**How to mitigate violations of representativeness in norm samples: Weighting in continuous norming**

Sebastian Gary1, Alexandra Lenhard1, Wolfgang Lenhard2\*, David Herzberg3

1 Test Development Center, Psychometrica, Dettelbach, Bavaria, Germany

2 Institute of Psychology, University of Würzburg, Bavaria, Germany

3 WPS, Torrance, CA, USA

\* Corresponding author:

E-Mail: wolfgang.lenhard@uni-wuerzburg.de

# Abstract

In this simulation study, we investigated whether the quality of normed test scores derived from non-demographically representative samples can be improved by applying compensatory weighting at the raw score level. To this end, we modeled a latent cognitive ability that showed an increasing developmental gradient within a reference population. We subsequently drew normative samples of limited size such that they systematically deviated from the reference population with regard to the distribution of three demographic variables: gender, migration status, and regional location. Next, we generated simulated test results for each person based on a one-parameter logistic IRT model. Using these simulated data, we applied continuous, regression-based norming techniques, both with and without compensatory weighting. The weighting technique was able to substantially reduce the bias of the normed scores, but could not fully compensate for the lack of representativeness. Deviations from unbiased normed scores mainly showed up at extreme person locations and when the lack of representativeness was large, but were generally small enough to be neglected in most practical applications.

# How to mitigate violations of representativeness in norm samples: Weighting in continuous norming

One prerequisite for normed scores of psychometric tests to be valid is that these normed scores are based on samples representative of the reference population. In fact, samples do not have to be representative in every single respect. Rather, representativeness is required only with respect to those variables that significantly correlate with the variable to be measured in the test. The first task of any test designer in planning the norming procedure for a psychometric test is therefore to identify such confounding variables. Typically, these include age, gender, or parental education. Usually, not all of these variables are treated the same way in the norming process. Rather, the handling depends on the specific purpose of the test respectively the specific variable to be measured. For example, during childhood and adolescence, the age of the test person has certainly the largest effect of all confounding variables on the test results, specifically on cognitive tests. Based on the suggestion of Wechsler (1939, Chapter 3), the reference populations of intelligence and achievement tests are therefore restricted to individuals of the same age only. Thus, to calculate IQ scores, the distributions of test results have to be established for each age level separately. As this task becomes very cost intensive when a lot of separate age groups are used, advanced mathematical methods have been developed in recent decades to model the continuous relationship between raw scores and normed scores across age with relatively small total sample sizes (e.g., Gorsuch, 1983, quoted from Zachary & Gorsuch, 1985; Cole, 1988; Cole & Green, 1992; A. Lenhard et al., 2018; W. Lenhard et al., 2018; Stasinopoulos et al., 2018; A. Lenhard et al., 2019; W. Lenhard & Lenhard, 2021). As a beneficial side effect of these models, normed scores do not necessarily refer to large age groups (i.e., 6-year-olds, 7-year-olds etc.) any-more but can be calculated as fine-grained as necessary, even down to the exact day.

For other norm-oriented tests or measures, it may even be useful to narrow down the reference population even more precisely. For example, when it is intended to determine whether a person's body mass index (BMI) is too high or too low, it may be useful to report norms not only for different age levels but also separately by gender, since the optimal BMI for women is lower than for men (Sang-Wook et al., 2015).

However, it is not always reasonable to provide separate norms for the different levels of a confounding variable. For example, girls have repeatedly been shown to have slightly higher reading skills than boys (e.g., W. Lenhard et al., 2017; Price-Mohr & Price, 2017). However, if a reading test is used to identify those children who need additional support - for example, the 7% worst readers - then it would be unfair to use gender-specific norms. Otherwise, some girls might receive additional training measures even though they perform better than boys that do not receive these measures. The second task of the test constructor is therefore to select appropriate reference groups.

## Stratification and post-stratification

Variables such as gender in the aforementioned example, which may have a significant effect on the variable being measured but for which it is not suitable to calculate separate norms, must be taken into account in a different way. The first option is to draw a random sample of very large size. The idea behind random sampling is that large random samples automatically result in representative samples with regard to all variables. However, collecting sufficiently large random samples involves a lot of work and high expenses. Therefore, most test constructors attempt to establish representativeness artificially by using stratification. In this method, the reference population is divided into different strata, that is, homogeneous subgroups, for example boys and girls. Subsequently, the norm sample is assembled in such a way that the proportions of the subgroups in the norm sample correspond to the proportions of the respective strata in the reference population. For example, if gender is a significant confounding variable, the norm sample should consist of or at least approach 50% males and 50% females. (For simplicity, we assume at this point that the number of individuals who identify as non-binary generally is negligibly small.) Each gender would represent a separate stratum in this case. Note that within each stratum, the sampling must again be random.

However, it is not always possible to collect data in such a way that the proportions of the subgroups in the norm sample correspond to the actual proportions of the respective strata in the reference population. One way to correct such distortions is to randomly remove data from the overrepresented strata until the norm sample is representative. However, this has the drawback of losing data. An alternative is therefore to go the other way round and to apply weights to the underrepresented strata when determining the cumulative distribution function. For example, if a sample consists of 100 boys and 50 girls, the data obtained for the girls could be given a weighting factor of 2. Each result of a girl would then be treated as if two girls had obtained such a result. A weight *w*k, that is assigned to an observation *x*i in subsample *k*, thus stands for the number of individuals this observation represents. The weights must therefore be chosen so that the proportion

corresponds to the proportion of stratum *k* in the reference population (with *nk* = size of subsample *k* in the norm sample). This procedure is referred to as post-stratification (Little, 1993; Park et al., 2004; Lumley, 2011, chapter 7).

Recently, it has been suggested to use multilevel regression combined with post-stratification to correct for non-representative samples in psychological intervention studies (Kennedy & Gelman, 2021). The authors suggest using this approach to adjust the means of non-representative samples to be able to compare them. However, the use of weights may not be as beneficial in establishing test norms, since the latter require determining not only the means but all achievable percentiles. From a mathematical perspective, taking weights into account when calculating percentiles is not rather complicated. As described above, each test result has simply to be treated as if this result was obtained by wk instead of only one individual. But the method bears the risk of high norming error, and specifically at the tails of the distributions. This is because the applied weights – although presumably correcting a biased sample mean relatively well – do not affect the variance of the raw scores in the respective subgroups, as would be the case if the size of the respective subgroup was actually changed. Therefore, the overall variance of the raw scores might be biased, and more so when the weights are high. As a consequence, the beneficial effects of weighting should decrease with increasing deviation from representativeness. Moreover, this effect should be specifically detrimental if a subgroup with very high or very low group mean is underrepresented, because at the tails of the distributions the number of observations is low anyway and therefore biased variance can distort the normed scores more easily at extreme locations. Unfortunately, results that are markedly above or below average are the ones for which precise normed scores are needed the most in practical diagnosis, because psychometric tests are usually applied to identify individuals with exceptional performance, for example mentally gifted or disabled individuals.

But the risk of high norming error is not even the only challenge that comes into play when post-stratification is applied to generate test norms. Especially for intelligence or achievement tests, there usually is a rather high number of confounding variables such as parental education, ethnicity, socioeconomic status, or region (e.g., federal state) that would have to be included as stratification variables (SVs). In addition, these variables often are interdependent. For example, areas with low socioeconomic status usually have lower education than areas with high socioeconomic status, but they also tend to have a lower proportion of non-white inhabitants. Therefore, the most accurate way of taking them into account would be to not only consider the distribution of every single variable for itself but to regard all cross-classifications as separate strata, that is, to consider the joint distributions. However, census data are often available only for the individual variables, but not necessarily for all cross-classifications. Besides, taking into account all possible cross-classifications quickly results in an immensely large number of strata. Imagine, for example, an intelligence test in which the SVs are gender (male; female), parental education (e.g., no high school; high school; college degree; bachelor's degree), ethnicity (e.g., white; black; hispanic; asian; native; other), and region (e.g., north; south; east; west). The complete cross-classification in this example would result in 192 cells for each of which at least a small amount of data would have to be collected. Above all, this would have to be done for every single age cohort. Given that the sample sizes of intelligence tests usually are limited to about 100 per age cohort (e.g., Kaufman & Kaufman, 2004; Wechsler, 2008; Wechsler, 2014), it is obvious that usually it will not be feasible to collect a random sample of sufficient size in which every single cell of the cross-classification is adequately represented in every single age cohort. It is therefore necessary to look for more economic approaches.

## Raking

The so-called raking (Ireland & Kullback, 1968; Kalton & Flores-Cervantes, 2003) is such an approach. Raking in fact does not draw on all cross-classifications. Instead, the weights are determined in an iterative process based on the marginal distributions of each SV, that is, the weights of the SVs are adjusted successively and, if necessary, repeatedly until they no longer change. This can be compared to smoothing out the soil in a garden bed by successively raking in different directions over and over again. It has been shown that the raking procedure is convergent and delivers best asymptotically normal estimates for the joint probabilities (Ireland & Kullback, 1968). Although raking is a widely used approach to correct for lack of representativeness in official statistics or political polls (Kalton-Flores-Cervantes, 2003), to our knowledge, this approach has only rarely if ever been applied to generate norms of psychometric tests. One reason for not using raking to establish test norms may be the very reason that prompted its development in the first place, namely that raking only considers the marginal distributions, but not the joint distributions. As a consequence, the errors caused by the weighting procedure are supposed to be even higher than with conventional post-stratification. Moreover, this drawback could be particularly detrimental for test norms, since - as described above - the typical SVs of psychometric tests are highly interdependent. At present, however, it can only be speculated whether the use of raking actually has a positive or negative effect on the accuracy of test norms, because to the best of our knowledge, the norming error that results from the application of raking has never been systematically studied.

