Sample size calculation and optimal design for regression-based norming of tests and questionnaires

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

© 2021 American Psychological Association
ISSN: 1082-989X

2023, Vol. 28, No. 1, 89–106
https://doi.org/10.1037/met0000394

Sample Size Calculation and Optimal Design for Regression-Based Norming of Tests and Questionnaires

Francesco Innocenti[1], Frans E. S. Tan[1], Math J. J. M. Candel[1], and Gerard J. P. van Breukelen[1, 2]

[1] Department of Methodology and Statistics, Care and Public Health Research Institute (CAPHRI), Maastricht University
[2] Department of Methodology and Statistics, Graduate School of Psychology and Neuroscience, Maastricht University

Abstract

To prevent mistakes in psychological assessment, the precision of test norms is important. This can be achieved by drawing a large normative sample and using regression-based norming. Based on that norming method, a procedure for sample size planning to make inference on Z-scores and percentile rank scores is proposed. Sampling variance formulas for these norm statistics are derived and used to obtain the optimal design, that is, the optimal predictor distribution, for the normative sample, thereby maximizing precision of estimation. This is done under five regression models with a quantitative and a categorical predictor, differing in whether they allow for interaction and nonlinearity. Efficient robust designs are given in case of uncertainty about the regression model. Furthermore, formulas are provided to compute the normative sample size such that individuals' positions relative to the derived norms can be assessed with prespecified power and precision.

Translational Abstract

Normative studies are needed to derive reference values (or norms) for tests and questionnaires, so that psychologists can use them to assess individuals. Specifically, norms allow psychologists to interpret individuals' score on a test by comparing it with the scores of their peers (e.g., individuals with the same sex, age, and educational level) in the reference population. Because norms are also used to make decisions on individuals, such as the assignment to clinical treatment or remedial teaching, it is important that norms are precise (i.e., not strongly affected by sampling error in the sample on which the norms are based). This article shows how this goal can be attained in three steps. First, norms are derived using the regression-based approach, which is more efficient than the traditional approach of splitting the sample into subgroups based on demographic factors and deriving norms per subgroup. Specifically, the regression-based approach allows researchers to identify the predictors (e.g., demographic factors) that affect the test score of interest, and to use the whole sample to derive norms. Second, the design of the normative study (e.g., which age groups to include) is chosen such that the precision of the norms is maximized for a given total sample size for norming. Third, this total sample size is computed such that a prespecified power and precision are obtained.

Keywords: normative data, optimal design, percentile rank score, sample size calculation, Z-score

Supplemental materials: https://doi.org/10.1037/met0000394.supp

Normative studies provide reference values, also known as norms, that psychologists can use to compare individuals with the reference population, for instance, to make decisions about clinical treatments, school admission or remedial teaching, or selection of candidates for job vacancies. Examples of normative studies are Goretti et al. (2014) and Parmenter et al. (2010), who have derived reference values for two batteries of neuropsychological tests to assess cognitive function in patients with multiple sclerosis, and Van der Elst et al. (2006), who have normed the Dutch version of three verbal fluency tests. Normative studies are of practical importance because they allow psychologists to interpret scores on the outcome variable of interest by comparing an individual's test score with the scores of his or her peers (e.g., individuals of the same age, sex, and educational level) in the reference population. For instance, knowing that a highly educated 75-year-old woman scored 11.5 on the profession naming verbal fluency test is in itself not informative on whether this score is within the normal range or exceptional. According to the normative data provided by Van der Elst et al., (2006, Table A.2), only 10% of her peers (i.e., women of the same age and educational level) have a test score equal

This article was published Online First August 12, 2021.

Francesco Innocenti https://orcid.org/0000-0001-6113-8992

Math J. J. M. Candel https://orcid.org/0000-0002-2229-1131

Gerard J. P. van Breukelen https://orcid.org/0000-0003-0949-0272

We have no conflict of interest to disclose. A summary of this study was presented at the 41st Annual Conference of the International Society for Clinical Biostatistics.

Correspondence concerning this article should be addressed to Francesco Innocenti, Department of Methodology and Statistics, Care and Public Health Research Institute (CAPHRI), Maastricht University, P.O. Box 616, 6200 MD, Maastricht, the Netherlands. Email: francesco.innocenti@maastrichtuniversity.nl

to or lower than 11.5, which indicates that her access to semantic memory is well below the average and might lead her psychologist to perform further tests, given that a poor performance on verbal fluency tests has been associated with Alzheimer's disease (Van der Elst, et al., 2006). This example shows that it is crucial to have precise norms to prevent mistakes in psychological assessment.

There are two approaches to norming: the traditional approach and the regression-based approach. The traditional approach consists of first splitting the sample drawn for norming into subgroups based on some relevant demographic factors like age and sex, and then computing the norm statistics of interest within each subgroup. Instead, in the regression-based approach, first, a regression of the test score on some relevant predictors is performed, and then norm statistics are estimated from the cumulative distribution of the standardized residuals obtained from the model (Oosterhuis et al., 2016; Van Breukelen & Vlaeyen, 2005). The regression-based approach has several advantages. First, it uses the whole sample to establish norms instead of norming per subgroup, thereby increasing the precision of the norms, that is, reducing the role of sampling error in estimating the norms. Second, it allows researchers to identify which independent variables (e.g., demographic factors) affect the test score, thereby increasing the validity of the norms. Third, under the (testable) assumption of a specific regression model for relating the test score to relevant predictors such as age and sex, it is possible to express the sampling error of the norm statistic of interest as a function of the joint distribution of the predictors, for instance the age distribution per sex, and the sex distribution, in the normative sample. Subsequently, one can then find that joint distribution of the predictors that minimizes this sampling error and thus maximizes the precision of the norms under the assumed regression model. This joint distribution will be called the optimal design for the normative study. In contrast, the traditional approach forces the researcher to choose between two evils: establishing a single set of norms for the whole population, thus ignoring effects of demographic factors, or establishing a separate set of norms per subgroup as defined by demographic factors, by splitting the normative sample into subgroups and thus reducing the sample size and precision of norming. For these reasons, the regression-based approach is adopted here. The validity of the regression model must be tested, of course.

This article provides the optimal design for estimating Z-scores and percentile rank scores under linear regression models that include a quantitative predictor (e.g., age) and a categorical predictor (e.g., sex), for which the residual errors (i.e., the differences between observed and predicted test scores) are assumed to be normally distributed and homoscedastic, possibly after a suitable transformation of the test score. Furthermore, it will be shown how to compute the total sample size for the normative study using the optimal design, and sample size requirements will be provided for the most relevant Z-scores and percentile rank scores used in practice.

To the best of the authors' knowledge, there is only one other article, Oosterhuis et al. (2016), that provides sample size requirements for normative studies. Specifically, Oosterhuis et al. (2016) give sample size guidelines for percentile estimation under both traditional and regression-based norming, and show that regression-based norming requires smaller sample sizes than traditional norming. However, these sample size requirements are based on a simulation study, and no equations are given in Oosterhuis et al. (2016) that can be used, for instance, to derive the optimal design. Oosterhuis et al. (2017) do provide variance formulas for several norm statistics, but predictors are not considered in the norming, and then no optimal design can be obtained from their formulas. In contrast, the variance formulas given in the present article are obtained under a regression model that includes predictors, thus allowing the derivation of the optimal design.

This article is organized as follows. First, the considered models are introduced, and first-order Taylor series approximations of the variances of an estimated Z-score and an estimated percentile rank score are derived. Second, the optimal designs (i.e., the optimal joint distribution of the predictors) for the considered models are shown, and it is discussed how to deal with uncertainty about the "true" model at the design stage. Third, a procedure is proposed to determine the required total sample size for the optimal design, given a desired power level for hypothesis testing, or a desired precision level for interval estimation. Fourth, the results of this article are illustrated with a real-life example. Finally, some concluding remarks are made. Online Supplement A presents the results of a literature review on regression-based normative studies, and of two simulation studies that assessed the bias of the variance approximations given in the next section and of their estimators. Furthermore, online Supplement A gives the derivations of the optimal and robust designs. Online Supplement B provides the R codes (R Core Team, 2019), and additional results of the simulation studies.

Models for Norming and Variances of the Norm Statistics

Models for Norming and Norm Statistics

A sample of N individuals is drawn from the reference population. This is called the normative sample. The normative sample allows researchers to identify which variables influence scores on the outcome variable of interest, to estimate the unknown model parameters, and to derive reference values, also known as norms. Once the norms are available, practitioners can use them to compare individuals' (e.g., patients', or students', or job applicants') scores with the reference population.

Let Y_i be the score on the outcome variable (e.g., score on a verbal fluency test) of individual i ($i = 1, \ldots, N$). Denote by X_1 a quantitative variable (e.g., age), and X_2 a categorical variable coded 0/1 (e.g., sex). The following multiple regression models can then be considered in order of increasing complexity and flexibility:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \tag{1}$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}^2 + \varepsilon_i, \tag{2}$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_4 X_{1i} X_{2i} + \varepsilon_i, \tag{3}$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}^2 + \beta_4 X_{1i} X_{2i} + \varepsilon_i, \tag{4}$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}^2 + \beta_4 X_{1i} X_{2i} + \beta_5 X_{1i}^2 X_{2i} + \varepsilon_i. \tag{5}$$

In all models (1–5) it is assumed that $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, that is, normality and homoscedasticity are assumed throughout the article.

In Model 3, the regression coefficient of the interaction between X_1 and X_2 is called β_4, instead of β_3, in order to unequivocally identify each regression parameter with a certain predictor throughout the text. Concerning the assumption of a linear or quadratic effect of X_1, note that any relation between Y and X_1 that can be described as part of a (inverted) U-shape, including a monotonic relation with accelerating or decelerating slope, can be fitted with a quadratic model.

Concerning the assumptions of normality and homoscedasticity, it is relevant to mention that in a literature review of 65 regression-based normative studies, involving 396 psychological tests, both normality and homoscedasticity appeared to be satisfied by conventional diagnostics in 71% of the models for which checks on these assumptions were reported (online Supplement A, Table S.A.1). Furthermore, the most common predictors were age (present in 88% of the models used for norming), sex (45%), and education (81%) (online Supplement A). In a literature review of 65 tests, Oosterhuis et al. (2016) have also found that these three predictors were the most relevant in defining norms, but age was used in 36.2% of the tests, sex in 33.3%, and educational level/job position in 30.4%. Nevertheless, education was not included in Models 1–5, because it would have been unfeasible to treat all possible models obtained from taking into account all possible degrees of interaction between the three predictors and the fact that education can be treated either as a categorical (i.e., low, medium, high level of education) or a quantitative variable (i.e., number of years in school). However, the variance formulas given in this section depend neither on the number of predictors, nor on their scale types.

In the notation, summarized in Table 1, it is important to make a distinction between, on the one hand, the normative sample from the reference population, from which the norms are derived, and, on the other hand, the individual (e.g., patient, student, job applicant) to whom the norms are applied.

Normative Sample

The five considered models for the reference population are all standard regression equations and can be expressed in the matrix form $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{y}$ is the $N \times 1$ vector of scores on the outcome variable, $\boldsymbol{X}$ is the $N \times (k+1)$ design matrix, $\boldsymbol{\beta}$ is the $(k+1) \times 1$ vector of regression coefficients, and $\boldsymbol{\varepsilon}$ is the $N \times 1$ vector of residual errors such that $\boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{X\beta}) \sim N(\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I}_N)$, where N is the number of individuals in the normative sample. The OLS estimators of the unknown parameters $\boldsymbol{\beta}$ and σ_ε^2 are, respectively, $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \sim N(\boldsymbol{\beta}, \sigma_\varepsilon^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$ and $\hat{\sigma}_\varepsilon^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{N-k-1}$ with $(N-k-1)\hat{\sigma}_\varepsilon^2 \sim \sigma_\varepsilon^2 \chi_{N-k-1}^2$ (Johnson & Wichern, 1998, pp. 389–390), so that $E(\hat{\sigma}_\varepsilon^2) = \sigma_\varepsilon^2$ and $V(\hat{\sigma}_\varepsilon^2) = \sigma_\varepsilon^4 \frac{2}{N-k-1}$, where $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

Individual to Whom the Norms Are Applied

Having estimated the unknown model parameters $\boldsymbol{\beta}$ and σ_ε^2, one can use these estimates to compare an individual with the reference population. This is done by converting the individual's outcome value into a Z-score. Denote by $\boldsymbol{x}_0$ the $(k+1) \times 1$ vector containing the individual's scores on the predictors. Let $Y_0 = \boldsymbol{x}_0'\boldsymbol{\beta} + \varepsilon_0$ be the observed individual's score on the outcome, and ε_0 the individual's residual error such that $\varepsilon_0 \sim N(0, \sigma_\varepsilon^2)$. The individual's Z-score is then defined as $Z_0 = \frac{\varepsilon_0}{\sigma_\varepsilon} = \sigma_\varepsilon^{-1}(Y_0 - \boldsymbol{x}_0'\boldsymbol{\beta})$ and is estimated by $\hat{Z}_0 = \frac{\hat{\varepsilon}_0}{\hat{\sigma}_\varepsilon} = \hat{\sigma}_\varepsilon^{-1}(Y_0 - \hat{Y}_0) = \hat{\sigma}_\varepsilon^{-1}(Y_0 - \boldsymbol{x}_0'\hat{\beta})$, where $\hat{\sigma}_\varepsilon$ and $\hat{\boldsymbol{\beta}}$ have been obtained from the normative sample. Under normality and homoscedasticity, this $\hat{Z}_0$ tells us how many standard deviations the individual's test score is below or above the average in the reference population, adjusted for the predictors in the model.

