong violation, respectively. Figure 4.4 shows the resulting curvilinear regression functions of Y_+ on X_1.

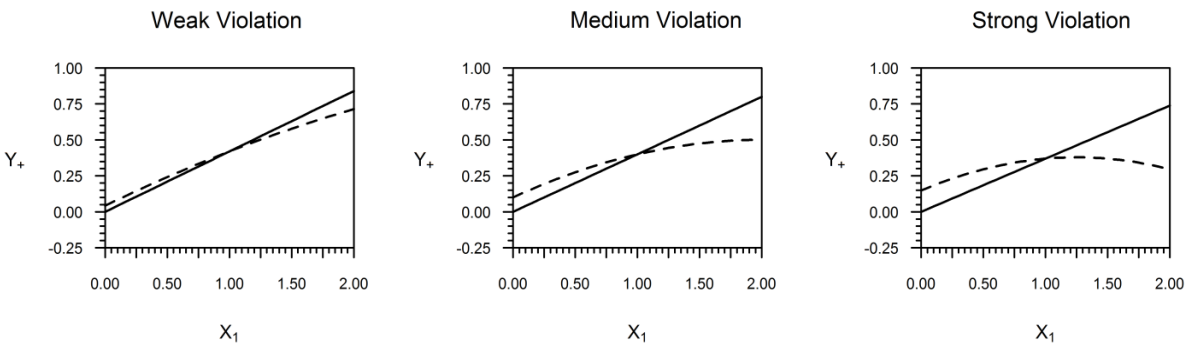

Figure 4.4. The regression of X_1 on Y_+ based on the population (dashed), and sample (solid) model for a violation of linearity.

Results and Discussion

Table 4.3 shows the effect sizes (η^2) corresponding to the ANOVAs.

Intercept and Slope. Table 4.3 shows that violation strength had a significant effect on the bias of B_0, and a marginally significant ($p < .10$) effect on the bias of B_1. B_0 was negatively biased, and bias increased as the violation of linearity became stronger (bias equal to -0.04, -0.10, and -0.15 for a weak, medium, and strong violation, respectively). Although B_1 was unbiased (bias equal to -.0001, -.0001, and -0.0003 for a weak, medium, and strong violation), it should be noted that the population regression equation contained a quadratic effect of X_1 that was omitted from the sample regression model. Hence, the linear part of the effect of X_1 was estimated without bias, but the quadratic part of the effect of X_1 was missing. As a result, not only the intercept but also the curvature of the regression line was estimated with bias (Figure 4.4). **Bias.** Figure 4.5 shows that the percentile estimates were unbiased for X_1 = 1, regardless of violation strength and percentile value. This was expected, because the sample regression line and population regression line coincide for this X_1-value (Figure 4.4). For X_1 = 0, the sample regression line was lower than the population regression line (Figure 4.1) and as a result the percentile estimates had substantial negative bias. For X_1 = 2, the sample regression line was higher than the population regression line (Figure 4.4) resulting in substantial positive bias in the percentile estimates. For both X_1 = 0 and X_1 = 2, bias increased as the violation of linearity became stronger and bias decreased as percentile value increased. The 99th percentile

estimates were unbiased, regardless of X_1- value. Bias of the percentile estimates was smaller for X_1 = 0 than for X_1 = 2, because the sample regression line was further away from the population regression line for the latter (Figure 4.4). Hence, a larger discrepancy between the sample and population regression line resulted in greater bias of the percentile estimates.

Table 4.3. Effect Sizes (η^2) Based on ANOVAs for a Violation of Linearity.

	Bias			RMSE			Coverage		
	X_1=0	X_1=1	X_1=2	X_1=0	X_1=1	X_1=2	X_1=0	X_1=1	X_1=2
Main effects									
VS	.240**	.415**	.201**	.002	.000	.018	.235**	.268**	.273**
N	.001	.060**	.001	.334**	.287**	.221**	.323**	.002	.268**
PV	.624**	.187**	.670**	.233**	.294**	.352**	.147**	.032	.081**
Interactions									
VS*N	.000	.000	.000	.000	.000	.005	.024*	.145**	.052*
VS*PV	.118**	.086**	.115**	.000	.000	.002	.023†	.008	.021
*N**PV	.000	.002	.000	.060*	.077**	.095**	.040*	.053†	.010
VS*N**PV	.000	.000	.000	.000	.000	.000	.004	.000	.003
Total	.983	.75	.987	.629	.658	.693	.796	.508	.708

Note. ANOVAs = analyses of variance. VS = violation strength. PV = percentile value. †p < .10. *p <.05. **p < .01.

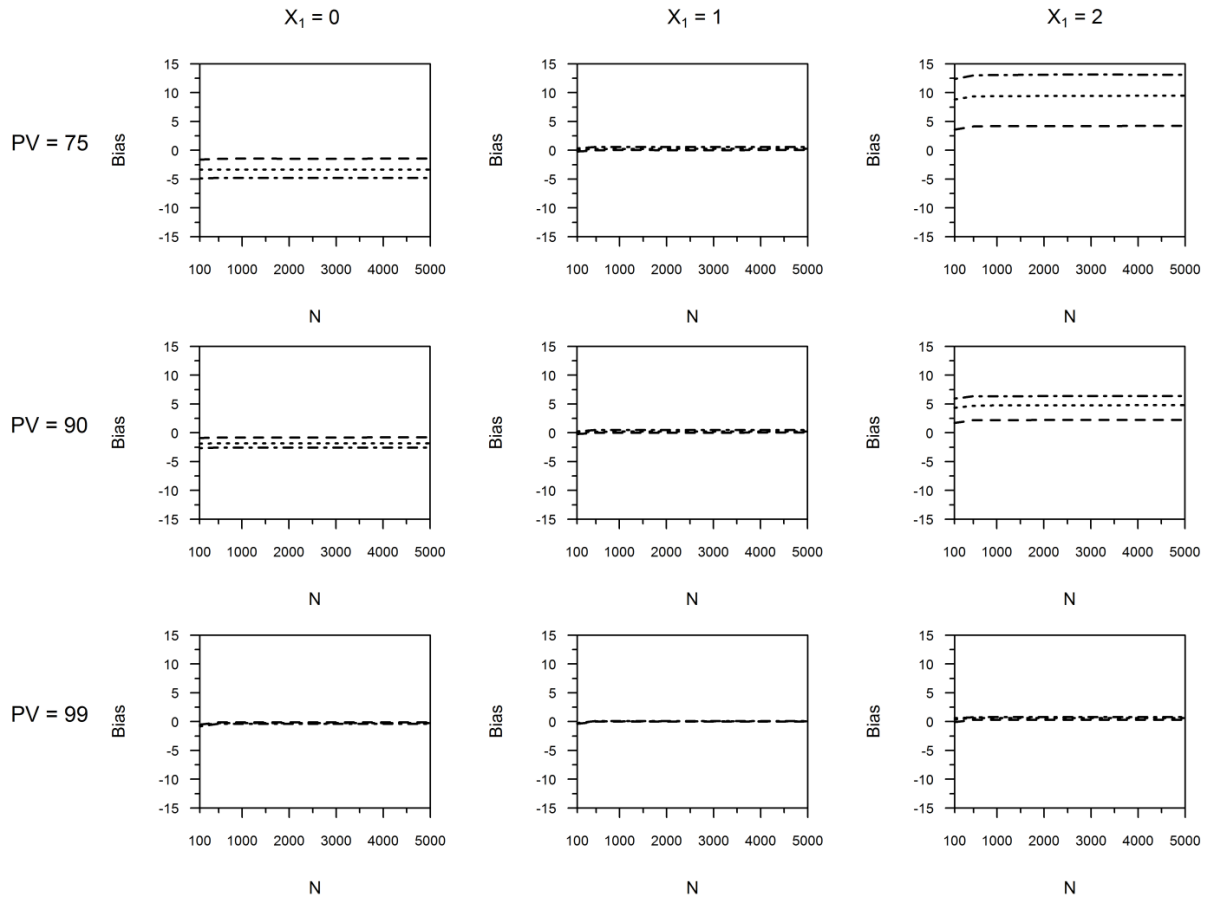

Figure 4.5. Bias of the estimates conditional on percentile value (rows) and X_1-value (columns) for weak (dashed), medium (dotted), and strong (dot-dash) violations of linearity.

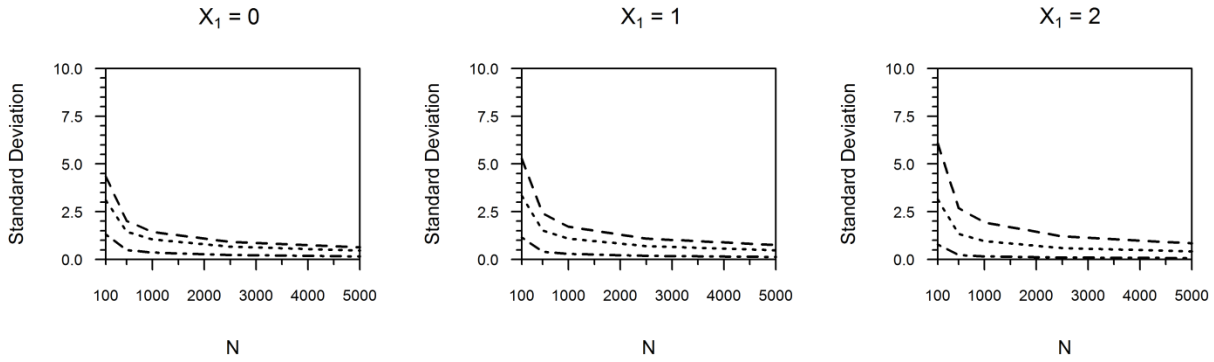

Figure 4.6. Standard deviation of the 75th (dashed), 90th (dotted), and 99th (dot-dash) percentile estimates conditional on X_1-value (columns) for a violation of linearity.

Precision. Figure 4.6 shows that the results were highly similar to the results of the model without assumption violations, which suggests that violations of linearity did not influence the precision of the percentile estimates.

Coverage. Figure 4.7 shows that, for X_1 = 1, the coverage of the 95% CIs for the percentile estimates was close to .95, regardless of violation strength and percentile value. This was expected, because the percentile estimates were unbiased for this X_1-value. For X_1 = 0 and X_1 = 2, coverage was substantially lower than .95, especially for larger *N*. Stronger violations of linearity and lower percentile values resulted in lower CI coverage. Hence, if the percentile estimates had greater bias, the corresponding CIs had lower coverage. Furthermore, the more the sample regression line differed from the population regression line (Figure 4.4), the greater the bias, and the lower the coverage of the corresponding 95% CIs.

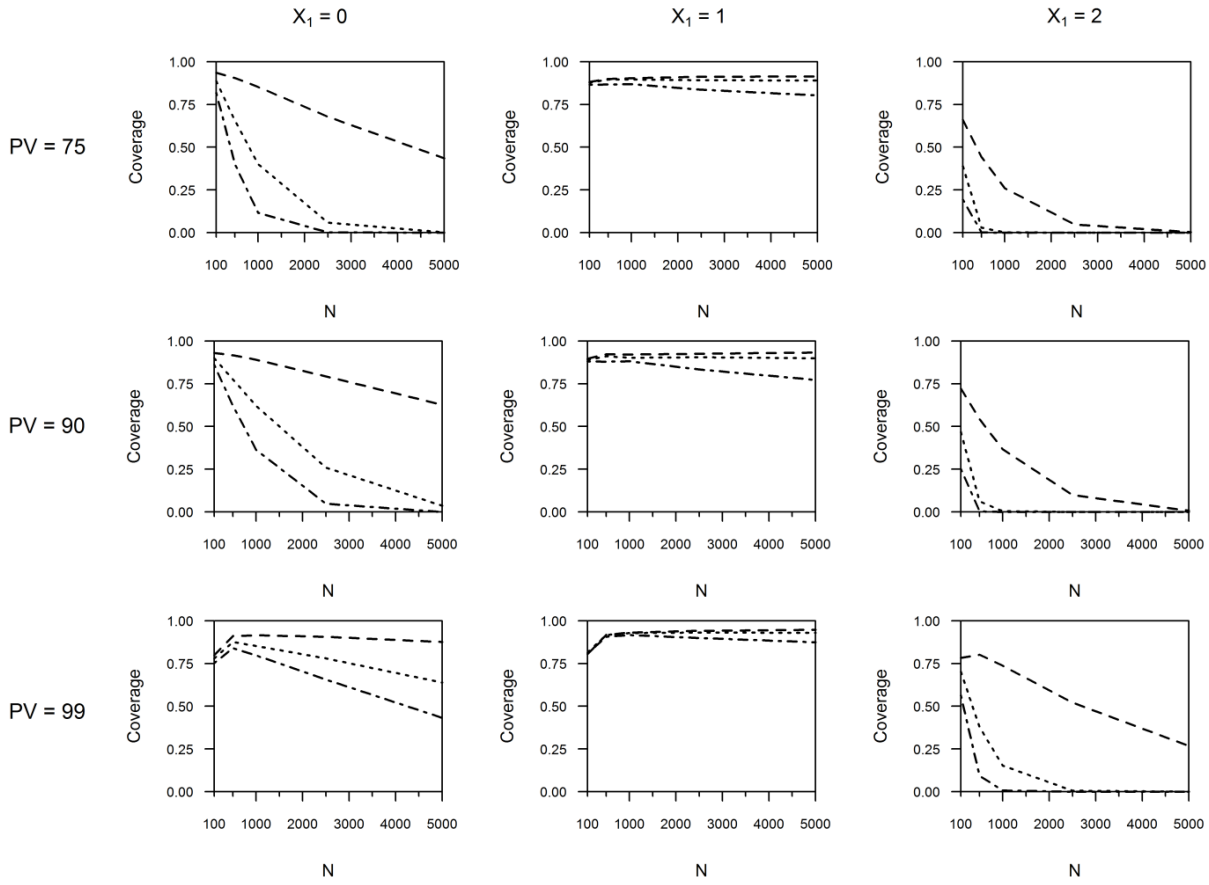

Figure 4.7. Coverage of 95% CIs of the estimates conditional on percentile value (rows) and X_1-value (columns) for weak (dashed), medium (dotted), and strong (dot-dash) violations of linearity.

4.8 Independence Violation

Method

Population regression model. The population model with a violation of independence contained two continuous covariates, denoted X_1 and X_2. The covariates followed a multivariate standard normal distribution with correlation ρ_{12}, and had equally large effects on Y_+ so that $\beta_2 = \beta_1$. Each of the N observations randomly received scores for X_1 and X_2, and for each observation a Y_+ score was computed so that

$$Y_+ = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \quad (25)$$

We chose R^2 =.18, which represents a medium effect of the covariate(s) on Y_+ (Cohen, 1992). As a result, $\beta_0 = 0$ and the values for β_1 and β_2, depended on ρ_{12}; that is, $\beta_1 = \beta_2 = \sqrt{.18/(2 + 2\rho_{12})}$ (see Appendix).

Population distribution of $Y_+|X_1$. For a violation of independence (Equation 25),

$$E(Y_+|X_1) = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \quad (26)$$

and $\sigma^2_{Y_+|X_1}$ was equal to the unconditional variance of the residuals (Equation 21), $\sigma^2_{Y_+|X_1} = \sigma^2_\varepsilon = .82$.

Violation strength. For violations of the independence assumption, ρ_{12}= .10, .30, or .50, represented a weak, medium, or strong violation, respectively. These ρ_{12}-values correspond to small, moderate, and large correlations, respectively (Cohen, 1992). Figure 4.8 shows the resulting regression functions of Y_+ on X_1, for $X_2 = 0$. We fixed X_2, because we were primarily interested in the effect of omitting X_2 from the regression model on the relationship between X_1 and Y_+.

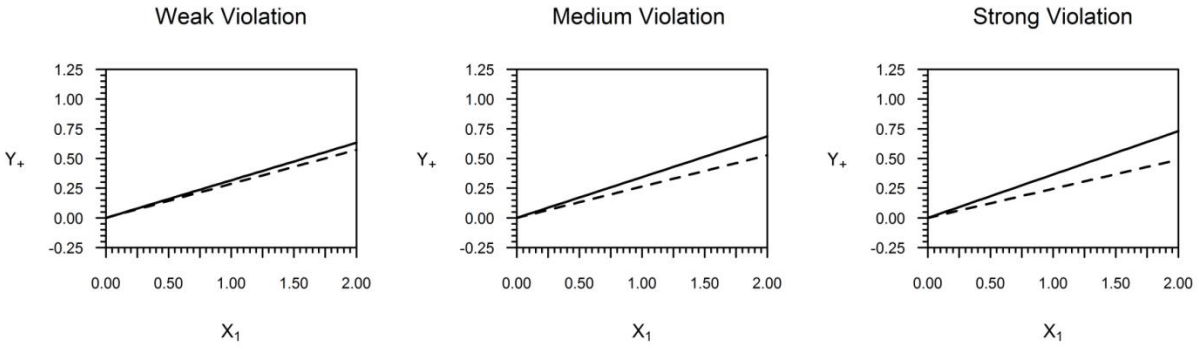

Figure 4.8. The regression of X_1 on Y_+ based on the population (dashed), and sample (solid) model for a violation of independence.

Table 4.4. For a Violation of Independence, Effect Sizes (η^2) Based on ANOVAs.