## Effects of continuous norming on non-representativeness samples

In fact, weighting is not the only way in which a lack of representativeness can be countered, at least hypothetically. Some of the distortions might also be mitigated by the continuous norming methods already mentioned above, making weighting at least partially unnecessary. A continuous norming method that has been demonstrated to deliver low norming error under various conditions, is the semi-parametric approach first suggested by A. Lenhard and colleagues (A. Lenhard et al., 2018; A. Lenhard et al., 2019; W. Lenhard & Lenhard, 2021). In the following, we will refer to this approach as SCN for semi-parametric continuous norming. One advantage of SCN is that it does not make specific assumptions about the raw score distributions and therefore can be applied to rather skewed scales or even scales with floor or ceiling effects. Secondly, it has been demonstrated to perform better than parametric approaches when applied on norm samples with the typical sample size of 100 per cohort, independent of the skewness of the raw score distributions (A. Lenhard et al., 2019). But there is yet another feature of SCN that might prove to be specifically beneficial when used together with samples that lack representativeness in some age groups, but not across the whole sample, namely that the trajectories of the percentiles across age are relatively rigid. This is because the implementation of SCN (i.e., the cNORM package in R, A. Lenhard et al., 2018) does not rely on splines to model these trajectories, as other approaches used to model continuous normed scores do (e.g., the GMLLS package in R, Stasinopoulos et al., 2018). As a result, error variance in the normed scores that is caused by a lack of representativeness in individual age groups can be compensated for by other, more representative age groups. In accordance with this assumption, it has been shown that the application of SCN results in much lower norming error than establishing test norms for each age group separately (W. Lenhard & Lenhard, 2021). It remains to be shown whether and to what extent the combination of raking and SCN can improve the accuracy of the normed scores even further.

## Focus and rationale of research in this simulation study

In summary, to counter the effects of non-representative norm samples ex post, raking and SCN could be applied either individually or in combination. However, to date, no research has investigated whether the combination of both methods (in the following referred to as *weighted continuous norming* WCN) further improves the accuracy of the normed scores as compared to SCN alone. In the best case, WCN could lead to a further reduction of the norming error, but in the worst case, unforeseen effects of raking or even interactions of both methods could occur, which might even deteriorate the quality of the normed scores at least under some conditions or for some performance ranges. Our overall goal, therefore, was to evaluate the potentials and the limits of raking when applied in addition to an established norming method, in this case, SCN as implemented in the cNORM package on R (A. Lenhard et al., 2018). To this end, we simulated the norming procedure of a typical cognitive variable that evolves across age. Furthermore, we modeled the effects of three different confounding variables on the cognitive variable, thereby exploiting the effect sizes of real confounding variables such as parental education, ethnicity, or region.

To be able to study the effects of WCN under different conditions and with different degrees of representativeness, we mimicked six different sampling scenarios, each with the same sample size. The first scenario served as a control condition with complete random sampling, that is, with minor deviations from representativeness resulting only from the random sampling process with limited sample size. In the second condition, because of misrepresentation of the SV with the largest effect, there was a moderate underrepresentation of individuals with above-average performance but overrepresentation of individuals with below-average performance, leading to altered means and variances of the raw score distributions at every single age level. The pattern in the third condition equaled condition two except that the deviations from representativeness were considerably larger this time. This way, we were able to investigate the effects of WCN depending on the degree of representativeness. In the fourth condition, both individuals with above-average performance and below-average performance were underrepresented across all age groups, resulting in unbiased means but slightly altered variances at each age level. This condition was of interest, because we assumed that weighting would be specifically problematic if the tails of the distributions were underrepresented. In the fifth condition, the random sampling was manipulated such that the marginal probabilities of all SVs on average matched the reference population, but the joint probabilities, i.e., the proportions of the overall 27 cross-classifications of the SVs, did not. In this specific condition, as in condition four, the means on average matched the reference population, but the variances did not. In contrast to condition four, however, not only subgroups at the tails of the distributions were underrepresented. Instead, there was a continuously alternating over- and underrepresentation of the 27 subgroups. And finally, in the sixth condition we modeled clustered sampling. In this condition, both the marginal and joint probabilities approximated the reference population, but only when averaged across all age cohorts. For each individual cohort, though, two thirds of the cross-classification cells were set to zero. Note that in conditions five and six, raking was supposed to barely contribute to compensating for the lack of representativeness, because the marginal probabilities of every single SV were unbiased in these cases.

Our pre-registered hypotheses about the outcomes of our simulation study were as follows:

1. Overall, we expected that WCN would lead to less biased estimates of the norm scores in terms of root mean square error (*RMSE*) and mean signed difference (*MSD*) as compared to SCN, hence improving the test norms.

2. We expected to find an interaction between the representativeness of the norm sample and the effectiveness of weighting. Specifically, we did not expect WCN to completely eliminate the lack of representativeness of the norm sample. Therefore, the norming error was supposed to increase with the bias of the norm sample, but to a lesser extend with WCN as compared to SCN.

3. We expected to find an interaction between the effectiveness of weighting and the estimated person location (as expressed in normed scores), with WCN being less effective at the tails of the distributions.

In addition, we assumed that WCN would not be equally efficient under conditions three, four, five and six, but we had no clear expectations about the direction of these effects. For example, we assumed that raking would not work under condition six, but that SCN would compensate for all or at least part of this shortcoming. Therefore, we can only provide post-hoc explanations for the observed effects, which will be elaborated on in the discussion section.

# Methods

## Data generation

### Overview

To assess the efficiency of WCN, we (a) first established a population model describing the development and distribution of a typical cognitive ability, (b) drew norm samples with different biases and (c) applied the retrieved normed scores to a large test dataset. The distribution of the cognitive ability within a population was specified as a function of age and three fictious SVs, that is, each individual in this population was characterized by age, a certain level of cognitive ability and belonged to a particular stratum spanned by the cross-classification of three SVs. Because the complexity of the simulation was relatively high already, we constrained the correlation between the different SVs to zero, that is, the joint probabilities matched the product of the marginal probabilities within the reference population. For sampling condition 1, that is, unbiased sampling, we compiled a population sample that exactly satisfied the population model and from which random norm samples of the size *N* = 600 (100 per cohort) were repeatedly drawn to serve as a benchmark for all other conditions. For the five remaining sampling conditions, we also compiled large population samples, from which random norm samples of the same size were subsequently drawn. The distribution of age and of the cognitive ability within each stratum of these population samples were exactly the same as for the unbiased population sample, but the SVs were distributed differently. Thus, it was possible to mimic the different sampling scenarios with varying types of violations of the representativeness and yet perform random sampling. After drawing the norm samples, test results for each individual were simulated by entering the individual’s *z*-standardized cognitive ability into a one-parameter logistic (1-PL) model. The fictitious test scale contained 31 test items, with the item difficulties covering the whole ability range. We subsequently applied both WCN and SCN to convert these test results into IQ scores that were or were not corrected for lack of representativeness. We will refer to this IQ scores as *IQ*WCN and *IQ*SCN in the following. Both kinds of IQ scores were finally compared to the best estimate of the person location that can possibly be obtained with a given test (in the following referred to as *IQ*best). The *IQ*best was derived from the distribution of test results of the entire reference population instead of small norm samples only. For each sampling condition, we repeated the process of drawing a random norm sample, simulating test scores, determining *IQ*WCN and *IQ*SCN and comparing them to *IQ*best 100 times.

### Population model

For the population model, we used six different age cohorts, i.e. six different age groups with the size of one year each. The cognitive ability increased with age in a slightly curved way, as is typical for the development of cognitive abilities during adolescence. The cognitive ability was influenced by three different SVs, each of which had three different levels with defined proportions in the reference population, as indicated in table 1. In all SVs, level 1 contained individuals with mean performance above the population average, level 2 represented average mean performance and level 3 represented below-average performance.

