university of groningen

University Medical Center Groningen

University of Groningen

Flexible regression-based norming of psychological tests

Voncken, Lieke

DOI:
10.33612/diss.124765653

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2020

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Voncken, L. (2020). *Flexible regression-based norming of psychological tests*. University of Groningen. https://doi.org/10.33612/diss.124765653

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

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 28-11-2021

Chapter 1

Introduction

Psychological tests are widely used to assess individuals in clinical, educational, and personnel contexts. Intelligence tests, developmental tests, personality tests, and neuropsychological tests are used for diagnosis, monitoring, assessment, and selection. Because the results of these tests are used to make important decisions about individuals, it is essential that the tests are of high quality (i.e., have high validity and reliability), and that the test scores can be meaningfully interpreted. Meaningful interpretation of a raw test score (e.g., the number of items correct on a test) is typically done via a reference point.

Flanagan (1939) distinguished four different reference points to interpret raw test scores (as cited in Mellenbergh, 2011, p. 346). First, a testee's test score can be compared to his/her score on other (sub)tests. This is referred to as *test-referenced* test score interpretation. For instance, given a constant test length, a testee's short-term memory test score can be compared to his/her long-term memory test score.

Second, the test score can be compared to the testee's score on the same test on different occasions. This is referred to as *occasion-referenced* test score interpretation. For instance, a testee's test score can be compared before treatment and after treatment.

Third, a testee's test score can be compared to an external criterion or standard. This is referred to as *criterion-referenced* or *standard-referenced* test score interpretation. This type of interpretation is mainly used in achievement testing (Mellenbergh, 2011, p. 369). For instance, teachers might believe that students have mastered the test materials when they have obtained a test score of at least 80% of the maximum test score.

Fourth, a testee's test score can be compared to the scores of other testees for the same test. This is referred to as *norm-referenced* test score interpretation. This type of interpretation makes sense for many psychological tests because one often wishes to compare the testee's score to the scores of a reference population. For instance, intelligence test scores are typically interpreted relative to scores of the community population. The focus of this thesis is on norm-referenced test scores.

Norm-referenced test scores

Norm-referenced test scores – referred to as normed scores – are typically created by transforming the raw test scores to another scale. There are three types of normed scores: percentile-based, distribution preserving, and normalized normed scores (Mellenbergh, 2011, pp.351-360). Percentile-based normed scores (e.g., percentiles, deciles, and stanines) are directly derived from the cumulative density distribution of the raw test scores in the reference population. The population percentile is the percentage of people in the reference population with the same score or below. For instance, percentile 72 indicates that 72% of the people in the reference population obtained the same score or below.

Distribution preserving normed scores have the same distribution as the raw test scores, and are obtained by linearly transforming the raw test score distribution to have a specific mean and standard deviation. Examples of distribution preserving normed scores are *IQ* scores ($M = 100$, $SD = 15$), Z scores ($M = 0$, $SD = 1$), Wechsler scores ($M = 10$, $SD = 3$), and T scores ($M = 50$, $SD = 10$). Normalized normed scores have a normal distribution with a specific mean and standard deviation, and are typically normalized versions of the distribution preserving normed scores mentioned above, such as normalized Z scores. Percentiles can be transformed to normalized Z scores via the inverse cumulative density function (CDF) of the normal distribution.

Transforming the raw test scores to normed scores allows for easy interpretation of the test score. For instance, a percentile of 50, which equals a normalized *IQ* score of 100, implies that 50% of the reference group obtained the same score or below. A normalized *IQ* score of 115 means that the testee scored one *SD* above average compared to the reference population. It is generally easy to go from one normed score type to another, but it can be difficult to determine the transformation from raw test score to normed test score. This transformation depends on the raw test score distribution in the reference population.

Because the test score distribution from the reference population is unknown and it is practically impossible to collect test scores from everyone within the population, a reference sample is used to make inferences about the reference population. That is, we estimate the raw test score distribution on the sample, and generalize this to the population. As we will see later, there are different approaches to estimate this (conditional) raw

score distribution.