	Bias			RMSE			Coverage		
	X_1=0	X_1=1	X_1=2	X_1=0	X_1=1	X_1=2	X_1=0	X_1=1	X_1=2
Main effects									
VS	.157**	.144**	.188**	.002	.002	.007	.146**	.100**	.166**
N	.019	.002	.002†	.295**	.260**	.233**	.531**	.673**	.494**
PV	.433**	.710**	.667**	.283**	.321**	.345**	.075**	.036**	.044*
Interactions									
VS*N	.000	.000	.000	.000	.001	.002	.122**	.000	.018
VS*PV	.021	.093**	.115**	.000	.000	.001	.006	.011	.001
*N*PV	.001	.000	.000	.072**	.085*	.092**	.009†	.001	.012
VS*N*PV	.000	.000	.000	.000	.000	.000	.005	.003	.003
Total	.631	.949	.972	.65	.669	.68	.894	.824	.738

Note. ANOVAs = analyses of variance. VS = violation strength. PV = percentile value. †p < .10. *p <.05. **p < .01.

Results and Discussion

Table 4.4 shows the effect sizes (η^2) corresponding to the ANOVAs that were performed.

Intercept and Slope. Table 4.4 shows that B_0 was not affected by violation strength or *N* and that it was unbiased (bias = -0.00006). On the other hand, violation strength had a significant effect on the bias of B_1; the bias was positive and increased as the violation of independence became stronger (bias equal to 0.03, 0.08, and 0.12 for a weak, medium, and strong violation, respectively). Given that the intercept of the sample regression line was unbiased, the average regression line in the sample coincided with the population regression line for X_1 = 0. However, as X_1 increased, the difference between the average sample regression line and the population regression line increased, because the average sample regression slope was steeper than the population regression slope (Figure 4.8).

Bias. Figure 4.9 shows that the percentile estimates were unbiased for X_1 = 0, regardless of violation strength and percentile value. This was expected, because the sample regression line and population regression line coincide for this X_1-value (Figure 4.8). For X_1 = 1 and X_1 = 2, the sample regression line was steeper than the population regression line (Figure 4.8) and as a result the percentile estimates had substantial positive bias. For both X_1 = 1 and X_1 = 2, bias increased as the violation of independence became stronger and bias decreased as percentile value increased. The 99th percentile estimates were unbiased, regardless of X_1- value and violation strength. Bias of the percentile

estimates was smaller for X_1 = 1 than for X_1 = 2, because the sample regression line was further away from the population regression line for the latter (Figure 4.8). Hence, a larger discrepancy between the sample and population regression line resulted in a greater amount of bias.

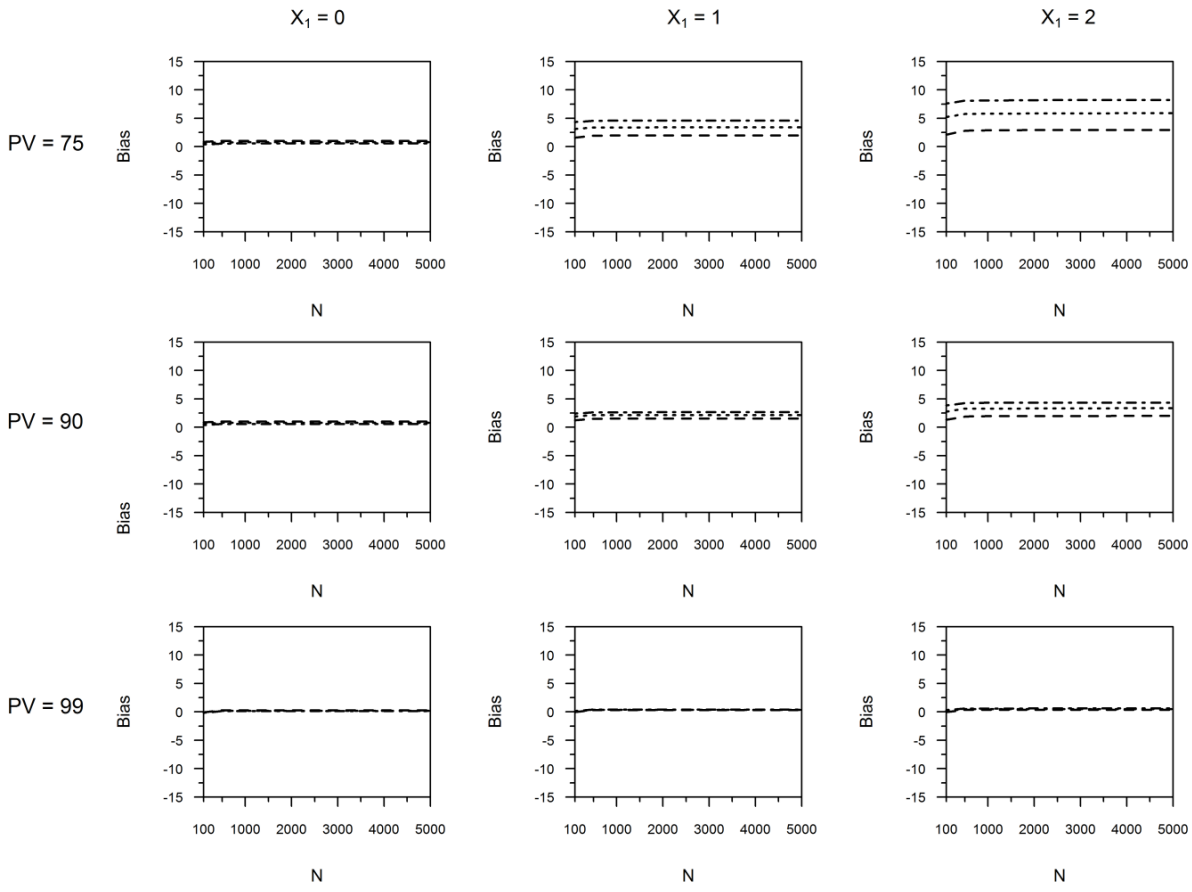

Figure 4.9. Bias of the estimates conditional on percentile value (rows) and X_1-value (columns) for weak (dashed), medium (dotted), and strong (dot-dash) violations of independence.

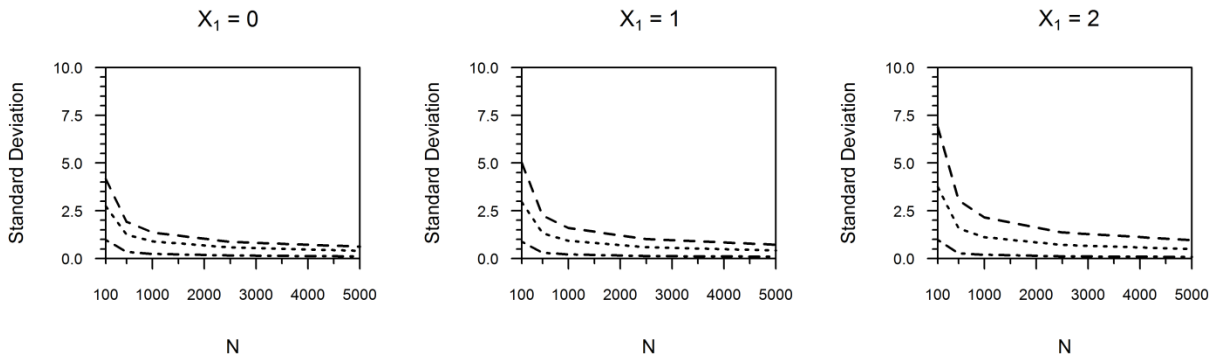

Figure 4.10. Standard deviation of the 75th (dashed), 90th (dotted), and 99th (dot-dash) percentile estimates conditional on X_1-value (columns) for violations of independence.

Precision. Figure 4.10 shows that the results were highly similar to the results of the model without assumption violations, which suggests that violations of independence did not influence the precision of the percentile estimates.

Coverage. Figure 4.11 shows that the coverage of the 95% CIs for the percentile estimates was substantially lower than .95 for almost all combinations of violation strength, percentile value, and X_1-value. For all X_1-values, larger N resulted in lower coverage of the CIs. For $X_1 = 0$, coverage decreased as violation strength decreased, whereas for $X_1 = 1$ and $X_1 = 2$, coverage decreased as violation strength increased. This was not unexpected, because, on average, for $X_1 = 0$ the average sample regression line coincided with the population regression line (Figure 4.8). Hence, if the percentile estimates had greater bias, the corresponding CIs had lower coverage. Furthermore, the more the average sample regression line differed from the population regression line (Figure 4.8), the greater the bias, and the lower the coverage of the corresponding 95% CIs.

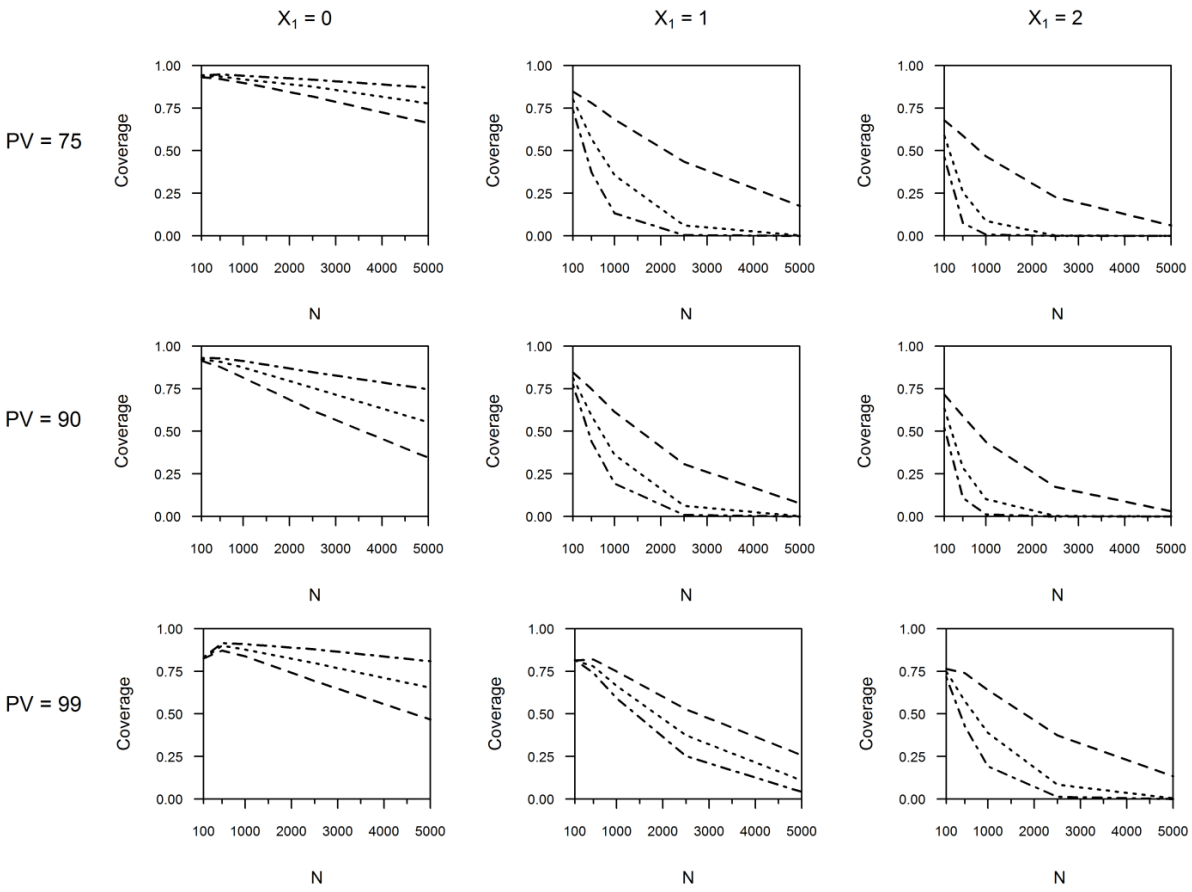

Figure 4.11. Coverage of 95% CIs of the estimates conditional on percentile value (rows) and X_1-value (columns) for weak (dashed), medium (dotted), and strong (dot-dash) violations of independence.

4.9 Homoscedasticity violation.

Method

Population regression model. The population model with a violation of homoscedasticity contained one standard normal continuous covariate, denoted X_1. Each of the *N* observations randomly received scores for X_1, and for each observation a Y_+ score was computed so that

$$Y_+ = \beta_0 + \beta_1 X_1 + (\alpha_1 + \alpha_2 X_1) \cdot \varepsilon, \quad (25)$$

where $\varepsilon | X_1 = (\alpha_1 + \alpha_2 X_1) \cdot \varepsilon$, and $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. The values for α_1 and α_2 depended on the strength of the violation of the assumption. We chose $R^2 = .18$, which indicates a medium effect of the covariate(s) on Y_+ (Cohen, 1992). As a result, $\beta_0 = 0$, and $\beta_1 = \sqrt{.18}$ (see Appendix).

Population distribution of $Y_+|X_1$. Based on the population model with a violation of homoscedasticity (Equation 25),

$$E(Y_+|X_1) = \beta_0 + \beta_1 X_1. \quad (26)$$

Furthermore, $\sigma^2_{Y_+|X_1}$ was equal to

$$\sigma^2_{Y_+|X_1} = \sigma^2_{\varepsilon|X_1} = (\alpha_1 + \alpha_2 X_1)^2 \cdot \sigma_\varepsilon^2 \quad (27)$$

(see Appendix).

Violation strength. Table 4.5 shows the values of α_1 and α_2 (Equation 27) and the resulting error variances for different levels of violation strength and X_1. For $X_1 = 0$, the unconditional variance of the residuals was equal to the conditional variance of the residuals, $\sigma_\varepsilon^2 = \sigma^2_{\varepsilon|X_1}$, whereas for $X_1 > 0$, $\sigma_\varepsilon^2 < \sigma^2_{\varepsilon|X_1}$. The difference between σ_ε^2 and $\sigma^2_{\varepsilon|X_1}$ increased as X_1 increased, and this difference was larger for stronger violations of homoscedasticity.

Table 4.5. $\sigma^2_{\varepsilon|X_1}$ and σ_ε^2 for Violations of Homoscedasticity of Differing Strengths.

			$\sigma_\varepsilon^2 \mid X_1$			
VS	α_1	α_2	$X_1 = 0$	$X_1 = 1$	$X_1 = 2$	σ_ε^2
Small	1	.10	0.812	0.982	1.169	0.812
Medium	1	.20	0.788	1.135	1.545	0.788
Large	1	.30	0.752	1.271	1.926	0.752

Note. VS = Violation Strength

Table 4.6. For a Violation of Homoscedasticity, Effect Sizes (η^2) Based on ANOVAs

	Bias			RMSE			Coverage		
	X_1=0	X_1=1	X_1=2	X_1=0	X_1=1	X_1 =2	X_1 =0	X_1=1	X_1=2
Main effects									
VS	.035**	.328**	.443**	.000	.003	.002	.170**	.152**	.136**
N	.003†	.001	.001	.297**	.357**	.392**	.127**	.489**	.329**
PV	.630**	.348**	.212**	.273**	.191**	.144**	.075*	.019	.001
Interactions									
VS*N	.000	.000	.000	.000	.001	.000	.107**	.018	.074*
VS*PV	.298**	.071**	.022	.001	.000	.005	.021	.016	.005
*N*PV	.002	.000	.000	.069*	.050*	.038†	.013	.003	.000
VS*N*PV	.000	.000	.000	.000	.000	.001	.004	.005	.002
Total	.968	.748	.678	.640	.602	.582	.517	.702	.547

Note. ANOVAs = analyses of variance. VS = violation strength. PV = percentile value. *p<.05. **p < .01.

Results and Discussion

Table 4.6 shows the effect sizes (η^2) corresponding to the ANOVAs that were performed.

Intercept and Slope. Table 4.6 shows that both B_0 and B_1 were independent of violation strength and *N* and were unbiased (bias = -0.00008 and .00018 for B_0 and B_1, respectively). This means that the average sample regression line coincided with the population regression line. Although the regression line was unbiased, Table 6 shows that $\sigma^2_{\varepsilon|X_1}$ was similar to σ^2_ε only for X_1 = 0, whereas for higher X_1-values, $\sigma^2_{\varepsilon|X_1} > \sigma^2_\varepsilon$.

Bias. Figure 4.12 shows that the percentile estimates had little to no bias for X_1 = 0, regardless of violation strength and percentile value. This is expected, because $\sigma^2_{\varepsilon|X_1} = \sigma^2_\varepsilon$ for this X_1-value (Table 4.5). For X_1 = 1 and X_1 = 2 , $\sigma^2_{\varepsilon|X_1} > \sigma^2_\varepsilon$ (Table 4.5), and as a result the percentile estimates had substantial negative bias. For both X_1 = 1 and X_1 = 2, bias increased as the violation of homoscedasticity became stronger and bias decreased as the percentile value increased. Bias of the percentile estimates was smaller for X_1 = 1 than for X_1 = 2, because the difference between $\sigma^2_{\varepsilon|X_1}$ and σ^2_ε was larger for the latter (Table 4.5). Hence, a larger discrepancy between $\sigma^2_{\varepsilon|X_1}$ and σ^2_ε resulted in greater bias.