Alternatively, one can compare the individual's test score Y_0 with the reference population by computing the percentile rank score PR_0 corresponding to the individual's $\hat{Z}_0$, that is, the percentage of individuals in the reference population with a Z-score equal to or lower than that $\hat{Z}_0$. Under the assumptions of normality and homoscedasticity, the percentile rank score can be estimated by:

$$PR(\hat{Z}_0) = \Phi(\hat{Z}_0) \times 100, \qquad (6)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. In the sequel, the subscript "0" will be used for symbols referring to the individual to whom the norms are applied (as opposed to symbols for the normative sample).

Variances of the Norm Statistics

The sampling variances of $\hat{Z}_0$ and $PR(\hat{Z}_0)$ are derived using the Delta method (Casella & Berger, 2002, p. 245). Oosterhuis et al. (2017) have also derived variance formulas for Z-scores and PR-scores using the Delta method. However, the results presented here

Table 1
Section in Which a Symbol is Used for the First Time

Section	Symbols
Models for norming and norm statistics	N, Y, X_1, X_2, β_0, β_1, β_2, β_3, β_4, β_5, ε, σ_ε^2, $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$, $\boldsymbol{\varepsilon}$, k, $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_\varepsilon^2$, $\hat{\boldsymbol{\varepsilon}}$, $\hat{\boldsymbol{y}}$, Y_0, $\boldsymbol{x}_0$, ε_0, Z_0, $\hat{Z}_0$, $\hat{\varepsilon}_0$, $\hat{Y}_0$, PR_0, $PR(\hat{Z}_0)$, Φ
Variances of the norm statistics	$V(\hat{Z}_0)$, $V(PR(\hat{Z}_0))$, ϕ, $\hat{V}(\hat{Z}_0)$, $\hat{V}(PR(\hat{Z}_0))$, ξ
Optimal and robust design	w, ξ^*, $d(\boldsymbol{X}, \xi)$, ξ_1^*, ξ_2^*, Q_2, w_1^*, w_2^*, w^*, $RE(\xi\ vs\ \xi^*)$
Sample size calculation	$V(\hat{Z}_0)^*$, $V(PR(\hat{Z}_0))^*$, Z_c, PR_c, Z_t, PR_t, δ, N^*, α, $1-\gamma$, $z_{1-\alpha}$, $z_{1-\alpha/2}$, $z_{1-\gamma}$, Z_{PR_c}, Z_{PR_t}, Φ^{-1}, τ

Note. Symbols with subscript "0" refer to the individual to whom the norms are applied (i.e., not a member of the normative sample).

differ from those of Oosterhuis et al., (2017) in two ways. First, Oosterhuis et al., (2017) did not consider predictors in the norming, thereby restricting application of their results to subgroups based on those predictors (the traditional norming approach), while here the regression-based approach is adopted by assuming Models 1–5 for norming. Second, here the residual error ε_i is assumed to be normally distributed, possibly after data transformation, whereas in Oosterhuis et al., (2017) the raw test score Y_i was assumed to follow a multinomial distribution because each Y_i value was treated as a possible outcome of a trial, over N independent trials. However, scores on a test are not necessarily integer values (see, for instance, Van der Elst, et al., 2006, Tables A.1–A.3), and treating test scores as continuous outcomes is a standard approach, as evidenced by the 65 regression-based normative studies reviewed in online Supplement A. As mentioned in the introduction, adopting the regression-based approach makes it possible not only to derive norms based on the whole sample yet adjusted for predictors, but also to derive the joint distribution of the predictors that maximizes the precision of norm statistics estimation, that is, the optimal design.

The sampling variances of $\hat{Z}_0$ and $PR(\hat{Z}_0)$ arise from the sampling error in $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2_\varepsilon$ in the normative sample, and not from measurement error in the outcome of the individual to whom the norms will be applied, so conditioning on Y_0. Measurement error in Y_0 itself is beyond the scope of this article, which is about the optimal design for the normative study (i.e., the optimal joint distribution of the predictors in the norming model). Note, however, that measurement error can be reduced by measuring the outcome repeatedly and averaging per individual in the normative sample as well as in applying norms to individuals. For $\hat{Z}_0$, the sampling variance is (for proofs, see Appendix A):

$$V(\hat{Z}_0) \approx \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0 + \frac{1}{2(N-k-1)}Z_0^2. \tag{7}$$

For $PR(\hat{Z}_0)$, the sampling variance is (for proofs, see Appendix A):

$$V(PR(\hat{Z}_0)) \approx 100^2\phi(Z_0)^2 V(\hat{Z}_0) = 100^2\phi(Z_0)^2 \times\left[\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0 + \frac{1}{2(N-k-1)}Z_0^2\right], \tag{8}$$

where $\phi(\cdot)$ is the probability density function of the standard normal distribution. Equations 7 and 8 can be estimated as follows

$$\hat{V}(\hat{Z}_0) \approx \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0 + \frac{1}{2(N-k-1)}\hat{Z}_0^2, \tag{9}$$

and

$$\hat{V}(PR(\hat{Z}_0)) \approx (100\times\phi(\hat{Z}_0))^2 \times\left[\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0 + \frac{1}{2(N-k-1)}\hat{Z}_0^2\right], \tag{10}$$

which differ from Equations 7 and 8 in replacing the unknown true Z_0 with the estimator $\hat{Z}_0$.

Equations 7 and 8, as well as Equations 9 and 10, are based on first-order Taylor series approximations, so their bias must be assessed. This was done through two simulation studies in which Equations 7–10 were compared with the true $V(\hat{Z}_0)$ and $V(PR(\hat{Z}_0))$, for which no analytical expressions were known, hence the need for simulations of true values. The design and the results of these simulation studies are thoroughly discussed in online Supplement A. A summary of the results of the simulation studies is given in Appendix B, and additional results are given in online Supplement B. In the first simulation study, the relative bias of Equations 7 and 8 for $V(\hat{Z}_0)$ and $V(PR(\hat{Z}_0))$, respectively, and the absolute bias of $\hat{Z}_0$ and Equation 6 for $PR(\hat{Z}_0)$ were assessed. In the second simulation study, the relative bias of Equations 9 and 10 for $\hat{V}(\hat{Z}_0)$ and $\hat{V}(PR(\hat{Z}_0))$, respectively, and the coverage of the 95% confidence interval obtained using Equations 9 and 10 were assessed. The smallest sample size considered in the simulation studies was $N = 338$. From the two simulation studies the following practical recommendations can be given:

- Equation 7 for $V(\hat{Z}_0)$ and Equation 9 for $\hat{V}(\hat{Z}_0)$ are accurate (i.e., relative bias $\in (-3\%, 3\%)$) even for small sample sizes such as $N = 338$. Likewise, the 95% confidence interval for Z_0, obtained using Equation 9, has good coverage (i.e., coverage $= 95\% \pm 0.5\%$) for $N \geq 338$.
- When the target of inference is *PR*, the sample size should be $N \geq 1{,}690$ in order to guarantee acceptable bias in Equation 6 for $PR(\hat{Z}_0)$ (i.e., absolute bias $\in [-0.1, 0.1]$), in Equation 8 for $V(PR(\hat{Z}_0))$ (i.e., relative bias $\in [-5\%, 3\%]$), and in Equation 10 for $\hat{V}(PR(\hat{Z}_0))$ (i.e., relative bias $\in [-3\%, 5\%]$), and good coverage (i.e., coverage $= 95\% \pm 1\%$) across all considered models (i.e., Equations 1–5). Under Model 1, the sample size for PR-scores can be $N \geq 676$, but not for the other models with the relative bias of Equations 8 and 10 exceeding 10%, and the coverage approaching 93%.

These lower-bounds for the sample size (i.e., 338 for Z-scores and 1,690 for PR-scores) are in line with typical sample sizes for normative studies (see Figure S.A.1, online Supplement A, and Oosterhuis et al., 2016).

In study planning, one wants to determine the required sample size to achieve the desired precision level for estimating (or power level for testing hypotheses on) a true Z-score or PR-score of interest. Hence, in sample size calculation one should use Equations 7–8 instead of Equations 9–10, because the former equations are functions of the true Z_0, while the latter equations are functions of the estimate $\hat{Z}_0$, which is available only after the study has been carried out. Clearly, taking N as large as feasible minimizes Equations 7 and 8, but is a waste of resources, when the desired precision of norm estimation can be obtained with a smaller sample size. The sample size can be minimized even further by the design ξ (i.e., the joint distribution of the predictors in the regression model for norming) of the normative sample. Such optimal designs will be presented for Models 1–5 in the next section. Next, in the Sample Size Calculation section, two approaches are proposed to determine the required size for the normative sample given an optimal design. The first approach ensures a desired power level for hypothesis testing, while the second approach ensures a desired margin of error for confidence interval estimation.

Optimal and Robust Design

Designs for Optimizing Precision of Z-Score and PR-Score Estimation

Theory

In this section, the optimal design that maximizes the precision of Z-score and PR-score estimation is presented. A design ξ is defined as a joint distribution of the predictors in the normative sample, given N. Each possible combination of the levels of the predictors (e.g., a 50-year-old male, or a 30-year-old female), is called a support point of ξ, and the proportion of the total N allocated to a support point is called design weight (w). The optimal design ξ^* is then defined as that joint distribution of the predictors (or equivalently, as that distribution of the total sample over the support points) that maximizes precision of Z-score and PR-score estimation, that is, that minimizes Equations 7 and 8 given N. This optimization is done over the design region, that is, over the set of all possible support points. In this optimization, $\frac{1}{2(N-k-1)}Z_0^2$ in Equations 7 and 8, and $100^2\phi(Z_0)^2$ in Equation 8 are kept fixed, because these terms do not depend on the normative sample (apart from its size N), but on the true Z_0 of the individual (e.g., patient, student, job applicant) to whom the norms are applied. A safe approach is to minimize Equations 7 and 8 for the worst-case scenario, that is, for that set of values of the predictors $\boldsymbol{x}_0$ for which $V(\hat{Z}_0)$ and $V(PR(\hat{Z}_0))$ are maximum, given N and Z_0. This in turn comes down to finding the D-optimal design, that is, to minimizing the determinant of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ (Atkinson et al., 2007; Berger & Wong, 2009; Goos & Jones, 2011). To see this, note that $\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0$, in Equations 7 and 8, is proportional to the standardized prediction variance for a given design ξ, $d(\boldsymbol{X},\xi) = N\sigma_\varepsilon^{-2}V(\hat{Y}_0) = N\sigma_\varepsilon^{-2}V(\boldsymbol{x}_0'\hat{\beta}) = N\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0$ (Atkinson et al., 2007, p. 55, Equation 5.32). Thus, the minimization of the maximum $V(\hat{Z}_0)$ and $V(PR(\hat{Z}_0))$ over the design region can be done by minimizing the maximum $d(\boldsymbol{X},\xi)$ over the design region, which is known as G-optimality (Atkinson et al., 2007; Berger & Wong, 2009; Goos & Jones, 2011). Under homoscedasticity of the residual (ε) distribution, the equivalence theorem states that G-optimality is equivalent to D-optimality. Hence, the D-optimality criterion will be used here, and the obtained designs will be both D- and G-optimal. The equivalence theorem also states that a design ξ is D-/G- optimal if and only if $d(\boldsymbol{X},\xi) = (k+1)$ at its support points (i.e., at each combination of predictor values included in that design) and $d(\boldsymbol{X},\xi) < (k+1)$ over the rest of the design region (Wong, 1995), where $(k+1)$ is the number of regression coefficients in the model. This latter result can be used to check whether a design is D-optimal by plotting the standardized prediction variance $d(\boldsymbol{X},\xi)$ over the design region, and will be used in the Sample Size Calculation section to derive an expression for $V(\hat{Z}_0)$ and $V(PR(\hat{Z}_0))$ under the optimal design.

In order to derive the optimal design analytically, it is helpful to categorize multiple regression models, such as Equations 1–5, based on the degree of interaction between predictors (Schwabe, 1996). There are three possible degrees of interaction and thus three types of models: complete interaction (that is, all possible interactions between predictors are included in the model, like in Models 3 and 5), no interaction (that is, Models 1 and 2), and partial interaction (i.e., not all possible interactions are present, like in Model 4). However, for all types of models the D-optimal design (ξ^*) is obtained by combining the optimal designs (ξ_1^* and ξ_2^*) for the marginal models, where the marginal models are the model with just X_1 (and possibly X_1^2) and the model with just X_2 as given in Table 2 (for details, see online Supplement A, pp. 37–39).