Representative sample

The population for whom the test is designed is referred to as the target population. For instance, the target population of the Dutch Wechsler Intelligence Scale for Children-V (WISC-V-NL; Wechsler, 2018) consists of all Dutch-speaking children between 6 and 17 years old. The test scores of the WISC-V-NL are interpreted relative to the test scores of the community population of the same age. Hence, the reference population consists of Dutch speaking people of a certain specific age within the range 6–17 years. Technically, the number of reference populations is infinite: There is one reference population for every exact age value in the range 6–17 years. The norm population is the combination of all reference populations. Note that the norm population does not have to be equivalent to the target population. For instance, the target population of a clinical test might consist of all people who are able to complete the test, while the norm population might consist of healthy people only. This would allow for comparison of the (possibly unhealthy) testee's test score to the test scores of healthy people.

In the test construction phase, one wants to collect a representative sample of the norm population, which means that the sample reflects the characteristics in the population. It is important that this sample is representative with respect to characteristics that are related to the test scores. For instance, if highly educated people are overrepresented in the normative sample of an intelligence test, the test scores in the sample are likely to be higher than the test scores in the reference population, and this will result in too high normed scores. As a result, the mean test score in the population would be interpreted as a below-average test score because the mean test score in the sample was higher.

In practice, it is very difficult to collect a representative sample. It is unclear beforehand which characteristics are related to the test scores. Theoretically, the best way to deal with this is to randomly sample from the population, through which all members of the population have the same probability of being included in the sample. This is referred to as simple random sampling. The larger the random sample, the larger the probability that the sample is representative with regard to all relevant characteristics in the population. Unfortunately, this is usually impossible in practice.

Random sampling from the norm population requires a list of the full norm population. However, these lists are typically not (publicly) available. For clinical populations, these lists do usually not even exist. For general populations, many countries have a population register, but privacy regulations like the General Data Protection Regulation (European Parliament and Council of the European Union, 2016) generally prohibit sharing data without explicit consent of the data subject. Even if lists of the norm population would be available, there is still the problem of nonresponse.

An alternative sampling method is cluster sampling, in which random clusters (e.g., schools) are chosen. This only requires a list of all clusters in the norm population, rather than all individuals. Unlike lists of individuals, list of clusters may be publicly available, as – for example – schools or hospitals. It is possible to only randomly select the clusters (i.e., one-stage clustering), or to randomly select both the clusters and the individuals within the clusters (i.e., two-stage clustering). Like in simple random sampling, there is still the problem of nonresponse. A disadvantage of this method is that the dependency between individuals within clusters makes the method less efficient than simple random sampling.

A more efficient sampling method than simple random sampling and cluster sampling is stratified sampling, in which samples are randomly drawn from subpopulations (strata) that are related to the test score. For instance, if it is known that the test score is related to education level, random samples are drawn from subpopulations relating to each education level, with the sample sizes proportional to the size of the subpopulations. If the proportions in the sample do not match those in the population, the observations can be weighted accordingly. In this way, there is no overrepresentation of one of the subpopulations in the normative sample. Unfortunately, as discussed before, random sampling is infeasible in practice, and it is unknown beforehand which characteristics are related to the test scores.

In practice, a non-random sampling technique of stratified sampling – judgmental sampling (Mellenbergh, 2011, p. 351) – is typically used, in which the subpopulations are typically based on easy-to-measure characteristics, like age, sex, education level, and region. The downsides of this approach are that it is unknown whether enough subpopulations are included, and that the proportions in each of the subpopulations are typically only assessed in a univariate way. That is, it is not assessed whether the proportion of

combinations of characteristics (e.g., highly-educated, elderly males) match those in the population. That is, the multivariate distribution is not assessed. Tacitly, it is thus assumed that the characteristics are independent – which more often than not will fail to hold.