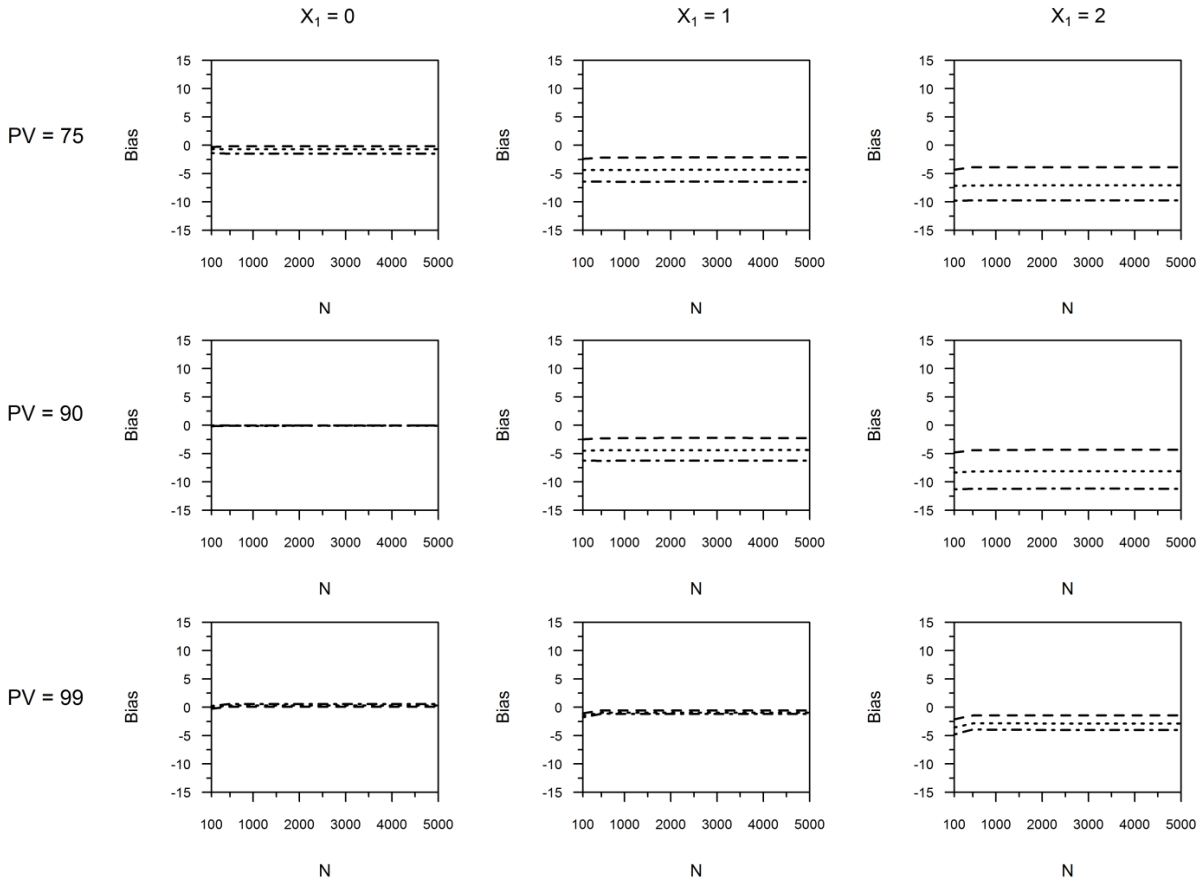

Figure 4.12. Bias of the estimates conditional on percentile value (rows) and X_1-value (columns) for weak (dashed), medium (dotted), and strong (dot-dash) violations of homoscedasticity.

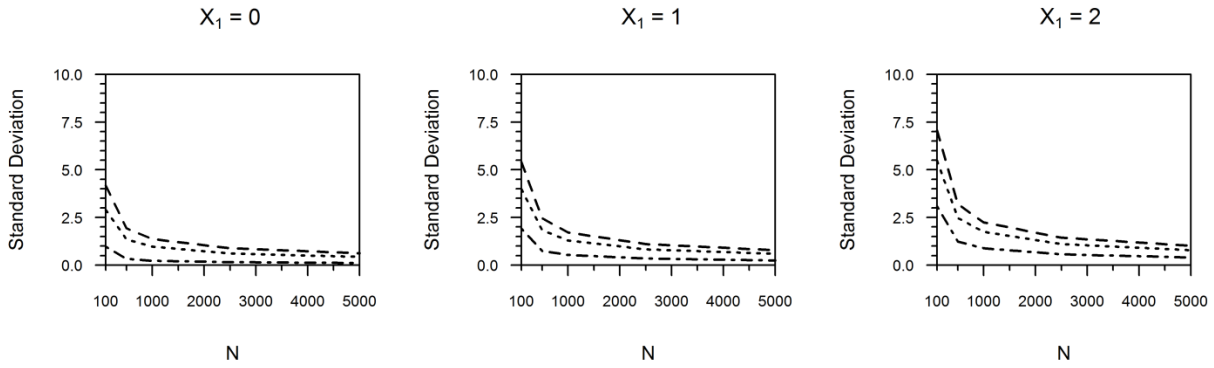

Figure 4.13. Standard deviation of the 75th (dashed), 90th (dotted), and 99th (dot-dash) percentile estimates conditional on X_1-value (columns) for violations of homoscedasticity.

Precision. Figure 4.13 shows that the results were highly similar to the results of the model without assumption violations, which suggests that violations of homoscedasticity did not influence the precision of the percentile estimates.

Coverage. Figure 4.14 shows that the coverage of the 95% CIs for the 90th percentile estimates was close to .95 for X_1 = 0, regardless of violation strength. This was expected, because the estimated percentiles were unbiased for this X_1-value. However, for X_1 = 0, the coverage of the 75th and 99th percentile estimates was lower than .95 if the violation of homoscedasticity was at least medium in strength, and coverage decreased as violation strength increased. For X_1 = 1 and X_1 = 2, coverage was substantially lower than .95, especially for larger *N*. Stronger violations of homoscedasticity resulted in lower CI coverage. Hence, if the percentile estimates had greater bias, the corresponding CIs had lower coverage. Furthermore, coverage was substantially lower for X_1 = 2 than for X_1 = 1. Furthermore, the larger the difference between $\sigma^2_{\varepsilon|X_1}$ and σ^2_{ε}, the greater the bias of the percentile estimates, and the lower the coverage of the corresponding 95% CIs.

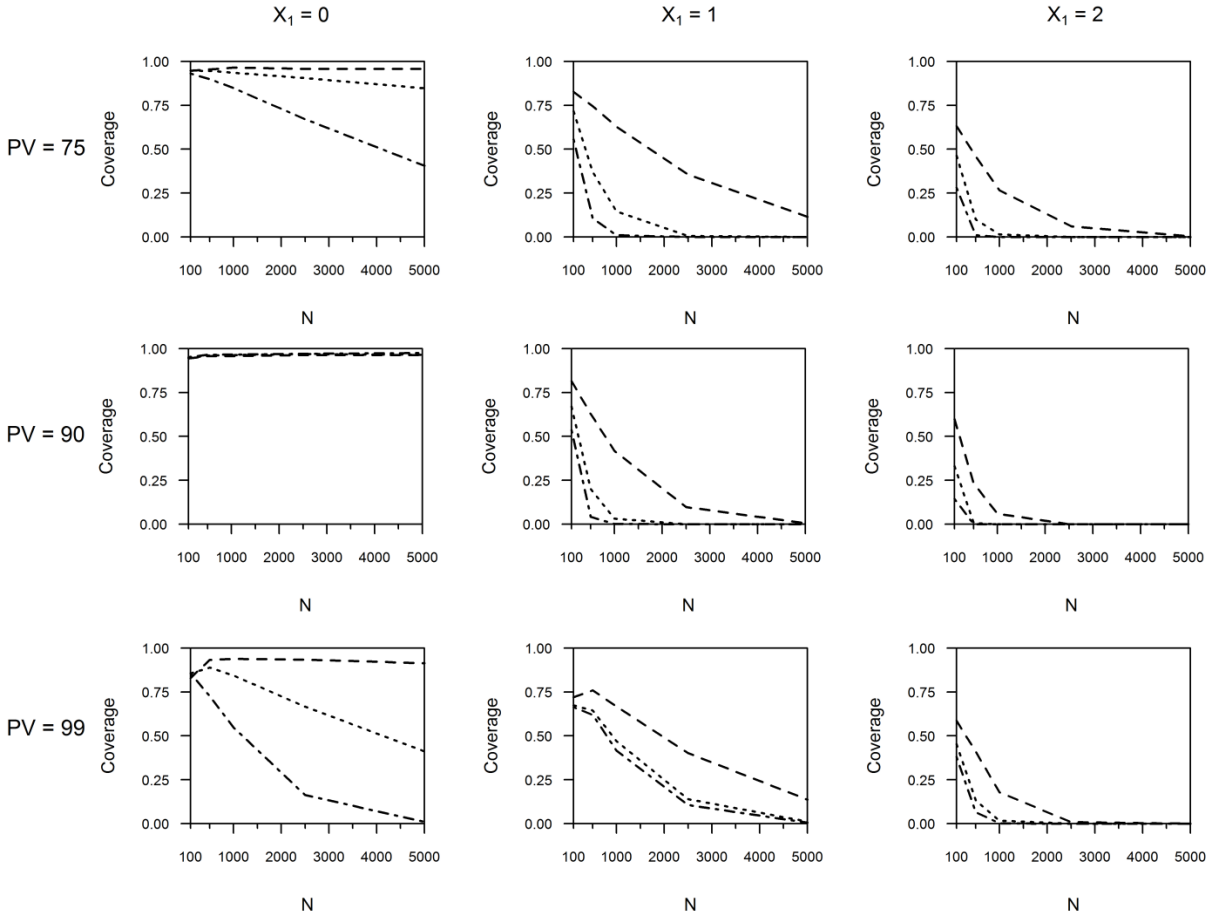

Figure 4.14. Coverage of 95% CIs of the estimates conditional on percentile value (rows) and X_1-value (columns) for weak (dashed), medium (dotted), and strong (dot-dash) violations of homoscedasticity.

4.10 General Discussion

We studied bias and precision of regression-based percentile estimates and coverage of corresponding CIs when the assumptions of linearity, independence, and homoscedasticity of the error variances were violated. We also examined the influence of assumption violation strength, sample size, and value of the estimated percentiles on the bias and the precision of the regression-based norms, and the corresponding CIs. Our results suggest that violations of linearity, independence, and homoscedasticity resulted in substantially biased percentile estimates and low coverage of the corresponding CIs, and these effects were stronger as assumption violations were stronger. On the other hand, assumption violations did not influence precision of the percentile estimates. Furthermore, sample size did not influence the bias of the percentile estimates, but an increase in sample size resulted in narrower CIs that had lower coverage and in more precise percentile estimates.

Given that violations of linearity, independence, and homoscedasticity resulted in biased norms with CIs that had low coverage, a comparison of test scores with norms could result in wrong conclusions, possibly entailing important consequences for individual test takers. Hence, if model assumptions are violated, the regression model does not fit the test data and either the traditional method or a model with weaker assumptions should be used. When estimating regression-based norms, test constructors are advised to investigate assumption violations, and to provide information about the methods used for this investigation, as well as the results. The assumptions of linearity and homoscedasticity can be investigated using a plot of the empirical residuals obtained from OLS regression (Tabachnick & Fidell, 2012, pp. 85-86, 97). The independence assumption, however, cannot be assessed using empirical analysis of the relationship between covariates and OLS regression residuals, because the least squares criterion causes the residuals to be uncorrelated with the covariates, regardless of the distribution of the unobserved population error term. As a result, the independence assumption can only be checked using underlying theory about the relationships between the covariates and the error terms in the population model. Oosterhuis et al. (2016b) provide an overview of the methods one can use to determine whether covariates influence the test scores, and sequential

regression can be used to determine whether inclusion or exclusion of certain covariates influences the effects of covariates already included in the model.

In practice, assumptions can be violated in multiple ways, and more than one assumption may be violated. Because for each of the assumptions we only examined one type of violation, consequences of assumption violations in real test data might be different. However, we expect the general results to be similar for all types of violations of linearity, independence, and homoscedasticity. Further research might focus on the consequences of assumption violations for other norm statistics, such as stanines and Z-scores, and the usefulness of models based on weaker assumptions about the data, such as non-parametric models.

4.11 Appendix

Population Model Parameters

The variance of Y_+ based on the population regression model equals

$$\sigma_{Y_+}^2 = \sigma_{\hat{Y}_+}^2 + \sigma_{\varepsilon_x}^2. \quad (28)$$

where $\sigma_{\hat{Y}_+}^2$ and σ_e^2 are the variance of $\hat{Y}_+$ and e, respectively. Then

$$R^2 = \frac{\sigma_{\hat{Y}_+}^2}{\sigma_{Y_+}^2} = 1 - \frac{\sigma_\varepsilon^2}{\sigma_{Y_+}^2} \quad (29)$$

(Equation 6). Let $R^2 = .18$. Equation 29 shows that σ_ε^2 is dependent on R^2,

$$\sigma_\varepsilon^2 = (1 - R^2) \cdot \sigma_{Y_+}^2 . \quad (30)$$

$$= (1 - .18) \cdot 1 = .82$$

No assumption violations. Assuming X_1 to be standard normal, linear algebra shows that the variance of Y_+ based on Equation 19 is equal to

$$\sigma_{Y_+}^2 = \sigma_{\beta_0+\beta_1 X_1+\varepsilon}^2 \quad (31)$$

$$= \beta_1^2 \sigma_{X_1}^2 + \sigma_\varepsilon^2$$

$$= \beta_1^2 \cdot 1 + \sigma_\varepsilon^2$$

$$= \beta_1^2 + \sigma_\varepsilon^2.$$

Next, Equation 31 implies that

$$\sigma_{Y_+}^2 - \sigma_\varepsilon^2 = \beta_1^2, \quad (32)$$

which, based on Equation 30, is equal to

$$R^2 \cdot \sigma_{Y_+}^2 = .18\sigma_{Y_+}^2 = \beta_1^2. \quad (33)$$

Equation 33 implies that

$$\beta_1 = \sqrt{(.18\sigma^2_{Y_+})}. \tag{34}$$
$$= \sqrt{.18}$$

Furthermore, basic Linear Algebra shows that the expected value of Y_+ equals

$$E(Y_+) = E(\beta_0) + \beta_1 \cdot E(X_1) + E(\varepsilon) \tag{35}$$
$$= \beta_0 + \beta_1 \cdot 0 + 0 = \beta_0.$$

And as a result,

$$\beta_0 = 0. \tag{36}$$

Violation of linearity. Let $\sigma_{X_1 X_1{}^2}$ be the covariance between X_1 and $X_1{}^2$. Linear algebra shows that for standard normal X, the variance of Y_+ (Equation 23) is equal to

$$\sigma^2_{Y_+} = \sigma^2_{\beta_0+\beta_1 X_1+\beta_2 X_1^2+\varepsilon} \tag{37}$$
$$= \beta_1^2\sigma^2_{X_1} + \beta_2^2\sigma^2_{X_1{}^2} + \sigma^2_\varepsilon + 2\beta_1\beta_2\sigma_{X_1 X_1{}^2}$$
$$= \beta_1^2 \cdot 1 + \beta_2^2 \cdot 2\sigma^4_{X_1} + \sigma^2_\varepsilon + 2\beta_1\beta_2 \cdot 0$$
$$= \beta_1^2 + 2\beta_2^2 + \sigma^2_\varepsilon.$$

Let f be an integer associated with the strength of the linearity assumption violation, and let β_2 depend on the value of β_1 so that

$$\beta_2 = f \cdot \beta_1. \tag{38}$$

If we insert Equation 38 into Equation 31, we obtain

$$\sigma^2_{Y_+} = \beta_1^2 + 2(f \cdot \beta_1)^2 + \sigma^2_\varepsilon, \tag{39}$$

which in turn implies that

$$\sigma^2_{Y_+} - \sigma^2_\varepsilon = \beta_1^2 + 2(f \cdot \beta_1)^2 \tag{40}$$
$$= (1 + 2f^2)\beta_1^2.$$

Based on equations 28 and 30,

$$\sigma^2_{Y_+} - \sigma^2_\varepsilon = \sigma^2_{\hat{Y}_+} = R^2 \cdot \sigma^2_{Y_+} = .18\sigma^2_{Y_+}. \tag{41}$$

Combining Equation 40 with Equation 41,

$$.18\sigma^2_{Y_+} = (1 + 2f^2)\beta_1^2, \tag{42}$$

which in turn implies that

$$\beta_1 = \sqrt{(.18\sigma^2_{Y_+})/(1 + 2f^2)}. \tag{43}$$

$$= \sqrt{.18/(1+2f^2)}$$

Furthermore, basic Linear Algebra shows that the expected value of Y_+ equals

$$E(Y_+) = E(\beta_0) + E(X_1)\cdot\beta_1 + E\left(X_1{}^2\right)\cdot\beta_2 + E(\varepsilon) \quad (44)$$
$$= \beta_0 + \beta_1\cdot 0 + \sigma_1^2\cdot\beta_2 + 0 = \beta_0 + \beta_2.$$

Based on Equation 44, the value of β_0 is given by

$$\beta_0 = E(Y_+) - \sigma_1^2\cdot\beta_2 = -\beta_2. \quad (45)$$

Violation of independence. Assuming covariates to be standard normal, linear algebra shows that the variance of Y_+ based on Equation 25 is equal to

$$\sigma_{Y_+}^2 = \sigma_{\beta_0+\beta_1X_1+\beta_2X_2+\varepsilon}^2 \quad (46)$$
$$= \beta_1^2\sigma_{X_1}^2 + \beta_2^2\sigma_{X_2}^2 + \sigma_\varepsilon^2 + 2\beta_1\beta_2\sigma_{X_1X_2}$$
$$= \beta_1^2\cdot 1 + \beta_2^2\cdot 1 + \sigma_\varepsilon^2 + 2\beta_1\beta_2\rho_{12}$$
$$= \beta_1^2 + \beta_2^2 + \sigma_\varepsilon^2 + 2\beta_1\beta_2\rho_{12}.$$

