UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

# Sample Size Requirements for Traditional and Regression-Based Norms

Oosterhuis, H.E.M.; van der Ark, L.A.; Sijtsma, K.

**DOI**
10.1177/1073191115580638

**Publication date**
2016

**Document Version**
Final published version

**Published in**
Assessment

Link to publication

**Citation for published version (APA):**
Oosterhuis, H. E. M., van der Ark, L. A., & Sijtsma, K. (2016). Sample Size Requirements for Traditional and Regression-Based Norms. *Assessment*, *23*(2), 191-202. https://doi.org/10.1177/1073191115580638

**General rights**
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

**Disclaimer/Complaints regulations**
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

*UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)*

Download date:02 Dec 2021

# Sample Size Requirements for Traditional and Regression-Based Norms

Assessment
2016, Vol. 23(2) 191–202
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1073191115580638
asm.sagepub.com
SAGE

**Hannah E. M. Oosterhuis[1], L. Andries van der Ark[2], and Klaas Sijtsma[1]**

**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.

**Keywords**

minimum sample size requirements for norms, norm distribution of test scores, precise test norms, regression-based norming, traditional norming

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 she or 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 (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, Hurks, & 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 with 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, pp. 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 with traditional norming, regression-based norming is expected to require a smaller sample to obtain equally precise norms (Bechger et al., 2009).

[1]Tilburg University, Tilburg, Netherlands
[2]University of Amsterdam, Amsterdam, Netherlands

**Corresponding Author:**
Hannah E. M. Oosterhuis, Department of Methodology and Statistics, Tilburg University, PO Box 90153, 5000 LE, Tilburg, Netherlands.
Email: h.e.m.oosterhuis@tilburguniversity.edu

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 payoff 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.

## 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.

### *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: (a) statistical significance, (b) effect-size assessment, (c) statistical significance and effect-size assessment, and (d) 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-square test can be used to select covariates (Aardoom, Dingemans, Slof Op't Landt, & Van Furth, 2012; Mond, Hay, Rodgers, & Owen, 2006; Pedraza et al., 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 (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 (Scheier, Carver, & Bridges, 1994). They only selected covariates that were statistically significant (ANOVA) and had at least a medium effect size (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 Hospital Anxiety and Depression Scale corresponding to each of the raw scores test takers can acquire. Also, refer to the Wechsler Individual Achievement Test–Third edition (Wechsler, 2009), the Wide Range Achievement Test–Third edition (Wilkinson, 1993), and the Bender Visual-Motor Gestalt Test–Second edition (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, 2010). 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.

## 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: (a) stepwise regression, (b) simultaneous regression, (c) correlational analysis, and (d) 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, $\alpha$ is the significance level), the predictor having the greatest $p$ value is deleted from the model. Third, the model including the remaining predictors is reestimated. Fourth, in the new model the predictor having the highest $p$ value greater than $\alpha$ 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 socioeconomic status 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. The authors performed several regression analyses using the Memory Compensation Questionnaire 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, and Garre-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*. Correlational 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 with 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 this 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 $X_1,...,X_K$ represent the 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 $\beta_0$ 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,

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

(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).

### 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). 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 the following: 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.

## 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 assessed between 2008 and 2012 so as to derive realistic approximations to the number of items, and so on. We used freely accessible test reviews from the Dutch Committee on Tests and Testing 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 ($M$ = 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), three or four ordered scores (26.2%), five ordered scores (23.3%), or more than five 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 $\theta$. Latent variable $\theta$ determined test score $Y_+$; see Figure 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