Choice of reference population

The choice of reference population depends first and foremost on the desired interpretation of the test score. In clinical tests, one typically wants to compare the testee's test score with the testee's "healthy" test score to assess whether the testee is now unhealthy. As the score of the healthy version of the testee is not available, the test score is compared to the scores of healthy people who are as similar as possible to the testee. In practice, this means that the testee is compared to people who are similar on easy-to-measure characteristics, like age, sex, and education level. For instance, Van Breukelen and Vlaeyen (2005) considered age, sex, education level, marital state, pain duration, diagnosis, geographic region, and type of medical center.

In intelligence tests, on the other hand, one typically wants to compare the testee's test score with the scores of a general population of people of the same age. Broader or narrower reference populations are typically uninformative for intelligence tests. For instance, if the reference populations consist of all people between 5 and 40 years old, it would be unsurprising to see that someone of age 5 scores below average because people generally score higher on intelligence tests as they get older. Also, a more narrow reference population would be uninformative because you are typically not interested in an interpretation like "You score exactly average for a 28-year-old woman, who wears glasses, is about 1.60m tall, and is aiming at obtaining her PhD at the University of Groningen on 14 May 2020". Rather, it would be informative to know how you scored relative to – for instance – people from the general population, or other PhD students, of your age.

Interestingly, the first version of the Groninger Intelligence Test (Snijders & Verhage, 1962) provided age scores, sex scores, and achievement scores. This allowed the test user to choose the interpretation of the test scores: relative to people of the same age, relative to people of the same sex, or relative to the full norm population.

Once the reference population is chosen based on the interpretation, it has to be investigated whether the chosen reference characteristics are related to the test score. If

the test scores are not related to the reference characteristics, it does not have added value to interpret the scores conditional on them. The test scores of intelligence and developmental tests are typically related to age. More specifically, the test scores typically increase strongly with age for young children, and this relationship diminishes or decreases from age 25–35 (Ferrer & McArdle, 2004; McArdle, Ferrer-Caja, Hamagami, & Woodcock, 2002).

Traditional and continuous norming

When the reference characteristics are measured with categorical variables (e.g., sex), the reference populations are defined for each category of the variable(s). Traditionally, the same approach was used for continuous variables (e.g., age) by discretizing them (e.g., Wechsler Intelligence Scale for Children-III, WISC-III; Wechsler, 1991). The empirical probability density function for a subgroup was used as estimate for the raw score distribution of that subgroup. This traditional norming approach is problematic because it is unrealistically assumed that the test score distribution is equal for everyone within a subgroup, and it can result in jumps in normed scores at the boundaries of the subgroups. It is assumed that the conditional test score distribution changes as a step function of the continuous variable(s), while theoretically it is more realistic that this relationship is smooth (Van Breukelen & Vlaeyen, 2005; Zachary & Gorsuch, 1985). A continuous function can be approximated by making the subgroups smaller, but this results in fewer observations per subgroup to estimate the raw score distribution.

These problems are solved in continuous test norming (Zachary & Gorsuch, 1985) – also referred to as regression-based norming – in which the raw score distribution is estimated as a continuous function of the reference characteristic(s) in a regression model. The continuous norming approach is more efficient than the traditional norming approach (Oosterhuis, Van der Ark, & Sijtsma, 2016) because all observations within the normative sample, rather than a subgroup, are used to estimate the raw score distribution. By using a regression model, information from the surrounding predictor values is used. In this way, the raw score distribution can be estimated for any predictor value within the predictor range, even if the specific predictor value itself is not observed in the normative sample. Still, it is important to have the full predictor range represented in the normative sam-

ple without large gaps. There are multiple sampling strategies for the predictor(s). For instance, the observations can be sampled in a uniform way across the predictor range, or more observations can be included around predictor ranges for which a stronger relationship between the predictor and the raw test score is expected. The normative sample does not have to be representative with respect to the chosen predictor(s). Rather, the reference sample has to be representative for the reference population, which is evaluated conditional on the predictor(s).