Assuming that X_1 and X_2 had equally strong effects on Y_+,

$$\beta_1 = \beta_2. \quad (47)$$

Inserting Equation 47 into Equation 46 results in

$$\sigma_{Y_+}^2 = (2+2\rho_{12})\beta_1^2 + \sigma_\varepsilon^2. \quad (48)$$

Next, Equation 48 implies that

$$\sigma_{Y_+}^2 - \sigma_\varepsilon^2 = (2+2\rho_{12})\beta_1^2, \quad (49)$$

which, based on Equation 35, is equal to

$$R^2\cdot\sigma_{Y_+}^2 = (2+2\rho_{12})\beta_1^2. \quad (50)$$

Equation 50 implies that

$$\beta_1 = \sqrt{\left(R^2\cdot\sigma_{Y_+}^2\right)/(2+2\rho_{12})}. \quad (51)$$
$$= \sqrt{(.18\cdot 1)/(2+2\rho_{12})}$$

Furthermore, basic Linear Algebra shows that the expected value of Y_+ equals

$$E(Y_+) = E(\beta_0) + \beta_1\cdot E(X_1) + \beta_2\cdot E(X_2) + E(\varepsilon) \quad (52)$$
$$= \beta_0 + \beta_1\cdot 0 + \beta_2\cdot 0 + 0 = \beta_0.$$

Violation of homoscedasticity. The sample regression model (Equation 27) assumed homoscedasticity of *e*, whereas the population error term $(\alpha_1 + \alpha_2X_1)\cdot\varepsilon$ was heteroscedastic. Considering that R^2 assumes a constant error term, we determined the

model parameters based on the sample value of R^2 instead of the population value. The sample estimate of the population error term was equal to

$$e = (\alpha_1 + \alpha_2 X_1) \cdot \varepsilon, \tag{53}$$

with variance

$$\sigma_e^2 = (1 - R^2) \cdot \sigma_{Y_+}^2, \tag{54}$$

which is equal to Equation 30. Linear algebra shows that the variance of Y_+ (Equation 27) is equal to

$$\sigma_{Y_+}^2 = \sigma_{(\beta_0 + \beta_1 X_1 + (\alpha_1 + \alpha_2 X_1) \cdot \varepsilon)}^2. \tag{55}$$

We determined the values of β_0 and β_1 based on the error term in the sample regression model (Equation 5), because this ensured that the expected value of R^2 for the sample model was equal to the desired R^2. Hence, if the population error term is replaced by the sample error term (Equation 53), Equation 55 is equal to

$$\begin{aligned} \sigma_{Y_+}^2 &= \sigma_{(\beta_0 + \beta_1 X_1 + e)}^2 \\ &= \beta_1^2 \sigma_{X_1}^2 + \sigma_e^2 + 2\beta_1 \sigma_{X_1 e} \\ &= \beta_1^2 \cdot 1 + \sigma_e^2 + 2\beta_1 \cdot 0 \\ &= \beta_1^2 + \sigma_e^2 \end{aligned} \tag{56}$$

In turn, Equation 56 implies that

$$\beta_1 = \sqrt{\sigma_{Y_+}^2 - \sigma_e^2} = \sqrt{1 - .82} = \sqrt{.18}. \tag{57}$$

Furthermore, linear algebra shows that the expected value of Y_+ equals

$$\begin{aligned} E(Y_+) &= E(\beta_0) + \beta_1 \cdot E(X_1) + E((\alpha_1 + \alpha_2 X_1) \cdot \varepsilon) \\ &= \beta_0 + \beta_1 \cdot E(X_1) + [\alpha_1 + \alpha_2 E(X_1)] \cdot E(\varepsilon) \\ &= \beta_0 + \beta_1 \cdot 0 + (\alpha_1 + \alpha_2 \cdot 0) \cdot 0 = \beta_0. \end{aligned} \tag{58}$$

Hence,

$$\beta_0 = E(Y_+) = 0.$$

To simulate scores using Equation 27, we also needed to determine the required value for σ_ε^2. Based on Equation 53 and Equation 54, the variance of the population error term (Equation 27) was equal to

$$\sigma_{(\alpha_1 + \alpha_2 X_1) \cdot \varepsilon}^2 = .18 \sigma_{Y_+}^2. \tag{59}$$

Linear algebra shows that Equation 50 can also be written as

$$\left(\alpha_1^2 + \alpha_2^2\sigma_{X_1}^2\right)\sigma_\varepsilon^2 = .18\sigma_{Y_+}^2 \tag{60}$$
$$= (\alpha_1^2 + \alpha_2^2 \cdot 1)\sigma_\varepsilon^2 = .18\sigma_{Y_+}^2.$$

In turn,

$$\sigma_\varepsilon^2 = \frac{\left[(1-R^2)\cdot\sigma_{Y_+}^2\right]}{\alpha_1^2+\alpha_2^2} = .18/(\alpha_1^2 + \alpha_2^2). \tag{61}$$

The conditional variance of residuals for a given value of X_1 denoted $\sigma_{\varepsilon|X_1}^2$, can be obtained by

$$\sigma_{\varepsilon|X_1}^2 = \sigma_{(\alpha_1+\alpha_2X_1)\cdot\varepsilon}^2 \tag{62}$$
$$= (\alpha_1 + \alpha_2X_1)^2 \cdot \sigma_\varepsilon^2.$$

Chapter 5

A Procedure for Estimating Regression-Based Norms Using a Real Data Example

Abstract

Using the Type-D Scale (DS-14) as an example, this article describes a procedure for estimating regression-based norms for psychological tests and questionnaires. The procedure consists of four steps: (a) selection of predictor variables, (b) checking model assumptions, (c) estimating precision of the norm estimates, and (d) the interpretation and presentation of the regression-based norms. First, we illustrate how test constructors can select covariates based on the statistical significance of their relationship with the test score, and we discuss the consequences of considering the covariate for test score interpretation. Second, because violations of the assumptions of linearity/additivity, independence of the covariates and the error term, and homoscedasticity can lead to substantially biased norm estimates, we discuss how to investigate the assumptions in real data. Third, to quantify estimation precision, we provide a procedure to estimate standard errors and confidence intervals for regression-based norms. Finally, we discuss how regression-based norms can be presented and interpreted.

5.1 Introduction

Norms are an integral part of psychological testing, because they help to interpret raw scores that would otherwise have no meaning. For example, the Type D Scale-14 (DS-14; Denollet, 2005) is a brief measure of Type-D personality, which is characterized by a combination of high negative affectivity (NA) and high social inhibition (SI). NA refers to the tendency to experience feelings of dysphoria, anxiety, and irritability, whereas SI refers to discomfort in social interactions, lack of social poise, and the tendency to avoid confrontation in social interaction leading to nonexpression (Denollet, 2005). NA and SI are both measured by means of seven items on a 5-point Likert scale (0 = false, 1 = rather false, 2 = neutral, 3 = rather true, 4 = true). Hence, the lowest possible score on each item is 0 and the highest score is 4. Given that the subtests for NA and SI each contain seven items, the minimum and maximum total raw scores on both subtests are 0 and 28, respectively. If a particular male test taker has a raw score of 15 on the items measuring NA and a raw score of 12 on the items measuring SI, these raw scores alone do not inform us about his level of NA and SI. Thus, Denollet (2005) provides a norm table, which shows that the NA raw score is high, and the SI raw score is above average compared to other males. Test constructors also frequently present norm statistics, including percentile ranks, linear standard scores such as Z-scores, and normalized standard scores such as stanines (Mertler, 2007, Module 6).

Traditionally, one or more covariates are used to define relevant subgroups in the norm population from which a stratified norm sample is drawn. Categorical covariates, such as gender and education level, can be used readily to subdivide the norm sample. For example, if gender is used as a covariate, the norm sample can be readily divided into two categories: males and females. On the other hand, if continuous covariates, such as age, are used, partly arbitrary categories have to be formed. For example, age can be categorized in years, months, or even days. Categorization of a continuous covariate can have important consequences for test takers who are close to the boundaries of the resulting categories. For example, suppose we have a mathematics test for children consisting of 15 questions. Norms are provided for age groups in years (i.e., 4 to 5 years old, 5 to 6 years old, 6 to 7 years old, etc.). Suppose Ethan is a five-year old boy whose birthday is on the 11th of

November. Ethan fills out the test on the 10th of November and scores 10 out of 15 questions correct. Considering that Ethan made the test prior to his birthday, his raw score is compared to the raw scores of children aged 5 to 6 years old. However, if Ethan were to fill out the test 2 days later, he would be 6 years old and compared to a completely different group of children (i.e., 6 to 7 year olds). It is possible that Ethan scores above average if compared to children age 5 to 6, but below average when compared to children 6 to 7 years old. Hence, even though Ethan's raw score is the same in both situations, and the test administrations are only 2 days apart, the interpretation of Ethan's test score might be completely different.

A solution would be to use narrower age categories. For example, we could determine norms for age categories a month long. In this case, Ethan would first be compared to children 5 years and 11 months to 6 years old. In the second scenario, Ethan would be compared to children exactly 6 years old to children 6 years and 1 month old. In this case, norms are expected to differ less between the age categories, because the age of the children in both categories differs less. However, the smaller the age range in each category, the larger the number of categories necessary to span a longer age range, and given a particular sample size, the smaller the number of observations within each category. Hence, although narrower age categories might reduce problems with interpretation due to categorization, a larger norm sample is required to maintain a desired precision level of the norms. The width of categories for continuous covariates is then determined by weighing bias in test score interpretation and the required size of the norm sample.

Zachary and Gorsuch (1985) proposed an alternative and increasingly popular method of estimating norms using a linear regression model, usually referred to as regression-based norming. Regression-based norming circumvents the need to arbitrarily categorize continuous covariates, because continuous covariates can be added to the regression model directly and categorical covariates can be added as dummy variables. Hence, the regression-based approach estimates norms using the entire norm sample while simultaneously taking the covariates into account, whereas the traditional approach to norming requires a subdivision of the total norm sample into smaller groups based on the covariates. This means the regression-based approach requires a smaller norm sample

compared to the traditional method to obtain equally precise norm estimates and is therefore more efficient (Oosterhuis, Van der Ark, & Sijtsma, 2016b). The greater efficiency of regression-based norming, as well as the straightforward incorporation of continuous covariates into the model, have caused the method to become popular in recent years (e.g., Goretti et al., 2014; Llinàs-Reglà,Vilalta-Franch, López-Pousa, Calvó-Perxas, & Garre-Olmo, 2013; Parmenter, Testa, Schretlen, Weinstock-Guttman, & Benedict, 2010; Roelofs et al., 2013a; Roelofs et al., 2013b; Van Breukelen & Vlaeyen, 2005; Van der Elst, Dekker, Hurks, & Jolles, 2012; Van der Elst, Hoogenhout, Dixon, De Groot, & Jolles, 2011; Vlahou et al., 2013). The aim of this chapter is to provide test constructors with a procedure to estimate regression-based norms, including (a) the selection of relevant covariates, (b) model assumptions, (c) precision of the norm estimates and (d) their presentation.

The procedure for estimating regression-based norms is illustrated using norm data for the DS14 (Denollet, 2005). The norm sample consisted of 532 test takers who completed all items (mean age 58.6 ± 10.5 years), including 65 women and 467 men. The norm data are used for illustration purposes only in this paper. For a substantive discussion of the DS-14 norms, we refer to Denollet (2005).

5.2 Selection of Covariates

In order to estimate regression-based norms, relevant covariates have to be selected and added to the regression model as predictors of the raw test scores in the norm sample. To find relevant covariates, test constructors might first investigate the association between the demographic variables that were used to stratify the norm sample and the test scores. Typically, a norm sample is stratified instead of randomly sampled to ensure that the sample is representative of the population of interest (Kline, 2000, pp. 53-59). Stratified sampling involves first subdividing the population into different subgroups, followed by random sampling from the separate subgroups**.** The variables that are used to stratify a sample are preferably related to score differences between individuals (Frick, Barry, & Kamphaus, 2005, pp. 30-33). Hence, for a norm sample, the stratification variables are expected to be correlated with the scores on the test under consideration (Kline, 2000, p. 54). For psychological tests, stratification variables often include gender, age, and social class (Frick et al. 2005, pp. 30-33; Kline, 2000, pp. 54-55; Oosterhuis et al., 2016b).

If too many stratification variables are used, the resulting norm sample has to be enormous to maintain accuracy, whereas if the wrong variables are used, the norm sample is not representative of the population the norms are intended for (Kline, 2000, pp. 54-56). Hence, it is critical that all important stratification variables are selected to form the norm sample, while the number of variables must be kept to a minimum. According to Kline (2000, pp. 55-56), four stratification variables are usually sufficient for norms intended for use in the general population, and additional variables have to be selected with care, because they inordinately increase the required norm sample size. The nature of the test determines which variables have to be selected. The correct stratification variables are those that correlate strongest with the test under consideration (Kline, 2000, pp. 53-59). For example, the stratification variables that were used to estimate community-specific norms for the Lorge-Thorndike Intelligence test (1957) were percentage of adult literacy, proportion of professional workers, percentage of home ownership, and median home rental value. On the other hand, norms for the Cattell 16PF test (Cattell, Eber, & Tatsuoka, 1970), a personality test, were estimated based on geographical area, population density, age group, and family income.

Instead of estimating norms for the general population, norms can also be estimated for a specific group of individuals (Kline, 2000, pp. 56-58). Similar to norms for the general population, stratification variables are selected based on their correlation with the test scores. Those variables that have the highest correlations with the test scores are used to stratify the sample. If norms are estimated for a specific group, the required size of the norm sample is much smaller, because the specific group is usually more homogenous. As a result, fewer stratification variables are required. It should be noted, however, that if a norm sample is very small, say, containing no more than 300 observations (Kline, 2000, p. 58), the resulting norms should be interpreted with care, because due to sampling error they might not be representative and might lead to an erroneous interpretation of individuals' test scores.

To determine which of the variables that were used to stratify the sample should be added to the regression model as covariates, the correlation between these variables and the test score has to be investigated. Usually, either correlational analysis, simultaneous regression, or stepwise regression is used for this purpose (Oosterhuis et al., 2016b). If

correlational analysis is used, a Pearson's correlation is computed for all continuous covariates that are hypothesized to be related to the test score (e.g., Kessels, Montagne, Hendriks, Perrett, & De Haan, 2014). If covariates are dichotomous (e.g., gender), a Mann-Whitney test (e.g., Cavaco et al., 2013; Goretti et al., 2014) or a Student's *t*-test (e.g., Goretti et al., 2014) can be performed for non-normal and normal data, respectively. If the covariate is discrete and consists of at least three categories, an analysis of variance (ANOVA) can be performed instead (e.g., Kessels et al., 2014). Next, all covariates that are significantly correlated with or have a significant effect on the raw test score ($p < \alpha$; p is the probability of exceedance, α is the significance level) are selected. Although this method is straightforward, univariate methods such as a Pearson's correlation or Student's *t*-test, require separate tests for each of the covariates, which might inflate the Type-I error. As a result, many test constructors prefer the use of regression analysis to assess the relationship between covariates and the test score.

For each covariate in the model, regression analysis provides a regression coefficient that describes the effect of the covariate on the test score while controlling for the effects of the other covariates in the model. If simultaneous regression is used, a regression model containing all covariates is used to simultaneously test the corresponding regression coefficients for significance (e.g., Conti, Bonazzi, Laiacona, Masina, & Coralli, 2014; Shi et al., 2014; Van der Elst et al., 2013; Yang et al., 2012). Only covariates that have significant regression coefficients ($p < \alpha$) are selected.

Instead of simultaneously testing the regression coefficients of all covariates, covariates can also be added to the regression model in steps, usually referred to as stepwise regression analysis. First, all covariates that are hypothesized to be related to the test score are simultaneously included in the regression model. Second, of all covariates that have insignificant regression coefficients ($p > \alpha$), the covariate having the greatest *p*-value is deleted form the regression model. Third, the regression model including the remaining covariates is reestimated. Fourth, in the new model, the covariate having the greatest insignificant *p*-value is deleted from the model. This procedure is repeated until all remaining covariates have regression weights significantly different from zero ($p < \alpha$).

When using simultaneous regression to select covariates, all covariates that have insignificant regression coefficients are simultaneously deleted from the model. Hence,

simultaneous regression does not take into account that deleting one or more covariates from the model might change the regression coefficients and significant tests of the remaining covariates. Stepwise regression, on the other hand, repeatedly tests the remaining regression coefficients after a covariate is deleted from the model. However, this method has two important drawbacks. First, considering that multiple comparisons have to be performed in each step of the procedure, the overall significance level cannot be controlled. Second, stepwise regression easily capitalizes on chance and thus likely produces results that are not replicable (Derksen & Keselman, 1992; Leigh, 1988). As a result, it is possible that stepwise regression leads to the selection of covariates that may not the best predictors of the test score.

Although covariates often are selected based on their relationship with the test score, test constructors should also consider the influence of the selected covariates on the interpretation of individuals' test scores. If the relationship is caused only by differences in symptomatology between groups of individuals, the covariate should be used to estimate the norms (Frick et al. 2005, pp. 30-33). For example, using gender to estimate norms means that the effect of gender differences on test performance is removed. Hence, roughly the same proportions of boys and girls are identified as having an exceptional score. To illustrate this, one may consider a test that measures attention deficit hyperactivity disorder (ADHD). Boys with ADHD are known to show more overt symptoms, such as hyperactivity, whereas girls with ADHD suffer more from covert symptoms, such as inattention (Biederman et al., 2005; Gaub & Carlson, 1997; Gershon, 2002). Traditionally, tests for ADHD have focused more on the overt symptoms of boys, and as a result, girls with ADHD are consistently underidentified (Skogli, Teicher, Andersen, Hovik, & Øie, 2013). Hence, to diagnose ADHD in both boys and girls correctly, norms for boys need to focus more on overt symptomatology, whereas norms for girls need to focus more on covert symptoms. In this case, separate norms for boys and girls would correctly identify a larger number of girls.

