A tutorial on regression-based norming of psychological tests with GAMLSS

Timmerman, Marieke E.; Voncken, Lieke; Albers, Casper J.

Published in:
Psychological Methods

DOI:
10.1037/met0000348

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

ISSN: 1082-989X

2021, Vol. 26, No. 3, 357-373
http://dx.doi.org/10.1037/met0000348

A Tutorial on Regression-Based Norming of Psychological Tests With GAMLSS

Marieke E. Timmerman, Lieke Voncken, and Casper J. Albers
University of Groningen

Abstract

A norm-referenced score expresses the position of an individual test taker in the reference population, thereby enabling a proper interpretation of the test score. Such normed scores are derived from test scores obtained from a sample of the reference population. Typically, multiple reference populations exist for a test, namely when the norm-referenced scores depend on individual characteristic(s), as age (and sex). To derive normed scores, regression-based norming has gained large popularity. The advantages of this method over traditional norming are its flexible nature, yielding potentially more realistic norms, and its efficiency, requiring potentially smaller sample sizes to achieve the same precision. In this tutorial, we introduce the reader to regression-based norming, using the generalized additive models for location, scale, and shape (GAMLSS). This approach has been useful in norm estimation of various psychological tests. We discuss the rationale of regression-based norming, theoretical properties of GAMLSS and their relationships to other regression-based norming models. Based on 6 steps, we describe how to: (a) design a normative study to gather proper normative sample data; (b) select a proper GAMLSS model for an empirical scale; (c) derive the desired normed scores for the scale from the fitted model, including those for a composite scale; and (d) visualize the results to achieve insight into the properties of the scale. Following these steps yields regression-based norms with GAMLSS for a psychological test, as we illustrate with normative data of the intelligence test IDS-2. The complete *R* code and data set is provided as online supplemental material.

Translational Abstract

Standardized psychological tests are widely used. Examples include intelligence, developmental, and neuropsychological tests. They are used for purposes as monitoring, selection, and diagnosing individuals. High-quality standardized tests have normed scores, like the well-known IQ scores for intelligence tests. Normed scores allow for properly interpreting an individual's test score. They are derived in the test construction phase, based on scores in a large normative sample. Normed scores express the position of an individual test taker in the reference population. The reference population for a test typically depends on individual characteristic(s), like age and possibly sex. This tutorial introduces the reader to a method to compute normed scores that depend on individual characteristic(s), making optimal use of all background knowledge and the scores in the whole normative sample. Therefore, the method yields potentially more realistic norms, and more precise norms than traditional methods, using the same amount of data. This is an important asset, because gathering sufficient data is difficult and costly. In this tutorial, we explain the technical background of the method, called regression-based norming with the generalized additive models for location, scale, and shape (GAMLSS), and explain how to apply it based on six steps. Following these steps yield regression-based norms with GAMLSS for a psychological test, as we illustrate with normative data of the intelligence test IDS-2. The complete R code and data set is provided as online supplemental material, so that test developers can apply the method to derive high-quality norms for their own test.

Keywords: continuous norming, norm-referenced scores, norm generation, relative norming

Supplemental materials: http://dx.doi.org/10.1037/met0000348.supp

This article was published Online First August 27, 2020.

Marieke E. Timmerman, Lieke Voncken, and Casper J. Albers, Department of Psychometrics and Statistics, University of Groningen.

The data described in this article are scores on Test 14 ("naming antonyms") and Test 7 ("naming categories") of the intelligence test IDS-2. The data stem from the normative samples of the Dutch IDS-2 (Grob, Hagmann-von Arx, Ruiter, Timmerman, & Visser, 2018) and the German IDS-2 (Grob & Hagmann-von Arx, 2018). The authors thank Alexander Grob and Antonia Hogrefe for providing them with the IDS-2 normative data. Preliminary ideas of those discussed in this article were presented at the 15th European Congress of Psychology (2017, Amsterdam, the Netherlands by Timmerman, M. E., Voncken, L. & Ruiter, S. A. J. High quality IQ-scores are based on Continuous norming).

Correspondence concerning this article should be addressed to Marieke E. Timmerman, Department of Psychometrics and Statistics, University of Groningen, Grote Kruisstraat 2/1, 9712TS Groningen, the Netherlands. E-mail: m.e.timmerman@rug.nl

Standardized psychological tests are widely used. They play a crucial role in individual assessments and in psychological research. Individual assessments take place in clinical, developmental, and personal psychology practice, with the aim of individual diagnosis, monitoring, or selection. Assessments can have a major impact on individuals' lives because important decisions, for example on clinical interventions, are based on them. Thus, it is essential to have high-quality tests.

A core feature of high-quality standardized tests is that they have well-normed scores. The majority of psychological tests have norm-referenced scores, including intelligence tests (e.g., Wechsler, 2008), developmental tests (e.g., Bayley, 2006), neuropsychological tests (e.g., Kaufman & Kaufman, 1994), and clinical tests (e.g., Goodman, 1997). In a norm-referenced test, the raw score is transformed into a norm-referenced score, which expresses an individual's performance relative to performances in the reference population of the test (e.g., Groth-Marnat & Wright, 2016, pp. 11–12; Mellenbergh, 2011, p. 348). This contrasts to a criterion-referenced test, in which the performance is compared against a predetermined standard. The vast majority of normed psychological tests make use of norm-referenced scores.

A salient feature of norm-referenced psychological tests is that the norms typically depend on age (e.g., in intelligence tests as Wechsler, 2014) and sometimes sex and/or educational level (e.g., in neuropsychological tests as Rommelse et al., 2018). This implies that there are in fact multiple reference populations for which norms are needed, which jointly make up the norm population of the test. The norms are established during the test construction phase, on the basis of scores collected in a sample of the norm population.

In this tutorial we focus on obtaining high-quality norms for a norm-referenced test, in which the norms may depend on individual characteristic(s) as age and sex. In what follows, we denote these variables on which norms depend as norm-predictors. In recent years, continuous norming (Zachary & Gorsuch, 1985) has been embraced by test constructors to compute their norms (e.g., van Baar, Steenis, Verhoeven, & Hessen, 2014; Wechsler, 2014), because of its favorable properties in terms of accuracy. The key of continuous norming is that one explicitly uses the information provided by the continuous (or ordered categorical) nature of the norm-predictor(s; like age or educational level) in computing the norms. Continuous norming can also be used when one or more discrete norm-predictors are involved. Even for a single discrete norm-predictor (e.g., sex only) continuous norming can be applied, but then the advantage over traditional norming becomes much smaller.

One can distinguish three types of continuous norming: semiparametric norming (Lenhard, Lenhard, Suggate, & Segerer, 2018; Snijders, Tellegen, & Laros, 1988), inferential norming (Angoff & Robertson, 1987; Zachary & Gorsuch, 1985; Zhu & Chen, 2011), and regression-based norming (Van Breukelen & Vlaeyen, 2005; Voncken, Albers, & Timmerman, 2019b). Having normed scores for a test means that one has the normed score available for each possible combination of raw score and norm-predictor value.

In semiparametric norming, one models the raw scores as a smooth, typically nonlinear function of the norm-predictor values (e.g., age) and the normed scores (e.g., percentiles). In the modeling process, one first discretizes the norm-predictor values into groups (e.g., per year of age) and then regresses the raw scores on the discretized norm-predictors.

In inferential norming, the raw score distribution is estimated using a two-step procedure: First, the mean, standard deviation, skewness, and sometimes kurtosis are separately fitted with polynomial regressions as a function of the norm predictor(s), using discretized norm-predictor values into groups; the polynomial functions are possibly manually adapted based on expert knowledge. Second, the estimated parameters are used as parameters for some parametric distribution that is deemed suitable. The resulting normed scores are hand smoothed, both within and between groups.

Regression-based norming has two great advantages over semiparametric norming and inferential norming: First, it has readily available statistical criteria for model selection and model assessment. Second, in the computations there is no need for discretizing a continuous variable as age, thereby avoiding the arbitrary and possibly influential decision on the interval width (see Lenhard et al., 2018; Zhu & Chen, 2011). We focus on regression-based norming using the generalized additive models for location, scale, and shape (GAMLSS; Rigby & Stasinopoulos, 2005; Stasinopoulos, Rigby, Heller, Voudouris, & De Bastiani, 2017). This highly versatile model family is suitable for a wide range of empirical norming cases and encompasses—as far as we know—all specific models used so far in regression-based norming.

To the best of our knowledge, there is no concise and clear introduction aimed at test constructors that explains regression-based norming. This article thus aims at (a) providing an overview of all relevant aspects in regression-based norming with GAMLSS for an empirical scale, deriving the norm-referenced scores from the model and visualizations to achieve insight into the properties of the scale; (b) outlining the commands used in *R* to achieve the regression-based norming; and (c) describing the most common issues encountered in norming practice. We do so on the basis of six steps, which are presented in Table 1. Following these steps yields appropriate regression-based norms with GAMLSS for a psychological test, as we will illustrate using an intelligence test. We further elaborate on the sensitivity to GAMLSS model spec-

Table 1
Six Steps to Arrive at Regression-Based Norms With GAMLSS for a Psychological Test

Step	Description
1	Define the test's reference population(s), norm population and target population
2	Design and carry out the study to gather the normative sample data
3	Select a candidate GAMLSS distribution as a conditional raw score model
4	Select candidate function(s) to relate the norm-predictor(s) to the GAMLSS distribution parameters
5	Carry out the model selection to arrive at the estimated GAMLSS model
6	Compute the normed scores for a scale based on the estimated GAMLSS model
(6.1)	(Compute the normed scores for a composite scale)

Note. GAMLSS = generalized additive models for location, scale, and shape.

ification. Before discussing the six steps and their background into detail, we introduce the core ideas of regression-based norming and compare it to traditional norming.

This tutorial builds on recent developments in GAMLSS for regression-based norming (Oosterhuis, van der Ark, & Sijtsma, 2016; Voncken, Kneib, Albers, Umlauf, & Timmerman, 2020; Voncken, Albers, & Timmerman, 2020; Voncken, Albers, & Timmerman, 2019a; Voncken et al., 2019b) and the experiences we gained in norming various psychological tests (Grob & Hagmann-von Arx, 2018; Grob et al., 2018; Rommelse et al., 2018; Tellegen & Laros, 2017; Voncken, Timmerman, Spikman, & Huitema, 2018).

Traditional Versus Regression-Based Norming

When norms depend on, for example, age, there are in fact multiple reference populations for which norms are needed. Per reference population (e.g., the general population at a specific age), the norms themselves are derived from the observed distribution of the raw test scores within the reference population concerned. The traditional norming approach is to define the various reference populations by categorizing the continuous variable age and computing the norms per age interval (and—if applicable—per sex and/or educational level; as e.g., in the Wechsler Intelligence Scale for Children-III, Wechsler, 2014). The traditional norming approach is problematic in that it easily yields jumps between norms at successive age intervals. This implies that the very same raw test score would be interpreted rather differently for two individuals who are in two successive age intervals, yet are of almost the same age (e.g., 12 years and 364 days, vs. 13 years and 1 day). These jumps are often unrealistic, namely when they conflict with theoretical knowledge on the construct measured (Van Breukelen & Vlaeyen, 2005; Zachary & Gorsuch, 1985), that is, that the test score distribution changes smoothly with increasing age.