Continuous norming approaches

The continuous norming approaches can be divided into three types (Emons, 2019): inferential norming (Wechsler, 2008; Zachary & Gorsuch, 1985; Zhu & Chen, 2011), moments regression-based norming (Oosterhuis, 2017; Van Breukelen & Vlaeyen, 2005), and non-parametric norming (Lenhard, Lenhard, Suggate, & Segerer, 2018; Tellegen & Laros, 2014).

In inferential norming, moments of the raw test score distributions are computed for subgroups of the normative sample, and these moments are regressed on subgroup-level predictor(s). For instance, the mean and standard deviation of the full scale score can each be regressed on the mean age in a subgroup. Then, the mean and standard deviation of the conditional (normal) distribution of the raw test scores can be predicted for each predictor value in the predictor range, which are used to transform the raw test scores to standardized scores. This procedure involves smoothing of the curves of the moments as a function of the predictor following suggestions by experts. This can be done using subjective "hand smoothing" (e.g., Zhu & Chen, 2011) or using a statistical model (e.g., Zachary & Gorsuch, 1985). While Zachary and Gorsuch (1985) only modelled the mean and standard deviation of the raw test score, Zhu and Chen (2011) also modelled the skewness and kurtosis, which allowed for capturing non-normality. The advantages of inferential norming compared to traditional norming are that information about the moments from all predictor groups is used, and that the normed scores are smooth across the predictor range(s). The main disadvantage of this approach is that the moments are estimated per subgroup, which results in estimates that are less precise, less efficient, and dependent on the exact subgroups. In addition, it is problematic that hand smoothing expresses individ-

ual beliefs about theoretical relationships. We expect that even for experts it is difficult to theorize how moments other than the mean depend on the predictor(s).

In moments regression-based norming, moments of interest are regressed on predictor(s) for the individual raw test score data, rather than for subgroup data. Van Breukelen and Vlaeyen (2005), and Oosterhuis (2017) used a standard regression model to estimate the mean of the raw test score distribution conditional on the predictor(s). Categorical predictors were included as dummy variables, and continuous predictors were included as linear and – possibly – quadratic terms. Residuals were calculated for each test taker, and these were transformed to standardized residuals (e.g., Z scores). Hereby, it was assumed that the Z scores were normally distributed conditional on the predictor(s), with a constant variance.

The main advantages of the moments regression-based approach are that no additional smoothing step is required, and that statistical criteria for model selection and model assessment are available. The disadvantage of the mean-regression based approach by Van Breukelen and Vlaeyen (2005), and Oosterhuis (2017) is that homoscedasticity and normality of the residuals are assumed. Oosterhuis (2017) argued that test constructors do not have to investigate this normality assumption "...because for sample size > 50 the central limit theorem ensures that the [standard] regression model is robust against violations of this assumption" (p. 128). This argument is valid only when estimating distribution preserving normed scores, but not when estimating normalized normed scores, because it only pertains to the (implied) model parameters (e.g., regression coefficients, mean, standard deviation), rather than the score distribution itself. In practice, test publishers typically report normalized normed scores rather than distribution preserving normed scores (e.g., Grob & Hagmann-von Arx, 2018; Tellegen & Laros, 2014).

For distribution preserving normed scores, only the mean and standard deviation of the conditional raw test score distribution have to be estimated. Given the central limit theorem (CLT), the sampling distributions of these parameters are (approximately) normal if the sample size is large enough, regardless of the shape of the conditional raw test score distribution. Note that the standard regression model assumes homoscedasticity. If this assumption is too strict, one needs the Gaussian model to estimate the standard deviation conditional upon the predictor(s).

For normalized normed scores, one needs to estimate the conditional raw test score distribution itself. If this distribution clearly deviates from normality, a regression model built upon the normality assumption cannot be used. The CLT pertains to the model parameters, and does not imply that the conditional raw test scores themselves are (approximately) normally distributed if the sample size would be large enough. For instance, floor- and ceiling effects will result in skewness of the conditional raw test score distribution, regardless of the used sample size.