If the relationship between covariates and the test score is caused by actual differences between groups with respect to the occurrence of the behavior or the trait, the covariate should not be used to estimate norms. It is possible that the number of individuals that actually suffer from a disorder or have a certain trait differs between

subgroups in a norm sample. For example, major depressive disorder has the lowest prevalence in individuals > 60 years old and the highest prevalence in 30-44 year old individuals (Kessler et al., 2005), which means we are likely to find a significant association between age and depression scores. If we use age as a covariate to estimate norms for a test measuring depression, an equal number of, for example, individuals 30-44 years old and individuals over 60 years old would be diagnosed with depression (Frick et al., 2005, pp. 30-33). This means that major depressive disorder is underidentified in individuals 30-44 years old and overidentified in individuals over the age of 60. In this case it might be better to estimate norms without taking age into account, even though age correlates with scores on the depression test.

Furthermore, in general norms for developmental tests (e.g., ability tests for children) should be age-dependent. The results of ability tests for children rarely make sense if compared to norms that are not age-dependent. For example, mathematics ability naturally increases as children grow older and have received more years of formal education (Eccles, 1999). As a result, older children are expected to have higher mean scores on these tests than younger children. If norms are provided independent of age, younger children would automatically have relatively poor scores, whereas older children would have relatively high scores. On the other hand, if age-dependent norms are provided, the same number of individuals in each age group are interpreted as low-scoring or high-scoring. Hence, age-dependent norms allow us to investigate the development of a child by answering how well the child performed compared to children of similar age. On the other hand, if the test administrator is interested in the absolute reading ability of a child, for example, to determine the appropriate difficulty level of reading materials, norms independent of age are suitable.

Hence, not only the strength and significance of the relationship between a covariate and the test score, but also the underlying reason for this relationship should be taken into account when discerning which covariates should be added to the regression model. If the norm sample consists of a special group instead of a sample from the general population, an individual's performance on the test is compared to individuals with known problems or a known diagnosis. For example, for a psychological test we can investigate whether the

symptoms shown by an individual are unusual compared to individuals that were referred for psychological evaluation (Frick et al. 2005, pp. 30-33).

Example. We performed correlational analysis, simultaneous regression and stepwise regression to investigate relations between covariates and scores on the DS-14. The covariates we considered were the stratification variables gender and age (measured in years), which are often used as covariates for psychological norm data (Oosterhuis et al., 2016b) .

Correlational analysis. The correlation between age and the total score on the DS-14 was equal to r = -0.095, and was significant, $t(530) = -2.19$, $p = .029$. This negative correlation coefficient indicates that higher age was associated with lower DS-14 scores. In addition to testing the statistical significance of an association, effect sizes should be investigated, because norm samples are usually large and might result in small effect sizes that are statistically significant. In this case, the correlation was small ($r > .10$ = small, $r >$ .30 = medium, and $r > .50$ = large; Cohen, 1988), which might be a reason not to select age as a covariate. Considering that gender was a dichotomous variable (0 = female, 1 = male), we performed an independent samples *t*-test. Females had a higher mean score on the DS-14 than males ($M = 19.91$ and $M = 18.63$, respectively), but this difference was non-significant, $t(530) = .933$, $p = .351$. Hence, based on the correlational analysis, age is selected as a covariate.

Simultaneous regression. Gender and age were used to predict raw scores on the DS-14. The model was not significant, $R^2 = .011$, $F(2,529) = 2.929$, $p = .054$. Hence, based on the simultaneous regression model, neither age nor gender are selected as covariates.

Stepwise regression. We estimated a model in which gender and age were included in the starting model. In the next step, gender was removed from the model, resulting in a model that included only age as a predictor of the DS-14 scores. Age had a negative effect on DS-14 scores, $t(1,530) = -2.19$, $p = .029$. A suitable effect size for individual regression coefficients is Cohen's f^2 (Cohen, 1988). Cohen's f^2 is a standardized measure of one variable's effect size within the context of a multivariate regression model (Selya, Rose, Dierker, Hedeker, & Mermelstein, 2012) and is defined as

$$f^2 = \frac{R^2_{AB} - R^2_A}{1 - R^2_{AB}}, \quad (1)$$

where B is the covariate for which the effect size is computed, A is the set of all other covariates included in the model, R^2_{AB} is the proportion of variance accounted for by A and B together, and R^2_A is the proportion of variance accounted for by A. This means that the numerator of Equation 1 corresponds to the variance uniquely accounted for by B (Cohen, 1988), whereas the denominator corresponds to the total variance not accounted for by all covariates in the model. The effect of age was very small, Cohen's f^2 = .009 ($f^2 \geq .02$ = small, $f^2 \geq .15$ = medium, and $f^2 \geq .35$ = large; Cohen, 1988). Considering that age had a significant effect on the DS-14 scores, it might be selected as a covariate, but given the small effect one might consider not using this covariate to estimate norms.

If we deem the effect size of age large enough to select it as a covariate, we need to consider next whether the age effect is caused by differences in symptomatology or differences in actual occurrence of Type-D personality. The norm sample consists of patients suffering from congenital heart conditions (CHC), and type-D personality is a known predictor of poor health outcomes in this patient group (Denollet, 2005). Hence, it might be possible that on average Type-D scores are slightly lower in the older age groups, because individuals with high DS-14 scores also have a worse health status. Due to the lower health status, individuals with high scores on the DS-14 might be less likely to cooperate with a scientific study, and might therefore be absent from the norm sample more often than individuals with lower DS-14 scores and better health status. In this case, the significant effect of age on DS-14 scores would be caused by an actual difference in the number of individuals with high DS-14 scores between age groups. This would mean that age need not be taken into account when using this special group to estimate norms for the DS-14. On the other hand, if the effect of age on DS-14 scores could be explained by a difference between younger and older individuals with respect to reporting style or symptomatology, age should be used as a covariate. For illustration purposes, in the remainder of this article we assume that age is an important covariate.

To be able to estimate regression-based norms, we first added Age to the linear regression model as a predictor of the dependent variable Y_+, the DS-14 scores. Next, we computed the predicted DS-14 scores based on the estimated regression weights,

$$\hat{Y}_+ = 24.257 - 0.093 * \text{Age}$$

For example, for a 30-year old test taker, $\hat{Y}_+ = 24.257 - 0.093 * 30 = 21.467$. Next, for each individual in the norm sample, the predicted test score can be compared to the observed test score. The resulting prediction error is often referred to as the residual. Let the 30-year old test taker have an observed DS-14 score of 25. The residual for this individual would then be equal to $e = Y_+ - \hat{Y}_+ = 25 - 21.467 = 3.533$, which means he scored 3.5 points higher than expected based on his age. Next, for each individual in the norm sample, the residual scores can be standardized using the standard error of the residuals,

$$S_e = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{N-K-1}}, = 10.33 \quad (2)$$

where *i* is used to index the individuals in the norm sample and *K* is equal to the number of covariates in the regression model (in this case *K* = 1). The standardized residual of the 30-year old test taker is then equal to $\frac{e}{S_e} = \frac{3.533}{10.33} = 0.342$. Hence, based on his age, the 30-year old test taker scored .342 standard deviations above his predicted score. The resulting distribution of standardized residuals in the norm sample can be used to estimate the regression-based norms, such as stanines or percentile ranks.

5.3 Assumption Violations

Before norms are estimated based on the linear regression model, the validity of the assumptions made by this model should first be investigated. The linear regression model is a statistical model, which means that it is an ideal and simplified representation of a complex reality and aids in the description and understanding of a specific process or to predict future outcomes (Von Davier, 2011, p. 3). Linear regression models make several assumptions, often referred to as the Gauss-Markov assumptions, about the truth. The validity of these assumptions determines whether the use of the regression model is justified or whether it might lead to incorrect inferences. The Gauss-Markov assumptions apply to the population regression model and include linearity and additivity, mean error term equal to zero, independence between error and covariates, independence between observations, homoscedasticity, and normality (Berry, 1993, p. 12). Let β_0 be the population intercept, let $\beta_1, \ldots, \beta_K$ be the population regression coefficients corresponding to the *K* covariates included in the model, and let ε be the population error term. This error

term encompasses both the effects of excluded variables that influence Y_+ and a random component characteristic human behavior. Then, the population model for the test score equals

$$Y_+ = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K + \varepsilon. \quad (3)$$

The assumption of linearity/additivity requires that the relationship between the covariates and the test score is linear and additive (i.e., no interaction effects). If the relationships in the model are non-linear or non-additive, norm estimates based on the model are biased (Oosterhuis, Sijtsma, & Van der Ark, 2016a). The remaining five assumptions involve the error term in the population regression model (Equation 3). First, the expected value of the error term in Equation 3 is assumed to be equal to zero. This assumption might be violated if the dependent variable in the model is measured with error or if the data are truncated (Berry & Feldman, 1985, p. 73). This assumption has the least important consequences for the model estimates, because a violation of this assumption only influences the value of the intercept, which absorbs the expected value of the error term (Berry & Feldman, 1985, p. 73). Hence, a violation of this assumption is inconsequential for norms based on the estimated regression model, and does not need to be investigated.

Second, the covariates and the error term in the population model are assumed to be independent. A violation of this assumption might occur if an excluded relevant covariate correlates with the included covariate(s) or if the model contains reciprocal causation (i.e., a feedback loop between a covariate and the dependent variable). If, for example, an important covariate is omitted from the model and this assumption is violated, the resulting norm estimates are biased (Oosterhuis et al., 2016a)

Third, observations in the norm sample are assumed to be independent of the other observations in the sample. As a result, the individual error terms are also required to be independent of each other. This assumption is often violated for longitudinal data and for data collected across spatial units (Neter, Kutner, Nachtsheim, & Wasserman, 1996. pp. 104-105). Hence, the nature of data collection can predict a violation of this assumption beforehand, and models are available that accommodate correlated errors (Neter et al., p. 125).

Fourth, the error term is assumed to be homoscedastic, which means that it is constant for all levels of the covariates (Fox, 1997, pp. 272-274). A violation of homoscedasticity occurs when a high value on a covariate is a necessary but not a sufficient condition for a high value of the test score. Other causes of a violation are model misspecification, heterogeneous data (i.e., large score-differences between individuals on the covariates or the test scores), data that contain error, have a large range, have skewed distributions, or suffer from outliers. If this assumption is violated, the resulting norm estimates are biased and corresponding confidence intervals (CIs) have low coverage (Oosterhuis et al., 2016a).

The fifth assumption is normality of the error term, which means that the errors are assumed to follow a normal distribution. This assumption also implies that the distribution of the test score conditional on the covariate scores is normal. The normality assumption is often justifiable in the social and behavioral sciences, because the error term is assumed to consist of random effects of excluded variables and random measurement error (Neter et al., pp. 29-30). If the normality assumption is violated, the ordinary least squares (OLS) estimators are still the best linear unbiased estimators (BLUE; Berry, 1993, pp. 18-22), but statistical hypothesis testing based on a normal distribution might not be valid. However, due to the central limit theorem, if the sample is larger than 50 observations (Casson & Farmer, 2014), the regression model usually is robust against violations of the normality assumption (Neter et al., pp. 29-30). Norm samples are typically much larger than 50 observations, which means that non-normality of the residuals usually does not have consequences for hypothesis tests of the regression model estimates. Hence, norms can be estimated regardless of the shape of the empirical distribution of the residuals, as long as the conditional distribution shape is constant.

Example. The linearity/additivity assumption was investigated by a plot of the (standardized) predicted values versus the (standardized) residual values (Tabachnick & Fidell, 2012, pp. 85-86, 97), which was made using *R* (*R* core team, 2016). The residuals in Figure 5.1 should be spread randomly around the horizontal zero-line, which means that trends should be absent. The dashed curve in Figure 1 is a scatterplot smoother, which shows the average value of the residuals for each fitted value. Hence, this smoothed curve indicates an absence of trends if it coincides with the dotted horizontal zero-line.

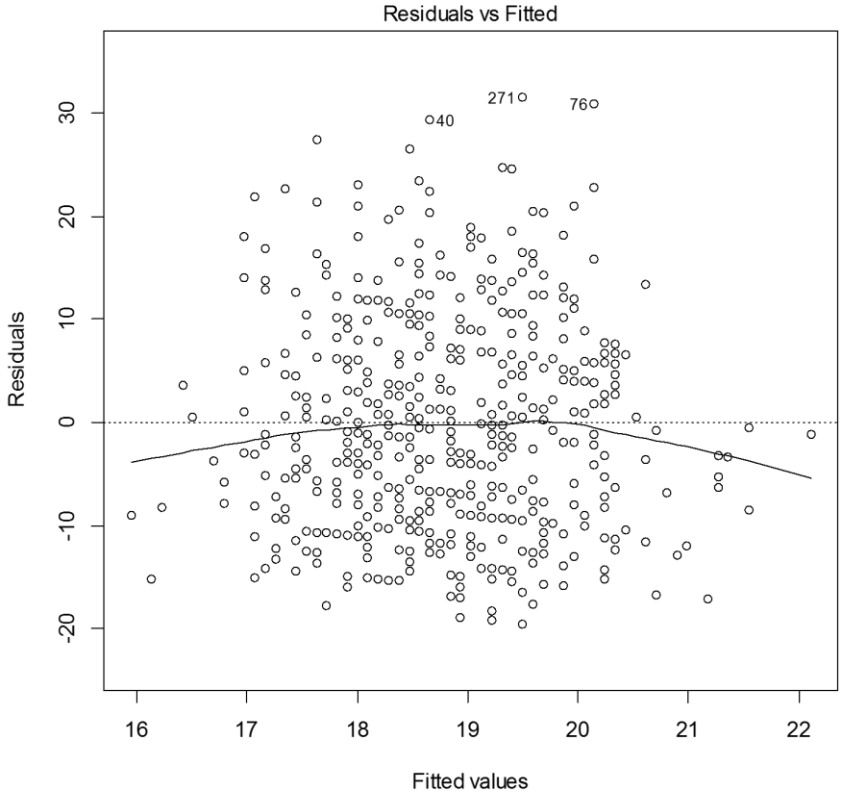

Figure 5.1. Residuals (y-axis) and fitted values (x-axis) based on linear regression model.

In this case, the dashed curve did not markedly deviate from the zero line which means that there was no suspected violation of linearity. The possibility of a non-linear effect can also be investigated by adding a quadratic term or other higher-order terms to the model. To check for this possibility, we added Age^2 to the regression model. We found that the quadratic effect was not significant, t (529) = 1.939, p = .053. Hence, this result provides further evidence that the assumption of linearity is not violated in the DS-14 data. Additivity only plays a role if the model contains more than one covariate, in which case an interaction term can be added to the model and can be tested for statistical significance.

Independence between covariates and the error term cannot be checked using the results from the estimated regression model, because the empirical model only allows a zero correlation between the covariates and the error term. This means the correlation between covariates and the error term in the empirical regression model is always equal to

zero, even though the correlation might be nonzero in the population regression model. Considering the population regression model is unknown, only scientific theory about the underlying relationships between both included and excluded covariates can be used to determine whether a violation of this assumption is likely. For example, scientific theory might indicate the possibility of a feedback loop between the dependent variable and one or more of the covariates. A violation of the independence assumption due to omission of an important covariate might be checked by adding the omitted covariate to the regression model. If this addition considerably changes the values of the regression coefficients of the covariates already included in the model, the suspected covariate should be retained in the regression model. On the other hand, if inclusion of the suspected covariate does not lead to changes of the other regression coefficients, it might be safe to omit the variable from the model.

The assumption of independence between observations in the norm sample can be checked by considering the method of data collection. In case of the DS-14 norm data, data were neither longitudinal nor collected across spatial units. Hence, a violation of this assumption was unlikely.

The assumption of homoscedasticity can also be checked using a plot of the (standardized) predicted values versus the (standardized) residual values (Tabachnick & Fidell, 2012, pp. 85-86, 97). Figure 1 shows that the residuals appeared to be spread evenly across the entire range of fitted values, which suggests there were no violations of homoscedasticity.

The assumption of normality was not checked, because the regression model was robust due to the large sample size ($N > 500$). Given that our evidence suggested that none of the other Gauss-Markov assumptions were violated in the DS-14 data, the sample regression model can be used to estimate norms.

5.4 Estimation Precision

Norm estimates are subject to sampling fluctuation, which means that if we draw multiple norm samples from the population, due to chance, norm estimates are expected to differ from sample to sample. This also means that decisions that are made by comparing test results to norms might differ from norm sample to norm sample. Considering that psychological tests and questionnaires are often used to make important decisions about

individuals, such differences in test score interpretation might have major consequences for individuals' lives. Hence, it is important that sampling fluctuation of the norms estimates is minimized as much as possible.

The larger the norm sample, the less sampling fluctuation, which means that the values of norm estimates are expected to fluctuate less from sample to sample. Crawford and Howell (1998) have argued that it is justifiable to treat norm statistics as population values if the sample is large enough. However, there are no clear size requirements for norm samples that can be used to determine whether a sample is large enough. The American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) have provided guidelines for test construction (AERA, APA, & NCME, 1999), but without mention of recommended norm sample sizes. On the other hand, Evers, Lucassen, Meijer, and Sijtsma (2009) have provided Dutch test constructors with practical guidelines for the size of norm groups, but these guidelines lack a sufficient statistical basis.