Results

Table 2 gives (from left to right): the full model (first column), the marginal models (second column), the optimal designs for the marginal models ξ_1^* and ξ_2^* (third and fourth columns), and the D-optimal design ξ^* for the full model (fifth column). All designs ξ_1^* and ξ_2^* in Table 2 (third and fourth columns) are D-optimal, with the only exception of ξ_1^* under the partial interaction Model 4, for technical reasons explained in online Supplement A (pp. 37–39). In Table 2, the range of values for the quantitative predictor X_1 (e.g., age) is rescaled to the interval $[-1, 1]$ (to emphasize that the results in Table 2 are valid for any age range, e.g., 20–80 or 50–90 years, where the range is chosen by the researcher, e.g., the age range for which the test to be normed is intended), and the categorical predictor X_2 (e.g., sex) is coded as 0/1. This does not affect the results in Table 2, because D-optimality is invariant with respect to linear transformations of the scale of the predictors (Atkinson et al., 2007, p. 152; Berger & Wong, 2009, p. 40; Schwabe, 1996, p. 22). A graphical illustration of the designs in Table 2 (third, fourth, and fifth columns) is given in Figure 1, where design weights are represented by dot size. As Figure 1 shows, the D-optimal design ξ^* in Table 2 consists of replicating the optimal design for the quantitative predictor X_1 (Table 2, third column) at each level of the categorical predictor X_2, which has only two levels in Figure 1, but the results in Table 2 hold for any number of levels of X_2 (e.g., for three levels if X_2 is education: low, medium, high). Two main results can be seen in Table 2 and Figure 1 First, the D-optimal design ξ^* has two age levels and these are at the boundaries of the age range, when the marginal model for age includes only a linear effect, and it has three equidistant age levels when a quadratic effect is present in the marginal model for age (Table 2, second and fifth column). Second, the D-optimal design ξ^* is balanced, that is, each support point of ξ^* (i.e., age-sex combination, such as $X_1 = -1$ and $X_2 = 1$) has the same design weight (i.e., the same sample size). An exception to the latter result is the D-optimal design for Model 4 that gives more weight to age levels -1 and 1 than it does to age level 0. This can be explained by noting that Model 4 combines Models 2 and 3, and that the D-optimal design under Model 4 is then a compromise between the D-optimal designs under Models 2 and 3. Indeed, this design gives equal weight to $X_1 = -1$ and $X_1 = +1$, which are needed to estimate the linear age effect (like for all models) as well as the interaction effect (like for Model 3), and a smaller weight to $X_1 = 0$, which is needed only to estimate the quadratic age effect (like for Model 2).

Including in a normative sample only two or three age levels over the whole age range, as suggested by Table 2, may sound counterintuitive, particularly if one compares this approach with the traditional norming that categorizes age into equidistant groups over the whole age range of interest (e.g., 20–30, 30–40, 50–60,

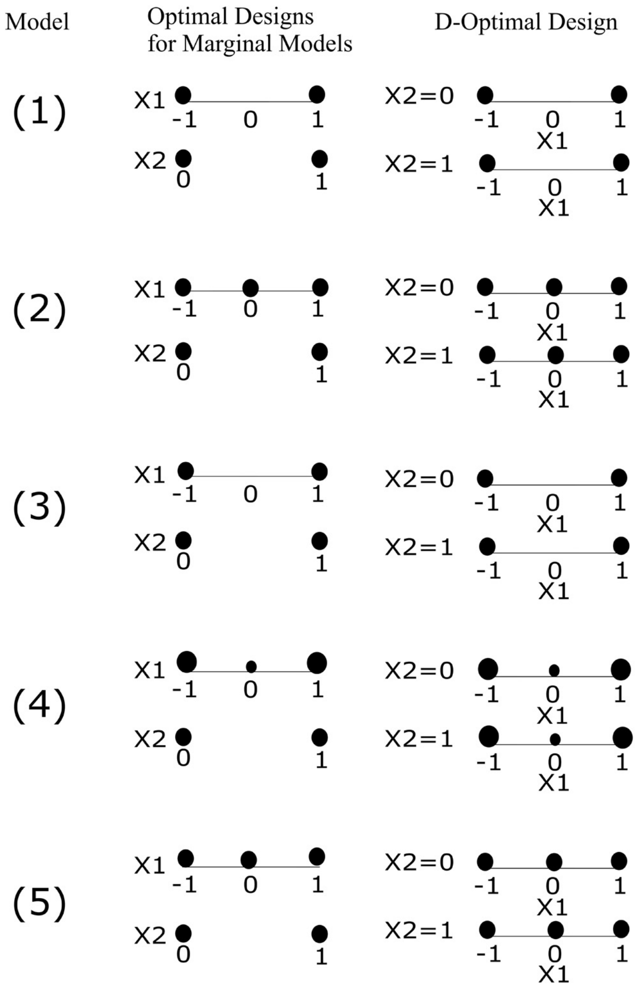
Table 2

D-Optimal Designs That Maximize Precision of Z-Score and PR-Score Estimation Under Models 1–5

Full model	Marginal models for X_1 (e.g., age) and X_2 (e.g., sex)	Optimal design for X_1 (ξ_1^*) on the interval $[-1, 1]$	Optimal design for X_2 (ξ_2^*) with Q_2 levels (e.g., $Q_2 = 2$ for sex)	D-optimal design (ξ^*) for the full model
Model 1	$E_1(Y) = \beta_0 + \beta_1 X_1$ for X_1, $E_2(Y) = \beta_2 X_2$ for X_2	Equal weight $w_1^* = \frac{1}{2}$ to -1 and 1 of X_1	Equal weight $w_2^* = \frac{1}{Q_2} = \frac{1}{2}$ to each level of X_2	Equal weight $w^* = w_1^* \times w_2^* = \frac{1}{2}\frac{1}{Q_2} = \frac{1}{4}$ to -1 and 1 of X_1 for each level of X_2
Model 2	$E_1(Y) = \beta_0 + \beta_1 X_1 + \beta_3 X_1^2$ for X_1, $E_2(Y) = \beta_2 X_2$ for X_2	Equal weight $w_1^* = \frac{1}{3}$ to -1, 0, and 1 of X_1	Equal weight $w_2^* = \frac{1}{Q_2} = \frac{1}{2}$ to each level of X_2	Equal weight $w^* = w_1^* \times w_2^* = \frac{1}{3}\frac{1}{Q_2} = \frac{1}{6}$ to -1, 0, and 1 of X_1 for each level of X_2
Model 3	$E_1(Y) = \beta_0 + \beta_1 X_1$ for X_1, $E_2(Y) = \beta_2 X_2$ for X_2	Equal weight $w_1^* = \frac{1}{2}$ to -1 and 1 of X_1	Equal weight $w_2^* = \frac{1}{Q_2} = \frac{1}{2}$ to each level of X_2	Equal weight $w^* = w_1^* \times w_2^* = \frac{1}{2}\frac{1}{Q_2} = \frac{1}{4}$ to -1 and 1 of X_1 for each level of X_2
Model 4	$E_1(Y) = \beta_0 + \beta_1 X_1 + \beta_3 X_1^2$ for X_1, $E_2(Y) = \beta_2 X_2$ for X_2	Equal weight $w_1^* = \frac{Q_2+1}{2(Q_2+2)} = \frac{3}{8}$ to -1 and 1 of X_1, and weight $1 - 2w_1^* = \frac{1}{Q_2+2} = \frac{1}{4}$ to 0 of X_1	Equal weight $w_2^* = \frac{1}{Q_2} = \frac{1}{2}$ to each level of X_2	For each level of X_2, equal weight $w_1^* \times w_2^* = \frac{Q_2+1}{2(Q_2+2)}\frac{1}{Q_2} = \frac{3}{16}$ to -1 and 1 of X_1, and weight $(1 - 2w_1^*) \times w_2^* = \frac{1}{Q_2+2}\frac{1}{Q_2} = \frac{1}{8}$ to 0 of X_1
Model 5	$E_1(Y) = \beta_0 + \beta_1 X_1 + \beta_3 X_1^2$ for X_1, $E_2(Y) = \beta_2 X_2$ for X_2	Equal weight $w_1^* = \frac{1}{3}$ to -1, 0, and 1 of X_1	Equal weight $w_2^* = \frac{1}{Q_2} = \frac{1}{2}$ to each level of X_2	Equal weight $w^* = w_1^* \times w_2^* = \frac{1}{3}\frac{1}{Q_2} = \frac{1}{6}$ to -1, 0, and 1 of X_1 for each level of X_2

Figure 1

Optimal Designs for the Marginal Models (Central Column), and D-Optimal Designs That Maximize Precision of Z-Score and PR-Score Estimation Under Models 1–5 (Rightmost Column)



Note. Dot size represents the design weights in Table 2 (for details, see online Supplement A, pp. 37–39). Recall that X_1 (e.g., age) $\in [-1, 1]$ and X_2 (e.g., sex) $\in \{0, 1\}$.

60–70, and 70–80 years). The key point to understand here is that the two/three age levels required by the optimal designs in Table 2 are a consequence of the assumption of a linear/quadratic age effect. If the linear/quadratic relation between age and Y is the "true" model (i.e., the best fitting polynomial), then including additional age levels into the normative sample yields, under the constraint of a fixed total N, a loss of statistical efficiency compared with the two/three age levels normative sample prescribed by Table 2 (as will be illustrated in the next section). However, the "true" model is often unknown and therefore the correctness of the specified model is uncertain. How to deal with uncertainty about the "true" model is the topic of the next section.

The results in Table 2 can be easily extended to models with three predictors, say age, sex and education, with complete or no interaction between age, sex, and education. If education is treated as a categorical predictor (e.g., low, medium, high level), the

D-optimal design is obtained by replicating the D-optimal design for age (Table 2, third column) at each combination of sex and education (i.e., with Q_2, now, as the number of combinations of sex and education). If education is treated as a quantitative predictor (e.g., number of years in school), the D-optimal design is obtained by replicating, at each level of sex, the D-optimal design for age and education. The latter, in turn, is obtained by crossing the D-optimal design for age (Table 2, third column) with that for education (Table 2, third column), with as design weights the product of the optimal weights of these two designs. For instance, if age and education have at most a quadratic effect, then the D-optimal design for the model with complete or no interaction crosses age values -1, 0, and 1 with the same values for education, with design weight $1/18$ for each of the nine age by education combinations which are to be replicated per level of sex. For optimal designs for higher-order polynomial effects, such as cubic, see Table 3.5 in Berger and Wong (2009).

Efficiency of the Optimal Design When There is Uncertainty About the "True" Model

The optimal designs in Table 2 depend on the assumed model, that is, they are optimal under the chosen model. At the design stage, however, the "true" model (i.e., the best fitting model) might be unknown. A solution to this issue is to find the design most robust against misspecification of the model. The most robust design can be obtained using two alternative criteria: the relative efficiency, and the efficiency. Under the relative efficiency criterion, the most robust design is defined as the design that guarantees the highest minimum relative efficiency (i.e., relative to the optimal design for a model) across all plausible models, and is called the *RE maximin design*. Under the efficiency criterion, instead, the most robust design is defined as the design that yields the highest minimum efficiency across all plausible models and is called the *absolute maximin design* (Van Breukelen & Candel, 2018). It will be shown that, when Models 1–5 are all equally plausible, the D-optimal design for Models 2 and 5 in Table 2 is both the RE maximin design and the absolute maximin design.

Relative Efficiency Criterion

Given a fixed total N, the relative efficiency of a design ξ versus the optimal design ξ^* is defined as the ratio of $V(\hat{Z}_0)$ or $V(PR(\hat{Z}_0))$ (that is, Equations 7 or 8) under ξ^* to $V(\hat{Z}_0)$ or $V(PR(\hat{Z}_0))$ under ξ, which reduces to the following expression (for proof, see online Supplement A, pp. 39–40):

$$RE(\xi \ vs \ \xi^*) \approx \frac{d(\boldsymbol{X}, \xi^*) + \frac{Z_0^2}{2}}{d(\boldsymbol{X}, \xi) + \frac{Z_0^2}{2}}, \qquad (11)$$

where $d(\boldsymbol{X}, \xi) = N\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0$ is the standardized prediction variance under design ξ. The RE maximin design is obtained in three steps:

1. For each design, find the lowest value of Equation 11 over $\boldsymbol{x}_0$, assuming each plausible model in turn to be the "true" model, given Z_0. This yields for each design and model combination the minimum RE.

2. For each design, find the lowest minimum RE among all minimum REs as obtained in Step 1, given Z_0. This gives, for each design, the lowest minimum RE across all plausible models.

3. Take the design with the highest lowest minimum RE across all designs considered.

The results of Steps 1–2, when Models 1–5 are all equally plausible, are shown in Table 3 for $Z_0 = 0$ and $Z_0 = \pm 2$ (for details, see online Supplement A, pp. 39–40). As can be seen in Table 3, the most robust design is the design in the second row (that is, the optimal design under Models 2 and 5), because that design has the highest lowest minimum relative efficiency across all designs considered. Further, a numerical evaluation has shown that what is the most robust design does not depend on Z_0 (for details, see online Supplement A, pp. 39–40). In its worst-case scenario (that is, true Model 3), the optimal design under Models 2 and 5 requires, for $Z_0 = 0$, $(RE^{-1} - 1)100\% = (0.80^{-1} - 1)100\% = 25\%$ more persons than the optimal design for that case (that is, for Model 3).