Traditional norming uses age intervals, thereby categorizing age. It thus relies on the (implicit) assumption that the test score distribution is constant within each age interval. In case of a smoothly changing test distribution with age, this assumption holds to a sufficient degree only when using narrow age intervals. This phenomenon is illustrated in Figure 1, which depicts the raw test scores as a function of age in an illustrative normative sample (N = 1,660) of Test 14 ("naming antonyms") of the intelligence test IDS-2. To ensure confidentiality of the original IDS-2 norms, this sample is composed of random samples (n = 830 each) of the normative sample of the Dutch IDS-2 (Grob et al., 2018) and the German IDS-2 (Grob & Hagmann-von Arx, 2018). Because the score distributions across age are rather similar in these normative samples, they may be combined, at least for illustration purposes. The horizontal line in Figure 1, panel A, indicates the estimated population median and the 95% confidence interval (CI) per age interval, for an interval width of four years. Especially in the lower age ranges, the estimated median appears to be a biased estimate at almost all ages within the interval. In panel B, the interval width is reduced to one year. As can be seen by comparing panels A and B, narrowing the interval widths reduces the bias as well as the jumps between different age categories, yet increases the width of the 95% CI considerably. Thus, using narrow intervals would substantially increase the total needed sample size to achieve sufficient precision for each age interval. The unrealistic jumps between norms in successive age intervals observed in traditional norming are thus due to model error and/or sampling fluctuations.

The approach used in regression-based norming is to model the raw test score distribution as a continuous function of age. This allows for the evaluation of the test score distribution at any specific age, thereby referring to the reference population of that age. The continuous function of age properly reflects the theoretical knowledge on the development measured. For example, the estimated population median of Test 14, based on a regression model with a continuous nonlinear function of age, smoothly increases with age (see Figure 1, panel C). Thus, this model for the median is more realistic than the one associated with traditional norming (Figure 1, panels A–B). Further, the model is estimated based on the total sample, rather than subsamples with ages in the same interval, implying that the estimation is statistically more efficient (i.e., requiring a smaller total sample size to achieve the same precision in estimation). Finally, in regression-based norming, the estimation takes place assuming a particular parametric distribution for the raw scores. Provided that the assumption holds, then the estimation is more efficient than with traditional norming, where the empirical distribution is being used.

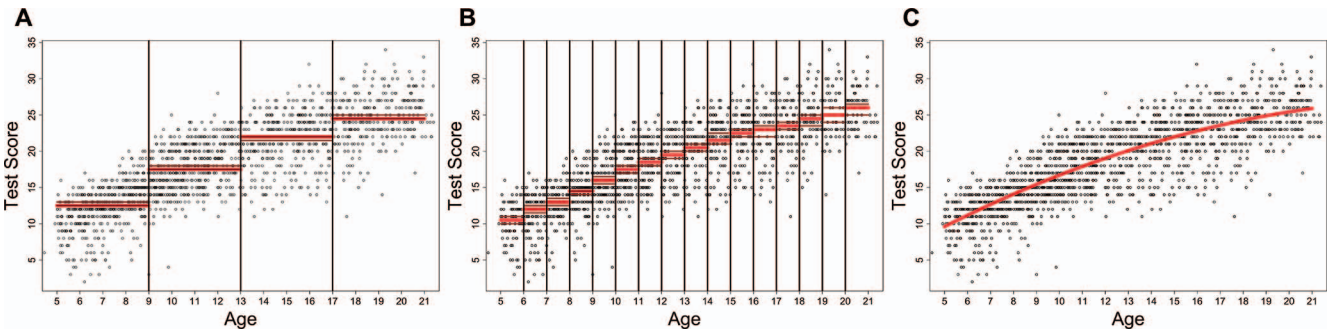

Figure 1. Scores of the an illustrative normative sample (N = 1,660) of Test 14 of the IDS-2 as a function of age, with the estimated median per age interval of 4 years (panel A) and 1 year (panel B) and the bounds of the 95% CIs of the median test score—as used in traditional norming, and as a continuous nonlinear function of age—as used in regression-based norming—(panel C). See the online article for the color version of this figure.

Regression-Based Norming With GAMLSS

Regression-based norming with GAMLSS can be carried according to the six steps presented in Table 1. In the next sections, we describe each of the six steps, their theoretical background and guidelines on how to implement these steps properly in empirical practice to obtain relative norms for a psychological test.

Step 1: Define the Test's Reference Population(s), Norm Population, and Target Population

Normed scores are expressed in comparison to the reference population of the test. For example, the reference population can be the general population of a country with the same age as the testee involved. All reference populations jointly (e.g., all ages within the age range of the target population of a test) make up the norm population of the test. The essence of regression-based norming is that one models the distribution of the raw test scores conditional upon the norm-predictor(s). The norm-predictor(s) are the predictor(s) that define the reference population(s). Only the norm-predictor(s) must be included in the norming model, even if other predictors exist that would be related to the test scores. This is so because otherwise the reference population would change and thus the interpretation of the normed scores would alter. Further, each norm-predictor should relate to the test score distribution, to increase the estimation efficiency.

For example, suppose that the reference population of a test is the community population of the same age and sex. Thus, the norm-predictors are age and sex. Suppose further that the test score distribution would be dependent on age and educational level and independent on sex. Then, only the norm-predictor age is to be used in the regression model. Including the predictor educational level would completely change the reference population to the community population of the same age, educational level, and sex, and thereby also fundamentally change the interpretation of the normed scores. Including the norm-predictor sex as a predictor in the norming model would not make a difference at the population level, but would make the estimation less efficient at the sample level. In regression-based norming, the relationship of predictors with the test score distribution received much attention (Oosterhuis et al., 2016; Van Breukelen & Vlaeyen, 2005); yet, the essential relationship with the reference population was neglected.

The reference population crucially depends on the test purpose. For example, intelligence test norms typically depend on age only, implying that the reference population is of the same age. The normed scores thus express the intellectual ability relative to individuals of the same age. In this way, the changes in test performances due to natural development across the life span are accounted for. This seems to be most useful for the test purposes of monitoring and selection. As a side-note: Raw intelligence test scores typically correlate with both age and educational level. Because only age is accounted for in the normed intelligence test scores, the latter will correlate with educational level.

Neuropsychological test norms typically depend on age, sex, and (for adults) educational level (Rommelse et al., 2018; e.g., Voncken et al., 2018). For clinical diagnosis, it is needed to detect normalities and abnormalities in performance across relevant performance domains; herewith (ab)normality is defined in comparison to healthy individuals. When test performance among healthy individuals is known to depend on age, sex, and/or educational level, this should be accounted for in the norms, to make the reference population as similar as possible as the patient under clinical assessment. This appears most useful for the test purpose of clinical diagnosis.

In certain cases, it is not eminently clear what yields the most informative norms, and different test constructors make different choices for seemingly related test purposes. For example, the occurrence of internalizing problems among children changes with age (Durbeej et al., 2019) and are displayed more frequently among girls than boys (Becker et al., 2018). The test score distribution of a test measuring internalizing problems thus relates to age and sex. Norms for such a test can thus be made age and/or sex specific, depending on the purpose of the test. Severity of internalizing problems can be investigated with the Child Behavior Checklist (CBCL; Achenbach, 1991) and the Strength and Difficulties Questionnaire (SDQ; Goodman, 1997). The norms of both tests depend on age, but interestingly sex specific norms are only available for the CBCL. Obviously, this choice has repercussions for the practical use of the test. For example, a screening with the SDQ on internalizing problems (e.g., select the top 10%) would result in the identification of more girls than boys at risk.

The target population of a test is the population for which a test is suitable. The norm population of a test (i.e., all reference populations of the test jointly) may be equal to the target population or may be a part of the target population. For example, an intelligence test aimed at the community population has the same target and norm populations. A neuropsychological test typically has as the norm population healthy individuals, while the target population of the test includes both healthy and unhealthy individuals.

Step 2: Design the Study to Gather the Normative Sample Data

A well-composed normative sample is crucial to achieve correct norms. Correct norms are unbiased and sufficiently precise. These can only be achieved with unbiased and sufficiently precise estimated distribution(s) of the raw test scores conditional upon the norm-predictor(s). We first consider the issue of achieving unbiased estimates and then the issue of estimation precision.

As is well-known, an unbiased estimate can be obtained based on a random sample from the population. In norming, random sampling from the population typically is an unattainable ideal. In many cases, the individual members of the population are unknown, rendering it impossible to draw a random sample. For example, members of a clinical population are not listed. In other cases, privacy regulations, like the European General Data Protection Regulation (European Parliament and Council of the European Union, 2016), preclude random sampling. For example, most countries have a population register that would be of use to sample from a community population, but the population register is not accessible to researchers. In some cases, random sampling seems feasible using hierarchical sampling. This is possible when the higher-level units are publicly known and thus can be sampled from. For example, organizations as schools and hospitals within a certain country are often neatly listed. However, even when a random sample could be approached, there remains a serious threat, namely nonresponse bias. In practice, it is common that a

substantial part of the invited participants does not participate. If the resulting missing data are missing not at random (see, e.g., van Buuren, 2018, pp. 8–9), then the estimates based on available data will be biased.

An alternative to random sampling that is practically feasible, is judgmental sampling (Mellenbergh, 2011, p. 351). Judgmental sampling is akin to stratified sampling, in that strata (subpopulations) are identified based on characteristics that relate to the test score and for which the population distribution is known. Then, sampling of individuals takes place such that the sampling distribution of the strata mimics those as known in the population. In judgmental sampling, unlike stratified sampling, the specific individuals selected result from the recruitment procedure applied, rather than from random sampling. Strata are often considered based on different variables, such as educational level, ethnicity, and region. Proper stratification would require the use of the joint distribution in the population, rather than their univariate distributions separately. Unfortunately, current norming practice seems to only consider the univariate distributions, thereby often leaving unclear whether there has been sampled under the assumption of independency (e.g., Wechsler, 2018). Though random sampling would be ideal from a theoretical perspective, a carefully designed and followed judgmental sampling scheme seems the best one could achieve in practice. Note that variables that determine the reference populations are of no use at all to define the strata. These variables are used as predictors in the regression model (i.e., the norm-predictors) and thus are conditioned upon. Because they are conditioned upon, this implies that a representative sample with respect to these variables is not needed. Note further that one should refrain from using the variables used to define the strata as norm-predictors (i.e., predictors in the norming model), because this would completely alter the nature of the reference populations and thus the interpretation of the normed scores.

Generally, the estimation precision depends crucially on the sample size. A larger normative sample thus results in more precision. In norming, the precision must be considered in terms of the estimate of the distributions conditional upon the predictors. Further, one needs observations across the full predictor space, to make sure that the model can be reliably estimated (i.e., the estimates are supported by data across the full predictor space). Currently, detailed knowledge is lacking on how large the sample size should be to achieve sufficient precision for regression-based norming. When standard linear regression is applicable, some guidance is available (Oosterhuis et al., 2016). In cases where more flexibility is required, larger samples sizes will be needed. It is highly desirable that generally applicable guidance becomes available, because typical norming models involve nonlinearity, non-normality, and heteroscedasticity. Such guidance should take into account the presumed nature of the model (e.g., models with more parameters typically require larger sample size to achieve the same precision) and minimally required precision.

For the time being, it seems wise to identify the minimum needed sample size based on standard linear regression (Oosterhuis et al., 2016) and use this as a lower boundary for the sample size needed. Furthermore, given that the model at the boundaries of predictor values suffer from limited precision, it is wise to collect some observations beyond these boundaries, provided that the test is suitable for the individuals involved. For example, the reference population of the IDS-2 is aged 5 to 20 years. Thus, to increase precision in norms at the age boundaries, one needs to include in the normative sample testees with an age slightly outside the boundaries. This is only useful in as far the IDS-2 is suitable for these individuals, which probably is more of a concern for the younger children than it is for the young adults.

Step 3: Select a Candidate GAMLSS Distribution as a Conditional Raw Score Model

The core step in regression-based norming is to get a proper model of the raw test score distribution as a function of the norm-predictor(s). The type of modeling required in norming is related to standard (linear) regression modeling, yet has two important differences. First, a standard linear regression model involves rather strict assumptions. The assumptions are that the relationships between predictor values and the distribution means are linear and that the residuals are independent and normally distributed with constant variance. The normality assumption of the residuals can also be described completely equivalently in terms of the normality assumption on the test score distribution conditional upon the predictors. In what follows, we adopt the terminology of distributions of test scores conditional upon predictor(s), because this complies perfectly with GAMLSS modeling to computed normed scores. In norming practice, the predictor(s) often are related in a nonlinear way to the test score distribution and the test score distribution conditional upon the predictors often deviates severely from homoscedastic normality. This thus calls for the use of a more flexible model than the standard linear regression model.