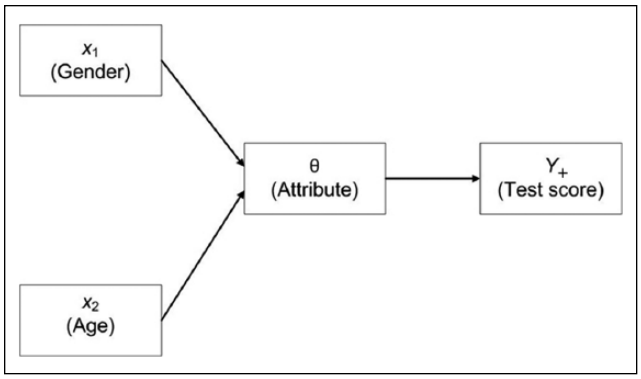


**Figure 1.** Population model for simulating test score ($Y_+$) based on latent variable ($\theta$) and covariates ($X_1, X_2$).

probability $p = .5$. Scores on $X_2$ were randomly sampled from the uniform distribution on the interval [4, 12].

Second, for each participant, a $\theta$ 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 $\theta$ depends on covariates $X_1$ and $X_2$. The regression parameters $\beta_0$, $\beta_1$, and $\beta_2$ were chosen such that the squared multiple correlation ($R^2$) between $\theta$ 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 $\theta$, respectively (Cohen, 1992; $.02 \le R^2 < .13$ is small, $.13 \le R^2 < .26$ is medium, and $R^2 \ge .26$ is large). The covariates were uncorrelated and explained an equal portion of the variance of $\theta$. 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 were discrete; hence, the resulting test scores were also discrete and had 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 $\alpha_j$ denote the discrimination parameter of item $j$, and let $\lambda_{jy}$ denote the location parameter of score $y$ of item $j$. The GRM is defined as

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

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

**Table 1.** Graded Response Model Parameters for Dichotomous and Polytomous Items.

| | | Dichotomous | Polytomous | | | |
|---|---|---|---|---|---|---|
| Item | $\alpha$ | $\lambda$ | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ |
| 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 |

Table 1 shows the values for item parameters $\alpha_j$ and $\lambda_{jy}$. The range and the mean of the values were based on parameter estimates obtained from real psychological test data (Embretson & Reise, 2000). 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, and so on. 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 $Y_+ = \sum_{j=1}^{J} Y_j$.

## 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 $\theta$ 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.** Summary of Simulated Test Scores ($N$ = 1,000).

| Population model | | | Test scores | | |
|---|---|---|---|---|---|
| Number of items | Item scores | $R^2$ | *M* | *SD* | Coefficient 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 |

Table 2 shows coefficient alpha (e.g., Cronbach, 1951) for each combination of test length, number of item scores, and size of covariate effect.

## 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 cutoff 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: $4 \le X_2 < 6$ (first category), $6 \le X_2 < 8$ (second category), and so on. Given that scores 0 and 1 on $X_1$ had equal probabilities and scores on $X_2$ were drawn 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 $6 \le X_2 < 8$ (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 ($Z_{Y_+} = \left(Y_+ - \overline{Y}_+\right) / S_{Y_+}$) 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 (*Z*-scores).

## 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 ($\eta^2$) 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 $\eta^2$ for the effect equals

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

**Table 3.** Effect Sizes ($\eta^2$) Based on ANOVAs Performed on IPR of Percentiles.

| | Percentiles | | | | |
|---|---|---|---|---|---|
| | 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.
*$p < .05$. **$p < .01$.

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 four times better than that of traditional norming.

## Results

### Analyses of Variance

For the 50th, 75th, 90th, 95th, and 99th percentile, Table 3 shows the effect size ($\eta^2$) corresponding to the main effects and the interaction effects on the IPRs.

*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 ($\eta^2 < .01$; see 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 to 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 to 6 show a similar relationship.