Note that Table 3 does not show all possible comparisons, because some of them are not feasible. Indeed, a design should include enough support points to ensure the estimation of all model parameters. For instance, a quadratic effect needs (at least) three support points to be identifiable, thus the relative efficiency of a design which has only two support points per sex level (that is, the optimal designs for Models 1 and 3) cannot be computed if the "true" model is Model 2, 4, or 5). Furthermore, Table 3 shows that including into the normative sample more age levels than prescribed by the optimal design under the "true" model, while keeping the total N fixed, is statistically inefficient. The $RE(\xi \, vs \, \xi^*)$ of equidistant age levels designs with four, five, six, and thirteen age levels, is shown in the last four rows of Table 3, and it decreases as the number of age levels increases. Specifically, the thirteen equidistant age levels design (i.e., that used in the simulation studies and perhaps the most appealing design for traditional norming if the age range is large, e.g., from 20 to 80 years) requires, depending on the "true" model, between $(0.6563^{-1} - 1)100\% \approx 52\%$ and $(0.4468^{-1} - 1)100\% \approx 124\%$ more persons than the optimal design, for $Z_0 = 0$, and between $(0.7609^{-1} - 1)100\% \approx 31\%$ and $(0.5185^{-1} - 1)100\% \approx 93\%$ more persons, for $Z_0 = \pm 2$.

Efficiency Criterion

Efficiency is defined as $(V(\hat{Z}_0))^{-1}$ or $(V(PR(\hat{Z}_0)))^{-1}$, depending on the norm statistic of interest. The absolute maximin design is obtained in three steps:

1. For each design, find the minimum efficiency (i.e., the maximum $V(\hat{Z}_0)$ or $V(PR(\hat{Z}_0))$) over $\boldsymbol{x}_0$, assuming each plausible model in turn to be the "true" model, given N and Z_0.

2. For each design, find the lowest minimum efficiency among all minimum efficiency values obtained in Step 1, given N and Z_0, thus obtaining, for each design, the lowest minimum efficiency across all plausible models.

Table 3

Minimum Relative Efficiency (i.e., Equation 11), Over $\boldsymbol{x}_0$, of Each Design (Row) Compared With the Optimal Design Under the "True" Model (Column) for $Z_0 = 0$, and for $Z_0 = \pm 2$ (Within Brackets)

	"True" model				
Design	Model 1	Model 2	Model 3	Model 4	Model 5
2 Age levels. Equal weight $w^* = \frac{1}{4}$ to -1 and 1 of $X_1 = \text{Age}$, for each level of $X_2 = \text{Sex}$	1.0 (1.0)		1.0 (1.0)		
3 Age levels. Equal weight $w^* = \frac{1}{6}$ to -1, 0, and 1 of $X_1 = \text{Age}$, for each level of $X_2 = \text{Sex}$	0.8571 (0.9091)	1.0 (1.0)	**0.8000 (0.8571)**	0.9091 (0.9333)	1.0 (1.0)
3 Age levels. Equal weight $\frac{3}{16}$ to -1 and 1 of $X_1 = \text{Age}$, and weight $\frac{1}{8}$ to 0 of $X_1 = \text{Age}$, for each level of $X_2 = \text{Sex}$	0.9000 (0.9375)	0.8000 (0.8571)	0.8571 (0.9000)	1.0 (1.0)	**0.7500 (0.8000)**
4 Age levels. Equal weight $w^* = \frac{1}{8}$ to -1, -0.333, 0.333 and 1 of $X_1 = \text{Age}$, for each level of $X_2 = \text{Sex}$	0.7895 (0.8621)	0.8333 (0.8824)	**0.7143 (0.7895)**	0.7576 (0.8140)	0.7895 (0.8333)
5 Age levels. Equal weight $w^* = \frac{1}{10}$ to -1, -0.5, 0, 0.5 and 1 of $X_1 = \text{Age}$, for each level of $X_2 = \text{Sex}$	0.7500 (0.8333)	0.7368 (0.8077)	**0.6667** (0.7500)	0.6731 (0.7424)	0.6774 (**0.7368**)
6 Age levels. Equal weight $w^* = \frac{1}{12}$ to -1, -0.6, -0.2, 0.2, 0.6 and 1 of $X_1 = \text{Age}$, for each level of $X_2 = \text{Sex}$	0.7241 (0.8140)	0.6747 (0.7568)	0.6364 (0.7241)	0.6195 (0.6950)	**0.6087 (0.6747)**
13 Age levels. Equal weight $w^* = \frac{1}{26}$ to $-1, -0.833, -0.667, -0.5, -0.333, -0.167,$ $0, 0.167, 0.333, 0.5, 0.667, 0.833,$ and 1 of $X_1 = \text{Age}$, for each level of $X_2 = \text{Sex}$	0.6563 (0.7609)	0.5185 (0.6176)	0.5600 (0.6563)	0.4861 (0.5698)	**0.4468 (0.5185)**

Note. Lowest minimum relative efficiency, across models, per design in boldface.

3. Take the design with the highest lowest minimum efficiency (i.e., the lowest highest maximum $V(\hat{Z}_0)$ or $V(PR(\hat{Z}_0)))$ across all designs considered.

When all Models 1–5 are plausible, Steps 1–3 yield the D-optimal design for Models 2 and 5 in Table 2 as the absolute maximin design (see online Supplement A, p. 41). Recall that this design is also the RE maximin design. Also note that designs with more than three age levels are not only inefficient in terms of their relative efficiencies (see Table 3), but also in terms of their efficiencies (online Supplement A, Table S.A.3), at least given the present set of plausible models.

Sample Size Calculation

Sample Size Calculation for Hypothesis Testing

The results of the previous sections are used in this section to determine the sample size N for a normative study under the optimal design (i.e., the optimal joint distribution of the predictors). Specifically, the required sample size can be determined using Equations 7 and 8 under the optimal design ξ^* as follows. First, note that $d(\boldsymbol{X}, \xi)$ (and thereby also $V(\hat{Z}_0)$ and $V(PR(\hat{Z}_0)))$, as a function of the predictor value $\boldsymbol{x}_0$, has its maximum at the support points of the optimal design (a result of the equivalence theorem, see also Figure 2, first row), which thus are safe starting points for sample size calculation, because other predictor values give smaller sampling variances. Second, at any of the support points of the optimal design ξ^*, Equations 7 and 8 can be rewritten for ξ^* as follows (with k = number of predictors, see Table 1):

$$V(\hat{Z}_0)^* \approx \frac{k+1}{N} + \frac{1}{2(N-k-1)} Z_0^2 \approx \frac{2(k+1)+Z_0^2}{2N}, \qquad (12)$$

and

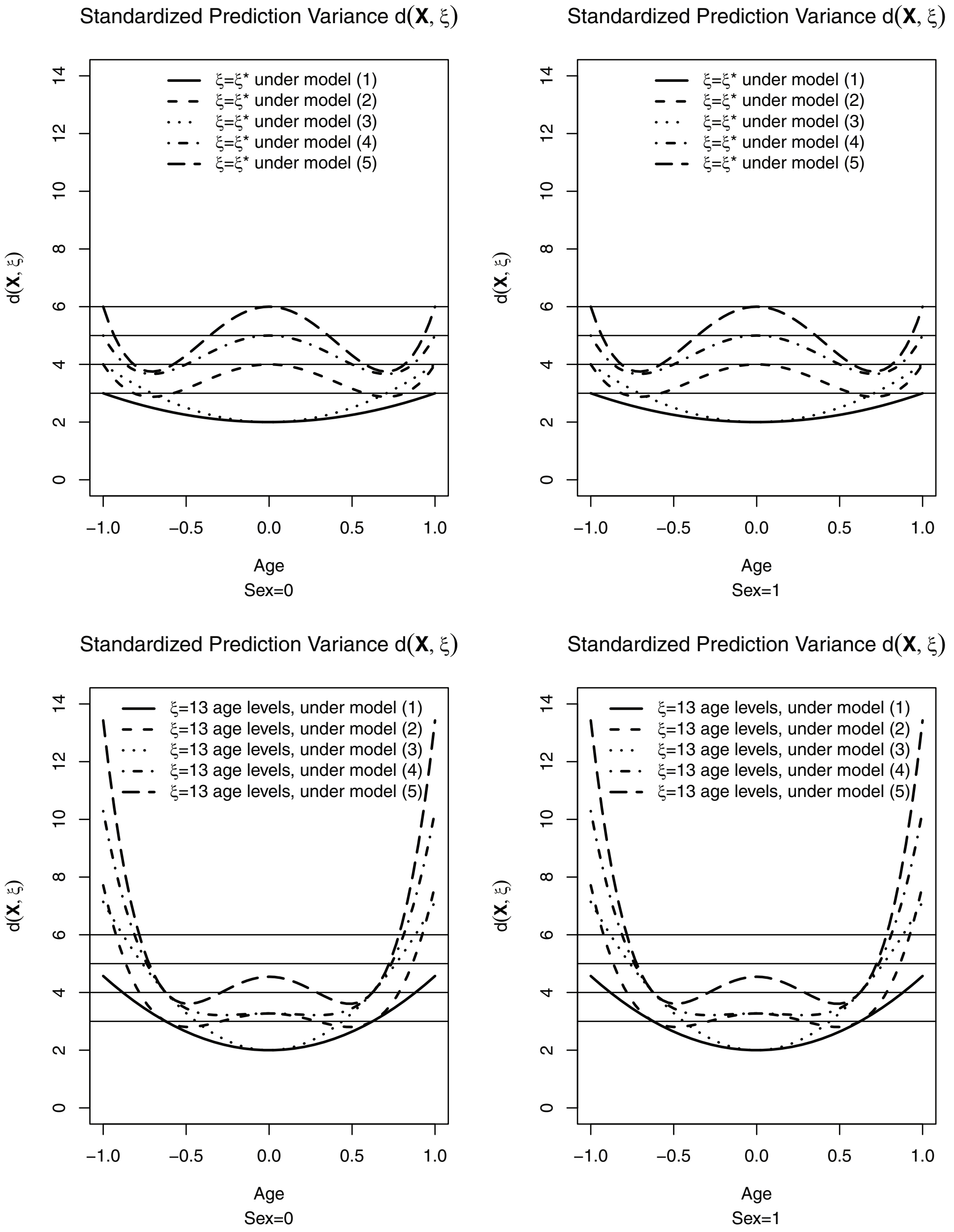
$$\begin{aligned} V(PR(\hat{Z}_0))^* &\approx 100^2 \phi(Z_0)^2 V(\hat{Z}_0)^* \\ &\approx 100^2 \phi(Z_0)^2 \left[\frac{2(k+1)+Z_0^2}{2N}\right], \end{aligned} \qquad (13)$$

by plugging into Equations 7 and 8 $\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0 = \frac{d(\boldsymbol{X},\xi)}{N}$ and by noting that $d(\boldsymbol{X}, \xi^*) = k + 1$ (a result of the equivalence theorem, see also Figure 2, first row). Regarding the first approximation in Equations 12 and 13, recall that Equation 7 is an accurate approximation of the true $V(\hat{Z}_0)$ for $N \geq 338$, and Equation 8 is an accurate approximation of the true $V(PR(\hat{Z}_0))$ for $N \geq 1{,}690$ (online Supplement A, pp. 7–33). For these lower-bounds for the sample size, $N - k - 1 \approx N$ which gives the second approximation in Equations 12 and 13. Hence, the required sample size cannot be smaller than these two to ensure that the approximations are satisfactory.

To infer sample size recommendations from Equations 12 and 13, one needs to define the objective of the norming study first and then determine the required sample size. In practice, the main use of reference values is to classify individual's performance relative to a chosen cut-off point as, say, "normal" (e.g., $-2 \leq Z_0 \leq +2$ or $5 \leq PR_0 \leq 95$) versus "too low" (e.g., $Z_0 < -2$ or $PR_0 < 5$) or

Figure 2

Standardized Prediction Variance $d(\mathbf{X}, \xi)$, as Function of Individual's Age and Sex (i.e., $\boldsymbol{x}_0$), Under the Five Optimal Designs ξ^* in Table 2 (Top Half of Figure 2), and Under the 13 Equidistant Age Levels Design in the Last Row of Table 3 (Bottom Half of Figure 2)



Note. The horizontal lines in each panel are located at $k+1$ (i.e., the number of regression coefficients) for each model. From the equivalence theorem (Atkinson et al., 2007); a design ξ is D/G-optimal if and only if $d(\boldsymbol{X}, \xi) = (k+1)$ at its support points and $d(\boldsymbol{X}, \xi) < (k+1)$ over the rest of the design region. Note that while the designs in Table 2 satisfy this condition (see top half of the figure), the 13 equidistant age levels design does not (see bottom half of the figure).