Van Breukelen and Vlaeyen (2005, p. 344) recommended to use scale transformations or normed scores based on deciles in the presence of non-normality, and to use the residual standard deviation for quartiles of the predicted scores in the presence of heteroscedasticity. Unfortunately, these alternatives require that the shape of the raw score distribution is equal for (subgroups of) the predictor or predicted score range, and they do not allow for local changes in the scale or shape of the distribution. Oosterhuis (2017, p. 93) recommended to use the traditional norming method or models with weaker assumptions in the presence of assumption violations. We believe that the best option is to use a continuous norming model with weaker assumptions in the presence of non-normality and heteroscedasticity.

In non-parametric norming, the relationship between the raw test scores, and normed scores and age is modelled using Taylor polynomials (Lenhard, Lenhard, & Gary, 2019; Lenhard et al., 2018; Tellegen & Laros, 2014). Lenhard et al. (2019) describe their norming approach in three steps. First, the normed scores are estimated based on the empirical cumulative distribution function. All powers of these normed scores, the age variable, and their interactions are calculated up to a predetermined degree. Second, the raw test scores are regressed on all calculated polynomials of the normed scores and age. Finally, the significant degrees of the polynomials of the second step are included in the final model. This model can be used to derive the normed scores for combinations of the raw test score and age. The advantage of this non-parametric approach is that it does not require assumptions about the conditional score distribution, and – thus – allows for modelling heteroscedasticity and non-normality. Tellegen and Laros (2014) argued that normality of raw test score distributions is rare, especially when subtests are designed for broad age ranges. Lenhard et al. (2019) argued that homoscedasticity and normality are only rarely

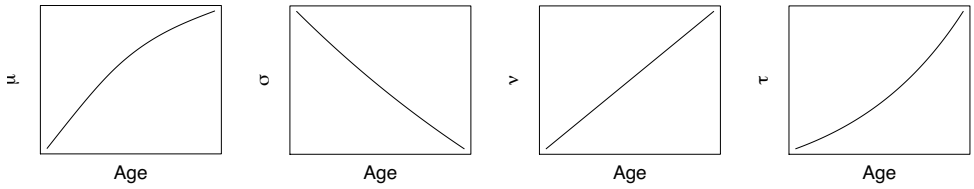
fulfilled in psychometric tests, and that non-normality is common because many tests contain floor and ceiling effects in at least some age ranges. A disadvantage of this flexibility is that the resulting percentile curves can intersect, which is impossible from a theoretical point of view. In addition, this approach requires discretization of the continuous predictor variable to estimate the normed scores.

Distributional regression

In this thesis, we investigate a flexible moments regression-based norming approach, namely using distributional regression (e.g., Rigby & Stasinopoulos, 2005; Umlauf, Klein, & Zeileis, 2018). This approach includes the approach by Van Breukelen and Vlaeyen (2005), and Oosterhuis (2017), and allows for many other distributions (Rigby, Stasinopoulos, Heller, & De Bastiani, 2019) and function types as well. In distributional regression, distributional characteristics (e.g., the mean, variance, skewness, and kurtosis) can be modelled as a continuous function of predictors, which allows for modelling heteroscedasticity and non-normality locally. Unlike the non-parametric norming approach, the undesired intersecting percentile curves are impossible to occur. We use both a frequentist and a Bayesian framework for distributional regression, namely the generalized additive models for location, scale, and shape (GAMLSS; Rigby & Stasinopoulos, 2005), and the Bayesian additive models for location, scale, and shape (and beyond) (BAMLSS; Umlauf et al., 2018).