Standard errors (SEs) provide a measure of the variability of a statistic due to sampling fluctuation. The SE corresponds to the standard deviation of the sampling distribution the statistic is assumed to be drawn from. The SE can also be used to construct a confidence interval (CI) for a statistic. CIs, of which Wald-based CIs are the most commonly used, are directly related to the SEs. Let $\hat{\theta}$ be the point estimate with $SE_{\hat{\theta}}$, let α be the two-tailed p-value and $z_{\alpha/2}$ the corresponding Z-score, then the limits of the 100(1-α)% Wald-based CI are

$$\hat{\theta} \pm z_{\alpha/2} \cdot SE_{\hat{\boldsymbol{\theta}}}. \quad (4)$$

Although the American Psychological Association (2010, p. 34) requires that point estimates are accompanied by an associated measure of variability (e.g., the standard error) and strongly recommends reporting CIs, SEs and CIs for norm statistics are rarely reported. This is mainly because for many norm statistics the SEs are unknown, difficult to derive, or if available not included in popular software. Crawford and Garthwaite (in press) discussed a procedure for obtaining CIs for z-scores and percentile norms, and have applied their procedure to self-report mood scales (Crawford, Cayley, Lovibond, Wilson, & Hartley, 2011). However, this procedure can only be applied to raw test scores (scores

need to be positive integers). Considering that regression-based norms are based on (standardized) residual scores, which can equal all real numbers, the procedure by Crawford et al. (2011) is not suitable for regression-based norms.

Under the mild assumption that scores follow a multinomial distribution, Oosterhuis et al. (in press) have derived SEs for the test-score standard deviation, percentiles rank scores, the boundaries of the stanines, and *Z*- scores. This procedure provides a point estimate, SE, and CI for each norm statistic. For percentile rank scores and z-scores, the procedure estimates percentile ranks and z-scores for each unique observed score. As a result, when using regression analysis, the number of estimated percentile ranks and z-scores can become extensive as sample size increases, because most residual scores are likely to be unique. Therefore, we suggest two remedies for this problem: (1) a limited number of percentile ranks and z-scores is selected from the results, or (2) specific percentile ranks and z-scores are selected beforehand and only for these specific values SEs and CIs are determined. Hence, the first approach entails estimating the percentile ranks that correspond to all observed unique raw test scores, whereas the second approach entails finding the observed raw test score that matches a predetermined percentile rank or z-score. The selection of these predetermined norm values is based on the intended use of the test and the test characteristics. For example, if the test is used to identify high-scoring individuals, the selected percentile ranks and z-scores should cover the area of the distribution in which these high-scoring individuals are located more extensively than the remainder of the distribution. In testing practice, predetermined percentile values of 50, 75, 90, 95, and 99 are commonly used as norms (Bride, 2007; Glaesmer et al., 2012; Krishnan et al., 2004; and Wizniter et al., 1992) or cutoff scores (Crawford & Henry, 2003; Crawford et al., 2001; Lee, Loring, & Martin, 1992; Mond et al., 2006; Murphy & Barkley, 1996; Posserud, Lundervold, & Gillberg, 2006; Van den Berg et al., 2009; Van Roy, Grøholt, Heyerdahl, & Clench-Aas, 2006; Wozencraft & Wagner, 1991).

If a small adaptation is made to the procedure by Oosterhuis et al. (in press), the second approach can be applied to *Z*-scores (Appendix), but this does not work for percentile ranks. Alternatively, the scores that correspond to the predetermined percentile ranks can be estimated with traditional quantile estimators, which are provided in *SPSS*

and *R* (available from the *quantile* function in *R*, *R* Core team, 2013), and the corresponding CIs can be based on the Marritz-Jarett SE estimator for quantiles (Wilcox, 2012, p 68.).

It is also possible to use the Harrell-Davis (H-D) quantile estimator (Harrel & Davis, 1982) to estimate scores corresponding to certain percentile ranks. The H-D estimator provides an exact bootstrap estimate based on a weighted linear combination of order statistics and gives the greatest weight to traditional nonparametric quantile estimators (Harrell & Davis, 1982). A great advantage of this estimator is that it is distribution-free and more efficient in small samples than the traditional quantile estimators that use only one or two order statistics. Furthermore, the H-D estimator and the traditional estimators are asymptotically equivalent. Exact bootstrap estimates of the SEs corresponding to the percentiles can be obtained using a jackknife variance estimator for quantiles (Hutson & Ernst, 2000), and can be used to construct CIs. The procedures to obtain quantile estimates and the corresponding SEs are available in the *R* package *Hmisc* (Harrell Jr., 2015).

Example. We used the approach by Oosterhuis et. al (in press) to estimate the standard deviation, stanine boundaries, percentile ranks and *Z*-scores for all 436 unique observed standardized residuals. The *R* syntax is available from the author. Table 5.1 shows the mean, standard deviation, and stanine boundaries, as well as their standard errors and 95% Wald-based CIs. Considering that a very large number of percentile ranks and *Z*-scores values was estimated, we used both remedies to reduce the number of unique percentile ranks and *Z*-scores to a smaller, more manageable number of norm values. First, we limited the number of percentiles and *Z*-scores that were presented. We decided to present only the standardized residuals that had percentile ranks closest to 50, 75, 90, 95, and 99 as well as the corresponding *Z*-values (Table 5.2). Table 2 also shows the 95% Wald-based CIs for the selected standardized residuals and corresponding percentile ranks and *Z*-scores.

The width of the CIs indicates the estimation precision of the norm statistics. For example, Table 5.2 shows that the standardized residual -0.097 corresponds to a percentile rank of 50.76 in the norm sample, but the CIs indicate that we are 95% sure that the true value of the percentile rank lies in the interval 45.9; 54.3. The smaller the width of the CI, the higher the estimation precision. If the intended use of a test is to make important decision about individuals' lives, the CI for norm statistics should be narrow. In this case,

estimation precision is not high enough to justify using the estimated norms for important decisions about individuals. If the test is used for less important decisions about individuals, the required precision of the norms is lower. In case of the DS-14, the intended use of the test is to find individuals who suffer from psychological risk factors for cardiac disease. Hence, the test is not used for important decisions and the required precision of the DS-14 norms is lower.

Table 5.1. Mean, Standard Deviation and Stanine Boundaries of the Standardized Residuals with Corresponding SEs and CIs.

Norm Statistic	PE	SE	95% CI
Mean	0.000	0.043	[-0.085; 0.085]
SD	1.000	0.027	[0.946; 1.052]
Stanine boundary			
1-2	-1.748	0.053	[-1.853; -1.644]
2-3	-1.249	0.046	[-1.339; -1.159]
3-4	-0.749	0.042	[-0.831; -0.668]
4-5	-0.250	0.042	[-0.331; -0.168]
5-6	0.250	0.046	[0.160; 0.340]
6-7	0.749	0.053	[0.645; 0.854]
7-8	1.249	0.063	[1.125; 1.372]
8-9	1.748	0.074	[1.604; 1.893]

Note. SD = standard deviation. PE = point estimate. SE = standard error. CI = confidence interval.

Table 5.2. Percentile Ranks and Z-Scores with Corresponding SEs and CIs.

	Percentile Rank			Z-score		
St. Residual	PE	SE	95% CI	PE	SE	95% CI
-0.097	50.76	2.166	[45.85; 54.34]	-0.10	0.043	[-0.18; -0.01]
0.654	75.09	1.873	[71.42; 78.76]	0.66	0.052	[0.55; 0.76]
1.370	89.94	1.301	[87.40; 92.49]	1.37	0.066	[1.24; 1.50]
1.742	95.02	0.939	[93.18; 96.86]	1.74	0.074	[1.60; 1.89]
3.049	99.91	0.094	[99.72; 100.09]	3.05	0.105	[2.85; 3.26]

Note. PE = point estimate. SE = standard error. CI = confidence interval.

Table 5.3. Standardized Residuals for Percentile Ranks and Z-Scores with Corresponding SEs and CIs.

PR	PE	SE	95% CI
50	-0.09	0.059	[-0.20; 0.02]
75	0.68	0.070	[0.54; 0.81]
90	1.37	0.064	[1.25; 1.50]
95	1.75	0.101	[1.25; 1.95]
99	2.50	0.185	[2.14; 2.87]
Z-score	**PE**	**SE**	**95% CI**
0	0.00	0.043	[-0.08; 0.08]
0.674	0.67	0.052	[0.57; 0.78]
1.282	1.28	0.064	[1.16; 1.41]
1.646	1.64	0.072	[1.50; 1.78]
2.326	2.32	0.087	[2.15; 2.50]

Note. PR = percentile rank. PE = point estimate. SE = standard error. CI = confidence interval.

We also adapted the procedure by Oosterhuis et al (in press) in order to estimate a limited number of predetermined *Z*-scores using the H-D quantile estimator. The predetermined percentile values were equal to 50, 75, 90, 95, and 99 and the predetermined *Z*-values were equal to 0, 0.674, 1.282, 1.645, and 1.96. These *Z*-values corresponded to the predetermined percentile values in a standard normal distribution. Table 5.3 shows the standardized residuals that correspond to the predetermined percentile ranks and *Z*-scores, as well as corresponding SEs and 95% CIs.

5.5 Interpretation and Presentation

Interpretation

After the norm statistics and the corresponding SEs and CIs have been estimated, they can be used to interpret the standardized residuals of individual test takers. To consider the estimation precision of the norms, the individual scores can be compared to the CIs of the norm statistics instead of the point estimates. If the score is not included in the CI, we infer that the individual score differs from the norm statistic.

Example. For a standardized residual equal to 1, the norm statistics in Table 5.1 inform us that this score is significantly higher than the mean, because 1 is not contained in the CI for the mean. Similarly, we conclude that this score is located in the 7th stanine,

because it is higher than the CI for the boundary between the 6th and 7th stanine and lower than the boundary between the 7th and 8th stanine. Furthermore, Table 5.2 shows that a standardized residual equal to 1 corresponds to a percentile rank between 71.4 (i.e., the lower boundary for the CI of a standardized residual is smaller than 1) and 92.5 (i.e., the upper boundary for the CI of a standardized residual is greater than 1). The CIs in Table 5.3 can be used to determine the standardized residual that corresponds to a particular percentile rank. For example, we can be reasonably sure that a standardized residual equal to 0.81 (i.e., the upper boundary for the CI of percentile rank 75) has a percentile rank that exceeds the 75th percentile, and that a standardized residual equal to 1.25 (i.e., the lower boundary for the CI of percentile rank 90) has a percentile rank lower than the 90th precentile.

Presentation

Usually, tests constructors provide the regression-based norm statistics in norm tables, similar to tables 5.1, 5.2, and 5.3. For the mean, standard deviation, and stanine boundaries, a norm table might be sufficient, because the number of statistics that have to be estimated is independent of norm sample size (e.g., there are always eight stanine boundaries regardless of sample size). However, for percentiles and *Z*-scores, norm tables might not be sufficient, because the number of statistics that have to be estimated is equal to the number of unique standardized residuals. The latter depends on norm sample size, which typically is large. As a result, norm tables have to be equally large to contain all the percentile ranks and *Z*-scores corresponding to the unique standardized residuals. As a solution, test constructors often present a limited number of statistics in the norms tables. For example, Kessels et al. (2014) provided the residual scores that correspond to 13 percentile ranks ranging from 2 to 98. Only presenting a limited number of the norm statistics limits the information available for the interpretation of individual test scores.

Furthermore, regression-based norming circumvents arbitrary categorization of continuous covariates, but continuous covariates are often categorized arbitrarily to be able to present the regression-based norms in tables. For example, Van Breukelen & Vlaeyen (2005) provided regression-based norm statistics for non-overlapping age categories containing ten years, and Conti et al. (2014) provided regression-based norm statistics for age categories containing five years. As a result, if the age of individual test

takers is close to the boundaries of these categories, the arbitrary categorization that is used might change the interpretation of their test scores. This means that presenting regression-based norms in norm tables might reintroduce the problem of arbitrary categorization, because only a limited number of norm statistics can be presented and continuous covariates are often arbitrarily categorized to achieve this.

To improve the comparison of test scores to regression-based norms and to avoid arbitrary selection of a limited number of norm statistics or arbitrary categorization of continuous covariates, test constructors are encouraged to provide a computer program and norm tables when presenting regression-based norms. For example, Crawford, Cullum, Garthwaite, Lycett, and Allsopp (2012) provided a computer program, which expresses a patient's raw score as a percentile rank. The advantage of a computer program over a norm table is that a computer program contains information for all unique residuals that were observed in the norm sample, whereas a norm table can only provide norm statistics for a limited number of raw scores or residuals. In addition to comparing the standardized residuals to the norm statistics, a computer program can also perform the computations that are necessary to obtain the standardized residual for an individual test taker's raw score, further facilitating the use of the test and its norms.

Example. The output from the procedure proposed by Oosterhuis et al (2016a) provides SEs and CIs for 436 percentile ranks and Z-scores. However, Table 5.2 only provides us with limited information, for example, that a standardized residual of 1 lies between the 75th and 90th percentile, which is a rather imprecise conclusion. It is possible to present a larger number of the norm statistics in the norm tables, but this increases the size of the tables and therefore might cause the norms be less comprehensible for test users. Alternatively, a computer program could be provided, which would provide the percentile rank and Z-score for the standardized residual in the norms sample that was closest to the value of 1. In this example, such a computer program would inform us that in the norm sample a standardized residual of 1 corresponded to a Z-score of 1, with SE = .058 and a 95% Wald-CI equal to [0.89; 1.12]. Similarly, the percentile rank for a standardized residual equal to 1 equaled 81.86, with SE = 1.67 and a 95% Wald-CI equal to [78.39; 85.13]. We conclude that we are relatively sure that the individual test taker who has a standardized residual equal to 1 had a Z-score between 0.89 and 1.12 and that, taking into

account age, he scored better than 78 to 85 % of the other test takers. This is more precise than the conclusion based on Table 2.5, namely that a standardized residual of 1 lies between the 75th and 90th percentile. Hence, compared to the use of only a norm table with a limited number of norm statistics, including all norm-sample data available into a computer program increases the precision of the conclusion regarding the score of individual test takers.

Epilogue

Regression-based norming has become increasingly popular in recent years as a solution to the norming problem that produces continuous test norms. I might speculate that the assumed efficiency of this method over the traditional method has not only been an added bonus for test constructors, but one of the primary incentives to use regression-based norming, given that norm estimation is both time consuming and costly. However, I think that many test constructors are unaware of the limitations of using a linear regression model to estimate norms, and perhaps understandably so because, in my opinion, little warning has been given about this method.

Similar to the traditional method of norm estimation, no well-founded sample size guidelines for precise regression-based norms exist. Although test constructors assume they require a smaller sample than in the traditional norming method, how much smaller is still largely unknown. In this dissertation, I have shown that the regression-based method is indeed more efficient than the traditional method of subdividing the norming sample into smaller groups based on relevant covariates. However, the greater estimation precision of regression-based norming was not clear-cut and depended on several factors, including the value of the norm statistic. As a result, I think that presenting minimum norm sample sizes for regression-based norming, similar to, for example, the guidelines provided by COTAN is not sufficient and can result in norms with less precision than expected.

To solve the problem of wishful thinking with respect to sample size, test constructors need a systematic method to investigate the precision of estimated norms that can replace rather arbitrary sample size guidelines. Hence, in this dissertation we derived standard errors (SEs) that quantify the estimation precision of norm statistics. These SEs, and corresponding confidence intervals (CIs) can be used to investigate whether the norms are precise enough for the intended use of the test. It is well known how measurement imprecision of individual scores can be quantified using, for example, the standard error of measurement, and this dissertation provides a method to quantify the sampling fluctuation of norm statistics. However, further research is required to combine the SEs or CIs when comparing individual scores to norms. Crawford and Howell (1998) have provided a modified *t*-test to compare an individual's score to the mean in the norm sample,

incorporating both measurement error and sampling fluctuation, but for other norm statistics no procedure is available yet.

I think that the most important limitation of regression-based norming is the assumptions of the linear regression model, which are pivotal for the validity of the resulting norms. These assumptions are often violated in test data, which renders the resulting regression-based norms useless. For example, data on an ability test might be skewed to the right for younger children (most children have similar, low scores, but few children perform exceptionally well), and skewed to the left for older children (most children have similar, high scores, but some children stay behind). Such a difference in distribution shape across age ranges is incompatible with the linear regression model. However, if test developers forget to mention or investigate the assumptions, the validity of the resulting regression-based norms remains disputable. To help test developers to investigate assumptions as well as norm estimation precision, I have provided a procedure in this dissertation that was based on a real-data example and includes these issues.

If one or more of the assumptions of the linear regression model is violated in the test data, the model is not suitable for norm estimation, and an alternative continuous solution to the norming solution is required. Several solutions were proposed recently. For example, Lenhard, Lenhard, Suggate, & Segerer (2016) described a model that estimates continuous norms without making assumptions about the distribution of the data. Another promising alternative might be the use of generalized additive models for location, scale, and shape (GAMLLS, originating from Rigby & Stasinopoulos, 2005), because they allow modeling of differences with respect to center, spread, skewness, and kurtosis as a function of continuous covariates such as age. Similarly, Sherwood, Zhou, Weintraub, and Wang (2016) have argued that quantile regression might be used to directly model conditional percentiles. I think further research into these models and other possible alternatives to regression-based norming is tantamount for the improvement of continuous norms.