"too high" (e.g., $Z_0 > +2$ or $PR_0 > 95$), in order to make decisions about, for instance, clinical treatments or remedial teaching. The size of the normative sample must then be sufficiently large to allow adequate classification of the individual. This classification problem can be expressed in terms of hypothesis testing, and the required sample size can be determined as a function of type I error rate, power, and effect size. Denote by Z_c or PR_c the cut-off point to be used for decision making, and denote by Z_t or PR_t the individual's true Z- or PR-score, and by δ the smallest "clinically relevant" difference between Z_t and Z_c, or between PR_t and PR_c. Depending on the norm statistic type of interest (i.e., Z-score or PR-score), the null hypothesis H_0 is defined as either $Z_t = Z_c$ or $PR_t = PR_c$, and the alternative hypothesis H_1 is one-sided, that is, either $Z_t < Z_c$ or $PR_t < PR_c$ (if $Z_c < 0$ or $PR_c < 50$), or $Z_t > Z_c$ or $PR_t > PR_c$ (if $Z_c > 0$ or $PR_c > 50$), because in practice psychologists are usually interested in distinguishing between "normal" and "too low" (for performance), or between "normal" and "too high" (for clinical symptoms). Thus, the required sample size is here defined as that size N^* of the normative sample that allows to detect the smallest "clinically relevant" difference (δ) between individual's true Z- or PR-score (Z_t or PR_t) and the cut-off point used for classifying individuals (Z_c or PR_c), given a prespecified type I error rate α and power $1-\gamma$. The required sample size N^* can be computed with the following procedure:

1. Choose the regression model for norming (thus fixing the number of predictors k), the norm statistic of interest (i.e., Z-score or PR-score), the cut-off point for classifying individuals (e.g., $Z_c = -2$ or $PR_c = 5$), the Type I error rate α, the power $1-\gamma$, and the smallest "clinically relevant" difference $\delta > 0$ between Z_t and Z_c, or between PR_t and PR_c, that one wants to be able to detect.

2. Compute the required sample size with one of the following equations (for proofs, see Appendix C). For Z-scores, the required sample size N^* is given by

$$N^* = \left[\frac{z_{1-\alpha}\left(k+1+\frac{Z_c^2}{2}\right)^{1/2} + z_{1-\gamma}\left(k+1+\frac{Z_t^2}{2}\right)^{1/2}}{\delta}\right]^2, \qquad (14)$$

where $z_{1-\alpha}$ and $z_{1-\gamma}$ are the $(1-\alpha)$th and $(1-\gamma)$th percentiles of the standard normal distribution (e.g., 1.65 and 1.28 if $\alpha = 0.05$ one-tailed and the power is 90%). For PR-scores, the required sample size N^* is given by

$$N^* = \left[\frac{z_{1-\alpha}\cdot 100\cdot \phi(Z_{PR_c})\left(k+1+\frac{Z_{PR_c}^2}{2}\right)^{1/2} + z_{1-\gamma}\cdot 100\cdot \phi(Z_{PR_t})\left(k+1+\frac{Z_{PR_t}^2}{2}\right)^{1/2}}{\delta}\right]^2, \qquad (15)$$

where $Z_{PR_c} = \Phi^{-1}\left(\frac{PR_c}{100}\right)$, $Z_{PR_t} = \Phi^{-1}\left(\frac{PR_t}{100}\right)$, and $\Phi^{-1}(\cdot)$ is the inverse function of the cumulative standard normal distribution.

3. If the sample size obtained from Equations 14 and 15 is smaller than the two lower-bounds 338 for $V(\hat{Z}_0)$ and 1,690 for $V(PR(\hat{Z}_0))$ (see Appendix B), researchers should be aware that the bias induced by the approximations in Equations 7–8 (on which Equations 14–15 are based) might be large (for details, see Appendix B), and should consider taking a larger sample size. For PR-scores, one can use the results of the simulation studies given in online Supplements A and B to decide upon the increase of N^*. For instance, if for Model 5 Equation 15 yields $N^* \in [676, 1690)$, one could increase N^* by 15%, which is the largest relative bias found for Equation 8 under Model 5 and $N = 676$ (see Figure S.A.8, online Supplement A).

Note that Equations 14 and 15 depend on the true Z- or PR-score (i.e., Z_t or PR_t). Given the chosen cut-off Z_c or PR_c and the chosen smallest clinically relevant difference $\delta > 0$, there are two possible values for the individual's true Z_t or PR_t: $Z_t = Z_c - \delta$ or $Z_t = Z_c + \delta$, and $PR_t = PR_c - \delta$ or $PR_t = PR_c + \delta$. Which of the two possible values of Z_t or PR_t should one plug into Equations 14 or 15? In practice, one is primarily interested in detecting extreme performance or symptoms, since these have important consequences for the individual (e.g., the assignment to a treatment), thus one should plug into Equations 14 and 15 a Z_t or PR_t more extreme than the chosen cut-off Z_c or PR_c. Hence, if $Z_c < 0$ take $Z_t = Z_c - \delta$, while if $Z_c > 0$ take $Z_t = Z_c + \delta$. Likewise, if $PR_c < 50$ take $PR_t = PR_c - \delta$, while if $PR_c > 50$ take $PR_t = PR_c + \delta$. Having established in which direction the true Z-score or PR-score should be assumed (i.e., always more extreme than the cut-off point for decision making), the choice of the exact value for Z_t or PR_t depends on the choice of the effect size δ. How small (or large) δ should be depends on the definition of "clinically relevant" for the specific test score of interest. Furthermore, note that δ is on the Z-score or the PR-score scale depending on the choice of the norm statistic in Step 1. For example, δ might be chosen to be 0.3 for Z-scores, or to be 2 for PR scores. How δ affects the required sample size will be discussed in the next paragraph.

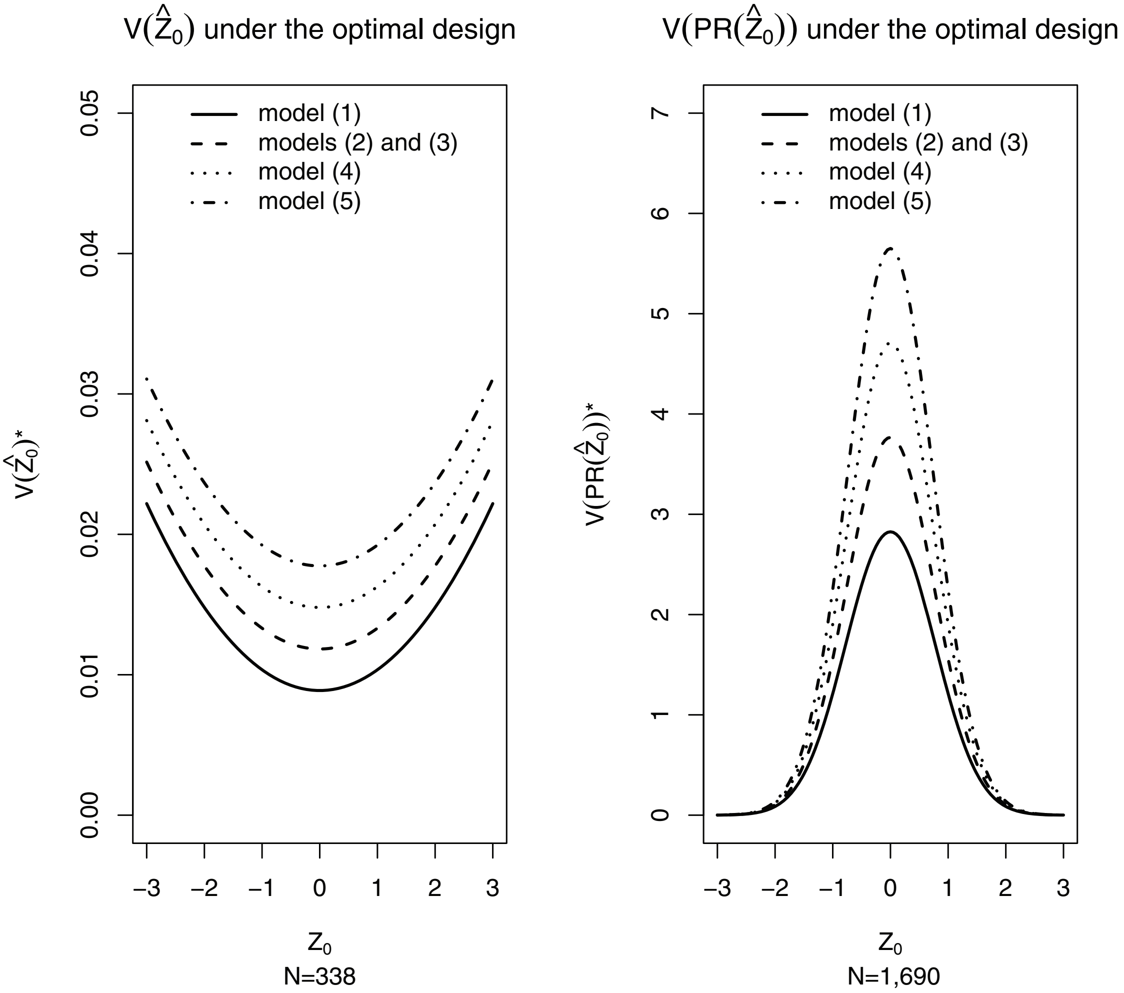
As can be seen in Equations 14 and 15, the required sample size N^* is an increasing function of the number of predictors k in the model and the statistical power $1-\gamma$ (the larger $1-\gamma$, the larger $z_{1-\gamma}$), and a decreasing function of the Type I error rate α (the larger α, the smaller $z_{1-\alpha}$). Furthermore, N^* for Z-scores increases as Z_c and Z_t move away from 0, while N^* for PR-scores increases as PR_c and PR_t move toward 50. This opposite pattern for Z-scores and PR-scores is explained by the presence of the factor $100 \times \phi(Z_0)$ in Equation 15 (but not in Equation 14), which increases rapidly as $|Z_0|$ decreases (i.e., PR_0 moves toward 50, see Figure 3). Finally, N^* is roughly proportional to δ^{-2} but not exactly so because δ is both at the denominator and the numerator of Equations 14 and 15 (recall that Z_t and PR_t are both functions of δ, given Z_c and PR_c, respectively). In Table 4, N^* is given for several cut-off points, three δ values per norm statistic type, different models (that is, Models 1–5), and two power levels. As shown in Table 4, N^* increases as δ decreases, both for Z-scores and PR-scores, but PR-scores tend to require a larger sample size than Z-scores.

Sample Size Calculation for Precision of Norms Estimation

An alternative approach to sample size calculation is to focus on the precision of norms estimation, instead of hypothesis testing (Maxwell et al., 2008). Under this approach, the required sample size N^* is defined as that size N^* of the normative sample that provides the desired margin of estimation error for the $(1-\alpha/2)100\%$ confidence interval for a Z- or PR-score of interest. Note that the width of a $(1-\alpha/2)100\%$ confidence interval equals $2\cdot z_{1-\alpha/2}(V(\hat{Z}_0))^{1/2}$ or $2\cdot z_{1-\alpha/2}(V(PR(\hat{Z}_0)))^{1/2}$, depending on the norm statistic type of interest, where $z_{1-\alpha/2}$ is the

Figure 3

Variances $V(\hat{Z}_0)^*$ and $V(PR(\hat{Z}_0))^*$ (i.e., Equations 12 and 13), as Functions of Z_0



$(1-\alpha/2)$th percentile of the standard normal distribution, $V(\hat{Z}_0)$ is Equation 7 or, under the optimal design, Equation 12, and $V(PR(\hat{Z}_0))$ is Equation 8 or, under the optimal design, Equation 13. Thus, the required sample size N^* can be obtained as that size of the normative sample such that half the confidence interval width equals the desired margin of estimation error. Then, for Z-scores the required N^* is given by

$$N^* = \left[\frac{z_{1-\alpha/2}(k+1+\frac{Z_0^2}{2})^{1/2}}{\tau}\right]^2, \qquad (16)$$

and for PR-scores the required N^* is given by

$$N^* = \left[\frac{z_{1-\alpha/2}\cdot 100\cdot \phi(Z_0)(k+1+\frac{Z_0^2}{2})^{1/2}}{\tau}\right]^2, \qquad (17)$$

where Z_0 is the Z-score of interest in Equation 16, and the Z-score corresponding to the PR-score of interest in Equation 17, and τ is the desired margin of estimation error (instead of the smallest "clinically relevant" effect size). Note that Equations 16 and 17 can be obtained by replacing $z_{1-\alpha}$ with $z_{1-\alpha/2}$, $z_{1-\gamma}=0$ (i.e., power $1-\gamma=50\%$), and δ with τ in Equations 14 and 15. Having replaced Equations 14 and 15 with Equations 16 and 17, one can still follow Steps 1–3 to determine N^*, which is now that size of the normative sample that yields sufficient precision of norms estimation (instead of that size that allows to detect the desired effect size). For example, under Model 1 (i.e., $k=2$), if the desired margin of estimation error for the 95% confidence interval for a $Z_0 = \pm 2$ of interest is 10% of Z_0, so $\tau = 0.20$, then using Equation 16 one obtains that $N^* = 480$.