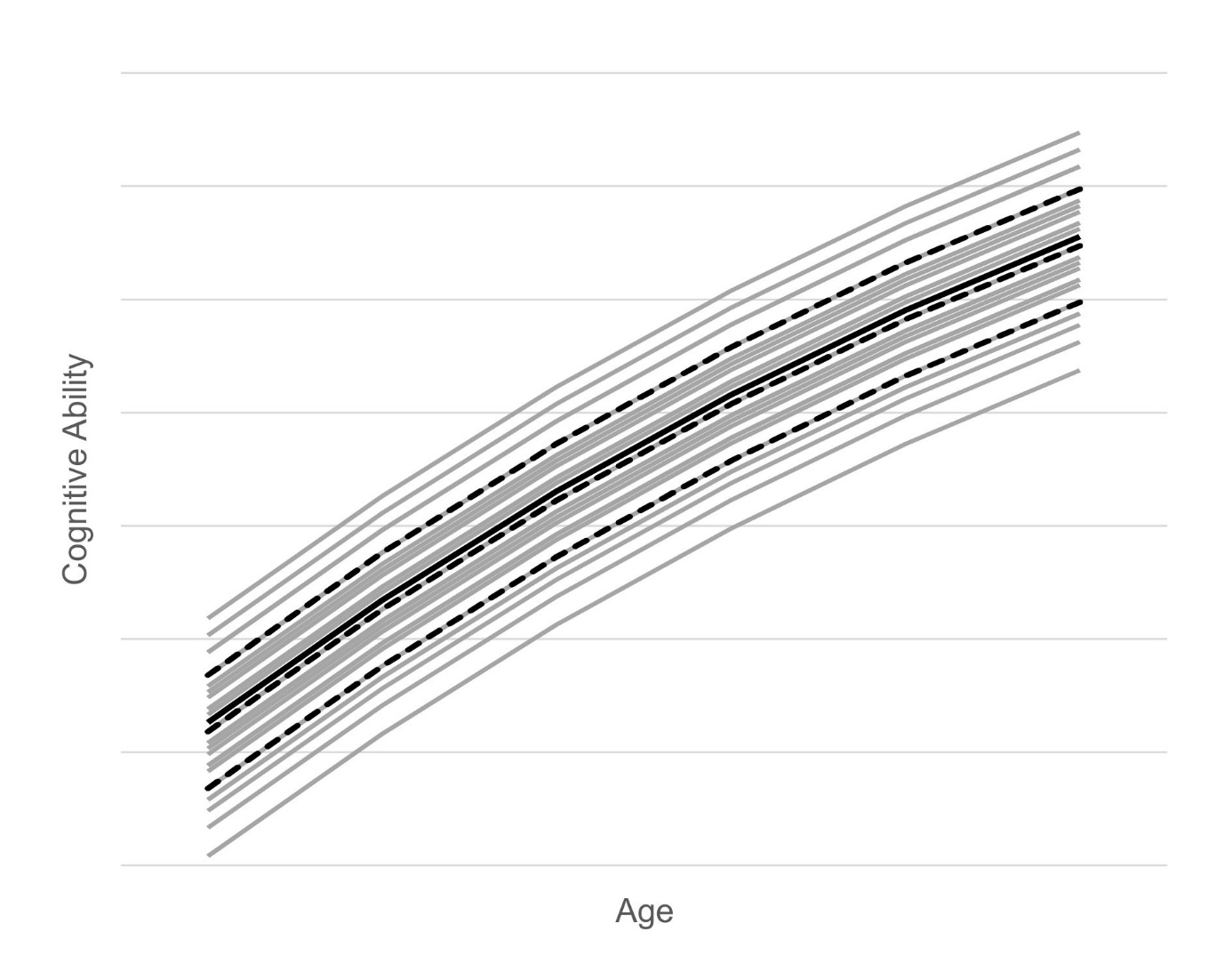
**Table 1**

*Proportions of the levels of the three stratification variables in the reference population*

|  |  |  |  |
| --- | --- | --- | --- |
| Stratification variable | Level 1 | Level 2 | Level 3 |
| 1 | 40 % | 20 % | 40 % |
| 2 | 30 % | 40 % | 30 % |
| 3 | 60 % | 20 % | 20 % |

Taking all cross-classifications into account, the population consisted of 3 x 3 x 3 = 27 strata that differed with regard to the group mean of the cognitive ability. Within each stratum and at each age level, the *SD* of the cognitive ability was constrained to 1. The relationship between the group mean of the cognitive ability, the levels of the SVs and age was described by the following polynomial:

(1)

****As a consequence, the SVs had different effects on the cognitive variable. The correlation of SV1 with the cognitive variable was *r* = -.78 at a given age level, that is, the effect was very large. The correlations of SV2 and SV3 were *r* = -.54 (medium effect) and *r* = -.31 (small effect), respectively. The population model is illustrated schematically in Figure 1.

***Figure 1.*** Population model: The mean cognitive ability increases across the six age cohorts. The black solid line represents the population mean. The black dotted lines represent level 1, 2 and 3 of SV1, that is, the SV with the largest effect on the cognitive ability. The grey lines represent the 27 strata with stratum1,1,1 being the best-performing group and stratum3,3,3 showing the lowest performance.

### Reference population

To simulate a reference population, we drew 24,000,000 pairs of random numbers (4,000,000 per age cohort), whereby each pair represented one individual in the population. The size of the population was roughly based on the corresponding cohort size of the US population. The first random number was uniformly distributed between 0 and 6 and represented age. The second one was normally distributed with *M* = 0 and *SD* = 1 and represented the cognitive ability of the individual with respect to other individuals of the same stratum and age. This random number was converted into the population-specific cognitive ability of the individual by adding the respective age- and stratum-specific mean of the cognitive ability (see formula 1). We thereby took into account the different proportions of the SVs in the reference population. For example, 7.2 % (= 40 % \* 30 % \* 60 %) of all individuals were assigned to stratum1,1,1 (SV1 = 1, SV2 = 1, SV3 = 1) and so on.

To be able to produce test results based on a 1-PL model later, we subsequently *z*-standardized all cognitive abilities using the mean and standard deviation of the entire population according to formula 2.

The resulting variable *θ*pop described a person’s latent ability with respect to the entire reference population.

### Population samples and norm samples

The reference population served as unbiased random sampling (i.e., sampling condition 1). To be able to mimic different biased sampling conditions, we generated five additional population samples in the same way, but this time, the proportions of the subgroups differed from the reference population with regard to their marginal or joint probabilities (cf. Rationale):

* Sampling condition 2: Level 1 of SV1 was underrepresented (28 % instead of 40 %) and level 3 of SV1 was overrepresented (52 % instead of 40 %), all other marginal probabilities were unbiased. This scenario reflects an overrepresentation of strata from underprivileged groups (e. g., low parental education).
* Sampling condition 3: The pattern of misrepresentation was the same as in sampling condition 1, but the degree of misrepresentation was higher (level 1 of SV1: 20 % instead of 40 %; level 3 of SV1: 60 % instead of 40 %)
* Sampling condition 4: Level 2 of SV1 was heavily overrepresented (40 % instead of 20 %) and the other two levels of SV1 were underrepresented (30 % instead of 40 %). The misrepresentations on average were as high as in sampling condition 3, but the effects on the raw score distributions were different: This manipulation reduced the variance of the raw score distributions because the tails of the distributions where underrepresented, but it did not alter the mean.
* Sampling condition 5: The joint probabilities showed large misrepresentations, with some of the 27 subgroups being less than half as large and others almost twice as large as they should be according to their respective joint probabilities in the reference population. Since the marginal probabilities nevertheless perfectly matched the reference population, this manipulation had little impact on the total mean, but it increased the variance of the raw score distributions.
* Sampling condition 6 (clustered sampling): In each separate age cohort, two thirds of the 27 strata were not represented at all. As compensation, the remaining subgroups were tripled in size. This was done in an alternating way across the cohorts, leaving both the marginal and joint distributions of the SVs unchanged.

The *z*-standardization of the cognitive abilities in all six population samples was performed with the mean and standard deviation of the entire unbiased reference population. Finally, we randomly drew 100 different norm samples with the size *N* = 600 (100 cases per age cohort) from each of the six population samples.

### Simulation of test results

We used a fictious test scale with 31 items to simulate test results for each of the 600 norm samples. The item difficulties *δ*i were generated randomly and varied uniformly between -3 and +3 (*M* = -0.04, *SD* = 1.64), so that the whole performance range was covered. The probability *p*k,i a person k with the latent ability *θ*pop\_k solved item *i* was given by the following 1-PL equation:

(3)

For every person *k* and item *i* a uniformly distributed random number between 0 and 1 was drawn and compared to *p*k,i. If *p*k,i exceeded the random number, the item was scored 1, otherwise it was scored 0. Finally, the scores of all 31 items were added up to that person's test raw score.

### Weighted and unweighted norming

To be able to compute *IQ*WCN and *IQ*SCN, each of the norm samples was processed twice: once with and once without additional weights. To determine the weights, we used the *rake* function from the *survey* package on R (Lumley, 2011), which is an implementation of the raking procedure described in the introduction section. Additionally, we standardized the weights by dividing every weight with the smallest weight in the respective norm sample. This transformation does not change the proportions of the strata, but it has the advantage that the weights can be interpreted more intuitively, with the most underrepresented subgroup always being assigned a weight of 1. Continuous norming was performed with the *cNORM* package on R (A. Lenhard et al., 2018). In case of WCN, the weights were entered twice in the norming process. First, they were used for the initial ranking process; second, they were used in the regression (for a detailed description of the norming process with the *cNORM* package, see A. Lenhard, Lenhard, & Gary, 2019). In case of SCN, no weights were used.

To perform the regression, cNORM draws on the *regsubsets* function of the leaps package on R (Lumley, 2017). This function is already prepared for using weights. By contrast, the ranking process of the cNORM package had to be modified accordingly. For example, a weight of 2 means that this raw score must be counted twice when determining the raw score distribution. However, since the weights were floating point numbers and not integers, all weights were multiplied by 106 and then rounded to whole numbers prior to the ranking process. This transformation made it possible to count the test results as many times as indicated by the relation between the weights. Subsequently, because of the high number of ties, the average rank was used for further processing, as is usual in the *cNORM* package.

### START HEREBest estimate of IQ score

To compute *IQ*best, the fictious test scale was applied to the entire reference population of 24,000,000 individuals. To convert the test results into age-specific IQ scores, we first divided the population into 365 x 6 distinct age groups. Thus, each reference group consisted of roughly 11,000 individuals with the same “birthday”. Subsequently, the test results were ranked and converted into IQ scores using rank-based inverse normal transformation for each age group separately. Since the entire population was used to determine these IQ scores, they can be regarded as the best estimate of the person location that can possibly be obtained with this specific test Note that each case was characterized by age, raw score and *IQ*best.

## Assessment of model fit

To be able to compare WCN to SCN, we computed *RMSE* and *MSD* of *IQ*WCN and *IQ*SCN with respect to *IQ*best for every resulting norm model. The *RMSE* reflects an overall measure of the model error, that is, it captures constant as well as variable errors of the used norming procedure (Lenhard & Lenhard, 2021). It was computed using the following formula:

where is the number of cases and *IQ*. stands for either *IQ*WCN or *IQ*SCN.