Regression-based norming can be readily performed using any computer program that is capable of linear regression, such as SPSS and SAS. To promote the use of possible alternative models, I think it is important that the availability of these alternative models is improved. Furthermore, models such as the GAMLSS might be promising, but this method is more difficult than the linear regression model, which I expect might discourage the use

of these models. Hence, comprehensive procedures for these models are required, as well as clear and well-founded recommendations and requirements for their use in estimating norms. Institutions like COTAN (Dutch Committee on Testing) currently only provide a limited number of recommendations for regression-based norms (Evers, Lucassen, Meijer, & Sijtsma, 2009), and I think this should be expanded to alternative methods to estimate continuous norms. To conclude, although regression-based norming is an efficient and straightforward continuous solution to the norming problem, test developers should tread carefully when using this method to estimate norms, because its usefulness is heavily dependent on the compatibility of the linear regression model and the test data.

Summary

Everyday psychological tests and questionnaires are used to make important decisions in the lives of individuals, such as whether or not to hospitalize a mental patient, admit a prospective student to an educational program, or hire an applicant for a job. Usually, norms are estimated based on the raw scores of a group of people, known as the norm sample, who take the test during the test's construction phase. It is well known that a norm sample should be both representative of the population and large enough to guarantee norms having small standard errors. Representativeness refers to the composition of the sample, which should accurately reflect the subgroup structure of the population of interest. For example, covariates such as age group, urbanization grade, educational level, socioeconomic status, and religious affiliation each provide a different subgroup structure. Norms are estimated for each of the subgroups separately. However, from a practical angle, collecting large and representative samples for each norm subgroup is difficult, time consuming, and costly, which is often resolved by limiting the number of subgroups.

Zachary and Gorsuch (1985; also see Bechger, Maris, & Hemker, 2009; Van Breukelen & Vlaeyen, 2005) proposed a more efficient norming procedure, called continuous norming or regression-based norming. They proposed using covariates, such as age and gender, as independent variables in a linear regression model to predict the raw test score. They used the corresponding empirical distribution of standardized residuals to estimate the norms. Given the overwhelming importance of norms in practical test use, it is surprising that so little research has been done to adequately balance the precision of regression-based norms and the size of the norm sample. Hence, a systematic investigation of the technical properties and precision of regression-based norming is required.

This dissertation was dedicated to improving norms for psychological test and questionnaires, focusing on regression-based norms. In the first simulation study (Chapter 2), we compared the sample-size requirements for traditional and regression-based norming by examining the 95% interpercentile ranges for percentile estimates as a function of sample size, norming method, size of covariate effects on the test score, test length, and number of answer categories in an item. Using the graded response model

(GRM), we simulated norm sample data for tests containing either 10, 50, or 100 items and items with either 2 or 5 answering categories. The range and the mean of the parameter values of the GRM were based on parameter estimates obtained from real psychological test data. The sample sizes of the simulated data ranged from 100 to 10,000, and were based on a literature study. We also simulated a dichotomous covariate (denoted X_1) representing gender, and a continuous covariate (denoted X_2). The two covariates either had no effect, or a small, medium, or large effect on the attribute that the test measured. For each combination of the independent variables, we simulated 10,000 replications and in each replication we used a linear regression model containing both covariates to predict the simulated norm data. Based on the sample regression model, we estimated the standardized residuals that corresponded to percentile ranks 75, 90, 95 and 99.

Provided the assumptions of the linear regression model are consistent with the data, for a subdivision of the total group into eight equal-size subgroups we found that regression-based norming required 18% to 40% of the sample size required for traditional norming. We also found that as the percentiles were further away from the median, the difference between the two norming methods was smaller. For both norming approaches, we also found that the interpercentile range (IPR) grew larger as the estimated percentiles lay further away from the mean. In general, estimating the tails of a distribution requires larger samples. Thus, in order to choose a sample size, test constructors first need to decide which percentiles are important for the use of the test, because more extreme percentiles require larger samples.

For both norming methods, the estimation of the 99th percentile was more precise for polytomous than dichotomous items. The explanation may be that in the highest score range, polytomous items provide more score diversity than dichotomous items, resulting in narrower IPRs for the norm estimates relative to the total scale length. Regression-based norming uses the relationship between covariates and the test score to adjust the discrete test scores, which results in a non-discrete distribution of residuals enabling distinguishing different extreme scores. If dichotomous items must be used and the test contains a limited number of items, regression-based norming enables high precision and also enables distinguishing different high-scoring individuals. We found that the value of the multiple correlation between covariates and test score did not affect the precision of norm

estimation, because the covariates influenced the mean test score of the norm groups but not the distribution shape. Finally, we presented sample-size requirements for each norming method, for different test lengths, and numbers of answer categories.

A norm statistic obtained from a normative sample should be viewed as a point estimate of the norm in the population, which means the norm estimate should be accompanied by an indication of estimation precision. However, norm constructors commonly fail to provide SEs or CIs to quantify estimation precision of the norms, because for many norm statistics the SEs are unknown, difficult to derive, or if available not included in popular software. Hence, in the second study (Chapter 3), we derived standard errors for four norm statistics (standard deviation, percentile ranks, stanine boundaries and *Z*-scores) under the mild assumption that the test scores are multinomially distributed. We used a general framework consisting of two steps to compute the SEs of the norm statistics. The first step was to write the norm statistic as a function of the frequencies of the raw scores, using generalized exp-log notation. In the second step, we used the delta method to approximate the variance of the norm statistic. The resulting standard errors can be used to investigate the sampling error of norm statistics, for example, by constructing 95% confidence intervals.

We performed a simulation study to investigate the bias and precision of the SEs we derived and of the coverage of the corresponding 95% Wald-based CIs. We simulated data using the 2-parameter logistic model (2PLM) for tests containing 10, 30, or 50 items and sample sizes equal to 500, 1,000, 1,500, 2,000 or 2,500. We found that, for all practical purposes, the SEs of the standard deviation, the stanine boundaries, and the Z-scores were unbiased. We also found that Wald-based CIs based on the SEs of the standard deviation, the stanines and the *Z*-scores had good coverage. Except when percentile ranks were close to 0 or 100, SEs of the percentile ranks were unbiased. For percentile ranks greater than 2.5% and smaller than 97.5%, we found that small samples (i.e., $N < 1{,}000$) produced unbiased SE estimates and CIs with good coverage. If a larger sample (i.e., $N \geq 1{,}000$) was used to obtain unbiased estimates, SEs or CIs for more extreme percentile ranks (e.g., the 99-th percentile rank) were unbiased and had good coverage.

The third study (Chapter 4) investigated the bias and the precision of regression-based percentile estimates, and the coverage of corresponding confidence intervals (CIs)

when the assumptions of the linear regression model were violated. The assumptions of the linear regression model are often referred to as Gauss-Markov assumptions and if these assumptions are met, OLS estimators are the best linear unbiased estimators. This means that the estimators are both unbiased and efficient, which are desirable properties. One of the assumptions states that the expected value of the error term in the population model is equal to zero, but this assumption is inconsequential for norm estimates, because the non-zero mean of the error term is incorporated in the intercept of the sample regression model. This means that test constructors need not investigate this assumption. The regression model also assumes that observations in the norm sample are independent, but a violation of this assumption can be easily predicted based on the data generation process. For example, dependence can be caused by proximity in time (e.g., longitudinal data), or when data have been collected across spatial units. If the assumption of independence between observations is violated, the linear regression model is not suitable for norm estimation, but models are available that accommodate correlated errors. Furthermore, it is not necessary that test constructors investigate the assumption of normality of the errors, because for sample size > 50, the central limit theorem ensures that the regression model is robust against violations of this assumption.

The assumptions of linearity, independence between population error term and covariates, and homoscedasticity of the error variances are expected to result in biased norm estimates and CIs with low coverage. We simulated data using a linear regression model that either contained no assumption violation, or a violation of linearity, independence between covariates and error term, or homoscedasticity. The results of our simulation study showed that the strength of the violations of these assumptions, sample size, and value of the estimated percentiles influenced the bias and coverage of corresponding 95% CIs. In particular, stronger violations resulted in greater bias and lower coverage of corresponding CIs. If sample size increased, the coverage of CIs for biased estimates decreased. Furthermore, as percentile value increased, bias decreased and coverage of the corresponding CIs increased. Although assumption violations did not influence estimation precision, larger samples and higher percentile values resulted in estimates that were more precise. Based on the results of the simulation study, we advise test constructors to investigate assumption violations when estimating regression-based

norms, because assumption violations can cause substantial bias in both norm estimates and corresponding CIs.

In Chapter 5, using the Type-D Scale (DS-14) as an example, we provided a straightforward procedure for estimating regression-based norms for psychological tests and questionnaires. This procedure incorporates the conclusions from the previous chapters and consists of 4 steps: (1) selection of predictor variables, (2) checking of model assumptions, (3) estimating precision of the norm estimates, and (4) interpretation and presentation of the regression-based norms. The procedure we provided first shows how to select covariates based on the statistical significance of the relationship with the test score, and the consequences of considering the covariate for test score interpretation. Second, we showed how to investigate violations of the regression model, especially linearity/additivity, independence of the covariates and the error term, and homoscedasticity, because violations of these assumptions could lead to substantially biased norm estimates. Third, to quantify estimation precision, we provided a procedure to estimate standard errors and confidence intervals for regression-based norms. Finally, we discussed how regression-based norms can be presented and interpreted.

References

Aardoom, J. J., Dingemans, A. E., Slof-Op't Landt, M. C. T., & Van Furth, E. F. (2012). Norms and discriminative validity of the Eating Disorder Examination Questionnaire (EDE-Q). *Eating Behaviors*, *13*, 305-309. doi:10.1016/j.eatbeh.2012.09.002

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

American Psychological Association (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Agresti, A. (2012). *Analysis of ordinal categorical data* (2nd ed.). Hoboken, NJ: Wiley

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken, NJ: Wiley

Agresti, A. & Min, Y. (2001). On small-sample confidence intervals for parameters in discrete distributions. *Biometrics, 57*, 963,971. doi: 10.1111/j.0006-341X.2001.00963.x

Ahn, S., & Fessler, A. (2003). Standard errors of mean, variance, and standard deviation estimators. Technical Report. Ann Arbor, MI: EECS Department, University of Michigan: July 2003. http://www.eecs.umich.edu/~fessler/papers/files/tr/stderr.pdf.

Bechger, T., Hemker, B., & Maris, G. (2009). *Over het gebruik van continue normering* [On the use of continuous norming]. Arnhem, Netherlands: Cito. Retrieved from http://www.cito.nl/Onderzoek%20en%20wetenschap/achtergrondinformatie/publicaties/research_notes.aspx

Bergsma, W. P. (1997). *Marginal models for categorical data*. Tilburg, Netherlands: Tilburg University Press.

Bergsma, W. P., Croon, M. A., & Hagenaars, J. A. (2009). *Marginal models for dependent, clustered and longitudinal categorical data.* New York, NY: Springer.

Berrigan, L. I., Fisk, J. D., Walker, L. A. S., Wojtowicz, M., Rees, L. M., Freedman, M. S., & Marrie, R. (2014). Reliability of regression-based normative data for the Oral Symbol Digit Modalities Test: An evaluation of demographic influences, construct validity,

and impairment classification rates in multiple sclerosis samples. *The Clinical Neuropsychologist*, *28*, 281-299. doi:10.1080/13854046.2013.871337

Berry, W. D. (1993). *Understanding regression assumptions* (Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-092). Newbury Park, CA: Sage.

Berry, W. D., & Feldman, S. (1985). *Multiple regression in practice*. Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07-050. Newbury Park, CA: Sage.

Biederman, J. (2005). Attention-deficit/hyperactivity disorder: a selective overview. *Biological Psychiatry, 57,* 1215-20.

Birnbaum, A. (1968). Some latent trait models and their use in inferring any examinee's ability. In F.M. Lord & M.R. Novick, *Statistical theories of mental test scores* (pp. 395–479). Reading, PA: Addison-Wesley.

Brannigan, G. G., & Decker, S. L. (2003). *Bender Visual-Motor Gestalt Test–Second edition*. Itasca, IL: Riverside.

Brennan, R. L. & Lee, W.-C. (1999). Conditional scale-score standard errors of measurement under binomial and compound binomial assumptions. *Educational and Psychological Measurement, 56,* 5-24. doi: 10.1177/0013164499591001

Bride, B. E. (2007). Prevalence of secondary traumatic stress among social workers. *Social Work*, *52*, 63-70. doi:10.1093/sw/52.1.63

Casson, R. J., & Farmer, L. D. M. (2014). Understanding and checking the assumptions of linear regression: a primer for medical researchers. *Clinical & Experimental Ophthalmology, 42*, 590-596.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). Handbook for the Sixteen Personality Factor Questionnaire (16PF). Champaign, IL: Institute for Personality and Ability Testing

Cavaco, S., Goncalves, A., Pinto, C., Almeida, E., Gomes, F., Moreira, I., . . . Teixeira-Pinto, A. (2013a). Semantic fluency and phonemic fluency: Regression-based norms for the Portuguese population. *Archives of Clinical Neuropsychology*, *28*, 262-271. doi:10.1093/arclin/act001

Cavaco, S., Goncalves, A., Pinto, C., Almeida, E., Gomes, F., Moreira, I., . . . Teixeira-Pinto, A. (2013b). Trail making test: Regression-based norms for the Portuguese population. *Archives of Clinical Neuropsychology*, *28*, 189-198. doi:10.1093/arclin/acs115

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

Cooch, E. & White, G. (2015). *Program MARK: A gentle introduction* (14th ed.). Fort Collins, CO: Colorado State University.

Conti, S., Bonazzi, S., Laiacona, M., Masina, M., & Coralli, M. V. (2014). Montreal Cognitive Assessment (MoCA)-Italian version: Regression based norms and equivalent scores. *Neurological Sciences*, *36*, 209-214. doi:10.1007/s10072-014-1921-3

Crawford, J., Cayley, C., Lovibond, P. F., Wilson, P. H., & Hartley, C. (2011). Percentile norms and accompanying interval estimates from an Australian general adult population sample for self-report mood scales (BAI, BDI, CRSD, CESD, DASS, DASS-21, STAI-X, STAI-Y, SRDS, and SRAS). *Australian Psychologist, 46,* 3-14. doi:10.1111/j.1742-9544.2010.00003.x

Crawford, J. R., Cullum, C. M., Garthwaite, P. H., Lycett, E., & Allsopp, K. J. (2012). Point and interval estimates of percentile ranks for scores on the Texas Functional Living Scale. *Clinical Neuropsychology, 26,* 1154-1165.

Crawford, J. R., Garthwaite, P. H., & Slick, D. J. (2009). On percentile norms in neuropsychology: Proposed reporting standards and methods for quantifying the uncertainty over the percentile ranks of test scores. *The Clinical Neuropsychologist, 23,* 1173–1195.

Crawford, J. R., & Henry, J. D. (2003). The Depression Anxiety Stress Scales (DASS): Normative data and latent structure in a large non-clinical sample. *British Journal of Clinical Psychology, 42,* 111-131. http://dx.doi.org/10.1348/014466503321903544

Crawford, J. R., Henry, J. D., Crombie, C., & Taylor, E. P. (2001). Brief report: Normative data for the HADS from a large non-clinical sample. *British Journal of Clinical Psychology, 40,* 429-434. http://dx.doi.org/10.1348/014466501163904

Crawford, J. R. & Howell, D. C. (1998). Comparing an individual's test score against norms derived from small samples. *The Clinical Neuropsychologist, 12*, 482-486. doi: 10.1076/clin.12.4.482.7241

Denollet, J. (2005). DS-14: standard assessment of negative affectivity, social inhibition, and Type D personality. *Psychosomatic Medicine, 67*, 89-97.

Derksen, S. & Keselman, J. C. (1992). Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *British Journal of Mathematic and Statistical Psychology, 45,* 265-282.

Eccles, J. (1999). The development of children ages 6 to 14. *The Future of Children, 9,* 30-44.

Egberink, I. J. L., Janssen, N. A. M., & Vermeulen, C. S. M. (2014). *COTAN Documentatie* [COTAN Documentation]. Amsterdam: Boom. Retrieved September 29, 2014, from www.cotandocumentatie.nl

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

Evers, A., Lucassen, W., Meijer, R. R., & Sijtsma, K. (2009). *COTAN beoordelingssysteem voor de kwaliteit van tests* [*COTAN assessment system for the quality of tests*]*.* Amsterdam: Nederlands Instituut van Psychologen.

Fox, J. (1997). *Regression diagnostics: An introduction* (Vol. 79). Newbury Park, CA: Sage.

Frick, P. J., Barry, C. T., & Kamphaus, R. W. (2005). *Clinical assessment of child and adolescent personality and behavior* (3rd ed.)*.* New York, NY: Springer.

Gaub, M. & Carlson, C. L. (1997). Gender differences in ADHD: a meta-analysis and critical review. *Journal of the American Academy of Child and Adolescent Psychiatry, 36,* 1036-45.

Gershon, J. (2002). A meta-analytic review of gender differences in ADHD. *Journal of Attention Disorders, 5,* 143-54.