Figures 1 and 2 illustrate how GAMLSS can be used to arrive at normed scores conditional on age for Dutch normative data of composite scale "IQ screening" of the intelligence test IDS-2 (N = 1,566) (Grob, Hagmann-von Arx, Ruiter, Timmerman, & Visser, 2018). Note that the process of test norming with BAMLSS is similar, but additionally requires a prior distribution for each of the model parameters. In short, one has to select the distribution and function type to model the conditional raw test score distribution as a function of age. In this illustration, we have chosen the Box-Cox Power Exponential (BCPE; Rigby & Stasinopoulos, 2004) distribution, which has four distributional parameters: μ, σ, ν, and τ, for the median, scale, skewness, and kurtosis, respectively. Identity link functions were used for μ and ν, and log link functions were used for σ and τ. The log link function prevented negative parameter estimates of σ and τ. We have selected a linear relation-

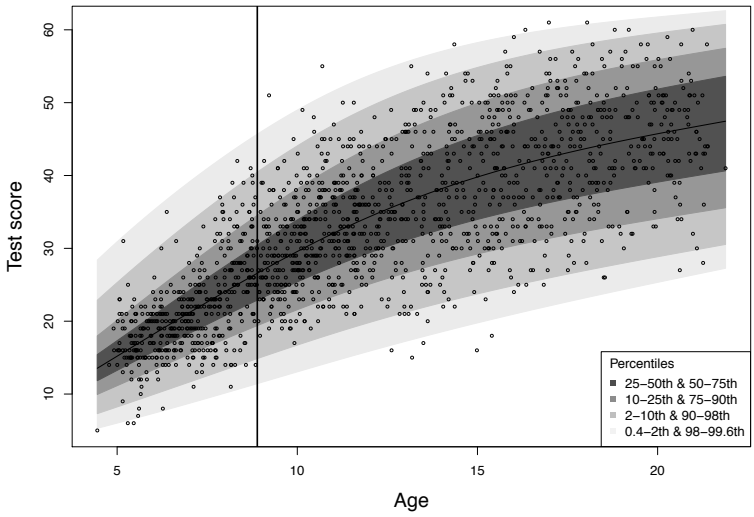
ship between age and ln(σ), ν, and ln(τ), respectively, and μ was modelled as a smooth function of age using monotonically increasing P-splines (Eilers & Marx, 1996). Figure 1 shows the estimated relationship between each of the distributional parameters of the BCPE distribution and age.



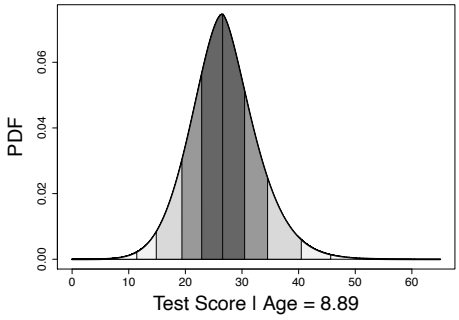
(a) μ and age (b) σ and age (c) ν and age (d) τ and age

Figure 1. Estimated relationship between each of the distributional parameters of the BCPE distribution (i.e., μ, σ, ν, and τ) and age for the Dutch normative data of composite scale "IQ Screening" of the IDS-2.

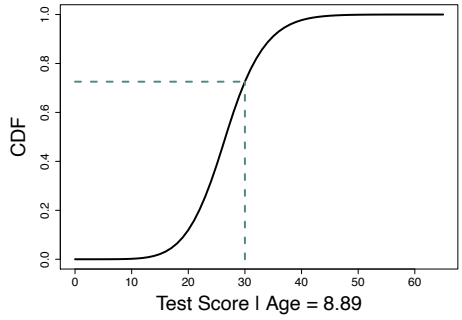
The estimated distributional parameters define the estimated conditional raw test score distribution for each age value in the age range. This information is used to transform raw test scores conditional on age to normed scores. Figure 2(a) shows the centile curves as derived from the estimated distributional parameters in Figure 1. The dots indicate the observations in the normative sample, the gray bands indicate percentile ranges, and the vertical line indicates the raw test score distribution evaluated at age 8.89. This precise age value is chosen to illustrate that the normed scores can be assessed for any age value in the age range. This conditional raw test score distribution corresponds to the probability density function (PDF) in Figure 2(b) and the cumulative density function (CDF) in Figure 2(c). Using the CDF, the raw test scores conditional on age can be transformed to percentiles. The dotted lines in panel (c) illustrate that the proportion of people of age 8.89 that obtained a raw test score of 30 or below equals 0.726, which means that the corresponding percentile is 72.6. This percentile of 72.6 equals to a normalized *Z* score of 0.6 and a normalized *IQ* score of 109.