The *MSD* was chosen as additional measure, since it reveals general biases of the normed scores, that is, general tendencies to overestimate (*MSD* > 0) or underestimate (*MSD* < 0) the person location. The formula used to calculate the *MSD* was:

Since we did not only want to investigate the overall effectiveness of weighting, but also its dependence on the performance level, the performance range was split up into 11 intervals of 7.5 IQ points each and *RMSE* and *MSD* were calculated for each of these intervals separately.

The resulting data were generally subjected to 6 x 11 x 2 mixed ANOVAs with *RMSE* or *MSD* as the dependent variables, sampling condition as between factor and person location and norming method (WCN vs. SCN) as within factors. The assumption of sphericity was tested and the degrees of freedom where corrected accordingly. Additionally, partial η2’s were computed as measures of the effect size. Because of the large statistical power of the analyses, the level of significance was set to *p* = .01. Furthermore, differences of less than 0.5 IQ points were interpreted as being too small to have any practical relevance.

# Results

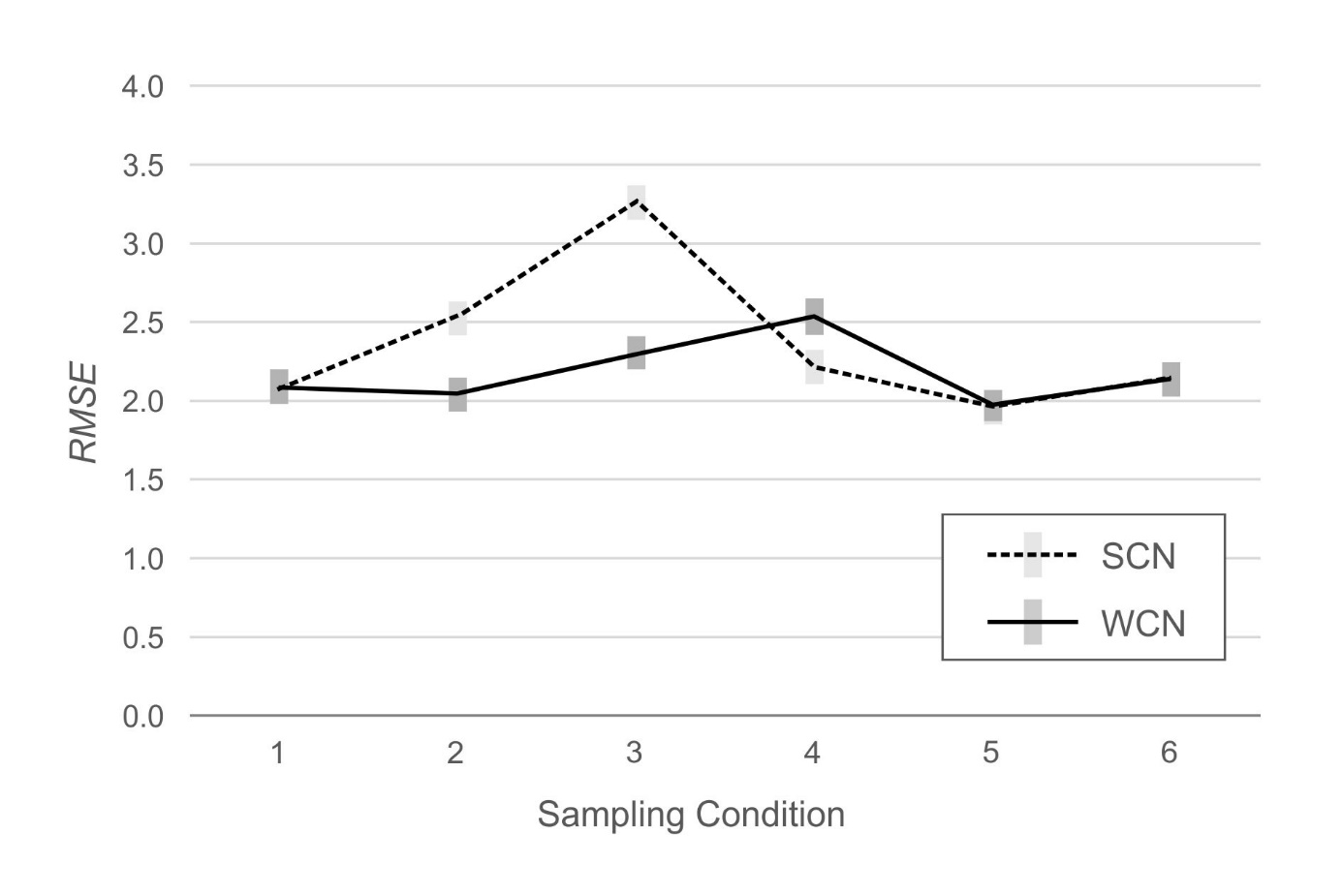
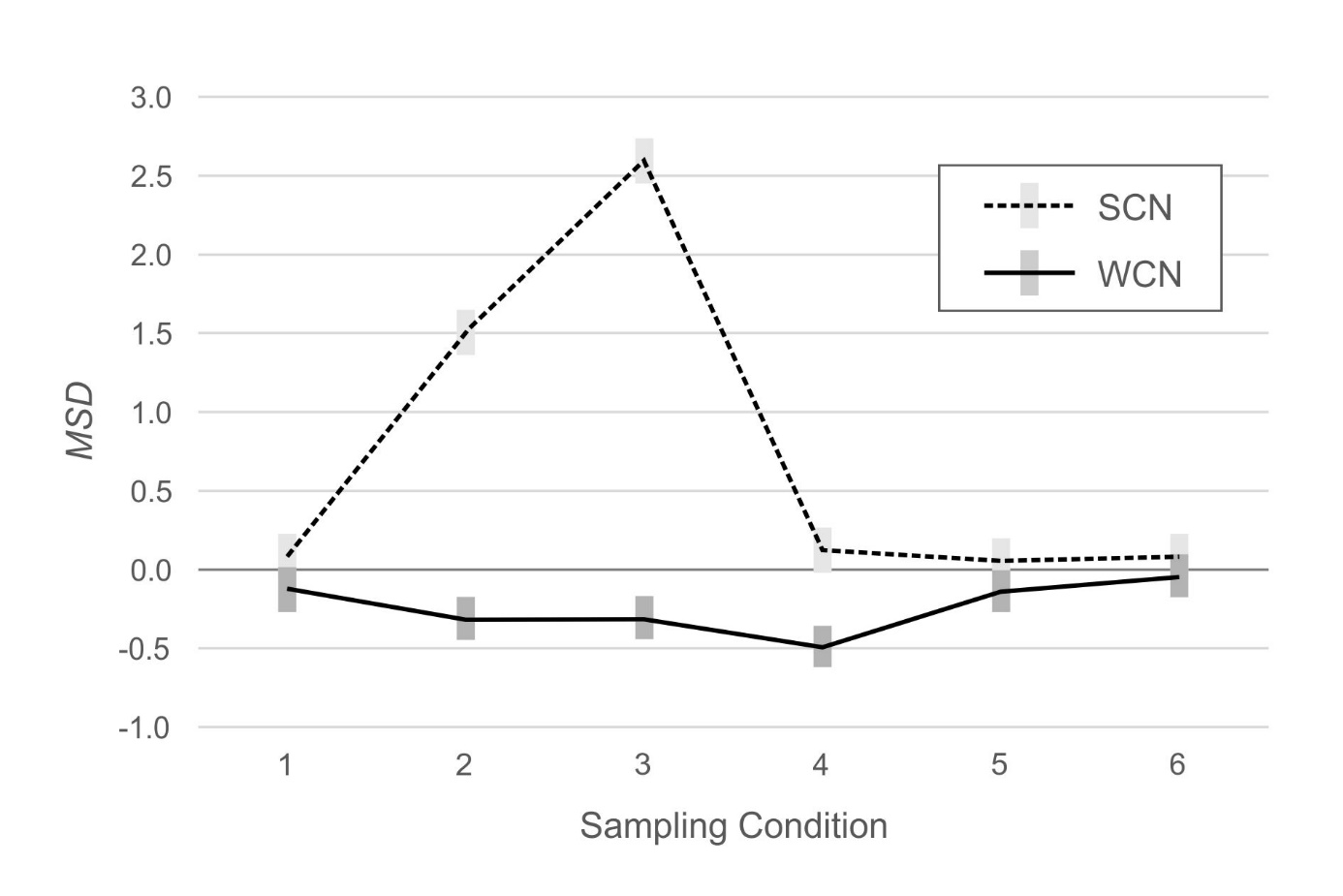
As indicated by Mauchly’s test, the sphericity assumptions were generally violated both for *RMSE* and *MSD*. Therefore, the degrees of freedom in all ANOVAs were corrected according to Greenhouse-Geisser.

Both 6 x 11 x 2 mixed ANOVAs yielded highly significant results for all main effects and all interactions (*p* < .001). We will therefore focus on the effects that are most important for deciding our hypotheses.

## Hypothesis 1

The first hypothesis assumed that WCN would generally lead to less biased estimates of the norm scores compared to SCN. The main effects of norming method indicated that this was generally the case, *RMSE*: *F*(1, 594) = 94.93, *p* < .001, *η*2 = .14, *MSD*: *F*(1, 594) = 3397.28, *p* < .001, *η*2 = .85. The *RMSE* was lower for WCN (*M* = 2.18, *SE* = .02) than for SCN (*M* = 2.36, *SE* = .02). The same was true for the MSD (WCN: *M* = 0.74, *SE* = .03; SCN: *M* = -0.24, *SE* = .03).