Second, in applying standard linear regression modeling, the typical goal is to either obtain a model that clarifies the effects of the predictors on the outcome variable for the population under study or to predict the value of future values of the predictor as accurately as possible. These goals imply that the interpretation of the specific model for the predictors is of key interest. Further, the explained variance needs to be substantial and thus the residual variance small. Both aspects sharply contrast to regression-based norming. There, the function of the predictor(s) in the model is solely to identify the reference populations (e.g., the community population of the very same age, across a certain age interval). Further, in regression-based norming, the spread in the distribution conditional upon the predictors needs to be substantial, because this spread expresses the individual differences in the construct to be measured by the psychological test. After all, a small spread would imply that the test distinguishes between individuals only to a minor extent.

To properly model the raw test score distribution in the normative sample, one needs a flexible modeling approach, which is found in GAMLSS. GAMLSS (Rigby & Stasinopoulos, 2005; Stasinopoulos et al., 2017) are univariate distributional regression models, in which all parameters of the assumed distribution can be modeled as additive functions of the predictor variable(s). For example, when using the normal distribution, one may model both the mean (μ) and standard deviation (σ) as a function of the predictor(s). GAMLSS can be fitted in *R* (R Core Team, 2019) using the *gamlss* package (Rigby & Stasinopoulos, 2005). The package includes over 100 continuous, discrete, and mixed classes of distributions for modeling the outcome variable. This offers great flexibility for modeling, making it rather likely that a prop-

erly fitting model for the raw test scores can be found. Further, also truncated versions of the distributions can be used, which allow for capturing floor and ceiling effects in raw test scores and taking into account theoretical minima and/or maxima in raw test scores. The flexibility makes GAMLSS eminently suitable for its use in norming.

To specify a GAMLSS model, one needs to select a likely fitting distribution. The distributions available and their properties are summarized in Stasinopoulos, Rigby, Heller, Voudouris, and De Bastiani (2017, pp. 58–63) and described in detail in Rigby, Stasinopoulos, Heller, and De Bastiani (2019). Suitable candidates for a distribution to use are identified by matching with the nature of the raw test scores, considering the score range, and considering their discrete or continuous character. In general, it is wise to take a distribution with as few parameters as possible, so as to avoid overfitting. Further, once the GAMLSS distribution has been selected, one needs to determine the so-called link function for each of the distributional parameters. The link function in combination with the distribution function completely determine the possible range of the parameter values. Highly popular functions are the identity link, which retains the parameter values in its original range, and the log link function, which restricts the parameter values to be nonnegative. Typically, the identity link function is used for the parameters μ (related to location) and ν (skewness), and the log link function for σ (scale) and τ (kurtosis). We now describe a few distributions that we found useful in GAMLSS modeling of test scores of psychological tests, thereby covering the most common scale types found in psychological tests.

A commonly occurring type of raw scale is ordered categorical, with fixed minimum and maximum scores. This type occurs when the test score is the sum of the scores on the items that make up the scale. Typical item scores are binary (e.g., 0 = false; 1 = correct), or multiple ordered categories, as in a Likert scale. For a limited range of scores conditional upon the predictor(s; say, maximally 25), we found the beta binomial (BB) distribution (i.e., $X \sim \text{BB}(bd\ \mu, \sigma)$), with X the observed score and bd the maximally theoretically possible score (i.e., binomial denominator) to fit rather well for many ordered categorical scales. For example, it has been used to model the scales of the nonverbal intelligence test SON-R 2–8 (Tellegen & Laros, 2017). The BB distribution is theoretically motivated if one has a scale composed of binary items where the distribution of the ability of the individuals follow a beta-distribution and the items are of equal difficulty (Wilcox, 1981), or where the difficulty of the items that make up the scale follow a beta-distribution and the ability of the individuals is equal (Albers, Vermue, de Wolff, & Beldhuis, 2018). Under either one of these assumptions, observed scales scores follow the BB distribution. In the norming models for scales composed of binary items it may be that both the item difficulty and the ability levels vary across items and individuals. The ability level variability will often be limited somewhat because one conditions on the norm-predictor(s). Further, even if these assumptions are not met, the BB distribution may offer a proper statistical fit. Yet, it remains important to evaluate model fit.

A continuous distribution is of use to model scales measured on a continuous scale, as well as to approximate categorical scales with a sufficiently large number of categories (say, more than 25). The Box-Cox power exponential (BCPE) distribution (Rigby & Stasinopoulos, 2004; i.e., $X \sim \text{BCPE}(\mu, \sigma, \nu, \tau)$) is a flexible, in practice often well-fitting continuous distribution. For example, the BCPE has been used to model the scales involving a reaction time (RT) as test score of the neuropsychological test COTAPP (Rommelse et al., 2018). The justification for using the BCPE distribution is of a statistical nature, in that it is a flexible distribution that can fit a broad range of continuous empirical distributions. As far as we know, there is no theoretical justification for modeling scales with the BCPE distribution. The BCPE distribution has four parameters, related to the location (μ, median), scale (σ, approximate coefficient of variation), skewness (ν, transformation to symmetry), and kurtosis (τ, power exponential parameter). Note that the symbols μ and σ in the context of a BCPE distribution refer to the median and scale, thereby bearing a different meaning than their default ones (i.e., mean and standard deviation). The normal distribution (i.e., $X \sim \text{NO}(\mu, \sigma)$) is a special instance of the BCPE distribution, namely when $\nu = 1$ and $\tau = 2$ (Voudouris, Gilchrist, Rigby, Sedgwick, & Stasinopoulos, 2012, p. 1283).

As an alternative to the BCPE distribution, one might consider the reparametrized version (Würtz, Chalabi, & Luksan, 2006) of the skew Student t distribution (Fernández & Steel, 1998; i.e., $X \sim SST(\mu, \sigma, \nu, \tau)$). This distribution has four parameters, namely mean μ, standard deviation σ, and ν and τ that relate to the skewness and kurtosis, respectively. It simplifies to the normal distribution when $\nu = 1$ and $\tau = \infty$. In a simulation study (Voncken et al., 2020), it appeared problematic to estimate the reparametrized version of the skew Student t distribution for simulated data sampled from a normal population, presumably because of extremely large estimated τ parameters. The distribution could be estimated well for skewed data. Further, the BCPE distribution did not suffer from estimation problems for normally distributed data. Combined with our experience that BCPE model appeared to fit generally well in norming empirical scales, we consider the BCPE distribution as the first choice for continuous scales and categorical scales with sufficiently many categories. However, the reparametrized version of the skew Student t distribution may yield proper fit as well in empirical practice.

Step 4: Select Candidate Function(s) to Relate the Norm-Predictor(s) to the GAMLSS Distribution Parameters

Once the candidate distribution for test scores has been identified, the parameters of the candidate distribution are to be modeled as additive functions of the norm-predictor(s). Modeling takes place as is common in regression modeling, taking into account the nature of the predictors, and in case of more than a single predictor, taking care of properly expressing any interactions between predictors. We refer to Cohen, Cohen, West, and Aiken (2003) as a great source for multiple regression modeling.

Norm-predictors can be categorical, ordered categorical, or continuous. Any categorical predictor needs to be modeled via dummy variables. Ordered categorical variables can be modeled via dummy variables, or treated as a continuous predictor—taking care that the resulting model fits the data well.

For a continuous predictor, a simple approach is to use a linear model for each of the parameters. For example, using a normal distribution would then result in both μ and σ being linearly dependent on the predictor(s). Note that if σ would be modeled

with an intercept only (i.e., taken as independent of any norm-predictor), the model would boil down to a standard linear regression model. Such a standard linear regression model is actually used in regression-based norming of psychological tests (e.g., Agelink van Rentergem, de Vent, Schmand, Murre, & Huizenga, 2018; Grober, Mowrey, Katz, Derby, & Lipton, 2015). However, nonlinear relationships with norm-predictors are found rather often in continuous norming practice, as Bechger, Hemker, and Maris (2009) indicated, and we saw confirmed in our norming of various psychological tests (e.g., Grob et al., 2018; Rommelse et al., 2018; Voncken et al., 2018), rendering the need for modeling nonlinearity.

Nonlinearity that involves smooth relationships between norm-predictor(s) and outcome variables, can be modeled using polynomials or splines. Polynomials are the simplest way, and pertain to adding to the linear equation one or more higher-order terms (e.g., age^2, age^3). To avoid estimation problems due to multicollinearity of the predictors, it is advised to use a centered version of the predictor and/or an orthogonalized version of the norm-predictor set.

Splines (for a review, see Perperoglou, Sauerbrei, Abrahamowicz, & Schmid, 2019) are piecewise polynomial functions, which are used to transform the norm-predictor(s). Using these transformed norm-predictors in the regression results in a smooth estimated function. Such a function can take any smooth functional form. The critical issue in using splines is the degree of smoothness required to achieve a model that represents the population well. Thus, as in any model, one needs to balance underfitting and overfitting. Different spline types manipulate smoothness in different ways. A popular type is P-splines (Eilers & Marx, 1996; Eilers & Marx, 2010), because of its favorable properties (e.g., numerically stable and easy to implement). Further, it requires only a single penalty parameter to manipulate the degree of smoothness of the complete function, making P-splines easy to apply. The monotonic P-spline variant is of use to achieve a more efficient estimation if it is known a priori that the smooth function is monotonically increasing (or decreasing). Such a monotonicity constraint is particularly useful when modeling test scores that are known to gradually increase with age, as for example intelligence scores in childhood. In the GAMLSS model this can be expressed by a monotonicity constraint on the location parameter (μ) as a function of age. Note that a monotonicity constraint on the other parameters, as spread and skewness, is typically not appropriate, because these typically do not show a monotonic pattern.

Both polynomial regression with higher order terms and splines yield a fair approximation to many types of relationships. Yet, polynomial regression is criticized, for its possibly undesired peaks and valleys in the estimation function (Harrell, 2015, p. 21) and the theoretically undesirable property that observed scores at a certain value or range of the predictor values may influence largely and undesirably the predicted scores at very different predictor values (Magee, 1998). In our experiences in continuous norming, we found both approaches to yield a proper and comparable fit in many instances; we also found instances for which P-splines showed local misfit, while polynomial regression yielded a proper fit overall and the other way around. For both approaches it holds that model selection is of key importance.

In modeling nonlinearity across the norm-predictor space (i.e., the complete range of observed norm-predictor values), discontinuities might occur. A typical case for which the piecewise approach seems to be favorable, is when a test uses different tasks for different age groups and the resulting score distributions show a jump at the boundary of the age groups. Then, piecewise functions might be useful to include. The core idea is to define various pieces of the predictor space (e.g., age $<$ 10, age $\geq$ 10) and estimate a model for each piece separately, using splines or polynomials. Restricting the models to first-order polynomials yields the well-known piecewise linear functions (e.g., Snijders & Bosker, 2012, pp. 268–270). Using piecewise linear functions adds flexibility—possibly yielding better fit—but also introduces additional boundaries, increasing the uncertainty in model fit at these boundaries. These sources of model fit should be balanced in deciding whether to adopt this approach. It may be useful to adopt a piecewise function for a subset of the distributional parameters only.

Step 5: Carry Out the Model Selection to Arrive at the Estimated GAMLSS Model

Fitting a function with either polynomials or P-splines requires model selection. This boils down to selecting either the polynomial(s) to include in the function in polynomial regression, or the value of the penalty parameter in P-splines. There is no general consensus on how to select these parameters. A popular approach is to select one or two promising candidate models from a range of presumably well-fitting models using statistical criteria. Of these candidate models the model fit of is then assessed via visual diagnostics.