For the 50th, 75th, 90th, and 95th percentiles (see figures 2 to 5), no effects other than norming method were

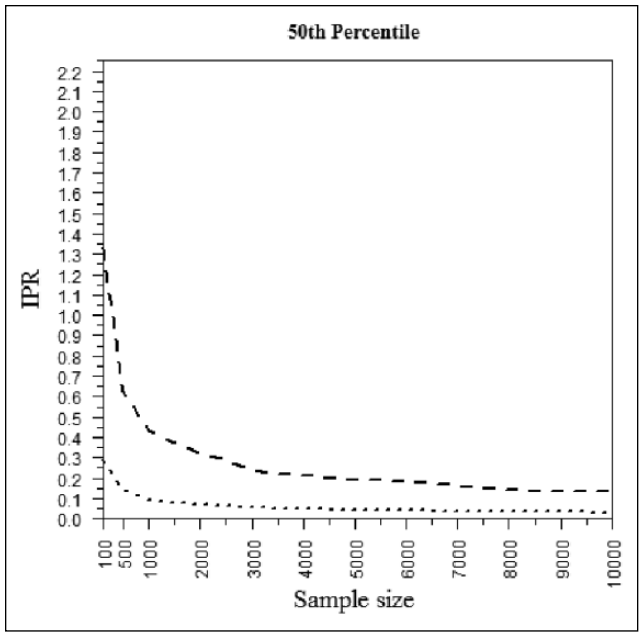


**Figure 2.** Interpercentile range for the 50th percentile estimate.
*Note.* IPR = interpercentile range; traditional norming (dashed) and regression-based norming (dotted).

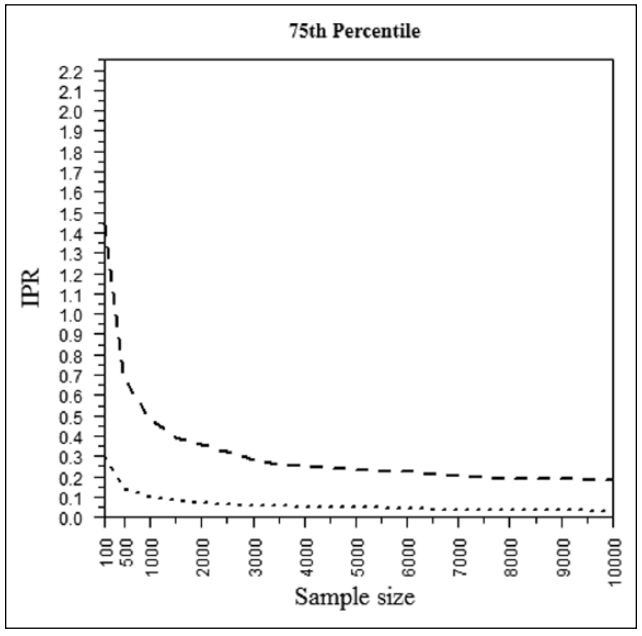


**Figure 3.** Interpercentile range for the 75th percentile estimate.
*Note.* IPR = interpercentile range; traditional norming (dashed) and regression-based norming (dotted).

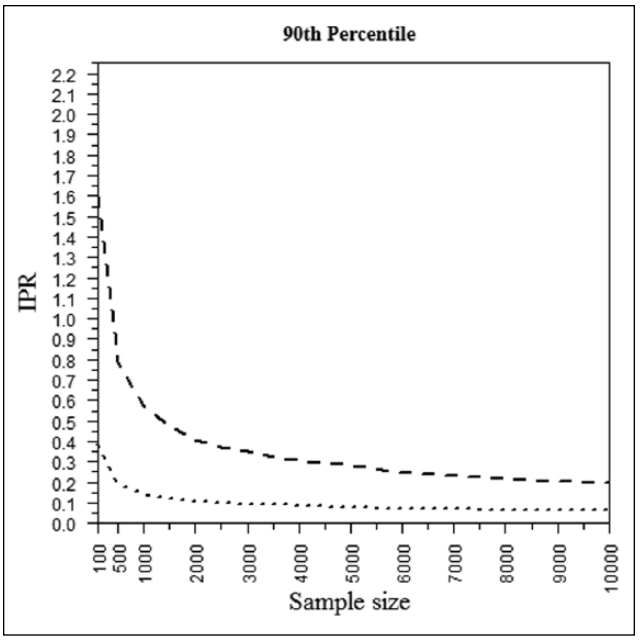


**Figure 4.** Interpercentile range for the 90th percentile estimate.
*Note.* IPR = interpercentile range; traditional norming (dashed) and regression-based norming (dotted).