But in both cases, the main effects were qualified by significant interactions with the sampling condition, indicating that weighting was not equally successful under every single sampling condition, *RMSE*: *F*(5, 594) = 98.98, *p* < .001, *η*2 = .45, *MSD*: *F*(5, 594) = 764.77, *p* < .001, *η*2 = .87. As can be seen in Figure 2 (*RMSE*) and Figure 3 (*MSD*), in two of the six sampling conditions, weighting improved the normed scores. In sampling conditions 2 and 3 the *RMSE* was reduced by 0.48 resp. 0.97 IQ points on average. In sampling conditions 1, 5 and 6, the difference between WCN and SCN with regard to *RMSE* approximated zero. But in sampling condition 4 weighting increased the *RMSE*, although the difference of

*****Figure 2.*** *RMSE* under different sampling conditions with (WCN) or without (SCN) weighting. The grey rectangles represent 95% confidence intervals.

***Figure 3.*** *MSD* under different sampling conditions with (WCN) or without (SCN) weighting. The grey rectangles represent 95% confidence intervals.

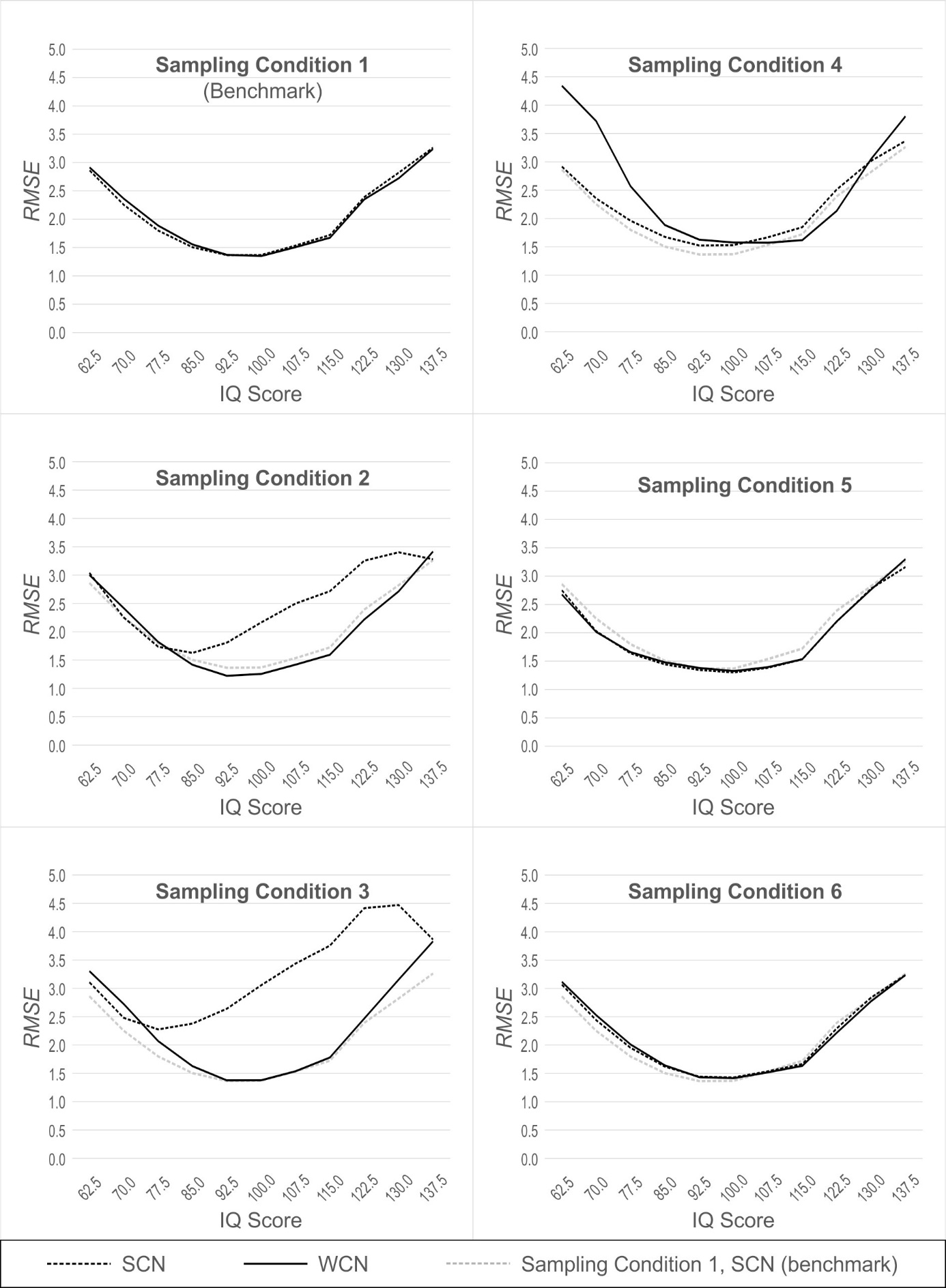
0.32 IQ points between WCN and SCN was below the defined level of practical relevance. The analysis of the *MSD* yielded equivalent results. In sampling conditions 2 and 3 the *MSD* was much closer to the ideal value of zero for WCN (both sampling conditions 2 and 3: -0.32 IQ points) as compared to SCN (sampling condition 2: 1.51 IQ points; sampling condition 3: 2.59 IQ points). In sampling conditions 1, 5 and 6, the *MSD* was generally close to zero, regardless of the norming method. But in sampling condition 4, the normed scores on average deviated slightly more from zero with WCN (-0.49 IQ points) as compared to SCN (0.13 IQ points), although, again, the difference between both methods was small.

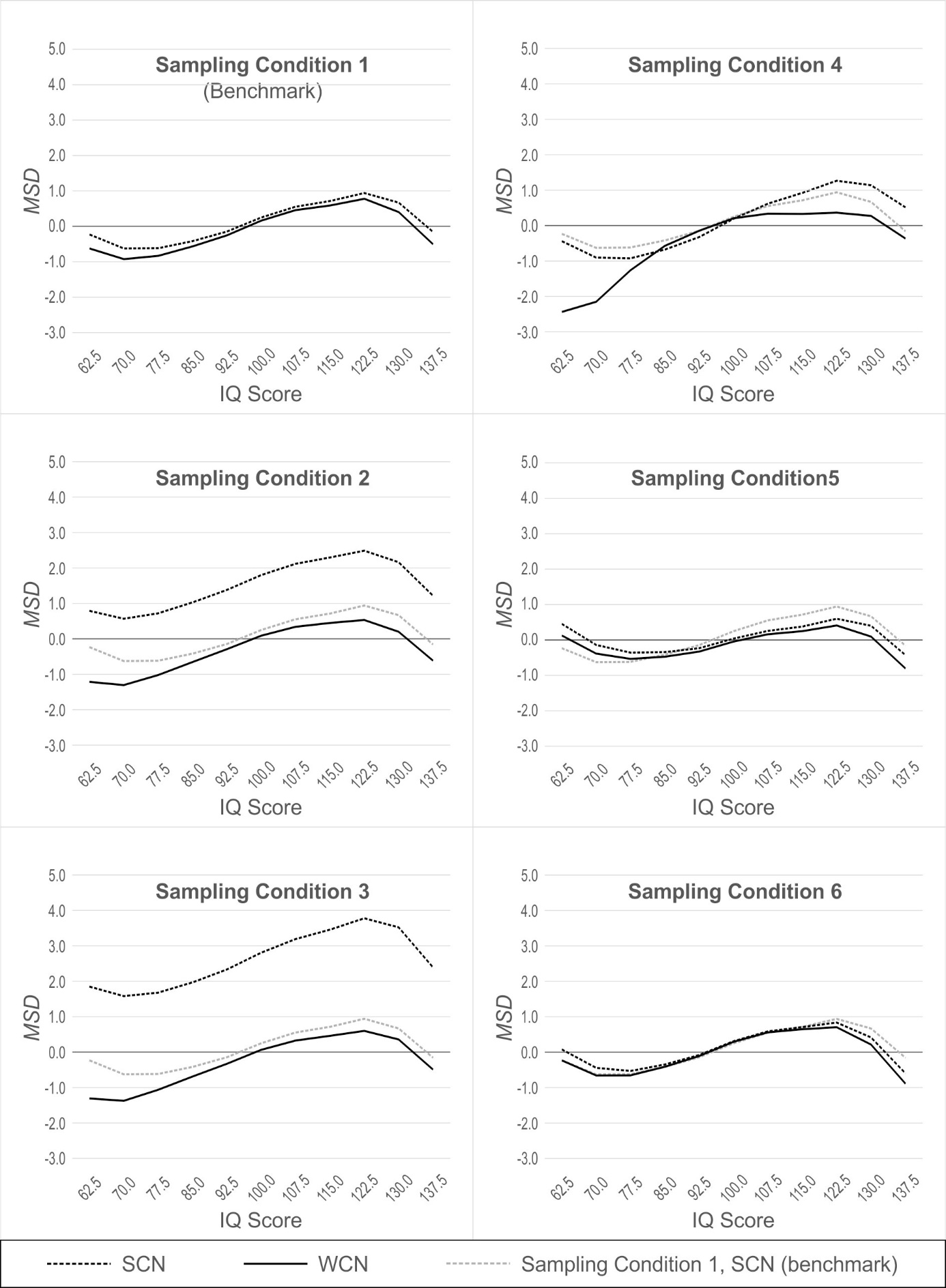
## Hypothesis 2

Hypothesis 2 stated that WCN would not be able to completely compensate for all of the norming error caused by the lack of representativeness. Therefore, the overall norming error was supposed to increase with the bias of the norm sample, but to a lesser extend with WCN as compared to SCN. To answer this hypothesis, we specifically compared sampling conditions 2 and 3. In these two conditions, the same groups were over- or underrepresented, but in sampling condition 3, the deviation from representativeness was considerably larger than in sampling condition 2. Therefore, we performed two additional ANOVAs that had the same within factors but concentrated on sampling conditions 2 and 3 only. The results of these analyses have been illustrated in Figures 2 and 3 already. In fact, both analyses yielded a significant interaction between norming method and sampling condition, *RMSE*: *F*(1, 198) = 26.01, *p* < .001, *η*2 = .12, *MSD*: *F*(1, 198) = 242.36, *p* < .001, *η*2 = .55, indicating that WCN countered the lack of representativeness. It should be noted, though, that the weights worked better with respect to the *MSD* than with respect to the *RMSE*, as can be concluded from the descriptive statistics as well as from the effect sizes. As already stated above, the *MSD* was the same for both sampling conditions, if weights were used (both sampling conditions 2 and 3: -0.32 IQ points). Moreover, in both cases it was close to zero, that is, weighting was able to prevent a general dislocation of the normed scores in one or the other direction. But with regard to the *RMSE*, the normed scores on average deviated a little bit more from *IQ*best in sampling condition 3, even when weights were used, *F*(1, 198) = 10.12, *p* = .002, *η*2 = .05. The average *RMSE* was 2.04 IQ points in sampling condition 2 and 2.29 IQ points in sampling condition 3. However, the difference of only 0.25 IQ points between both sampling conditions seems to be too small to be of any practical relevance.