There are two types of statistical model selection criteria available. Both aim at properly fitting the sample data while preventing overfitting, to ensure generalizability of the model to the population. In cross-validation, this is done directly by evaluating the quality of the model's prediction of new data. In generalized Akaike information criteria (GAIC; Akaike, 1983), this is done indirectly by penalizing the number of parameters to include in the model. Different GAIC variants exist that differ in their degree of penalizing, such as the Akaike information criterion (AIC), GAIC(3), and Bayesian information criterion (BIC), with penalties on the number of parameters equal to 2, 3, and ln(N), respectively. There is no general knowledge on which degree of penalty is needed for what type of model selection problem. For all GAIC criteria it holds that the model with the lowest GAIC value among the range of models considered, is favored.

When defining the range of presumably fitting models it is important to note that most GAMLSS distributions have dependent parameters. That is, a change in the regression model for one parameter (e.g., μ in the BCPE distribution) affects the estimates in the regression models for the other parameters (i.e., σ, ν, τ). To avoid the heavy computational load needed to consider all possibly useful combinations of number of polynomials for all four parameters of a BCPE distribution, an efficient algorithm has been developed and implemented for one continuous predictor (Voncken et al., 2019b). This so-called free order procedure searches for the model favored according to a specified model selection criterion. In a simulation study, the comparative performance of the AIC, GAIC(3), BIC, and cross-validation in identifying the best fitting polynomial models for the four parameters of the BCPE distribution has been assessed (Voncken et al., 2019b). Generally, cross-validation performed worse than the three, about

equally performing, GAIC criteria. As of the three, the BIC favors the simplest model. Therefore, we recommend the BIC as the primary statistical criterion to identify the candidate model(s) and assess the consistency in favored models with another GAIC criterion (e.g., AIC).

Model fit inspection of the candidate model(s) can take place in two ways. First, a worm plot (van Buuren & Fredriks, 2001) shows the relationship between the empirical quantiles and the model-implied quantiles. These detrended Q-Q plots include 95% bands that enable to assess to what extent observed deviations may be due to sampling fluctuations. Under the theoretical model, it is expected that 95% of the deviations lie within the confidence bands. It is useful to assess local model fit, by considering worm plots for multiple levels and/or ranges of the norm-predictor(s). Figure 2 shows example worm plots of two normative data models, to be discussed into detail in the Illustrative Example section.

Second, a percentile plot shows the model implied percentiles and the observed scores, as a function of a continuous predictor. If more than a single predictor (continuous and/or categorical) is used, centile plots are to be made for each possible combination of predictors (e.g., age per sex). This provides additional insight into the local fit of the model. Further, it gives an impression of the amount of data that supports the various regions of the estimated centiles. Figures 3 and 4 show examples of percentile plots, to be discussed further in the Illustrative Example section.

If the worm plots and centile plot(s) clearly suggest misfit, one proceeds by adapting the model to remedy the misfit. Adaptations can be in trying a different distribution for the outcome variable, and/or using different function(s) to model the relationship(s) with the distribution parameters. Otherwise, one can proceed to deriving the normed scores.

Step 6: Compute the Normed Scores for a Scale Based on the Estimated GAMLSS Model

The transformation of the raw test scores into any type of desired norm score is done based on the cumulative distribution function (CDF). From an estimated GAMLSS model, one can obtain a model-implied CDF for each value of the predictor (or, in case of multiple predictors, any combination of predictor values). The norms can thus be derived for any desired reference population. Note that extrapolating beyond the boundaries of observed data is discouraged.

One can distinguish three types of normed scores: percentile-based, distribution preserving, and normalized normed scores (Mellenbergh, 2011). Percentile-based normed scores include deciles, percentiles, and stanines. Because these type of normed scores directly relate to percentiles, they are readily derived from the CDF.

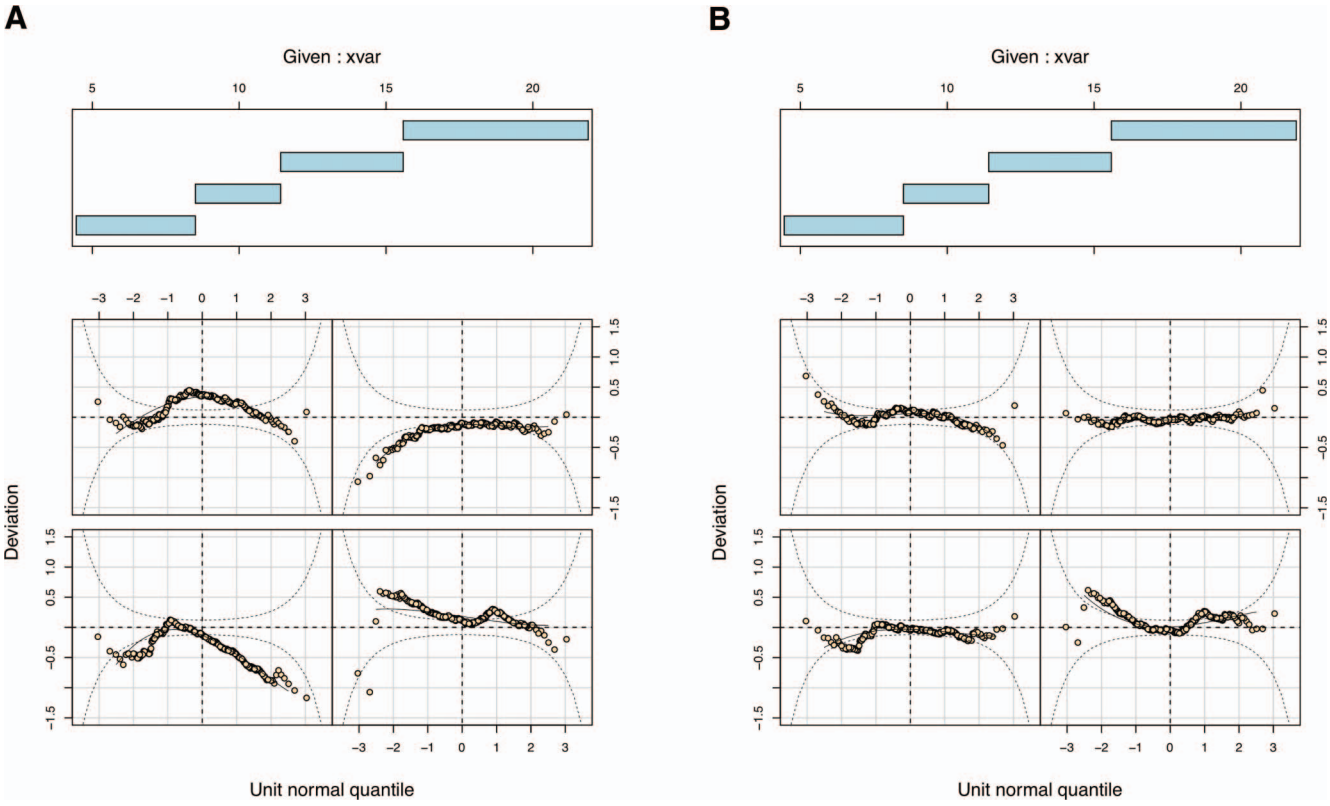

Figure 2. Worm plots for the normal model (panel A) and the default BCPE model with splines (panel B) for the illustrative normative sample of Test 14, with the predictor (xvar) age. The blue bars above the worm plots indicate the predictor range of age in each panel, which are ordered row-wise from the bottom left to the top right panel. See the online article for the color version of this figure.

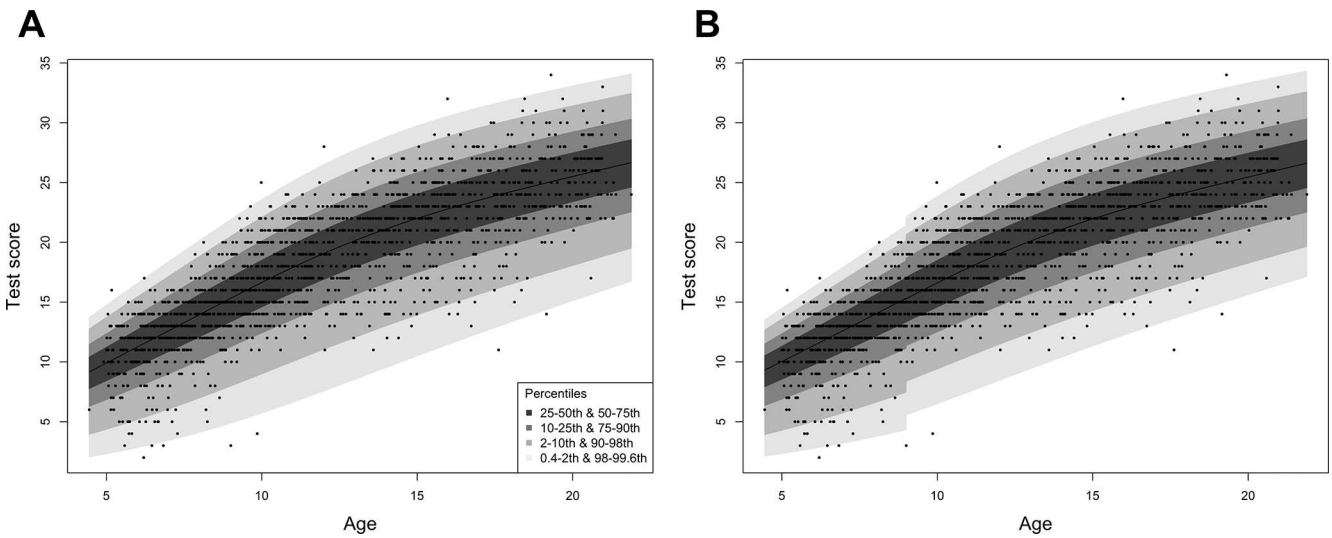

Figure 3. Test scores as a function of age for the illustrative normative sample of Test 14, and the percentile curves (.4, 2, 10, 25, 50, 75, 90, 98, 99.6 percentile) of the default BCPE model with splines (panel A) and the default BCPE model with splines and for the skewness a separate intercept for age ≥ 9 (panel B).

Distribution preserving normed scores are standardized to have a specific mean and standard deviation. The most often used forms are *Z*-scores ($\mu = 0, \sigma = 1$), Wechsler scaled scores ($\mu = 10, \sigma = 3$), *T* scores ($\mu = 50, \sigma = 10$), and *IQ*-scores ($\mu = 100, \sigma = 15$). These can be computed by linearly transforming the raw scores, using the CDF's mean and standard deviation. Some GAMLSS distributions have explicit expressions for their mean and standard deviation (e.g., normal and BB-distributions), others require numerical approximation (e.g., BCPE distribution, for which the μ and σ parameter express the median and scale, respectively).