Determining the required sample size with Equations 14–17 has several advantages. First, it is a safe approach because it guarantees the desired precision level/power level for the worst-case predictor values, and at least the same precision level/power level for all other predictor values for which the sampling variance is smaller (i.e., a result of drawing a normative sample as prescribed by the G-optimal design). Second, Equations 14–17 allow researchers to analytically compute the required sample size given either the desired margin of error for interval estimation, or the desired power level and effect size for hypothesis testing. Third, the proposed approach is not restricted to a specific regression model (for example, Equation 1),

Table 4

Size N^* of the Normative Sample, Under the Optimal Designs for Models 1–5, That Allows to Detect the Smallest "Clinically Relevant" Difference (δ) Between Individual's True Z- or PR-Score (Z_t or PR_t) and the Cut-Off Point Used for Classifying Individuals (Z_c or PR_c)

		Z_c			PR_c		
Power	Model	±2.5	±2	±1.5	2.5 or 97.5	5 or 95	10 or 90
		δ = 0.4			δ = 2		
0.8	1	**250**	**204**	**168**	**149**	**574**	**1,657**
	2 and 3	**289**	**243**	**207**	**178**	**702**	2,085
	4	**328**	**282**	**245**	**207**	**831**	2,513
	5	366	**320**	**284**	**237**	**960**	2,941
0.9	1	353	**288**	**236**	**169**	**741**	2,231
	2 and 3	406	341	**290**	**202**	**906**	2,805
	4	460	395	344	**235**	**1,071**	3,379
	5	513	449	397	**268**	**1,236**	3,954
		δ = 0.3			δ = 1.5		
0.8	1	439	358	**295**	**316**	**1,085**	3,020
	2 and 3	508	427	363	**379**	**1,330**	3,802
	4	576	496	432	**441**	**1,574**	4,585
	5	645	564	501	**504**	1,819	5,368
0.9	1	615	502	413	**386**	**1,430**	4,098
	2 and 3	711	597	508	**462**	1,751	5,157
	4	806	693	603	**537**	2,072	6,215
	5	901	788	698	**613**	2,393	7,274
		δ = 0.2			δ = 1		
0.8	1	974	794	654	**824**	2,584	6,958
	2 and 3	1,128	949	809	**988**	3,171	8,767
	4	1,283	1,104	963	**1,152**	3,757	10,576
	5	1,437	1,258	1,118	**1,315**	4,344	12,386
0.9	1	1,360	1,109	913	**1,060**	3,470	9,510
	2 and 3	1,574	1,324	1,127	**1,269**	4,254	11,978
	4	1,788	1,538	1,341	**1,479**	5,039	14,446
	5	2,002	1,752	1,555	**1,688**	5,823	16,914

Note. The Type I error rate α is 0.05. The individual's true Z- or PR-score is assumed to be more extreme than the cut-off (i.e., $Z_t < Z_c < 0$ or $Z_t > Z_c > 0$, and $PR_t < PR_c < 50$ or $PR_t > PR_c > 50$). Sample sizes in boldface are below the lower-bounds $N^* = 338$ for Z-scores, and $N^* = 1{,}690$ for PR-scores, which are required to ensure a good approximation by the equations used (for details, see Appendix B).

because Equations 12 and 13 (used to derive Equations 14–17) are restricted neither to a specific number of predictors nor to specific scale types of the predictors. Relatedly, from Equations 14–17 it can be seen that if a categorical predictor is included into the regression model by dummy coding, the required sample size increases with the number of categories Q_2, because the higher Q_2, the higher the number of predictors k in the model. Increasing Q_2 also affects the optimal design, because the optimal design for the quantitative predictor (i.e., X_1 in Figure 1, central column) is then replicated for each level of the categorical predictor (i.e., X_2 in Figure 1, central column), thus increasing the total number of support points and decreasing the design weight per support point proportionally (see Table 2, rightmost column).

Application

The results of this article are illustrated using Van der Elst et al.'s (2006) normative study of three verbal fluency tests (VFT). In a VFT, participants are asked to name as many words as possible belonging to a specific category (e.g., professions, animals) in 60 s. Hence, the raw score of a VFT is the total number of correct, nonrepeated words (so the higher the raw score, the better). The normative sample was composed of $N = 1{,}825$ healthy individuals belonging to 12 equidistant age groups within the range 24–81 years and of approximately equal size (except the eldest group, which was the smallest one), and the male-female ratio per age group was (fairly) balanced (Van der Elst, et al., 2006; Table 1). The distribution of education within each age group, instead, was not balanced. Van der Elst et al. (2006) fitted three separate multiple linear regression models, one for each VFT, but the focus is here on the profession naming VFT, which was also used in the simulation studies. The predictors were age, age^2, sex, and education (i.e., low, medium, high), including all possible two-way interactions, but in the final model no interaction was present (Van der Elst, et al., 2006, Table 2). The assumptions of normality and homoscedasticity of residuals were satisfied as far as regression diagnostics results could tell.

Suppose a researcher wants to derive new normative data for the profession naming VFT, focusing on men and women in the age range 55–85 years, and leaving education out from the study for simplicity now. The normative study can be planned in three steps:

1. Choice of the model: Based on Van der Elst et al. (2006, Figure 1 and Table 2), it is reasonable to assume a quadratic age effect on the score of the profession naming VFT. Further, even though an interaction between sex and age, and between sex and age^2 were not found, it is prudent to assume them both because a Type II error cannot be ruled out in Van der Elst et al. (2006). So, Model 5 is assumed.

2. Choice of the design: As shown in Figure 1, the optimal design for Model 5 has six support points: $sex = 0$ and $age = 55$, $sex = 0$ and $age = 70$, $sex = 0$ and $age = 85$, $sex = 1$ and $age = 55$, $sex = 1$ and $age = 70$, $sex = 1$ and $age = 85$. Furthermore, the same number of subjects must be sampled for each support point (see Table 2). This design yields maximum efficiency under Model 5 (and Model 2), but Model 5 might not be the "true" model. To prevent a great loss of efficiency due to model misspecification, the most robust design should be chosen instead. In this case, the optimal design for Model 5 is also the most robust design, both in terms of relative efficiency and of efficiency (see Table 3, and Table S.A.3, online Supplement A), at least assuming the age effect to be linear or quadratic, but not of higher order.

3. Sample size calculation: Suppose that the researcher wants to provide normative data in terms of Z-scores. Hence, to determine the required total sample size either Equation 14 or 16 should be used. Equation 14 should be used if the researcher wants to use the derived norms to make decisions about individuals (e.g. assignment to a treatment). Instead, Equation 16 should be used when the primary interest is in assessing within which range of values the true Z-score of an individual lies. In any case, it is prudent to target extreme Z-scores because they require a larger sample size (see Figure 3). Suppose that the

researcher wants to have a sample such that half the width of the 95% confidence interval for $Z_0 = -1.64$ (i.e., the 5th percentile, below which an individual's performance is considered "abnormal") is $\tau = 0.18$ (which is half the distance between the 10th and 5th percentile of the Z distribution). Plugging these values into Equation 16:

$$N^* = \left[\frac{z_{1-\alpha/2}\left(k+1+\frac{Z_0^2}{2}\right)^{1/2}}{\tau}\right]^2 = \left[\frac{1.96\left(5+1+\frac{(-1.64)^2}{2}\right)^{1/2}}{0.18}\right]^2,$$

it then follows that the researcher should sample 870 individuals, so 145 per support point, assuming Model 5 (i.e., $k = 5$ predictors).

Discussion

The aim of this article was to illustrate how to design regression-based normative studies for which the norm statistics of interest are Z-scores and percentile rank (*PR*) scores. The sampling variances of these norm statistics were derived under the assumptions of normality and homoscedasticity of the residual errors. Because these variances were based on approximations, a simulation study was performed to investigate the bias induced by these approximations. A second simulation study investigated the bias in the estimators of these variances as well as the coverage of 95% confidence intervals for an individual's true Z-score or PR-score obtained with these variance estimators. From these simulation studies, it can be concluded that for the sampling variance equation and its estimator to be accurate, and for the coverage to be close to the nominal value, the sample size should be at least $N = 338$ for Z-scores, and $N = 1{,}690$ for PR-scores.

Five regression models with a quantitative predictor (e.g., age) and a categorical predictor (e.g., sex), differing in whether they allowed for interaction and for nonlinearity, were considered. For each of these models the optimal design, that is, the joint distribution of the predictors which maximizes precision of norms estimation under the constraint of a fixed sample size, was presented. Extensions to the inclusion of a third predictor (e.g., education) were also discussed. Furthermore, the robustness of the optimal design against misspecification of the model was investigated. It turned out that the optimal design for Models 2 and 5 guarantees the highest minimum relative efficiency (i.e., relative to the optimal design under the "true" model) across all considered models (Tables 2 and 3), and the highest minimum efficiency (that is, under the model that maximizes the sampling variance of the norm statistic for all considered designs, which is Model 5, see Table S.A.3 in online Supplement A). Thus, in presence of high uncertainty about the best model among the five models considered, the optimal design for Models 2 and 5 is recommended.

Equidistant age levels designs with intervals of 5–20 years, typically used in traditional norming, were compared with the optimal design. For efficient estimation, it turned out that the normative sample need not be representative of the reference population with respect to the distribution of the predictors. If the "true" model is known, then maximum efficiency is achieved by drawing a normative sample as prescribed by the optimal design (see Table 3). If knowledge of the "true" model may sound as a strong prerequisite, it should be noted that normative studies often deal with tests that have already been normed in the past or in other countries/languages (see, for instance, Mitrushina et al., 2005). These previous studies can help researchers in formulating (at least) an educated guess of the best fitting polynomial that helps to improve the design of the normative study, thus saving resources. In the 65 regression-based normative studies reviewed in online Supplement A, either a linear or a quadratic age effect was observed. If researchers do not trust the linear/quadratic model and suspect that a polynomial of higher order is the "true" model, they can include intermediate age levels into the design of the normative study. Also for such models an optimal design yielding maximum efficiency can be derived. Specifically, the optimal design for an age effect of polynomial order h (e.g., $h = 3$ for cubic trend) consists of $h + 1$ age levels only (see Berger & Wong, 2009, Table 3.5, . 67).

A procedure was proposed that allows researchers to analytically compute the required sample size for estimating norm statistics (Z-scores or PR-scores) that gives sufficient power for hypotheses testing, or sufficient precision of interval estimation of an individual's position relative to reference values. Hypothesis testing is of interest when reference values are used to make binary decisions about whether to treat a person or not (e.g., remedial teaching in school, cognitive training in revalidation, medication or psychotherapy for depression). Instead, when the main interest is in assessing within which range the person's true Z-score or PR-score lies, interval estimation is preferable. Both for interval estimation and hypothesis testing, researchers need to choose the norm statistic type of interest (i.e., either Z- or PR-score), and a regression model. For interval estimation, researchers need to specify the coverage probability, the value of the norm statistic (e.g., $Z_0 = 2$, or $PR_0 = 95$), and the desired margin of error. For hypothesis testing, researchers need to specify a cut-off value to be used for decision making (e.g., $Z_c = -2$ or $PR_c = 5$, for distinguishing between "normal" and "too low" performance), the Type I error rate, the desired power level, and the smallest "clinically relevant" difference between the cut-off and the individual's position that one wants to be able to detect. The choice of the norm statistic type has important consequences for sample size calculation, because PR-scores tend to require larger sample sizes than Z-scores. Finally, when researchers are interested in several cut-off points (e.g., for distinguishing between "too low," "low," and "normal" performance), Equations 16 and 17 for confidence intervals could be used in sample size planning and the most extreme Z-score and the least extreme PR-score of interest should be targeted. This approach guarantees not only the desired margin of error for the most extreme Z-score and the least extreme PR-score, but also (at least) the same precision level for the less extreme Z-scores and the more extreme PR-scores (see Figure 3). An alternative approach is sequential hypothesis testing (e.g., first testing whether a performance is "normal" or "low" and, only if classified as "low," whether it is "low" or "too low"), but this approach is complicated due to multiple testing issues. How to control for familywise Type I and II error rates and thus adapt Equations 14 and 15 in sequential hypothesis testing for norming could be a topic for future research.

A limitation of the regression-based approach is that the validity of the norms depends on whether the model assumptions are met. For instance, if homoscedasticity is violated in Models 1–5, $\hat{Z}_0 = \frac{\hat{\varepsilon}_0}{\hat{\sigma}_\varepsilon}$ is biased, and thereby $PR(\hat{Z}_0)$ as well (see Equation 6). For this reason, it is important that researchers report the results of regression diagnostics when providing normative data (note that in the literature review on regression-based normative studies in online Supplement A, these checks were reported only for 42% of the models). If both homoscedasticity and normality are violated, a simple solution can be to compute the standard deviation of the residuals per quartile (or per decile) of the predicted values, first, then to standardize the residuals with these standard deviations and, finally, to estimate percentiles from the empirical distribution of the standardized residuals. A more sophisticated solution could be to derive norms using generalized additive models for location, scale, and shape (see Voncken et al., 2019a, 2019b); which allow researchers to model a wide range of test score distributions. Future research could extend the results of this article to heteroscedasticity and/or non-normality. Furthermore, the simulation studies in online Supplement A, and the results of Tables 3 and 4, could be extended to the presence of additional predictors (e.g., education). Other extensions could be the derivation of the optimal design for regression-based norming for repeated cognitive assessment (see Van der Elst et al., 2013), or, because normative studies often involve several outcomes (online Supplement A), for multivariate regression-based norming (see Van der Elst et al., 2017).

References

Atkinson, A. C., Donev, A. N., & Tobias, R. D. (2007). *Optimum experimental designs, with SAS*. Oxford University Press.

Berger, M. P. F., & Wong, W. K. (2009). *An introduction to optimal designs for social and biomedical research*. Wiley. https://doi.org/10.1002/9780470746912

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Duxbury.