Glaesmer, H., Rief, W., Martin, A., Mewes, R., Brähler, E., Zenger, M., & Hinz, A. (2012). Psychometric properties and population-based norms of the Life Orientation Test Revised (LOT-R). *British Journal of Health Psychology, 17,* 432-445. http://dx.doi.org/10.1111/j.2044-8287.2011.02046.x

Goretti, B., Niccolai, C., Hakiki, B., Sturchio, A., Falautano, M., Eleonora, M., …, Amato, M. (2014). The Brief International Cognitive Assessment for Multiple Sclerosis (BICAMS): normative values with gender, age and education corrections in the Italian population. *BMC Neurology, 14,* 171-176. doi: 10.1186/s12883-014-0171-6

Grande, G., Romppel, M., Glaesmer, H., Petrowski, K., & Herrmann-Lingen, C. (2010). The type-D scale (DS14) – Norms and prevalence of type-D personality in a population-based representative sample in Germany. *Personality and Individual Differences, 48,* 935-939. http://dx.doi.org/10.1016/j.paid.2010.02.026

Grizzle, J. E., Starmer, C. F., & Koch, G. G. (1969). Analysis of categorical data for linear models. *Biometrics, 2*5*,* 489–504. doi: 10.2307/2528901

Hardy, M. A. (1993). *Regression with dummy variables.* Newbury Park, CA: Sage.

Harrell, F. E. & Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika, 69*, 635-640. doi: 10.1093/biomet/69.3.635

Harrell Jr., F. E. (2015). Hmisc: Harrell Miscellaneous. R package version 3.15-0 [computer software]. http://CRAN.R-project.org/package=Hmisc

Hayes, A. F., & Cai, L. (2007). Using heteroscedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods, 39*, 709-722.

Heaton, R. K., Avitable, N., Grant, I., & Matthews, C. G. (1999). Further crossvalidation of regression-based neuropsychological norms with an update for the Boston Naming Test. *Journal of Clinical and Experimental Neuropsychology, 21,* 572-582. http://dx.doi.org/10.1076/jcen.21.4.572.882

Henry, J. D., & Crawford, J. R. (2005). The short-form version of the Depression Anxiety Stress Scales (DASS-21): construct validity and normative data in a large non-clinical sample. *The British Journal of Clinical Psychology, 44,* 227-39.

Hunt, M., Auriemma, J., Cashaw, A. C. A. (2003). Self-report bias and underreporting of depression on the BDI-II. *Journal of Personality Assessment, 80,* 26-30. http://dx.doi.org/10.1207/S15327752JPA8001_10

Hutson, A. D. & Ernst, M. D. (2000). The exact bootstrap mean and variance of an L-estimator. *Journal of the Royal Statistical Society, 62*, 89-94.

Kendall, M. & Stuart, A. (1977). *The advanced theory of statistics, Vol 1: Distribution theory* (4th ed.). New York, NY: Macmillan.

Kessels, R. P., Montagne, B., Hendriks, A. W., Perrett, D. I., & De Haan, E. H. (2014). Assessment of perception of morphed facial expression using the Emotion

Recognition Task: normative data from healthy participants aged 8-75. *Journal of Neuropsyhology, 8,* 75-93. doi: 0.1111/jnp.12009.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry,62,* 593-602.

Kline, P. (2000). *Handbook of psychological testing* (2nd ed.). London: Routledge.

Krishnan, E., Sokka, T., Häkkinen, A., Hubert, H., & Hannonen, P. (2004). Normative values for the Health Assessment Questionnaire Disability Index. *Arthritis & Rheumatism, 50,* 953-960. http://dx.doi.org/10.1002/art.20048

Kritzer, H. M. (1977). Analyzing measures of association derived from contingency tables. *Sociological Methods and Research, 5*, 35-50. doi: 10.1177/004912417700500401

Kuijpers, R. E., Van der Ark, L. A., & Croon, M. A. (2013a). Standard errors and confidence intervals for scalability coefficients in Mokken scale analysis using marginal models. *Sociological Methodology, 43,* 42-69. doi: 10.1177/0081175013481958

Kuijpers, R. E., Van der Ark, L. A., & Croon, M. A. (2013b). Testing hypotheses involving Cronbach's alpha using marginal models. *British Journal of Mathematical and Statistical Psychology, 66*, 503-520. doi: 10.1111/bmsp.12010

Lang, J. B. (2008). Score and profile likelihood confidence intervals for contingency table parameters, *Statistics in Medicine, 27*, 5975 - 5990. doi: 10.1002/sim.3391

Larson, R. & Edwards, B. (20134). *Calculus* (10th ed.). Boston, MA: Cengage Learning, Brooks/Cole.

Lee, G. P., Loring, D. W., & Martin, R. C. (1992). Rey's 15-item visual memory test for the detection of malingering: Normative observations on patients with neurological disorders. *Psychological Assessment, 4,* 43-46. http://dx.doi.org/10.1037/1040-3590.4.1.43

Lee, W.-C., Brennan, R.L., & Kolen, M. J. (2000). Estimators of conditional scale-score standard errors of measurement: a simulation study. *Journal of Educational Measurement, 37,* 1-20. doi: 10.1111/j.1745-3984.2000.tb01073.x

Lehtonen, R., & Pahkinen, E. (2004). *Practical methods for design and analysis of complex surveys* (2nd ed.). West Sussex, England: Wiley.

Leigh, J. P. (1988). Assessing the importance of an independent variable in multiple regression: is stepwise unwise? *Journal of Clinical Epidemiology, 41,* 669-77.

Lenhard, A., Lenhard, W., Suggate, S., & Segerer, R. (2016). A continuous solution to the norming problem. *Assessment*. Advance online publication. doi: 10.1177/1073191116656437

Llinàs-Reglà, J., Vilalta-Franch, J., López-Pousa, S., Calvó-Perxas, L., & Garre-Olmo, J. (2013). Demographically adjusted norms for Catalan older adults on the Stroop Color and Word Test. *Archives of Clinical Neuropsychology, 28,* 282-96. http://dx.doi.org/10.1093/arclin/act003

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*(3), 217-224.

Lorge, I., & Thorndike, R. L. (1957). *The Lorge-Thorndike Intelligence Tests technical manual.* Boston, MA: Houghton Mifflin.

Merrell, K. W. (1994). *Preschool and Kindergarten Behavior Scales. Test manual.* Brandon, VT: Clinical Psychology Publishing Company.

Mertler, C. A. (2007). *Interpreting standardized test scores: Strategies for data-driven instructional decision making.* Thousand Oaks, CA: Sage.

Mond, J. M., Hay, P. J., Rodgers, B., & Owen, C. (2006). Eating Disorder Examination Questionnaire (EDE-Q): Norms for young adult women. *Behaviour Research and Therapy, 44,* 53-62. http://dx.doi.org/10.1016/j.brat.2004.12.003

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics* (3rd ed.). New York, NY: McGraw-Hill.

Murphy, K., & Barkley, R. A. (1996). Prevalence of DSM-IV symptoms of ADHD in adult licensed drivers: Implications for clinical diagnosis. *Journal of Attention Disorders, 1,* 147-161. http://dx.doi.org/10.1177/108705479600100303

Neter, J., Kutner, M. H., Nachtsheim, C., J., & Wasserman, W. (1996). *Applied linear statistical models.* Chicago, IL: Irwin.

Oosterhuis, H. E. M., Sijtsma, K. & Van der Ark, L. A. (2016a). The effect of assumption violations on regression-based norms. *Manuscript under review.*

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

Oosterhuis, H. E. M., Van der Ark, L. A., & Sijtsma, K. (in press). Standard errors and confidence intervals of norm statistics for educational and psychological tests. *Psychometrika.*

Palomo, R., Casals-Coll, M., Sánchez-Benavides, G., Quintana, M., Manero, R. M., Rognoni, T., …, Peña-Casanova, J. (2011). Spanish normative studies in young adults (NEURONORMA young adults project): Norms for the Rey-Osterrieth Complex Figure (copy and memory) and Free and Cued Selective Reminding Test. *Neurologiá, 28,* 226-235. doi: 10.1016/j.nrl.2012.03.008

Parmenter, B. A., Testa, S. M., Schretlen, D. J., Weinstock-Guttman, B., & Benedict, R. H. (2010). The utility of regression-based norms in interpreting the minimal assessment of cognitive function in multiple sclerosis (MACFIMS). *Journal of the International Neuropsychological Society, 16,* 6-16. http://dx.doi.org/10.1017/S1355617709990750

Pedraza, O, Lucas, J. A., Smith, G. E., Petersen, R. C., Graff-Radford, N. R., & Ivnik, R. J. (2010). Robust and expanded norms for the Dementia Rating Scale. *Archives of Clinical Neuropsychology, 25,* 347-358. http://dx.doi.org/10.1093/arclin/acq030

Posserud, M.-B., Lundervold, A. J., & Gillberg, C. (2006). Autistic features in a total population of 7-9 year old children assessed by the ASSQ (Autism Spectrum Screening Questionnaire). *Journal of Child Psychology and Psychiatry, 47,* 167-175. http://dx.doi.org/10.1111/j.1469-7610.2005.01462.x

Rao, R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York, NY: Wiley.

R Core Team (2016). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, Retrieved from http://www.R-project.org/.

Rigby, R. A., & Stasinopoulus, D. M. (2005). Generalized additive models for location, scale and shape. *Applied Statistics, 54,* 507-554. doi: 10.1111/j.1467-9876.2005.00510.x

Roelofs, J., Braet, C., Rood, L., Timbremont, B., Van Vlierberghe, L., Goossens, L., & Van Breukelen, G. J. P. (2013a). Norms and screening utility of the Dutch version of the children's depression inventory in clinical and nonclinical youths. *Psychological Assessment, 22,* 866-877. http://dx.doi.org/10.1037/a0020593

Roelofs, J., Van Breukelen, G. J. P. , De Graaf, E., Beck, A. T., Arntz, A., & Huibers, M. J. H. (2013b). Norms for the Beck Depression Inventory (BDI-II) in a large Dutch community sample. *Journal of Psychopathology and Behavioral Assessment, 35,* 93-98. http://dx.doi.org/10.1007/s10862-012-9309-2

Samejima, F. (1969). *Estimation of Latent Ability Using a Response Pattern of Graded Scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved, September 10, 2013, from http://www.psychometrika.org/journal/online/MN17.pdf

Sartorio, F., Bravini, E., Vercelli, S., Ferriero, G., Plebani, G., Foti, C., & Franchignoni, F. (2013). The Functional Dexterity Test: Test-retest reliability analysis and up-to-date reference norms. *Journal of Hand Therapy, 26,* 62-68. doi: 10.1016/j.jht.2012.08.001

Selya, A. S., Rose, J. S., Dierker, L. C., Hedeker, D., & Mermelstein, R. J. (2012). A practical guide to calculating cohen's f^2, a mixed measure of local effect size, from PROC MIXED. *Frontiers in Psychology, 3, 111.*

Semel, E., Wiig, E. H., & Secord, W. A. (2004). *Clinical evaluation of language fundamentals, fourth edition—Screening test (CELF-4 screening test)*. Toronto, Canada: The Psychological Corporation/A Harcourt Assessment Company.

Sherwood, B., Zhou, A. X., Weintraub, S., & Wang, L. (2016). Using quantile regression to create baseline norms for neuropsychological tests. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 2,* 12-18.

Shi, J., Wei, M., Tian, J., Snowden, J., Zhang, X., Ni, J., …, & Wang, Y. (2014). The Chinese version of story recall: a useful screening tool for mild cognitive impairment and Alzheimer's disease in the elderly. *BMC Psychiatry, 14*, 71-81. http://dx.doi.org/10.1186/1471-244X-14-71

Skogli, E. W., Teicher, M. H., Andersen, P. N., Hovik, K.T & Øie, M., (2013). *BMC Psychiatry, 13,* 298.

Smerbeck, A. M., Parrish, J., Yeh, E. A., Hoogs, M., Krupp, L. B., Weinstock-Guttman, B., & Benedict, R. H. B. (2011). Regression-based pediatric norms for the Brief Visuospatial

Memory Test- Revised and the Symbol Digit Modalities Test. *The Clinical Neuropsychologist, 25*, 402-412. http://dx.doi.org/10.1080/13854046.2011.554445

Smerbeck, A. M., Parrish, J., Yeh, E. A., Weinstock-Gutmann, B. W., Hoogs, M., Serafin, D., ..., & Benedict, R. H. B. (2012). Regression-based norms improve the sensitivity of the National MS Society Consensus Neuropsychological Battery for Pediactric Multiple Sclerosis (NBPMS). *The Clinical Neuropsychologist, 26,* 985-1002. http://dx.doi.org/10.1080/13854046.2012.704074

Tabachnick, B. G., & Fidell, L. S. (2012). *Using Multivariate Statistics: International Edition (6th ed.)*. Boston: Pearson.

Tellegen, P., & Laros, J. A. (2011). *SON-R 6-40: Snijders-Oomen Niet-Verbale Intelligentietest: Deel I Verantwoording* [Snijders-Oomen non-verbal intelligence test: Part I Motivation.]. Amsterdam: Hogrefe.

Van Belle, G. (2003). *Statistical rules of thumb* (2nd ed.). Hoboken, NJ: Wiley.

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

Van den Berg, E., Nys, G. M. S., Brands, C. R., Ruis, C., Van Zandvoort, M. J. E., & Kessels, R. P. (2009). The Brixton Spatial Anticipation Test as a test for executive function: Validity in patient groups and norms for older adults. *Journal of the International Neuropsychological Society, 15,* 695-703. http://dx.doi.org/10.1017/S1355617709990269

Van der Ark, L. A. (2012). New developments in Mokken Scale Analysis in R. *Journal of Statistical Software, 48,* 1-27. doi: 10.18637/jss.v048.i05

Van der Ark, L. A., Croon, M. A., & Sijtsma, K. (2008). Mokken scale analysis for dichotomous items using marginal models. *Psychometrika, 73*, 183-208. doi: 10.1007/s11336-007-9034-z.

Van der Elst, W., Dekker, S., Hurks, P., & Jolles, J. (2012). The Letter Digit Substitution Test: Demographic influences and regression-based normative data for school-aged children. *Archives of Clinical Neuropsychology, 27,* 433-439. http://dx.doi.org/10.1093/arclin/acs045

Van der Elst, W., Hoogenhout, E.M., Dixon, R. A., De Groot, R. H. M., & Jolles, J. (2011). The Dutch Memory Compensation Questionnaire: Psychometric properties and regression-based norms. *Assessment, 18,* 517-528. http://dx.doi.org/10.1177/1073191110370116

Van der Elst, W., Ouwehand, C., Van der Werf, G., Kuyper, H., Lee, N., & Jolles, J. (2012). The Amsterdam Executive Function Inventory (AEFI): Psychometric properties and demographically corrected normative data for adolescents aged between 15 and 18 years. *Journal of Clinical and Experimental Neuropsychology, 34,* 160-171. http://dx.doi.org/10.1080/13803395.2011.625353

Van der Elst, W., Ouwehand, C., Van Rijn, P., Lee, N., Van Boxtel, M., & Jolles, J. (2013). The Shortened Raven Standard Progressive Matrices: item response theory-based psychometric analyses and normative data. *Assessment, 20*, 48-59. http://dx.doi.org/10.1177/1073191111415999

Van der Linden, W. J. & Hambleton, R. K. (1997). *Handbook of modern Item Response Theory*. New York, NY: Springer.

Van Roy, B., Grøholt, B., Heyerdahl, S., & Clench-Aas, J. (2006). Self-reported strengths and difficulties in a large Norwegian population 10-19 years. *European Child & Adolescent Psychiatry, 15,* 189-198. http://dx.doi.org/10.1007/s00787-005-0521-4

Vlahou, C. H., Kosmidis, M. H., Dardagani, A., Tsotsi, S., Giannakou, M., Giazkoulidou, A., …, Pontikakis, N. (2013). Development of the Greek Learning Test: reliability, construct validity, and normative standards. *Archives of Clinical Neuropsychology, 28,* 52-64. http://dx.doi.org/10.1093/arclin/acs099

Von Davier, A. A. (2011). *Statistical models for test equating, scaling, and linking.* New York, NY: Springer.

Warlaumont, A. S., Richards, J. A., Gilkerson, J., & Oller, D. K. (2014). A social feedback loop for speech development and its reduction in autism. *Psychological Science, 25,* 1314-24. doi: 10.1177/0956797614531023

Wechsler, D. (2005). *Wechsler Individual Achievement Test 2nd Edition (WIAT II).* London: The Psychological Corp.

Wilkinson G. S. (1993). *The Wide Range Achievement Test 3rd Edition: Manual*. Wilmington, DE: Wide Range.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing*. Waltham, MA: Academic Press.

Wizniter, M., Verhulst, F. C., Van den Brink, W., Van der Ende, J., Giel, R., & Koot, H. M. (1992). Detecting psychopathology in young adults: The Young Adult Self Report, the General Health Questionnaire and the Symptom Checklist as screening instruments. *Acta Psychiatrica Scandinavica 86,* 32-37. http://dx.doi.org/10.1111/j.1600-0447.1992.tb03221.x

Wozencraft, T., & Wagner, W. (1991). Depression and suicidal ideation in sexually abused children. *Child Abuse & Neglect, 15,* 505-511. http://dx.doi.org/10.1016/0145-2134(91)90034-B

Yang, L., Unverzagt, F. W., Jin, Y., Hendrie, H. C., Liang, C., Hall, K. S., … Gao, S. (2012). Normative data for neuropsychological tests in a rural elderly Chinese cohort. *The Clinical Neuropsychologist, 26,* 641-653. http://dx.doi.org/10.1080/13854046.2012.666266

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

Zigmond, A. S., & Snaith, R. P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica, 67,* 361-370. http://dx.doi.org/10.1111/j.1600-0447.1983.tb09716.x