Normalized normed scores have a normal distribution, with a specific mean and standard deviation. The most often used forms are the same as the ones mentioned under the distribution preserving normed scores, yielding, for example, normalized *IQ*-scores. In

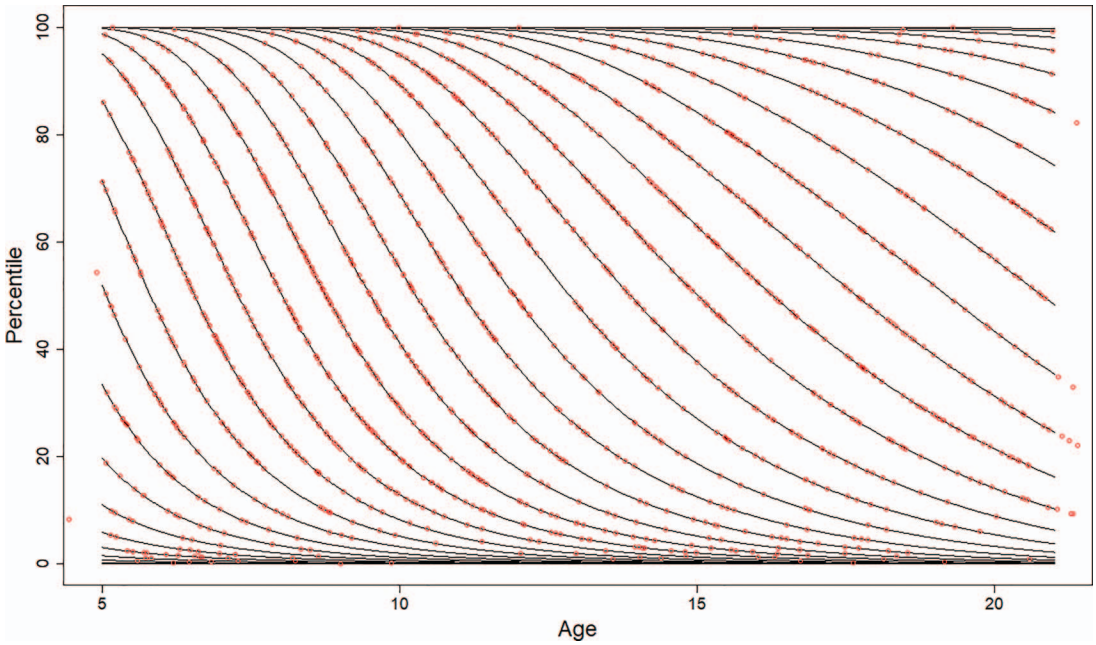

Figure 4. Percentile curves as a function of age for the default BCPE model with splines for the illustrative normative sample of Test 14, and the observed scores indicated as dots. Each line in the plot is associated with an observed score (ranging from 0 to 34), and in increasing order, with the lowest line pertaining to score 0, and the highest to 34. See the online article for the color version of this figure.

test manuals, one typically uses the term *IQ*-score (e.g., Wechsler, 2018), rendering it necessary to carefully check in test practice whether the normed scores are normalized or follow the raw score distribution. Normalized normed scores are obtained by applying the appropriate linear transformation to the standard scores (i.e., *Z*-scores under the normal distribution), which are derived from the model-implied CDF.

For both distribution preserving and normalized normed scores, it is common practice to truncate the continuous distribution at 3 *SD*s around the mean (e.g., Tellegen & Laros, 2017). The idea behind this is that observations in the tails are so scarce that the tails cannot be estimated with sufficient reliability, and that distinguishing individuals within these tails is not of relevance.

The estimated CDF is based on sample data, rather than data of the full population. Therefore, it is of use to express the uncertainty in the normed score due to sampling variability. CIs for norms based on a GAMLSS model with polynomials can be reliably obtained using posterior simulation (Voncken et al., 2019b). We recommend to be cautious in using this procedure for spline-based models because it requires the parameter's variance-covariance matrix and it is not known to what extent this one can reliably be estimated for splines. Note that these CIs only express the uncertainty due to sampling variability, and not uncertainty due to the unreliability of the test.

Step 6.1: Compute the Normed Scores for a Composite Scale

A composite scale is based on a linear combination of tests. The weights can be identified beforehand, typically using weights equal to one (as, e.g., in the composite scale "verbal skills" of the IDS-2; Grob & Hagmann-von Arx, 2018). Alternatively, the weights can be based on factor analysis. Herewith, one expresses differences in degree to which individual tests contribute to measuring the construct associated with the composite scale (as, e.g., in the "response speed" scale of the COTAPP; Rommelse et al., 2018). The linear combination is commonly based on *Z*-scores of the individual tests, rather than the raw test scores, precluding that arbitrary differences in mean and standard deviation between tests play a role. If also distributional differences between the tests are considered to be arbitrary, one needs to use the standard scores (i.e., normalized *Z*-scores) of the individual tests.

The composite score of a test (Z_{comp}) is thus a linear combination of (possibly normalized) *Z*-scores of tests. To achieve the desired normed composite score (e.g., *Z*-score), the composite score itself needs to be linearly transformed, where the transformation may depend on the predictors (i.e., $Z^{norm}_{comp} = (Z_{comp} - \hat{\mu}_{comp})/\hat{\sigma}_{comp}$, with $\hat{\mu}_{comp}$ and $\hat{\sigma}_{comp}$ the estimated population mean and standard deviation of the composite score, conditional upon the norm-predictor(s)). This linear transformation is needed because the variance of the composite scores is not known. That is, the composite score variance conditional upon the norm-predictor(s) depends on the variances of the *Z*-scores of the tests, and the covariances between the Z-scores of the tests, both conditional upon the norm-predictor(s). While the variances of the Z-scores conditional upon the predictor(s) are known (i.e., one), the covariances conditional upon the norm-predictor(s) are unknown, and they may depend on the predictor values. Thus, one needs the estimated CDF of the composite scores, taking into account the predictors. The estimated CDF can be obtained by fitting a GAMLSS model to the composite scores, including the predictors. For composite scores based on normalized *Z*-scores, one fits the normal distribution (as a linear combination of normally distributed variables is also normally distributed). For composite scores based on *Z*-scores, one needs to select a potentially suitable continuous distribution.

Thus, the transformation of the raw test scores into a norm score on the composite scale essentially takes place in three steps: First, compute the (normalized) *Z*-scores of tests involved. Second, compute the composite score as the (possibly weighted) sum of the (normalized) *Z*-scores. Third, compute the normed composite score, via a suitable linear transformation.

For tests that make up a composite scale and that are not perfectly correlated, one may note a peculiarity in comparing the normed scores of the individual tests and of the composite. That is, when a testee has normed scores on the individual tests that are consistently higher (or lower) than the mean, then the normed composite will be further away from the mean than the average normed score across the individual tests. Further, this effect will be larger for individuals scoring more extreme (either high or low). This effect may be surprising for a test administrator. However, it is just an expression of regression to the mean.

Illustrative Example

In the illustrative example, we describe how to arrive at normed scores for scales of the IDS-2 intelligence test applying GAMLSS. The composition of the illustrative normative data ($N = 1{,}660$) were described in the Traditional Versus Regression-Based Norming section. We will present the example along the six steps outlined in Table 1.

Step 1 involves defining the various populations. The reference populations for the Dutch IDS-2 are the general population in the Netherlands of the same age as the testee involved, in the age range 5; 0–20; 11 years. For the German IDS-2, the reference populations are analogous, yet with Austria, Germany, and the German speaking part of Switzerland as the target countries. The norm population then is the general population in the respective target country or countries in the age range 5; 0–20; 11. For the IDS-2, the norm population equals the target population of the test.

Step 2 involves designing and carrying out the study to gather the normative sample data. For both the German and Dutch IDS-2, judgmental sampling was used to achieve a representative sample. For example, the Dutch sample was stratified on sex, education level of the mother, migration status, urbanization degree, country region, education type (for those attending high school), and clinical status. The target figures were obtained from the Dutch central agency for statistics. Further, to achieve decent sample sizes across the full range of the norm-predictor age, minimal sample sizes per year of age were set (i.e., 100 for age 5; 0–12; 0, and 50 for 12; 0–20; 11). The minimal numbers were exceeded considerably, resulting in a total sample of $N = 1{,}665$. Given that age is used as a norm-predictor, and that this sample is properly stratified, exceeding the minimally set sample sizes only has a positive effect, namely leading to more precise norm estimation. For details on the sampling procedure, we refer to the Dutch 2 (Grob et al., 2018) and German manuals of the IDS-2 (Grob & Hagmann-von Arx, 2018).

In the next sections, we will illustrate Steps 3 to 6 to arrive at the normed scores with GAMLSS, for a test and a composite scale that are normed conditional upon age. We do so for Test 14 ("naming antonyms") and the composite scale "verbal skills," which is composed of Test 7 ("naming categories") and Test 14. We further illustrate how to compute CIs around the normed scores, which express the uncertainty due to sampling variability (Voncken et al., 2019a).

The illustrative data of Tests 7 and 14 and the *R* code to carry out the complete norming as described here, can be found in the online supplemental materials. The analyses were performed using *R* Version 3.6.1 (R Core Team, 2019), *gamlss* Version 5.1–4 (Rigby & Stasinopoulos, 2005), and *gamlss.tr* Version 5.1–0 (Stasinopoulos & Rigby, 2018).

For both Test 7 and Test 14, the raw test scores equal the number of correct items; each with a theoretical score range from 0 to 34. Both tests are administered according to predefined starting and stopping rules, to keep the administration time as short as possible. The items are ordered in difficulty. The starting item of a test is determined based on age, where 5- to 8-year-olds start with Item 1 for all tests. For Test 7, 9- to 12-year-olds start with Item 5, and 13- to 20-year-olds with Item 10. For Test 14, 9- to 20-year-olds start with Item 11. In case a testee gives an incorrect answer to the starting item, the item set before the starting item is administered; otherwise the items succeeding the starting item are administered. The nonadministered items before the start item are scored as being correct, thereby assuming that the noncompleted previous items would be answered correctly.

Norming Test 14 With GAMLSS

Figure 1 depicts the raw test scores of Test 14 in the normative sample as a function of age. The figure shows a clear positive relationship between the raw test scores and age. A notable observation is the extremely small number of observed raw test scores (i.e., dots) below 10, as of the age of 9. This is likely due to the starting rule, which essentially implies that a testee of 9-years-old receives a minimum score of 10, unless the testee answers Item 11 incorrectly.

Step 3 involves identifying one or a few candidate distribution types for the raw test scores. The raw test scores have an ordered categorical scale with a maximum theoretically possible score (i.e., 34). The range in the observed raw test scores conditional upon each age, within the age interval observed, is about 15. We consider three candidate distribution types. First, the BB distribution, because the test scores are ordered categorical with a limited range (i.e., lower than 25) and the BB distribution is the theoretical distribution of the test score when the difficulties of items stem from a beta distribution. Second, the BCPE distribution, because the range of test scores might be large enough to approximate the ordered categorical scale with a continuous distribution and the four parameters of the BCPE offer flexibility in fitting the shape of the distribution. Third, the right-truncated BCPE distribution, for the same reasons as the continuous BCPE distribution, plus accounting for the maximally theoretically possible score (at test score 34). Note that a BCPE distribution is by definition left-truncated at zero. To avoid test scores equal to zero, one must add a very small constant (i.e., 0.0001) to the raw test scores in any BCPE model estimation. The normed scores must be created for these adjusted raw scores and then the test scores must be transformed back to the original raw scores by subtracting the constant.

Step 4 is to select a candidate model to relate the norm-predictor variable (i.e., age) to the parameters of each of the three candidate distributions. For each distribution, we chose to select two types of functions to relate the predictor age to the model parameters, namely one using polynomials and one using P-splines. For both, the BIC was used as the statistical criterion to select the optimal model from a range of possible models. The AIC was considered as well to assess consistency in model selection. For the polynomials, we considered orthogonal polynomials of age, with polynomial degrees from 0 up to 4, for all distributional parameters (e.g., μ, σ, ν, τ for the BCPE distributions). Thus, five potential models were considered for each distributional parameter of the distribution involved. Herewith, a polynomial with degree 0 is the simplest model, involving an intercept only. A polynomial function up to degree 4 thus has five terms, typically offering sufficient flexibility in modeling the relation between the predictor and the distributional parameter involved. For the BB distribution, we fitted all 25 combinations of models (i.e., 5^2, for μ, σ). For the BCPE distributions, we used the free order selection procedure (Voncken et al., 2019b). For the P-spline models, we modeled with splines the parameters that appeared to be nonlinear in the polynomial model—we refrained from using splines for parameters that appeared linearly related to, or independent of age because splines for these kinds of effects may easily result in estimation problem. We used monotonically increasing P-splines to model for μ, thereby expressing that test scores generally increase with age, and regular P-splines for the other distributional parameter(s).