Goos, P., & Jones, B. (2011). *Optimal design of experiments: A case study approach*. Wiley. https://doi.org/10.1002/9781119974017

Goretti, B., Niccolai, C., Hakiki, B., Sturchio, A., Falautano, M., Minacapelli, E., Martinelli, V., Incerti, C., Nocentini, U., Murgia, M., Fenu, G., Cocco, E., Marrosu, M. G., Garofalo, E., Ambra, F. I., Maddestra, M., Consalvo, M., Viterbo, R. G., Trojano, M., . . . Amato, M. P. (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), 1–6. https://doi.org/10.1186/s12883-014-0171-6

Johnson, R. A., & Wichern, D. W. (1998). *Applied multivariate statistical analysis* (4th ed.). Pearson Prentice Hall.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*(1), 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed.). Oxford University Press.

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

Oosterhuis, H. E. M., Van der Ark, L. A., & Sijtsma, K. (2017). Standard errors and confidence intervals of norm statistics for educational and psychological tests. *Psychometrika*, *82*(3), 559–588. https://doi.org/10.1007/s11336-016-9535-8

Parmenter, B. A., Testa, S. M., Schretlen, D. J., Weinstock-Guttman, B., & Benedict, R. H. B. (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*(1), 6–16. https://doi.org/10.1017/S1355617709990750

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Schwabe, R. (1996). *Optimum designs for multi-factor models*. Springer-Verlag. https://doi.org/10.1007/978-1-4612-4038-9

Van Breukelen, G. J. P., & Candel, M. J. J. M. (2018). Efficient design of cluster randomized trials with treatment-dependent costs and treatment-dependent unknown variances. *Statistics in Medicine*, *37*(21), 3027–3046. https://doi.org/10.1002/sim.7824

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

Van der Elst, W., Molenberghs, G., Van Boxtel, M. P. J., & Jolles, J. (2013). Establishing normative data for repeated cognitive assessment: A comparison of different statistical methods. *Behavior Research Methods*, *45*(4), 1073–1086. https://doi.org/10.3758/s13428-012-0305-y

Van der Elst, W., Molenberghs, G., Van Tetering, M., & Jolles, J. (2017). Establishing normative data for multi-trial memory tests: The multivariate regression-based approach. *The Clinical Neuropsychologist*, *31*(6-7), 1173–1187. https://doi.org/10.1080/13854046.2017.1294202

Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2006). Normative data for the animal, profession and letter M naming verbal fluency tests for Dutch speaking participants and the effects of age, education, and sex. *Journal of the International Neuropsychological Society*, *12*(1), 80–89. https://doi.org/10.1017/S1355617706060115

Voncken, L., Albers, C. J., & Timmerman, M. E. (2019a). Improving confidence intervals for normed test scores: Include uncertainty due to sampling variability. *Behavior Research Methods*, *51*(2), 826–839. https://doi.org/10.3758/s13428-018-1122-8

Voncken, L., Albers, C. J., & Timmerman, M. E. (2019b). Model selection in continuous test norming with GAMLSS. *Assessment*, *26*(7), 1329–1346. https://doi.org/10.1177/1073191117715113

Wong, W. K. (1995). On the equivalence of D and G-optimal designs in heteroscedastic models. *Statistics & Probability Letters*, *25*(4), 317–321. https://doi.org/10.1016/0167-7152(94)00236-1

(*Appendices follow*)

Appendix A

Derivations of the Sampling Variances of the Z-Score and PR-Score Estimators

Sampling Variance of $\hat{Z}_0$

To derive the sampling variance of $\hat{Z}_0 = \hat{\sigma}_\varepsilon^{-1}(Y_0 - \boldsymbol{x}_0'\hat{\beta})$ with the Delta method (Casella & Berger, 2002, p. 245), one needs to derive the first-order derivatives of $\hat{Z}_0$ with respect to $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_\varepsilon^2$ evaluated at their expectations $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_\varepsilon^2\,)'$:

$$\left.\frac{\partial \hat{Z}_0}{\partial \hat{\boldsymbol{\beta}}}\right|_{\hat{\boldsymbol{\beta}}=\boldsymbol{\beta},\hat{\sigma}_\varepsilon^2=\sigma_\varepsilon^2} = -\hat{\sigma}_\varepsilon^{-1}\boldsymbol{x}_0'\Big|_{\hat{\boldsymbol{\beta}}=\boldsymbol{\beta},\hat{\sigma}_\varepsilon^2=\sigma_\varepsilon^2} = -\sigma_\varepsilon^{-1}\boldsymbol{x}_0',$$

$$\left.\frac{\partial \hat{Z}_0}{\partial \hat{\sigma}_\varepsilon^2}\right|_{\hat{\boldsymbol{\beta}}=\boldsymbol{\beta},\hat{\sigma}_\varepsilon^2=\sigma_\varepsilon^2} = -\frac{(\hat{\sigma}_\varepsilon^2)^{-\frac{3}{2}}}{2}(Y_0 - \boldsymbol{x}_0'\hat{\boldsymbol{\beta}})\Big|_{\hat{\boldsymbol{\beta}}=\boldsymbol{\beta},\hat{\sigma}_\varepsilon^2=\sigma_\varepsilon^2}$$

$$= -\frac{(\sigma_\varepsilon^2)^{-\frac{3}{2}}}{2}(Y_0 - \boldsymbol{x}_0'\boldsymbol{\beta}).$$

Then, $\hat{Z}_0$ can be approximated with a first-order Taylor series as follows

$$\hat{Z}_0 \approx g(\boldsymbol{\beta}, \sigma_\varepsilon^2) + \left.\frac{\partial \hat{Z}_0}{\partial \hat{\boldsymbol{\beta}}}\right|_{\hat{\boldsymbol{\beta}}=\boldsymbol{\beta},\hat{\sigma}_\varepsilon^2=\sigma_\varepsilon^2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

$$+ \left.\frac{\partial \hat{Z}_0}{\partial \hat{\sigma}_\varepsilon^2}\right|_{\hat{\boldsymbol{\beta}}=\boldsymbol{\beta},\hat{\sigma}_\varepsilon^2=\sigma_\varepsilon^2}(\hat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2) = \sigma_\varepsilon^{-1}(Y_0 - \boldsymbol{x}_0'\boldsymbol{\beta}) - \sigma_\varepsilon^{-1}\boldsymbol{x}_0'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

$$-\frac{(\sigma_\varepsilon^2)^{-\frac{3}{2}}}{2}(Y_0 - \boldsymbol{x}_0'\boldsymbol{\beta})(\hat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2).$$

Finally, taking the expectation and variance of the previous expression with respect to $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_\varepsilon^2$, one obtains

$$E(\hat{Z}_0) \approx$$

$$E\left(\sigma_\varepsilon^{-1}(Y_0 - \boldsymbol{x}_0{}'\boldsymbol{\beta}) - \sigma_\varepsilon^{-1}\boldsymbol{x}_0'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \frac{(\sigma_\varepsilon^2)^{-\frac{3}{2}}}{2}(Y_0 - \boldsymbol{x}_0'\boldsymbol{\beta})(\hat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2)\right)$$

$$= \sigma_\varepsilon^{-1}(Y_0 - \boldsymbol{x}_0{}'\boldsymbol{\beta}) = Z_0,$$

and

$$V(\hat{Z}_0) \approx$$

$$V\left(\sigma_\varepsilon^{-1}(Y_0 - \boldsymbol{x}_0{}'\boldsymbol{\beta}) - \sigma_\varepsilon^{-1}\boldsymbol{x}_0'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \frac{(\sigma_\varepsilon^2)^{-\frac{3}{2}}}{2}(Y_0 - \boldsymbol{x}_0'\boldsymbol{\beta})(\hat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2)\right)$$

$$= \sigma_\varepsilon^{-2}\boldsymbol{x}_0'V(\hat{\boldsymbol{\beta}})\boldsymbol{x}_0 + \frac{\left((\sigma_\varepsilon^2)^{-\frac{3}{2}}\right)^2}{4}(Y_0 - \boldsymbol{x}_0'\boldsymbol{\beta})^2 V(\hat{\sigma}_\varepsilon^2)$$

$$+ \sigma_\varepsilon^{-1}(\sigma_\varepsilon^2)^{-\frac{3}{2}}(Y_0 - \boldsymbol{x}_0{}'\boldsymbol{\beta})\boldsymbol{x}_0'Cov(\hat{\boldsymbol{\beta}}, \hat{\sigma}_\varepsilon^2).$$

Note that $Cov(\hat{\boldsymbol{\beta}}, \hat{\sigma}_\varepsilon^2) = 0$ for the following reasons: (a) $Cov(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\varepsilon}}) = 0$ (Johnson & Wichern, 1998, p. 387); (b) $\hat{\sigma}_\varepsilon^2$ is a function of $\hat{\boldsymbol{\varepsilon}}$; and (c) Theorem 4.3.5, p. 161 in Casella and Berger (2002); that is, if two random variables X and Y are independent (here, $X = \hat{\boldsymbol{\beta}}$ and $Y = \hat{\boldsymbol{\varepsilon}}$) and $f(\cdot)$ and $g(\cdot)$ are two functions, then $f(X)$ and $g(Y)$ are independent (here, $f(\cdot)$ is the identity function, and $g(\cdot)$ is $\hat{\sigma}_\varepsilon^2$). Furthermore, recall that $V(\hat{\boldsymbol{\beta}}) = \sigma_\varepsilon^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $V(\hat{\sigma}_\varepsilon^2) = \sigma_\varepsilon^4\frac{2}{N-k-1}$ (Johnson & Wichern, 1998, pp. 389–390). Thus, $V(\hat{Z}_0)$ can be simplified as follows

$$V(\hat{Z}_0) \approx \sigma_\varepsilon^{-2}\boldsymbol{x}_0'V(\hat{\boldsymbol{\beta}})\boldsymbol{x}_0 + \frac{\left((\sigma_\varepsilon^2)^{-\frac{3}{2}}\right)^2}{4}(Y_0 - \boldsymbol{x}_0'\boldsymbol{\beta})^2 V(\hat{\sigma}_\varepsilon^2)$$

$$= \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0 + \frac{1}{\sigma_\varepsilon^2}\frac{1}{2(N-k-1)}(Y_0 - \boldsymbol{x}_0'\boldsymbol{\beta})^2$$

$$= \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0 + \frac{1}{2(N-k-1)}Z_0^2.$$

which is Equation 7.

Sampling Variance of $PR(\hat{Z}_0)$

The sampling variance of Equation 6 is derived as follows. First, note that $\frac{\partial PR(\hat{Z}_0)}{\partial \hat{Z}_0}\Big|_{\hat{Z}_0=E(\hat{Z}_0)} = 100 \times \phi(\hat{Z}_0)\Big|_{\hat{Z}_0=E(\hat{Z}_0)} \approx 100 \times \phi(Z_0)$, where $\phi(\cdot)$ is the probability density function of the standard normal distribution, and the last equality follows from $E(\hat{Z}_0) \approx Z_0$ (see previous paragraph). Second, Equation 6 can be approximated with the following first-order Taylor series evaluated at $E(\hat{Z}_0)$

$$PR(\hat{Z}_0) = g(E(\hat{Z}_0)) + \left.\frac{\partial PR(\hat{Z}_0)}{\partial \hat{Z}_0}\right|_{\hat{Z}_0=E(\hat{Z}_0)}(\hat{Z}_0 - E(\hat{Z}_0))$$

$$\approx 100 \times \left[\Phi(Z_0) + \phi(Z_0)(\hat{Z}_0 - Z_0)\right].$$

Taking the expectation and variance of the previous expression with respect to $\hat{Z}_0$, one obtains

$$E(PR(\hat{Z}_0)) \approx \Phi(Z_0) \times 100,$$

and the sampling variance of $PR(\hat{Z}_0)$, approximating $V(\hat{Z}_0)$ by Equation 7, is

$$V(PR(\hat{Z}_0)) \approx 100^2\phi(Z_0)^2 V(\hat{Z}_0)$$

$$= 100^2\phi(Z_0)^2\left[\boldsymbol{x}_0{}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0 + \frac{1}{2(N-k-1)}Z_0^2\right],$$

which is Equation 8.

(*Appendices continue*)

Appendix B

Summary of the Results of the Simulation Studies

Simulation Design

A simulation study was needed because the true $V(\hat{Z}_0)$ and $V(PR(\hat{Z}_0))$ were unknown. The true $V(\hat{Z}_0)$ and $V(PR(\hat{Z}_0))$ were generated in three steps: (a) a normative sample was drawn from the reference population, (b) the model parameters were estimated using the normative sample, and (c) the raw outcome of an individual (not belonging to the normative sample) was translated into an estimated Z-score and a PR-score. These three steps were repeated $20,000$ times, and the true $V(\hat{Z}_0)$ and $V(PR(\hat{Z}_0))$ were computed as the variance of $\hat{Z}_0$ and $PR(\hat{Z}_0)$ over the $20,000$ generated normative samples. Having generated the true $V(\hat{Z}_0)$ and $V(PR(\hat{Z}_0))$, the relative bias of Equations 7–10 could be computed (for details, see online Supplement A). The factors involved in the simulation studies were:

- The reference population as expressed by the regression model used for norming: The five considered models were Equations 1–5.
- The true values of the Z-scores and PR-scores: The considered true Z-scores were all values from -3 to 3 with increment 0.5, yielding 13 values. The considered PR-scores were not those corresponding to the 13 Z-scores but $PR_0 \in \{1, 2.5, 5, 10, 90, 95, 97.5, 99\}$, because the latter set of values is more often used in practice.
- The distribution of the predictors (i.e., age and sex) in the normative sample: For age all the values from 20 to 80, with Step 5, were considered, giving $13 \times 2 = 26$ combinations of age and sex.
- The sample size N: Four values were considered $N = \{338, 676, 1{,}690, 3{,}380\}$. The first value $N = 338$ was obtained by combining each Z-score value with each age-sex combination. The other three values were obtained by replicating each age-sex-Z-score combination by a factor of two, five, and ten, respectively.
- The true regression parameter values: Three sets of true values were chosen (see Table S.A.2, online Supplement A). It turned out that these true values did not affect the bias of Equations 7–8. Hence, only one set of values was used in the simulation study for assessing the bias in Equations 9–10.