## Hypothesis 3 and exploratory hypotheses

Hypothesis 3 stated that WCN would be less effective at the tails of the distributions. As already mentioned above, the original 6 x 11 x 2 mixed ANOVAs yielded threefold interactions of sampling condition, norming method and person location with considerable effect sizes, which indicated that the efficiency of weighting did not only differ with regard to the different sampling conditions but that this relation was further qualified by the person location. Therefore, to answer the question under which specific combination of sampling condition and person location WCN yields suboptimal results, we analyzed the different sampling conditions separately. We did this with two different kinds of ANOVAs. Firstly, we conducted 11 x 2 ANOVAs with person location and norming method (WCN vs. SCN) as within factors and *RMSE* respectively *MSD* as dependent variables. Secondly, in case of biased norm samples, we also compared WCN directly to the benchmark condition, that is, SCN applied to unbiased norm samples. In this case, the second factor in the ANOVA was sampling condition as between factor. The results of these analyses are illustrated in Figure 4 (*RMSE*) and Figure 5 (*MSD*). Because of the large number of results, we will only report main effects and interactions, if at least one of the differences considered in the respective analysis exceeds 0.5 IQ points.

***Figure 4.*** *RMSE* under different sampling conditions with (WCN) or without (SCN) weighting as a function of person location. The dotted grey line represents SCN with unbiased norm samples (benchmark).

***Figure 5.*** *MSD* under different sampling conditions with (WCN) or without (SCN) weighting as a function of person location. The dotted grey line represents SCN with unbiased norm samples (benchmark).

### Sampling condition 1

In the unbiased condition, the ANOVAs for both *RMSE* and *MSD* yielded a significant main effect of person location, *RMSE*: *F*(2.68, 264.87) = 70.68, *p* < .001, *η*2 = .42, *MSD*: *F*(2.34, 231.65) = 35.54, *p* < .001, *η*2 = .26. The *RMSE* generally increased with growing distance from the average IQ of 100. This is a well-known effect (cf. A. Lenhard et al., 2019) that also could be observed in all other sampling conditions. With respect to the *MSD*, the main effect of person location exhibited a sinusoidal pattern. Again, this effect reoccurred in all other sampling conditions. Since both main effects were not specific to weighting but rather to the continuous norming method in general, they do not contribute to answering the main hypotheses of this article. We will therefore refrain from reporting the respective main effects of person location for the remaining sampling conditions. We will, however, come back to the unexpected sinusoidal pattern of the *MSD* in the discussion. All other effects in simulation condition 1 were too small to be of practical relevance. To put it differently, WCN basically yielded the same results as SCN when the norm samples were unbiased.

### Sampling condition 2

The first ANOVA yielded main effects of the norming method, *RMSE*: *F*(1, 99) = 74.94, *p* < .001, *η*2 = .43, *MSD*: *F*(1, 99) = 1924.64, *p* < .001, *η*2 = .95. They indicated that WCN generally improved the quality of the normed scores. Only for the *RMSE*, an additional interaction between person location and norming method showed up, *F*(2.72, 268.76) = 41.72, *p* < .001, *η*2 = .30. With WCN, the *RMSE* was mainly reduced in the above-average range, that is, the performance range that had been underrepresented in the respective norm samples.

The second ANOVA showed that the *RMSE* produced with WCN did not even differ significantly from the benchmark, that is, WCN fully compensated for the bias in the norm samples. With respect to the *MSD*, a main effect of the simulation condition, *F*(1, 198) = 15,37, *p* < .001, *η*2 = .07, and a marginally significant interaction between simulation condition and person location, *F*(2.31, 456.93) = 2.91, *p* = .048, *η*2 = .01, indicated that WCN performed slightly worse than the benchmark at very low person locations, with an absolute difference between WCN and the benchmark (in the following referred to as |∆RMSE|) of 0.68 IQ points at IQ 70 and 0.99 IQ points at IQ 62.5.

### Sampling condition 3

The first ANOVA yielded significant main effects of the norming method, *RMSE*: *F*(1, 99) = 155.76, *p* < .001, *η*2 = .61, *MSD*: *F*(1, 99) = 2707.76, *p* < .001, *η*2 = .97, and significant interactions between norming method and person location, *RMSE*: *F*(2.79, 276.13) = 71.89, *p* < .001, *η*2 = .42, *MSD*: *F*(2.63, 260.05) = 6.58, *p* = .001, *η*2 = .06, for both *RMSE* and *MSD*. The pattern essentially replicated that of simulation condition 2, except that this time the effects were larger, that is, WCN was able to compensate for the larger biases under sampling condition 3, too.

The comparison to the benchmark in this case yielded significant main effects of simulation condition, *RMSE*: *F*(1, 198) = 9.52, *p* = .002, *η*2 = .05, *MSD*: *F*(1, 198) = 12.52, *p* = .001, *η*2 = .06, and marginally significant interactions between simulation condition and person location, *RMSE*: *F*(2.81, 557.01) = 3.06, *p* = .031, *η*2 = .02, *MSD*: *F*(2.26, 447.76) = 3.28, *p* = .033, *η*2 = .02. These effects again indicated that WCN performed slightly worse than the benchmark at very low and this time also at very high person locations. But a relevant difference for the *RMSE* only occurred at IQ 137.5 (|∆*RMSE*| = 0.57 IQ points). As far as the *MSD* was concerned, WCN performed slightly worse than the benchmark at IQ 62.5 (|∆*MSD*| = 0.68 IQ points).

### Sampling condition 4

The first ANOVA again yielded significant main effects of the norming method, *RMSE*: *F*(1, 99) = 39.02, *p* < .001, *η*2 = .28, *MSD*: *F*(1, 99) = 164.63, *p* < .001, *η*2 = .62, and significant interactions of norming method and person location, *RMSE*: *F*(3.02, 299.01) = 42.43, *p* < .001, *η*2 = .30, *MSD*: *F*(2.53, 250.47) = 42.44, *p* = .001, *η*2 = .30. But surprisingly, this time WCN did not generally provide better quality than SCN. Instead, the *RMSE* for WCN was higher at extremely low person locations. The difference between WCN and SCN exceeded 0.5 IQ points at IQ 77.5 (|∆*RMSE*| = 0.61 IQ points), IQ 70.0 (|∆*RMSE*| = 1.37 IQ points) and IQ 62.5 (|∆*RMSE*| = 1.43 IQ points). For the *MSD* the pattern was mixed, with SCN outperforming WCN at low person locations (IQ 70.0: |∆*MSD*| = 1.25 IQ points; IQ 62.5: |∆*MSD*| = 2.00 IQ points) but WCN outperforming SCN at high person locations (IQ 115.0: |∆*MSD*| = 0.60 IQ points; IQ 122.5: |∆*MSD*| = 0.90 IQ points; IQ 130.0: |∆*MSD*| = 0.87 IQ points; IQ 137.5: |∆*MSD*| = 0.88 IQ points).

Because this time, it was SCN (applied on the biased norm samples) that showed almost no relevant deviation from the benchmark, the comparison with the benchmark and WCN yielded almost exactly the same results as the first ANOVA, with main effects of simulation condition, *RMSE*: *F*(1, 198) = 38.46, *p* < .001, *η*2 = .16, *MSD*: *F*(1, 198) = 31.10, *p* < .001, *η*2 = .14, and significant interactions between simulation condition and person location, *RMSE*: *F*(2.83, 560.75) = 22.19, *p* < .001, *η*2 = .10, *MSD*: *F*(2.55, 505.02) = 19.44, *p* < .001, *η*2 = .09. Only with regard to the MSD and only for high person locations, the benchmark was at little bit closer to WCN than SCN was, so that the difference between WCN and the benchmark disappeared at person locations above IQ 122.5. At IQ 122.5, the difference was 0.58 IQ points, with WCN being slightly closer to the optimal value of zero than the benchmark.

### Sampling conditions 5 and 6

Both for sampling condition 5 and sampling condition 6, the *RMSE* was approximately the same for WCN, SCN and also for the benchmark, that is, no single difference exceeded 0.5 IQ points. With regard to the *MSD*, the second ANOVA for simulation condition 5 yielded a marginally significant main effect of the simulation condition, *F*(1, 198) = 5.02, *p* = .026, *η*2 = .03, and an interaction between simulation condition and person location, *F*(2.29, 453.98) = 6.25, *p* = .001, *η*2 = .03. For some person locations, the differences between WCN and the benchmark in fact exceeded 0.5 IQ points. However, there was no clear advantage of one condition over the other, because at IQ 122.5 and IQ 130.0, WCN was closer to zero, but at IQ 137.5 the benchmark performed better. Moreover, the difference between WCN and the benchmark reached a maximum of |∆*MSD*| = 0.73 at IQ 137.5, that is, although exceeding 0.5 IQ points, it remained small.