Apart from these six (i.e., 3 [candidate distributions] $\times$ 2 [type of function]) potentially well-fitting candidate models, we fitted a standard linear regression model, using a normal distribution with constant variance. This normal model can be expected to fit poorly and is presented here for illustrational purposes.

Step 5 is to carry out the model selection to arrive at the estimated GAMLSS model. The model fit, as expressed by the BIC and AIC, and the degrees of the polynomials of the six candidate models and the normal model is presented in Table 2. Among the seven models considered, both the BIC and AIC favor

Table 2
BIC and AIC Values of the Selected Models of Test 14, as Fit With the BCPE, Right-Truncated BCPE, BB Distributions Using Polynomials and Splines, and With the Normal Distribution for the Polynomial Models, the Degrees for the Parameters Are Indicated

	Polynomial			P-splines	
Model distribution	degrees for (μ, σ) or (μ, σ, ν, τ)	BIC	AIC	BIC	AIC
Default BCPE	(2, 1, 0, 0)	8404	8367	**8403**	**8361**
Right-truncated BCPE	(2, 1, 0, 0)	8456	8429	8456	8426
BB	(2, 2)	8463	8430	8529	8507
Normal	(1, 0)	8555	8538		

Note. BCPE = Box-Cox power exponential; BIC = Bayesian information criterion; AIC = Akaike information criterion; BB = beta binomial. Lowest BIC and AIC values among the seven models presented here are indicated in bold face.

the default BCPE model with P-splines. Therefore, we selected this model as the candidate model of which the model fit is to be inspected visually. The normal model fits worst, as was expected.

In Figure 2, worm plots for the default BCPE model with P-splines (panel B) and the normal model (panel A) are presented. The worm plot is a series of detrended Q-Q plots, split by subranges if the norm-predictor values (here: age). The worm plot visualizes how well the statistical model fits the data, for finding locations at which the fit can be improved, and for comparing the fit of different models. The blue bars above the worm plots indicate the predictor range in each panel, which are ordered row-wise from the bottom left to the top right panel. The worm plots of the BCPE model (panel B) show that most observations (i.e., dots) are within the 95% bands, suggesting a decent fit of the model for all age ranges, except for the age range 8 to 12 (bottom right plot). The local misfit might be due to the extremely small number of observed raw test scores below 10, as of the age of 9, which was already noted in Figure 1. The worm plots of the normal model (panel A) suggest considerable misfit at all ages ranges, as was expected.

Figure 3, panel A, depicts the observed scores as a function of age and the estimated percentile curves of the default BCPE model with P-splines. Figure 3 supports the interpretation of the worm plots: The percentile curves seem to fit generally well, except for the age 9–12. Indeed, the lack of observations below 10, as of the age of 9 seems responsible for the misfit. As indicated, we deem this likely due to the starting rule applied, yielding an overoptimistic view on the performance of some low-scoring children. If such a situation is detected in practice, the test constructors might consider to adapt the starting rule a bit, for example by administering Items 1 to 10 also when Items 12 and/or 13 are answered incorrectly. We expect that this would yield an increase in observed raw scores in the range 5 to 10, for children just above the age of 9. The current BCPE model seems to be in line with this scenario.

Alternatively, we may try to capture the discontinuity in the lowest part of the distribution at the age of 9. We did so by adding piecewise functions to the selected default BCPE model with P-splines, with the pieces pertaining to the age below and above 9. That is, we fitted all combinations of models with piecewise linear functions (i.e., intercept and/or slope may differ for age < 9 and age ≥ 9) for the scale, skewness and/or kurtosis (i.e., σ, ν, τ); as no discontinuities are to be expected for the median, we keep the monotonically increasing P-splines. Of all these possible combinations, the BIC appeared lowest for the default BCPE model with P-splines extended with a separate intercept for age ≥ 9 for skewness. The estimated percentile curves are represented in Figure 3, panel B. This figure clearly illustrates the discontinuity in the lowest part of the distribution, as captured by the model. Note that actually no observations are available in the age 9 to 12 for the estimated the .4 to second percentile—implying that this part of the model is completely supported by the surrounding observations and the models assumptions (i.e., related to the score distribution and smoothness). The BIC of this model is 8405, which indicates a slightly worse model fit than the default BCPE model with P-splines (depicted in panel A), which is 8403. Therefore, we proceeded with the default BCPE model with P-splines.

Figure 4 shows the percentile curves as a function of age associated with each possible score (i.e., in the range 0 to 34) and the observed scores. It can be seen that for each possible score, the associated percentile curve is monotonically nonincreasing. This is in line with what would be expected from a theoretical point of view. Further, these estimated curves are decently supported with observed scores throughout the complete estimated region, both in age and scores. For some curves, support is offered outside the age region, which will increase the precision of the estimates close to the age boundaries of the test. Finally, the steps in percentile curves between successive scores can be seen directly: Smaller steps offer a more fine-grained distinction between testees.

Step 6 is to compute the normed scores based on the estimated GAMLSS model. This selected normative model defines the CDF for each age value. Using these CDFs, the percentiles can be obtained for each possible raw test score conditional on each exact age value in the age range. In the illustration, the percentiles were calculated for each possible raw score conditional on 1,000 equally spaced age values in the range of 5 to 21 years (i.e., about one age value per week). These percentiles were transformed to normalized *Z*-scores via the inverse CDF of the normal distribution and then truncated to the range $[-3; +3]$.

CIs around the normed scores that express sampling variability can be computed using posterior simulation (Voncken et al., 2019a). We illustrate the computation based on the polynomial BCPE model, rather than its P-spline version, because the procedure's performance is only known to be proper for the polynomial models. This can be done here, as both models showed about equal fit. We simulated 5,000 sets of model parameters from a multivariate normal distribution defined by the point estimates of the parameters and the corresponding variance-covariance matrix. For all 5,000 simulated sets, we computed the normed scores for the raw test scores conditional on the age values of interest. As an example, we computed the normed scores for raw test scores 15 and 20, conditional on age 7 and 17, respectively. Then, using the percentile CI method, the 2.5th to 97.5th percentiles of the resulting simulated normed score distribution defined the boundaries of the 95% CI. The 95% CI expressing uncertainty due to sampling variability in the percentiles of someone of age 7 with test score 15 was [82.98, 86.55] and for someone of age 17 with test score 20 it was [12.55, 16.62].

Norming the Composite Test "Verbal Skills"

Step 6.1 deals with norming a composite scale. The composite scale "verbal skills" is composed of Tests 7 and 14, with both tests weighing equally. To norm a composite test, one needs the normed scores of its parts. The norming of Test 14 is described in the previous section. The norming of Test 7 was carried out analogously and will not be described further. A scatterplot of the normalized *Z*-scores of Tests 7 and 14 is provided in Figure 5, panel B. It can be seen that there is a positive linear relationship between the scores on both tests.

We computed the unweighted sum of the normalized *Z*-scores of Tests 7 and 14 in the normative sample. Then, a normal model was estimated, with possibly μ and/or σ depending on age. The relationship of age with each parameter (i.e., μ, σ) was modeled with an intercept only and a linear effect, resulting in four estimated models. Using the BIC as selection criterion, the model with age independent μ and σ was favored. To illustrate this model, Figure

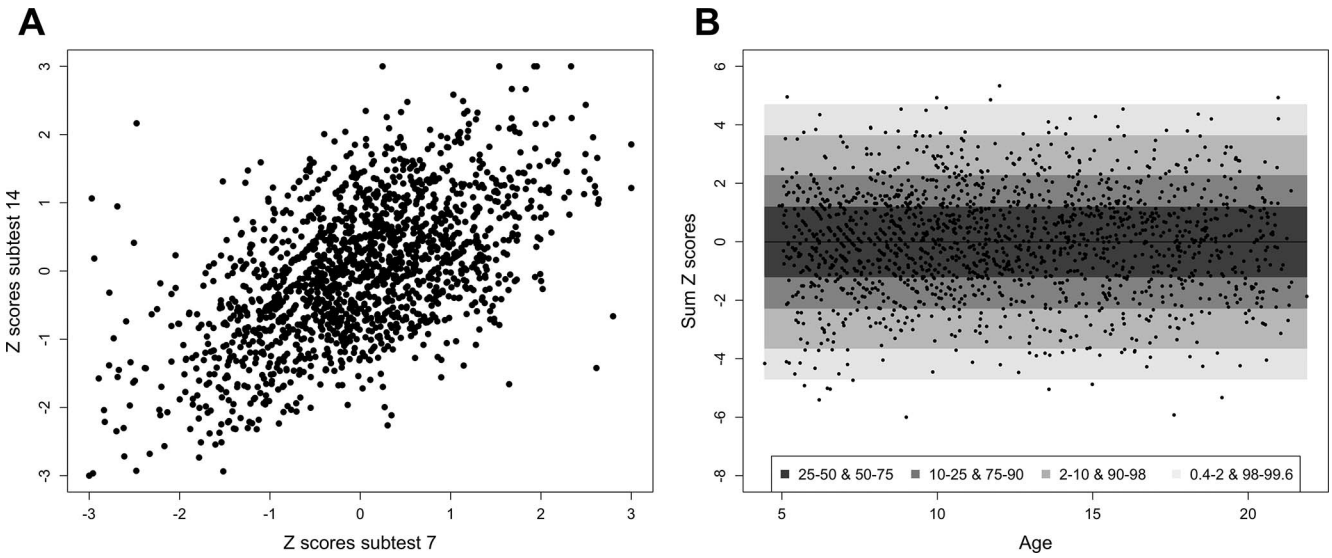

Figure 5. Scatterplot of the *Z* scores of Subtest 7 and Subtest 14 (panel A); sum of *Z* scores of Subtests 7 and 14, as a function of age, and the percentile curves (.4, 2, 10, 25, 50, 75, 90, 98, 99.6 percentile) of the normal distribution with constant mean and variance across age (panel B).

5, panel B, depicts the estimated percentile curves and the observed sum of *Z*-scores as a function of age.

The estimated μ and σ were used to linearly transform the sum of normalized *Z*-scores into the normed composite scores. An additional check revealed that the resulting composite scores had an estimated age independent mean < 0.001 and age independent standard deviation of 0.998. We deem this sufficiently close to the desired 0 and 1, respectively.

Reflection on Step 5: Sensitivity to Model Specification

The norms are determined completely on the basis of the estimated GAMLSS model. This means that it is essential to have an estimated GAMLSS model that properly fits the raw score population distribution as a function of the norm-predictors. Because the population model is unknown and the estimation takes place on the basis of sample data, an important question is how sensitive the estimates are to model misspecification. In the current context, model misspecification may stem from two sources, namely from the candidate GAMLSS distribution and the candidate model to relate the norm-predictors to the GAMLSS distribution parameters. Further, model misspecification can occur in two forms, namely a too strict or a too flexible model, meaning that the model either has too few parameters or more parameters than strictly needed, respectively, to adequately capture the population characteristics. Generally, a too flexible model can be expected to have smaller bias than its too restricted nonfitting version, yet it has larger sampling variability.

The sensitivity of certain GAMLSS norm models (i.e., involving normal, skew Student *t*, and BCPE distributions) to different forms and sources of model misspecification has been examined in a simulation study (Voncken et al., 2020). This study showed that models with too strict distributional assumptions yield biased estimates, whereas too flexible models yield increased variance. Based on the findings, it is recommended to use the BCPE distribution rather than the skew Student *t* distribution (to avoid estimation problems for normally distributed data) and to select the specific model parameters (e.g., degrees of the polynomials for the GAMLSS model parameters) using a criterion that properly penalizes the model complexity (e.g., using the BIC).