(a) Centile curves



(b) PDF



(c) CDF

Figure 2. Estimated centile curves (panel a), and for age 8.89 the corresponding PDF (panel b) and CDF (panel c) for the Dutch normative data of composite scale "IQ Screening" of the IDS-2. The dots in panel (a) indicate the observations in the normative sample. The gray bands in panels (a) and (b) indicate percentile ranges. The dotted line in panel (c) indicates the cumulative density corresponding to raw test score 30.

Aims and overview of this thesis

The availability of many different models makes norming with distributional regression flexible, but this flexibility also comes with challenges. The large availability of models makes the model selection difficult. It is practically impossible to compare all possible models, which makes it very useful to have a well-performing automated model selection procedure. So far, it is unknown how well automated model selection procedures perform in the context of distributional regression. In addition, it is unknown how much model flexibility is optimal. A flexible model that fits at the population generally has smaller bias, but also has larger sampling variability than more restricted model versions. Larger sampling variability results in more uncertainty in the parameter estimates, and thus also in the normed scores. This type of uncertainty in normed scores is typically ignored in practice. The sampling variability can be decreased by increasing the size of the normative sample, but this is costly and not always possible in practice.

In this thesis, we address the challenges related to model selection and sampling variability in test norming using distributional regression. In Chapter 2, an automated model selection procedure is developed for the flexible BCPE distribution, and its performance is compared to the performance of an existing procedure. In Chapter 3, the bias-variance trade-off in GAMLSS models is explored. It is investigated what the costs are of using a too strict model (i.e., bias) versus the costs of using a too flexible model (i.e., variance). This is important for guiding model selection. In Chapter 4, a procedure to create confidence intervals that express the uncertainty in normed scores due to sampling variability is investigated. In Chapter 5, it is investigated whether norm estimation can be made more efficient (i.e., requiring a smaller sample size to obtain the same norm precision) by using prior information via Bayesian Gaussian distributional regression. Finally, a general discussion is provided in Chapter 6.

To evaluate the procedures and models, we make extensive use of simulation studies. To make the simulation studies realistic and the conditions empirically relevant, we make use of empirical normative data of psychological tests. The Dutch/German normative data of the Snijders-Oomen non-verbal intelligence test 6-40 (SON-R 6-40; Tellegen & Laros, 2014) is used in Chapters 2 and 4. The Dutch normative data of the Cognitive

Test Application (COTAPP; Rommelse et al., 2018) is used in Chapter 3. The Dutch normative data of the Ekman 60 Faces Test of the Facial Expressions of Emotion - Stimuli and Tests (FEEST; Voncken, Timmerman, Spikman, & Huitema, 2018) is used in Chapter 4. Finally, the German and Dutch normative data of the Intelligence and Developmental Scales 2 (IDS-2 Grob & Hagmann-von Arx, 2018; Grob et al., 2018) are used in Chapters 3 (German) and 5 (German and Dutch).[1]

All computations in this thesis were performed in R (R Core Team, 2019). The R code used in this thesis is available from `https://osf.io/52nzt/` (Chapter 2), `https://osf.io/k6fzn/` (Chapter 3), `https://osf.io/z62xm/` (Chapter 4), and `https://osf.io/cjx3v/` (Chapter 5).

[1] We thank Peter Tellegen and Jacob Laros for providing us with the SON-R 6-40 normative data, we thank Nanda Rommelse and the other authors of the COTAPP for providing us with the COTAPP normative data, we thank Joke Spikman for providing us with the FEEST normative data, and we thank Alexander Grob and the other authors of the German and Dutch IDS-2 for providing us with the IDS-2 normative data.