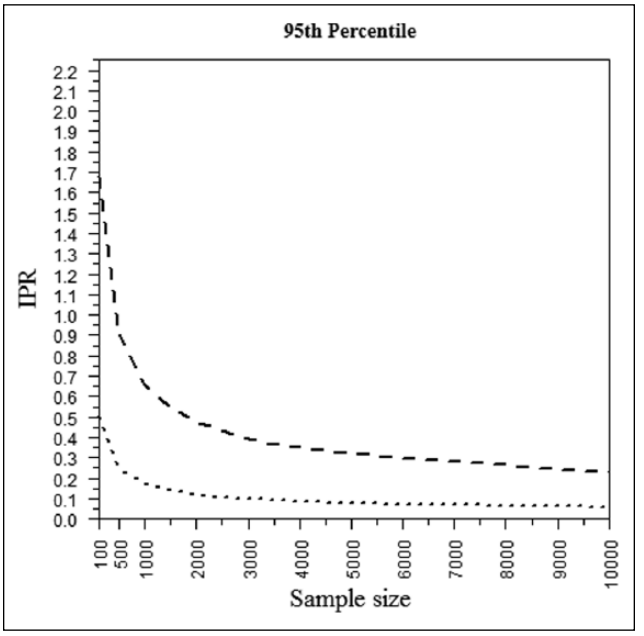


**Figure 5.** Interpercentile range for the 95th percentile estimate.
*Note.* IPR = interpercentile range; traditional norming (dashed) and regression-based norming (dotted).

included, resulting in two curves. For the 99th percentile (see Figure 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.

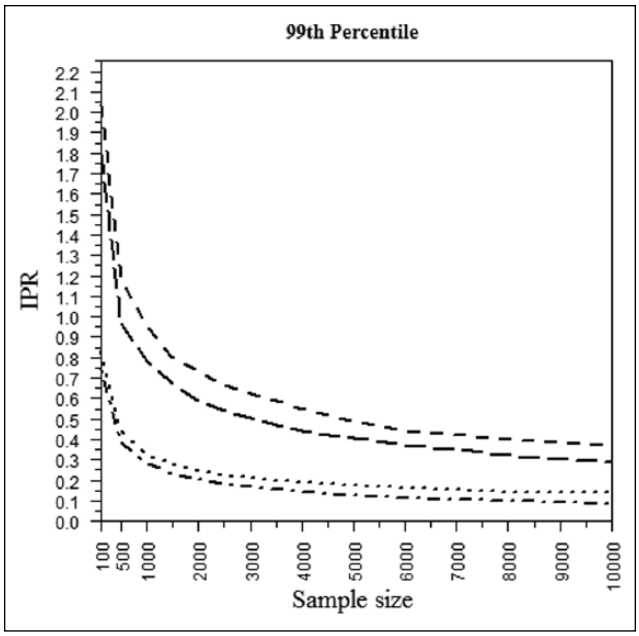


**Figure 6.** Interpercentile range for the 99th percentile estimate.
*Note.* IPR = interpercentile range; traditional norming with dichotomous items (dashed), traditional norming with polytomous items (long dashed), regression-based norming with dichotomous items (dotted), and regression-based norming with polytomous items (dotted-dashed).

### *IPR Ratio of Traditional Norming Versus Regression-Based Norming*

Table 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.

## 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.

**Table 4.** Summary of Ratio Between IPR of Traditional and Regression-Based Norming for Given $N$.

| | IPR ratio | | |
|---|---|---|---|
| Percentile | Min. | Max. | *M* (*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) |

*Note.* IPR = interpercentile range.