For details on the simulation study and their interpretation, we refer to Voncken, Albers, and Timmerman (2020). Here, we provide an illustration of the sensitivity of model estimation to model specification. From the simulation study by Voncken et al. (2020), we selected a specific condition to highlight effects of estimating the model that fits at the population level, a too strict model, and a too flexible model. Specifically, we drew a sample of $N = 1{,}000$ from a normal population model, with a linear relationship between the norm-predictor age and its expected value, and heteroscedasticity (denoted as the Li-HeNo condition in Voncken et al., 2020). On this sample, we estimated five models: (a) true model (i.e., linear normal model, with heteroscedasticity); (b) strict model (i.e., linear normal model, yet imposing homoscedasticity); (c) too flexible distribution, with a far too small penalty on model complexity (i.e., BCPE model with P-splines, with GAIC(0.1) criterion, denoted as BCPE(GAIC(0.1))); (d) same model as c., yet penalizing model complexity based on the AIC (denoted as BCPE(AIC)); (e) same model as c and d, yet penalizing model complexity based on the BIC (denoted as BCPE(BIC)).

In Figure 6, for each of the five estimated models, we present in the left column nine estimated percentile curves (i.e., 1, 5, 10, 25, 50, 75, 90, 95, 99 percentiles; dashed lines), population percentile curves (straight lines) and the sample scores (gray dots) as a function of age; in the right column the associated worm plots of the estimated models are depicted. The plots in the left column indicate how close the model estimated percentile curves (dashed lines) are to the population curves (straight lines). The plots in the right column indicate to what extent the sample scores comply with the estimated model. In practice, we only have the informa-

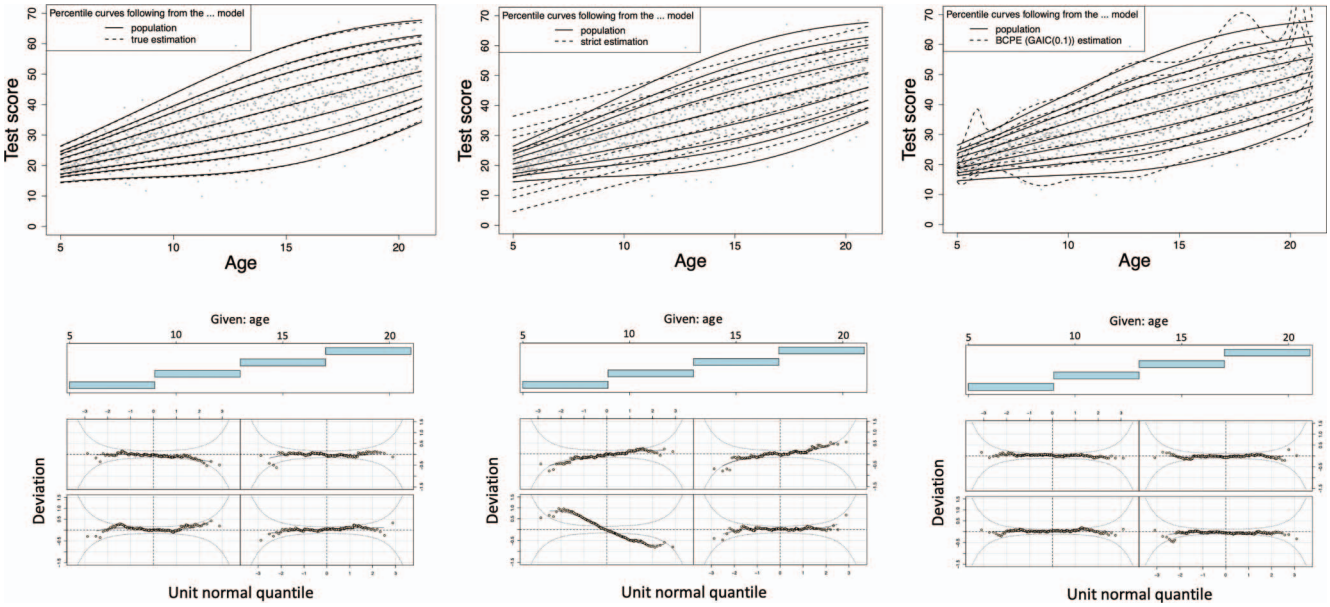

Figure 6. The five rows pertain to five estimated model; from top to bottom these are true model, strict model, BCPE(GAIC(0.1)), BCPE(AIC) and BCPE(BIC). Each plot in the left column: Nine percentile curves as based on the estimated model (at 1, 5, 10, 25, 50, 75, 90, 95, 99 percentiles; dashed lines), population percentile curves (straight lines) and sample scores (gray dots) as a function of age. Each plot in the right column: worm plots associated with the estimated models. See the online article for the color version of this figure. (*Figure continues on next page.*)

tion from the right column, while we aim at having close population fit, as depicted in the left column.

In the left column plots in Figure 6, it can be seen that the estimated and population percentile curves are rather close for the true and BCPE(AIC) estimation models. The good model fit to the population is also reflected in the worm plots, which show the fit at the sample level. The estimated and population percentile curves deviate considerably for the strict model (as the estimated model is restricted to linear percentile curves), the BCPE (GAIC(0.1)) model (the estimated percentile curves toward the percentile boundaries are too wiggly) and somewhat for the BCPE(BIC) model (the estimated percentiles toward the percentile boundaries at the age boundaries are too straight). The bad model fit to the population is also reflected in the worm plots for the strict model and the BCPE(BIC) model. In contrast, the worm plot of the BCPE(GAIC(.1)) model indicate very good model at the sample level. This is not surprising as the model complexity is hardly penalized, thereby overfitting the sample data. In practice, one needs to safeguard oneself to overfitting, by using a model selection criterion that penalizes model complexity reasonably well, and by examining the percentile curves in relation to what would be expected on a theoretical basis.

The example illustrates four important findings, which are corroborated by extensive simulation studies (Voncken et al., 2019b; Voncken et al., 2020): (a) both a too strict and a too flexible model yield misfit to the population model; (b) both a true model and a properly penalized flexible model yield estimated models that are well fitting to the population; (c) misfit with underfitting models can be diagnosed with worm plots, which are sample-based only; (d) misfit with overfitting models cannot be diagnosed with worm plots, yet can be precluded by applying a model selection criterion penalizing model complexity (as the AIC and BIC) and judging the estimated percentile curves based on theoretical knowledge. These findings provide important support for the use of GAMLSS modeling, provided that proper model selection takes place, to compute norms.

Discussion

This article serves as an introduction to regression-based norming with GAMLSS: a parametric modeling approach to estimate norm-referenced scores that depend on one or more individual characteristics. We have discussed the background of continuous norming and presented the important issues for regression-based norming with GAMLSS in six steps. Steps 1 and 2 pertain to issues in designing and carrying out the norming study, to gather the normative sample data. We further stressed the importance of identifying the norm-predictors that comply with the reference population, so as to ensure norms that express a comparison to the intended population. These issues are important for norming, irrespective of the specific method to actually estimate the norms. Steps 3 to 5 pertain to proper ways to arrive at a well-fitting GAMLSS model for the normative data. We discussed the importance of a proper model selection, where a flexible distribution can be used when combined with some penalty for model complexity in the model selection. We illustrated that a proper model selection appears possible based on sample data. Step 6 pertains to computing the normed scores based on the selected GAMLSS model for the normative data; Step 6.1 deals with normed scores for com-

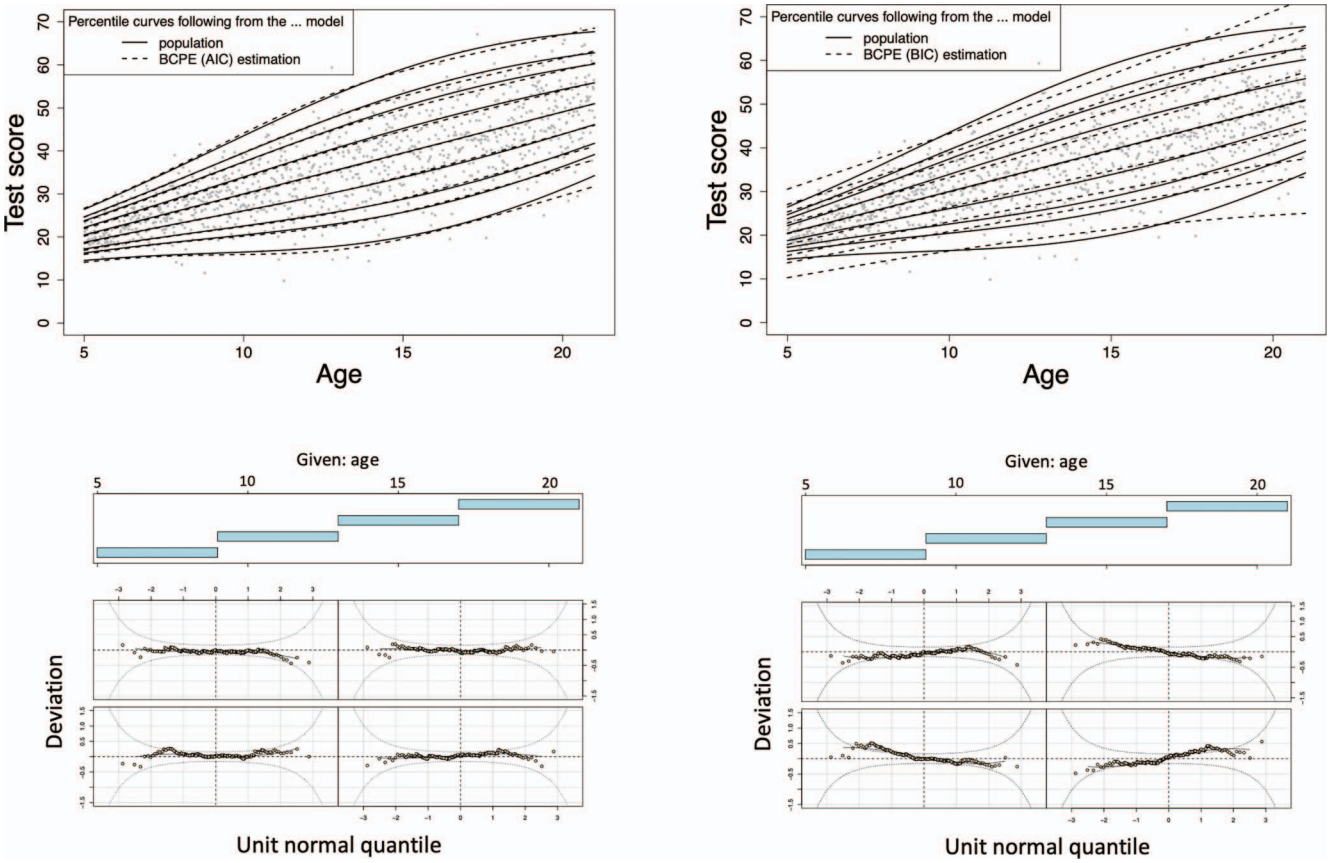

Figure 6. (continued)

posite scales. The principles described apply generally to any continuous norming method.

We illustrated the steps explained using an example normative data set, in computing the norms for a single scale and a composite scale. Further, we showed a useful way to visualize the norms of a test and their associated test scores related to age. The percentile plot (as in Figure 4) reveals the nature of the relationship between test scores and norms and shows whether all parts are supported by empirical data. The annotated *R* code and the example normative data set that we offer, allow readers to gain "hands on" experience in regression-based norming with GAMLSS and may serve as a basis for their own norming endeavor.

We focused on regression-based norming, primarily because of their statistically well-founded criteria for model selection and model assessment. These criteria are lacking in the other available continuous norming methods, inferential norming and semiparametric norming. Regression-based norming with GAMLSS offers large flexibility, implying that one likely finds a proper fitting model for the normative data. Note that a proper fitting model is essential to achieve high-quality norms, because the norms completely depend on it. Therefore, one needs to follow a good model selection procedure, fitting the sample while guarding against overfitting, and carefully assessing the estimated model quality via visual inspection. In case no fitting GAMLSS model could be identified, an alternative continuous norming method could be of use. In these conditions, we consider semiparametric norming to be most promising, because inferential norming (Zhu & Chen, 2011) still relies on distributional assumptions.