# Discussion

## Summary of results

The present simulation study aimed at evaluating whether a specific combination of weighting and continuous norming would be able to improve the quality of normed scores derived from non-representative norm samples. To this end, we mimicked a norming process under six different sampling conditions.

First of all, our first hypothesis that WCN leads to less biased estimates of the normed scores in terms of *RMSE* and *MSD* could be confirmed largely, but not universally. What can be stated, though, is that in five out of six different sampling conditions, the application of WCN was either beneficial for the quality of the normed scores or it was at least neutral. Only in one out of the six sampling conditions, it deteriorated the quality of the normed scores for some, but not all person locations.

In sampling condition 1, that is, when the norm samples were drawn randomly from a completely representative population, the application of WCN was neither beneficial nor detrimental. Thus, an application is unnecessary in this case, but also poses no problems. One might expect that WCN would at least level out small deviations from representativeness caused by the limited sample size. Obviously, though, these deviations are too small to measurably affect the normed scores.

In sampling condition 2, the distribution of the SV with the highest effect size deviated from the respective distribution in the reference population, with level 1 (the high performing subgroup) being moderately underrepresented and level 3 (the low performing subgroup) being moderately overrepresented. In this case, WCN clearly improved the normed scores with regard to both *RMSE* and *MSD*. In fact, the weighting method worked so well under this condition, that the normed scores were even as precise as normed scores derived from completely unbiased norm samples. Moreover, this was the case for every single person location.

When the deviation from representativeness followed the same pattern as in sampling condition 2 but was markedly higher (i.e., sampling condition 3), WCN still improved the normed scores considerably. However, this time, the normed scores did not fully reach the quality of normed scores derived from representative samples. This result generally confirms hypothesis 2. It should be noted, though, that only at extreme person locations, both *RMSE* and *MSD* tended to be larger than in the benchmark condition. Fortunately, the impairments were unexpectedly small and did not even exceed 1 IQ point at any single person location. This is overall an encouraging finding, since the impact of the stratification variable on the test results was very high and so was the bias of the norm sample.

In sampling condition 4, the average performing subgroup was overrepresented and the above respectively below performing subgroups were underrepresented, leading to unbiased means but reduced variances of the raw score distributions. Without weighting, the impact of this manipulation on the normed scores was much less detrimental than in conditions 2 and 3. In fact, the quality of the normed scores generated without weighting almost remained unaffected as compared to the quality of normed scores derived from unbiased samples. Note that in sampling condition 4 the average misrepresentation of the different groups was as large as the misrepresentation in sampling condition 3. The only difference between these conditions was, which groups were over- and which were underrepresented. In this specific sampling condition, weighting did not improve the normed scores, but it even deteriorated them. As predicted in hypothesis 3, the deterioration was not observable across all person locations, but increased with growing distance from the population mean (i.e., IQ 100), and was particularly high at very low person locations. This result confirms that weighting can have negative effects specifically at the tails of the distributions. As already explained in the introduction section, weighting does not affect the variance in the respective subgroups, as would usually be the case with increased subsample size. Imagine, for example, that a certain subgroup of the norm sample contains only 5 instead of 10 individuals. Since this is a very small subgroup, the mean and variance within this subgroup could by chance deviate considerably from the true mean and variance of this subgroup within the population. So counting every single case twice would also double the error this subgroup contributes to the norming error. If it is an average performing subgroup, then there are probably many observations from other subgroups that can at least partially compensate this bias, because the center of the normal distribution is also the location with the highest density of observations. If, however, the subgroup is located at one or the other tail of the distribution, the error can barely be compensated, because the number of observations is generally small at the tails. Therefore, although the median of the raw score distribution might be adjusted adequately by using weights, the percentiles at the tails are not.

In sampling condition 5, the marginal probabilities of the SVs in the norm samples approximately matched the probabilities in the reference population, but the joint probabilities did not. Theoretically, this condition seemed to be interesting, because we assumed that the raw score distributions would be biased without raking being able to correct for this bias. Admittedly, despite our best efforts, we did not fully succeed in establishing the joint probabilities such that the means and variances of the raw score distributions were strongly flawed. Clearly, in practical use cases, fixing the marginal probabilities with raking will not always fix the joint probabilities. Yet, it seems to be very unlikely that the marginal probabilities of the SVs match the target probabilities, but the normed scores are still highly biased. As a caveat, we must admit that in this specific simulation study, the SVs were constructed such that one of them had a relatively large effect, one had a medium effect, and one had only a small effect with an additional small interaction between the latter two. Thus, the SV with the large effect mainly determined the outcome, and the marginal probabilities of this variable matched the reference population. Although we tried to model the effects of the SVs according to actual effects of SVs (education, ethnicity and region) from a German vocabulary test (A. Lenhard et al., 2015), the effects of SVs on other real-life variables may differ depending on the specific circumstances. For example, there might be more than one SV with a large effect or stronger interactions between the variables. In such cases, neglecting the joint probabilities could have more detrimental effects. How this problem can be treated in practice is outlined in more detail below.

Finally, the clustered sampling simulated in sampling condition 6 also did not negatively affect the normed scores, that is, neither with WCN nor with SCN the normed scores deviated measurably from the benchmark. In this condition, the marginal and joint probabilities matched the target probabilities across all age cohorts, but not within each single cohort. As a consequence, the raw score distributions were in fact significantly biased in every single cohort. Raking cannot have compensated for this distortion, because the marginal probabilities entered in the norming procedure were not specified for each cohort separately, but only across all cohorts. Nevertheless, neither the *RMSE* nor the *MSD* were significantly impaired as compared to the benchmark condition, regardless of whether weighting was applied or not. We assume that in this case, the semi-parametric continuous norming method that was used together with raking was responsible for leveling out the lack of representativeness in the different age cohorts. As already described in the introduction, this method has the feature that the modeled curves are not overly flexible, that is, this method would rather not result in wavy iso-percentiles with this method. In the case of cognitive variables, this feature does not negatively affect the model fit, because the percentiles of cognitive variables usually develop monotonically – not wavy – across age. In the case of clustered sampling, as expected, the stiffness of the method even helps to level out lacks of representativeness limited to single cohorts. Therefore, the combination of raking and semi-parametric continuous norming as implemented in the cNORM package (A. Lenhard et al., 2018) seems to be well suited for modeling norms under clustered sampling conditions, that is, with high representativeness across all age groups but not within each separate age group.

## Implications for the use of WCN in test norming

In summary, we showed that the combination of raking and semi-parametric continuous norming improves rather than deteriorates test norms. However, we were also surprised at how robust the test norms were to violations of representativeness. Even with the largest violation of representativeness, the difference of the *RMSE* between uncorrected normed scores and normed scores derived from representative samples of the same size was only 2 IQ points at most. This was the case even though the correlation between the biased SV and the dependent variable was unusually high. This result certainly would not be obtained with every norming method, especially not with conventional norming per age group, because with the latter method, deviations from representativeness that are limited to single age groups cannot be compensated by other age groups, as is the case with continuous norming. Consistent with this view, we have demonstrated previously that with conventional norming per age group, the *RMSE* is on average about twice as high as with semi-parametric continuous norming, even with representative random samples (W. Lenhard & Lenhard, 2021). The selection of an appropriate norming method is therefore probably the most important prerequisite for high quality test norms, regardless of whether this procedure is used with or without weighting. By contrast, the size of the norm sample only plays a minor role, if continuous norming is used. For example, in other simulation studies we could demonstrate that with continuous norming the *RMSE* was only slightly reduced when the sample size was increased to 250 instead of only 100 per age group (A. Lenhard et al., 2019).

Certainly, representativeness should nevertheless not be neglected as a factor for high quality test norms. First and foremost, it is important to use random samples to establish representativeness as compared to samples that have accrued in a clinical context, since the latter are likely to be heavily biased toward low performance. Second, as shown in this simulation study, weights can compensate moderate deviations from representativeness, but they cannot compensate insufficient variance. Care must therefore be taken to ensure that the tails of the distributions are not heavily underrepresented in the sample. This is not necessarily important with regard to all possible SVs, but at least with regard to SVs with high effect sizes. If the outlying subgroups are adequately represented, we believe it is very unlikely that the norms will be deteriorated by the use of raking.

In our own simulation study, joint probabilities only played a minor role and in most cases, taking them into account is not even an option, simply because they are only rarely available. Nevertheless, in cases where they are available, they could in fact be used together with raking, if the data were prepared accordingly. The only requirement is that each cross-classification is treated as a separate stratum. For example, if the joint probabilities of two different SVs with three levels each are given, the nine different joint probabilities can be treated as marginal probabilities of one SV with nine different levels. Yet, we generally recommend not to use too many strata, because this would bear the risk that individual strata are heavily underrepresented. As a consequence, the according weights would become very high and the sampling error contained in the respective subgroups would also be multiplied when using these weights. In order to reduce the number of strata, we would recommend combining different subgroups into a single one, if these different subgroups do not differ significantly with regard to their mean performance. This merging of subgroups is not necessarily limited to cases where joint probabilities are used, but it can generally be applied to reduce the number of strata. And finally, the deviations from representativeness should not be too high. In our simulations, the efficiency of weighting was reduced, when the size of one of the three subgroups of SV 1 was 20 % instead of 40 %, that is, the subgroup size was only half as large as would have been necessary to establish representativeness.