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 0.1 standard deviations. In our study, for a 50-item dichotomous test, 0.1 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) 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 11 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 nondiscrete 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 $\theta$, 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 on request from the first author.

Further research may investigate the effect of failure of the assumptions of the regression model, which are heteroscedasticity, nonlinearity, and nonnormality, 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.

## Declaration of Conflicting Interests

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

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: For this research Hannah E. M. Oosterhuis received funding from the Netherlands Organisation for Scientific Research, Grant 406-12-013.

## 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.

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

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

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

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

Cavaco, S., Gonçalves, 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., Gonçalves, 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

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, CES-D, 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., & 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. doi: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. doi:10.1348/014466501163904

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

Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, *45*, 265-282. doi: 10.1111/j.2044-8317.1992.tb00992.x

Egberink, I. J. L., Janssen, N. A. M., & Vermeulen, C. S. M. (2014). *COTAN Documentatie* [COTAN Documentation]. Amsterdam, Netherlands: Boom. Retrieved from www.cotan-documentatie.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, Netherlands: Nederlands Instituut van Psychologen.

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. doi: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. doi:10.1016/j.paid.2010.02.026

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

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. doi:10.1076/jcen.21.4.572.882

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. doi:10.1207/S15327752JPA8001_10

Jolles, J., Houx, P. J., Van Boxtel, M. P. J., & Ponds, R. W. H. M. (1995). *Maastricht Aging Study: Determinants of cognitive aging. Maastricht, the Netherlands: Neuropsych*. doi: 10.1093/geronb/60.1.P57

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*, 282-296. doi:10.1093/arclin/act003

Kessels, R. P. C, Montagne, B., Hendriks, A. W., Perrett, D. I., & De Haan, E. H. F. (2014). Assessment of perception of morphed facial expressions using the Emotion Recognition Task: Normative data from healthy participants aged 8-75. *Journal of Neuropsychology*, *8*, 75-93. doi:10.1111/jnp.12009

Kline, P. (2000). *Handbook of psychological testing* (2nd ed.). London, England: 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. doi:10.1002/art.20048

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. doi:10.1037/1040-3590.4.1.43

Leigh, J. P. (1988). Assessing the importance of an independent variable in multiple regression: Is stepwise unwise? *Journal of Clinical Epidemiology*, *41*, 669-677. doi: 10.1016/0895-4356(88)90119-9%20

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. doi: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. doi:10.1177/108705479600100303

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. doi: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. doi: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. doi:10.1111/j.1469-7610.2005.01462.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. doi: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. doi: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.

Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A re-evaluation of the Life Orientation Test. *Journal of Personality and Social Psychology*, *67*, 1063-1078. doi: 10.1037/0022-3514.67.6.1063

Semel, E., Wiig, E. H., & Secord, W. A. (2004). *Clinical evaluation of language fundamentals, fourth edition—Screening test (CELF-4 screening test)*. Toronto, Ontario, Canada: Psychological Corporation.

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. doi:10.1186/1471-244X-14-71

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. doi: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. doi:10.1080/13854046.2012.704074

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics: International edition* (6th ed.). Boston, MA: 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, the Netherlands: Hogrefe.

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. doi: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. doi:10.1017/S1355617709990269

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. doi: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. doi: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. doi: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. doi:10.1177/1073191111415999

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. doi: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. doi:10.1093/arclin/acs099

Wechsler, D. (2009). Wechsler Individual Achievement Test–Third edition (WIAT III). San Antonio, TX: Psychological Corporation.

Wilkinson, G. S. (1993). *The Wide Range Achievement Test–Third edition: Manual*. Wilmington, DE: Wide Range.

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. doi: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. doi: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. doi: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.

Zigmond, A. S., & Snaith, R. P. (1983). The Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinavica*, *67*, 361-370. doi:10.1111/j.1600-0447.1983.tb09716.x