One might ask whether a semiparametric method would be preferred in general, because of its loose assumptions. We conjecture that generally regression-based norming is statistically more efficient than semiparametric norming, implying that one reaches a higher level of estimation precision using the same sample size. We conjecture this because generally regression-based norming requires a smaller number of parameters to estimate and there is no need to discretize the norm-predictors (which possibly introduces imprecision). As far as we know, there is only one study that examined this issue (Lenhard, Lenhard, & Gary, 2019) in a norming context. Results of the simulation study partly support our conjecture, with the exception in the extreme conditions (i.e., easy and difficult test scales). We conjecture that the poor performance of the regression-based norming could be attributable to poor model fit. In their simulation study, only continuous GAMLSS distributions (i.e., normal, Box-Cox, truncated Box-Cox and Box-Cox Power Exponential distribution) were included, despite the simulated data being ordered categorical, with a relatively small range. This implies that it needs to be studied further how the performance of both methods relate under various conditions. This must offer insight which of the methods is preferable under which conditions.

Via this tutorial, we aim to stimulate the application and investigation of proper continuous norming methods. Despite the availability of publications and *R* code that is of use to carry out continuous norming, more research is needed to serve empirical practice. In our view, the main challenges involve guidance on the use of parametric versus parametric methods, guidance on planning a sampling scheme and extending the method to multilevel data (for cases in which hierarchical sampling is used).

References

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist, 4–18 and 1991 profile*. Burlington, Vermont: University of Vermont, Department of Psychiatry.

Agelink van Rentergem, J. A., de Vent, N. R., Schmand, B. A., Murre, J. M. J., & Huizenga, H. M. (2018). Multivariate normative comparisons for neuropsychological assessment by a multilevel factor structure or multiple imputation approach. *Psychological Assessment, 30,* 436–449. http://dx.doi.org/10.1037/pas0000489

Akaike, H. (1983). *Statistical inference and measurement of entropy* (G. E. P. Box, T. Leonard, & C. Wu, Eds.). New York, NY: Academic Press. http://dx.doi.org/10.1016/B978-0-12-121160-8.50015-6

Albers, C. J., Vermue, C., de Wolff, T., & Beldhuis, H. (2018). *Model-based academic dismissal policies; a case-study from the Netherlands*. Retrieved from https://psyarxiv.com/6a9cz/

Angoff, W. H., & Robertson, G. J. (1987). A procedure for standardizing individually administered tests, normed by age or grade level. *Applied Psychological Measurement, 11,* 33–46. http://dx.doi.org/10.1177/014662168701100102

Bayley, N. (2006). *Bayley scales of infant and toddler development—3rd ed. Technical manual*. San Antonio, TX: Harcourt Assessment, Inc.

Bechger, T., Hemker, B. T., & Maris, G. (2009). *Over het gebruik van continue normering* [On the use of continuous norming]. Arnhem, the Netherlands: Cito. Retrieved from https://docplayer.nl/40799177-Over-het-gebruik-van-continue-normering-timo-bechger-bas-hemker-gunter-maris.html

Becker, A., Wang, B., Kunze, B., Otto, C., Schlack, R., Hölling, H., . . . the BELLA Study Group. (2018). Normative data of the self-report version of the German strengths and difficulties questionnaire in an epidemiological setting. *Zeitschrift fur Kinder- und Jugendpsychiatrie und Psychotherapie, 46,* 523–533. http://dx.doi.org/10.1024/1422-4917/a000589

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Durbeej, N., Sörman, K., Norén Selinus, E., Lundström, S., Lichtenstein, P., Hellner, C., & Halldner, L. (2019). Trends in childhood and adolescent internalizing symptoms: Results from Swedish population based twin cohorts. *BMC Psychology, 7,* 50. http://dx.doi.org/10.1186/s40359-019-0326-8

Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science, 11,* 89–102. http://dx.doi.org/10.1214/ss/1038425655

Eilers, P. H. C., & Marx, B. D. (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics, 2,* 637–653. http://dx.doi.org/10.1002/wics.125

European Parliament and Council of the European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. Retrieved from http://data.europa.eu/eli/reg/2016/679/oj

Fernández, C., & Steel, M. F. J. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association, 93,* 359–371.

Goodman, R. (1997). The strengths and difficulties questionnaire: A research note. *Journal of Child Psychology and Psychiatry, 38,* 581–586. http://dx.doi.org/10.1111/j.1469-7610.1997.tb01545.x

Grob, A., & Hagmann-von Arx, P. (2018). *IDS-2. Intelligenz- und Entwicklungsskalen für Kinder und Jugenliche* [IDS-2: Intelligence- and developmental scales for children and youth]. Bern, Switzerland: Hogrefe.

Grob, A., Hagmann-von Arx, P., Ruiter, S. A. J., Timmerman, M. E., & Visser, L. (2018). *IDS-2: Intelligentie- en ontwikkelingsschalen voor kinderen en jongeren* [IDS-2: Intelligence- and developmental scales for children and youth]. Amsterdam, the Netherlands: Hogrefe.

Grober, E., Mowrey, W., Katz, M., Derby, C., & Lipton, R. B. (2015). Conventional and robust norming in identifying preclinical dementia. *Journal of Clinical and Experimental Neuropsychology, 37,* 1098–1106. http://dx.doi.org/10.1080/13803395.2015.1078779

Groth-Marnat, G., & Wright, A. J. (2016). *Handbook of psychological assessment* (6th ed.). New York, NY: Wiley.

Harrell, F. (2015). *Regression modeling strategies*. New York, NY: Springer. http://dx.doi.org/10.1007/978-3-319-19425-7

Kaufman, A. S., & Kaufman, N. L. (1994). *K-SNAP: Kaufman short neuropsychological assessment procedure*. Circle Pines, MN: American Guidance Service.

Lenhard, A., Lenhard, W., & Gary, S. (2019). Continuous norming of psychometric tests: A simulation study of parametric and semi-parametric approaches. *PLoS ONE, 14,* e0222279. http://dx.doi.org/10.1371/journal.pone.0222279

Lenhard, A., Lenhard, W., Suggate, S., & Segerer, R. (2018). A continuous solution to the norming problem. *Assessment, 25,* 112–125. http://dx.doi.org/10.1177/1073191116656437

Magee, L. (1998). Nonlocal behavior in polynomial regressions. *The American Statistician, 52,* 20–22.

Mellenbergh, G. J. (2011). *A conceptual introduction to psychometrics: Development, analysis and application of psychological and educational tests*. The Hague, the Netherlands: Eleven International Publishing.

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

Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., & Schmid, M. (2019). A review of spline function procedures in R. *BMC Medical Research Methodology, 19,* 46. http://dx.doi.org/10.1186/s12874-019-0666-3

R Core Team. (2019). *R: A language and environment for statistical computing* [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Rigby, R. A., & Stasinopoulos, D. M. (2004). Smooth centile curves for skew and kurtotic data modelled using the Box-Cox power exponential distribution. *Statistics in Medicine, 23,* 3053–3076. http://dx.doi.org/10.1002/sim.1861

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society Series C, Applied Statistics, 54,* 507–554. http://dx.doi.org/10.1111/j.1467-9876.2005.00510.x

Rigby, R. A., Stasinopoulos, D. M., Heller, G., & De Bastiani, F. (2019). *Distributions for modeling location, scale, and shape: Using GAMLSS in R*. Boca Raton, FL: CRC/Chapman & Hall. http://dx.doi.org/10.1201/9780429298547

Rommelse, N., Hartman, C. A., Brinkman, A., Slaats-Willemse, D., de Zeeuw, P., & Luman, M. (2018). *COTAPP handleiding* [COTAPP manual]. Amsterdam, the Netherlands: Boom.

Snijders, J. T., Tellegen, P. J., & Laros, J. A. (1988). *Verantwoording en handleiding van de SON-R 5.5–17* [Account and manual of the SON-R 5.5–17]. Göttingen, Germany: Hogrefe.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis. an introduction to basic and advanced multilevel modelling* (2nd ed.). London, UK: Sage.

Stasinopoulos, D. M., & Rigby, R. A. (2018). Gamlss.tr: Generating and fitting truncated 'gamlss.family' distributions (R package Version 5.1–0) [Computer software]. Retrieved from https://CRAN.R-project.org/package=gamlss.tr

Stasinopoulos, D. M., Rigby, R. A., Heller, G., Voudouris, V., & De Bastiani, F. (2017). *Flexible regression and smoothing*. New York, NY: Chapman and Hall/CRC. http://dx.doi.org/10.1201/b21973

Tellegen, P. J., & Laros, J. A. (2017). *SON-R 2–8: Snijders-Oomen niet-verbale intelligentietest* [SON-R 2–8: Snijders-Oomen nonverbal intelligence test]. Amsterdam, the Netherlands: Hogrefe.

van Baar, A. L., Steenis, L. J. P., Verhoeven, M., & Hessen, D. J. (2014). *Bayley-III. technische handleiding* [Bayley-III. technical manual]. Amsterdam, the Netherlands: Pearson Assessment and Information.

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

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). New York, NY: Chapman and Hall/CRC. http://dx.doi.org/10.1201/9780429492259

van Buuren, S., & Fredriks, M. (2001). Worm plot: A simple diagnostic device for modelling growth reference curves. *Statistics in Medicine, 20,* 1259–1277. http://dx.doi.org/10.1002/sim.746

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,* 826–839. http://dx.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,* 1329–1346. http://dx.doi.org/10.1177/1073191117715113

Voncken, L., Albers, C. J., & Timmerman, M. E. (2020). Bias-variance trade-off in continuous test norming. *Assessment*. Retrieved from https://psyarxiv.com/cz8k3/

Voncken, L., Kneib, T., Albers, C. J., Umlauf, N., & Timmerman, M. E. (2020). Bayesian Gaussian distributional regression models for more efficient norm estimation. *British Journal of Mathematical and Statistical Psychology*. Retrieved from https://psyarxiv.com/7j8ym/

Voncken, L., Timmerman, M. E., Spikman, J. M., & Huitema, R. (2018). Beschrijving van de nieuwe, Nederlandse normering van de Ekman 60 Faces Test (EFT), onderdeel van de FEEST [Description of the new, Dutch norming of the Ekman 60 Faces Test (EFT), part of the FEEST]. *Tijdschrift Voor Neuropsychologie, 13,* 143–151.

Voudouris, V., Gilchrist, R., Rigby, R., Sedgwick, J., & Stasinopoulos, D. (2012). Modelling skewness and kurtosis with the BCPE density in GAMLSS. *Journal of Applied Statistics, 39,* 1279–1293. http://dx.doi.org/10.1080/02664763.2011.644530

Wechsler, D. (2008). *WAIS-IV: Technical and interpretative manual*. San Antonio, TX: Pearson.

Wechsler, D. (2014). *Wechsler intelligence scale for children* (5th ed.). Bloomington, MN: Pearson.

Wechsler, D. (2018). *WISC-V-NL. Wechsler intelligence scale for children - Nederlandstalige bewerking* [Dutch ed.] (5th ed.). Amsterdam, the Netherlands: Pearson.

Wilcox, R. R. (1981). A closed sequential procedure for comparing the binomial distribution to a standard. *British Journal of Mathematical and Statistical Psychology, 34,* 238–242. http://dx.doi.org/10.1111/j.2044-8317.1981.tb00633.x

Würtz, D., Chalabi, Y., & Luksan, L. (2006). Parameter estimation of ARMA models with GARCH/APARCH errors. An R and SPlus software implementation. *Journal of Statistical Software, 55,* 28–33.

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

Zhu, J., & Chen, H. (2011). Utility of inferential norming with smaller sample sizes. *Journal of Psychoeducational Assessment, 29,* 570–580. http://dx.doi.org/10.1177/0734282910396323

Received November 7, 2019
Revision received June 29, 2020
Accepted July 1, 2020 ■