## Limitations of the study

First of all, in our simulation studies, we compared the normed scores generated with raking to normed scores generated without weighting, that is, we did not compare raking to other methods of post-stratification. Specifically, we did not compare it to fully cross-classified post-stratification. Moreover, in sampling condition 5, which was specifically designed to test the weakness of raking compared to fully cross-classified post-stratification, not even the normed scores generated without raking deviated significantly from the benchmark. Therefore, the misrepresentations of the different strata applied in sampling condition 5 might have been too small to uncover the shortcoming of raking, which consists in the consideration of only marginal probabilities instead of joint probabilities. As discussed above, we assume that neglecting joint probabilities would only rarely degrade the quality of test norms. We did not even manage to manipulate the joint probabilities in such a way that the raw score distributions were significantly biased without also affecting the marginal probabilities. Admittedly, under certain conditions violations of the joint probabilities might nevertheless play a larger role, for example, in the case of stronger interactions of the SVs. On the other hand, the effect sizes of the SVs in our study certainly represented the upper edge of what realistically occurs. Therefore, the biases of normed scores that can be expected due to non-representative samples are probably even smaller than in this simulation study.

Second, we only used one fixed population model, that is, the correlations of the three different SVs with the dependent variable and its development across age were the same in all simulations. Specifically, we mimicked the development of a cognitive ability or achievement during adolescence and the according effects of typical SVs. However, a lot of variables measured with psychometric tests, for example, personality traits such as the big five (Donnellan & Lucas, 2008) do not display the same dependency on age or do not have the same confounding variables. Since the development across age is usually less pronounced for such personality traits, a weighting method that is not combined with continuous norming might be more appropriate in these cases.

A third limitation could be the fact that the maximum misrepresentations in the norm samples we used in this simulation study were large in our own eyes, but perhaps still not realistically large given the huge quality range of psychometric tests. We would therefore like to point out here that the possibility of at least partially correcting bias in representativeness by means of mathematical methods does not release test constructors from the obligation to carefully assemble norm samples.

Fourth, the cNORM package usually requires a careful selection of appropriate model parameters as, for example, the exponents of location and age. Since hundreds of models must be selected in simulation studies like the present one, the norming models must be selected automatically using algorithms. The criteria used to select the models in this study were rather simple. For example, we used a default setting k = 4 for both exponents. This can, however, lead to systematic distortions of the norms. It is possible that the sinusoidal shape of the *MSD* was a consequence of these default settings. Thus, with more optimal selection criteria, perhaps the differences between the individual conditions could have been identified more clearly. Further simulation studies not reported in this paper have shown that an exponent of 3 for age and 5 for person location used in the cNORM package can further reduce the norming error. We therefore recommend to use these exponents by default and have even changed the corresponding default setting in the cNORM package.

Finally, in the context of weighting methods, we did not compare the semi-parametric continuous norming approach implemented in the cNORM package with other continuous norming approaches such as parametric continuous norming (e.g., Stasinopoulos et al., 2018). Therefore, we can only speculate about how other continuous methods would perform if combined with raking. As already discussed, from a theoretical point of view, the stiffness of the cNORM approach seems to be better suited for clustered sampling than other, more flexible approaches, for example, approaches that draw on splines to model the age development of the measured variable. Moreover, we have already shown that the cNORM approach seems to be more favorable than parametric continuous norming when the sample size is limited to 150 per cohort and below or when the raw score distributions are markedly skewed (A. Lenhard, 2019). Yet, the efficiency of post-stratification techniques combined with parametric continuous norming remains to be investigated.

## Concluding Remarks and Outlook

The enhancement of norming techniques is a relatively recent and not very widespread scientific field. Therefore, the number of unanswered questions in this area is high and can certainly not be dealt with in one single article or simulation study. During the software implementation and analysis of the simulated data, we came across a lot of unresolved problems and had many ideas that we would like to explore in more detail in the future. For example, we used the weights generated by the raking method twice: Once for the ranking of the raw scores and once for the subsequent regression performed in the cNORM package. But the application of the weights in the regression could possibly be of little benefit or even detrimental because the raw score distributions might be less reliable at underrepresented locations, that is, in groups with high weights.

In analyzing the data, it also became clear that combining different mathematical methods is too complex to predict the benefits or harms of such combinations under different conditions. We therefore consider it necessary to subject each and every combination of mathematical methods to close scrutiny before practical application.

# Acknowledgements

We want to thank WPS publisher in Torrance, CA, for funding the simulation studies.

# References

Cole T. (1988). Fitting smoothed centile curves to Reference Data. *Journal of the Royal Statistical Society Series A (Statistics in Society), 151*(3), 385.

Cole, T. J., & Green P. J. (1992). Smoothing reference centile curves: The lms method and penalized likelihood. *Statistics in Medicine, 11*(10), 1305–19.

Donnellan, M. B., & Lucas, R. E. (2008). Age differences in the Big Five across the life span: evidence from two national samples. *Psychology and aging, 23*(3), 558–566. https://doi.org/10.1037/a0012897

Ireland, C. T., & Kullback, S. (1968). Contingency tables with given marginals. *Biometrika, 55*(1), 179–188. https://doi.org/10.1093/biomet/55.1.179

Kalton, G., & Flores-Cervantes, I. (2003). Weighting Methods. *Journal of Official Statistics, 19*(2), 81-97.

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Assessment Battery for Children Second Edition.* San Antonio: Pearson Clinical Assessment.

Kennedy, L., & Gelman, A. (2021). Know your population and know your model: Using model-based regression and poststratification to generalize findings beyond the observed sample. *Psychological Methods, 26*(5), 547–558. https://doi.org/10.1037/met0000362

Lenhard, A. & Lenhard, W. & Gary, S. (2018). cNORM. (Continuous Norming). *The Comprehensive R Network.* https://cran.r-project.org/web/packages/cNORM/index.html. doi:10.1177/1073191116656437

Lenhard, A., Lenhard, W. & Gary, S. (2019). Continuous norming of psychometric tests: A simulation study of parametric and semi-parametric approaches. *PLOS ONE, 14*(9), e0222279. https://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*(1), 112-125. doi.org/10.1177/1073191116656437

Lenhard, A., Lenhard, W., Segerer, R. & Suggate, S. (2015). *Peabody Picture Vocabulary Test - Revision 4 (PPVT-4)*, German adaptation. Pearson Assessment.

Lenhard, W. & Lenhard, A. (2021). Improvement of Norm Score Quality via Regression-Based Continuous Norming. *Educational and psychological measurement, 81*(2), 229-261. https://doi.org/10.1177/0013164420928457

Lenhard, W., Lenhard, A. & Schneider, W. (2017). *ELFE II - Ein Leseverständnistest für Erst- bis Siebtklässler.* nHogrefe.

Little, R. J. (1993). Post-stratification: A modeler’s perspective. *Journal of the American Statistical Association, 88*(423), 1001-1012. https://doi.org/10.2307/2290792

Lumley, T. (2011). *Complex Surveys – A Guide to Analyses Using R*. Wiley.

Lumley, T. (2017). leaps: Regression Subset Selection. *The Comprehensive R Network.* https://cran.r-project.org/web/packages/leaps/index.html

Park, D. K., Gelman, A., & Bafumi, J. (2004). Bayesian multilevel estimation with post-stratification: State-level estimates from national polls. *Political Analysis, 12*(4), 375-385.

Price-Mohr, R., Price, C. (2017). Gender Differences in Early Reading Strategies: A Comparison of Synthetic Phonics Only with a Mixed Approach to Teaching Reading to 4–5 Year-Old Children. *Early Childhood Education Journal, 45*, 613–620. https://doi.org/10.1007/s10643-016-0813-y

Sang-Wook Y., Heechoul O., Soon-Ae S., & Jee-Jeon Y. (2015). Sex-age-specific association of body mass index with all-cause mortality among 12.8 million Korean adults: a prospective cohort study, International Journal of Epidemiology*, 44*(5), 1696–1705. https://doi.org/10.1093/ije/dyv138

Stasinopoulos M. D., Rigby, R. A., Voudouris, V., Akantziliotou, C., Enea, M. & Kiose, D. (2018). gamlss: Generalised Additive Models for Location Scale and Shape. *The Comprehensive R Network.* https://cran.r-project.org/web/packages/gamlss/index.html

Wechsler, D. (1939). The measurement of adult intelligence. Williams & Wilkins Co. https://doi.org/10.1037/10020-000

Wechsler, D. (2008). *WAIS-IV Technical and interpretive manual*. Pearson.

Wechsler, D. (2014). *WISC-V Technical and interpretive manual.* Pearson.

Zachary, R. A., & Gorsuch, R. L. (1985). Continuous norming: Implications for the WAIS-R. *Journal of Clinical Psychology, 41*(1), 86–94.

# Figure Captions

***Figure 1.*** Population model: The mean cognitive ability increases across the six age cohorts. The black solid line represents the population mean. The black dotted lines represent level 1, 2 and 3 of SV1, that is, the SV with the largest effect on the cognitive ability. The grey lines represent the 27 strata with stratum1\_1\_1 being the best-performing group and stratum3\_3\_3 showing the lowest performance.

***Figure 2.*** *RMSE* under different sampling conditions with (WCN) or without (SCN) weighting. The grey rectangles represent 95% confidence intervals.

***Figure 3.*** *MSD* under different sampling conditions with (WCN) or without (SCN) weighting. The grey rectangles represent 95% confidence intervals.

Figure 4. *RMSE* under different sampling conditions with (WCN) or without (SCN) weighting as a function of person location. The dotted grey line represents SCN with unbiased norm samples (benchmark).

Figure 5. *MSD* under different sampling conditions with (WCN) or without (SCN) weighting as a function of person location. The dotted grey line represents SCN with unbiased norm samples (benchmark).