Simulation Results

Z-Scores

In all the considered scenarios, the relative bias of $V(\hat{Z}_0)$ (that is, Equation 7) was always within the interval $(-3\%, 3\%)$, showing no clear patterns with respect to the predictors or the sample size (see Figures S.A.3-S.A.4, online Supplement A). Likewise, the relative bias of $\hat{V}(\hat{Z}_0)$ (that is, Equation 9) was always within the range $(-3\%, 3\%)$ in all scenarios (see Figures S.A.11-S.A.12, online Supplement A). Furthermore, the absolute bias of $\hat{Z}$ was within the interval $(-0.01, 0.01)$ for any Z-score, age by sex combination, sample size, and model (see Figures S.A.5-S.A.6, online Supplement A). Finally, in all the considered scenarios, the coverage of 95% confidence intervals for Z_0 obtained using $\hat{V}(\hat{Z}_0)$ was close to the nominal coverage probability (i.e., $95\% \pm 0.5\%$) for any Z-score (see Figures S.A.13-S.A.14, online Supplement A). Given these good results, $N = 338$ might be a conservative lower-bound for Equations 7 and 9, and one might expect that smaller sample sizes still might yield acceptable bias (i.e., bias $\in [-5\%, 5\%]$) and coverage (i.e., nominal coverage probability $\pm 1\%$). However, note that most of the required sample sizes for detecting the smallest "clinically relevant" difference in Table 4 are larger than $N = 338$, and the few sample sizes smaller than $N = 338$ are, in most cases, close to this lower-bound. So, considering sample sizes much smaller than $N = 338$ (say, $N < 280$) was not of interest.

PR-Scores

The relative bias of $V(PR(\hat{Z}_0))$ (that is, Equation 8) was a decreasing function of the sample size, an increasing function of model complexity (that is, from Model 1 to Model 5), and was acceptable (i.e., relative bias $\in [-5\%, 3\%]$) if $N \geq 1{,}690$ for Models 3–5 or if $N \geq 676$ for Models 1–2, except for combinations of extreme values of age (i.e., age = 20 and 80) with extreme PR-scores (PR = 1 or 99) (see Figures S.A.7-S.A.8, online Supplement A). The relative bias of $V(PR(\hat{Z}_0))$ increased as age and/or the true PR-score became more extreme, but was not affected by sex. The relative bias of $\hat{V}(PR(\hat{Z}_0))$ (that is, Equation 10) was also a decreasing function of sample size, and was always within the interval $[-3\%, 10\%]$ for $N \geq 676$ (see Figures S.A.15-S.A.16, online Supplement A). The bias was within the interval $[-3\%, 5\%]$ under Model 1 for $N \geq 676$, and under all other models for $N \geq 1{,}690$. The absolute bias of $PR(\hat{Z}_0)$ (that is, Equation 6) was a decreasing function of the sample size, and was always within the interval $[-0.1, 0.1]$, on the scale from 0 to 100, if $N \geq 1{,}690$ (see Figures S.A.9-S.A.10, online Supplement A). Furthermore, $PR(\hat{Z}_0)$ tended to overestimate $PR(Z_0)$ if $PR_0 < 50$ and underestimate it if $PR_0 > 50$. Finally, the coverage of the 95% confidence interval for PR_0 obtained using Equation 10 was a decreasing function of model complexity, and was always within the interval [93%, 96%] for $N \geq 676$, and within the interval [94%, 96%] for $N \geq 1{,}690$ (see Figures S.A.17-S.A.18, online Supplement A).

(*Appendices continue*)

Appendix C

Derivations of the Sample Size Calculation Formulas (That Is, Equations 14 and 15)

In this appendix, the formulas to determine the required sample size N^* to detect the smallest "clinically relevant" difference between individual's true Z- or PR-score (Z_t or PR_t) and the cut-off (Z_c or PR_c), δ, are derived as functions of the type I error rate α, and the statistical power $1 - \gamma$ (i.e., to reject the null hypothesis H_0 that $Z_t = Z_c$, for Z-scores, and $PR_t = PR_c$, for PR-scores). These formulas are derived under the optimal design for the normative sample, that is, Equations 12 and 13 are used.

Z-Scores

The true Z-score Z_t differs from the cut-off Z_c by $\delta > 0$, that is, either $Z_t = Z_c - \delta$ or $Z_t = Z_c + \delta$. From the Delta method it follows that the sampling distribution of $\hat{Z}_0$ is (approximately) normal with mean Z_0 (with $Z_0 = Z_c$ under H_0, and $Z_0 = Z_t$ under H_1), and variance as given in Equation 7 (Casella & Berger, 2002, p. 245). Under the optimal design, the sampling variance of $\hat{Z}_0$ can be approximated with Equation 12 which is denoted by $V(\hat{Z}_0|Z_c)^*$ if $Z_0 = Z_c$, and $V(\hat{Z}_0|Z_t)^*$ if $Z_0 = Z_t$. Note that $\frac{\hat{Z}_0 - Z_c}{(V(\hat{Z}_0|Z_c)^*)^{1/2}}$ follows (approximately) a standard normal distribution if $Z_0 = Z_c$ and likewise for $\frac{\hat{Z}_0 - Z_t}{(V(\hat{Z}_0|Z_t)^*)^{1/2}}$ if $Z_0 = Z_t$. For the sake of brevity, the case of $Z_t = Z_c + \delta$ is considered (but the same result, that is, Equation 14, is obtained for $Z_t = Z_c - \delta$). Denote by z_α and $z_{1-\alpha}$ the $100(\alpha)\%$ and $100(1-\alpha)\%$ percentile of the standard normal distribution, respectively. The power $1 - \gamma$ of rejecting the null hypothesis H_0 that $Z_t = Z_c$ is defined as follows

Step 1.

$$1 - \gamma = P(\text{reject } H_0 \,|\, Z_t = Z_c + \delta)$$
$$= P\left(\frac{\hat{Z}_0 - Z_c}{(V(\hat{Z}_0 \,|\, Z_c)^*)^{1/2}} > z_{1-\alpha} \,\middle|\, Z_t = Z_c + \delta\right).$$

Step 2.

$$P\left(\frac{\hat{Z}_0 - Z_c}{(V(\hat{Z}_0 \,|\, Z_c)^*)^{1/2}} > z_{1-\alpha} \,\middle|\, Z_t = Z_c + \delta\right)$$

$$= P\left(\frac{\hat{Z}_0 - Z_t + Z_t - Z_c}{(V(\hat{Z}_0 \,|\, Z_t)^*)^{1/2}} > z_{1-\alpha}\frac{(V(\hat{Z}_0 \,|\, Z_c)^*)^{1/2}}{(V(\hat{Z}_0 \,|\, Z_t)^*)^{1/2}} \,\middle|\, Z_t = Z_c + \delta\right)$$

$$= P\left(\frac{\hat{Z}_0 - Z_t}{(V(\hat{Z}_0 \,|\, Z_t)^*)^{1/2}} > z_{1-\alpha}\frac{(V(\hat{Z}_0 \,|\, Z_c)^*)^{1/2}}{(V(\hat{Z}_0 \,|\, Z_t)^*)^{1/2}} - \frac{\delta}{(V(\hat{Z}_0 \,|\, Z_t)^*)^{1/2}} \,\middle|\, Z_t = Z_c + \delta\right),$$

where the last equality follows from $\delta = Z_t - Z_c$.

Step 3. The desired power level $1 - \gamma$ is obtained when

$$-z_{1-\gamma} = z_{1-\alpha}\frac{(V(\hat{Z}_0 \,|\, Z_c)^*)^{\frac{1}{2}}}{(V(\hat{Z}_0 \,|\, Z_t)^*)^{\frac{1}{2}}} - \frac{\delta}{(V(\hat{Z}_0 \,|\, Z_t)^*)^{1/2}},$$

where $z_{1-\gamma}$ is the $100(1-\gamma)\%$ percentile of the standard normal distribution. This equality in turn can be rewritten using Equation 12 as follows

$$-z_{1-\gamma} = z_{1-\alpha}\frac{\left(\frac{k+1+\frac{Z_c^2}{2}}{N}\right)^{\frac{1}{2}}}{\left(\frac{k+1+\frac{Z_t^2}{2}}{N}\right)^{\frac{1}{2}}} - \frac{\delta}{\left(\frac{k+1+\frac{Z_t^2}{2}}{N}\right)^{1/2}},$$

from which the following expression for the required sample size N^* (that is, Equation 14) is obtained

$$N^* = \left[\frac{z_{1-\alpha}\left(k+1+\frac{Z_c^2}{2}\right)^{1/2} + z_{1-\gamma}\left(k+1+\frac{Z_t^2}{2}\right)^{1/2}}{\delta}\right]^2,$$

which holds for $Z_t = Z_c + \delta$ and can also be shown to hold for $Z_t = Z_c - \delta$.

PR-Scores

The true PR-score PR_t differs from the cut-off PR_c by $\delta > 0$, that is, either $PR_t = PR_c - \delta$ or $PR_t = PR_c + \delta$. From the Delta method it follows that the sampling distribution of $PR(\hat{Z}_0)$ is (approximately) normal with mean $PR(Z_0)$ (with $PR(Z_0) = PR_c$ under H_0, and $PR(Z_0) = PR_t$ under H_1), and variance as given in Equation 8. Under the optimal design, the sampling variance of $PR(\hat{Z}_0)$ can be approximated with Equation 13 which is denoted by $V(PR(\hat{Z}_0) \,|\, PR_c)^*$ if $Z_0 = Z_{PR_c} = \Phi^{-1}(\frac{PR_c}{100})$ (i.e., the Z-score corresponding to PR_c), and $V(PR(\hat{Z}_0) \,|\, PR_t)^*$ if $Z_0 = Z_{PR_t} = \Phi^{-1}(\frac{PR_t}{100})$ (i.e., the Z-score corresponding to PR_t), where $\Phi^{-1}(\cdot)$ is the inverse of the cumulative standard normal distribution. Note that $\frac{PR(\hat{Z}_0) - PR_c}{(V(PR(\hat{Z}_0) \,|\, PR_c)^*)^{1/2}}$ follows (approximately) a standard normal distribution if $PR(Z_0) = PR_c$ and likewise for $\frac{PR(\hat{Z}_0) - PR_t}{(V(PR(\hat{Z}_0) \,|\, PR_t)^*)^{1/2}}$ if $PR(Z_0) = PR_t$. Only the case of $PR_t = PR_c + \delta$ is considered (but the same result, that is, Equation 15,

(Appendices continue)

is obtained for $PR_t = PR_c - \delta$). Following the same Steps 1–3 as for Z-scores, but now using Equations 8 and 13 instead of Equations 7 and 12, respectively, one gets that the power $1 - \gamma$ of rejecting the null hypothesis H_0 that $PR_t = PR_c$ is obtained when

$$-z_{1-\gamma} = z_{1-\alpha}\frac{(V(PR(\hat{Z}_0)\,|\,PR_c)^*)^{1/2}}{(V(PR(\hat{Z}_0)\,|\,PR_t)^*)^{1/2}} - \frac{\delta}{(V(PR(\hat{Z}_0)\,|\,PR_t)^*)^{1/2}},$$

which can be rewritten using Equation 13 as follows

$$-z_{1-\gamma} = z_{1-\alpha}\frac{\left(100^2\phi(Z_{PR_c})^2\left[\frac{k+1+\frac{Z^2_{PR_c}}{2}}{N}\right]\right)^{\frac{1}{2}}}{\left(100^2\phi(Z_{PR_t})^2\left[\frac{k+1+\frac{Z^2_{PR_t}}{2}}{N}\right]\right)^{\frac{1}{2}}} - \frac{\delta}{\left(100^2\phi(Z_{PR_t})^2\left[\frac{k+1+\frac{Z^2_{PR_t}}{2}}{N}\right]\right)^{1/2}},$$

from which the following expression for the required sample size N^* (that is, Equation 15) is obtained

$$N^* = \left[\frac{z_{1-\alpha}\cdot 100\cdot\phi(Z_{PR_c})\left(k+1+\frac{Z^2_{PR_c}}{2}\right)^{1/2} + z_{1-\gamma}\cdot 100\cdot\phi(Z_{PR_t})\left(k+1+\frac{Z^2_{PR_t}}{2}\right)^{1/2}}{\delta}\right]^2.$$

which holds for $PR_t = PR_c + \delta$ and can also be shown to hold for $PR_t = PR_c - \delta$.

Received July 7, 2020
Revision received December 8, 2020
Accepted January 4, 2021 ■