Reducing the bias of norm scores in non-representative samples: Weighting as an adjunct to continuous norming methods

Sebastian Gary[1], Alexandra Lenhard[1], Wolfgang Lenhard[2*], David S. Herzberg[3]

[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 accuracy of normed test scores derived from non-demographically representative samples can be improved by combining semi-parametric continuous norming methods with compensatory weighting at the raw score level. In a simulated reference population, we modeled a latent cognitive ability with a typical developmental gradient, along with three demographic variables that were correlated to varying degrees with the latent ability. We then simulated five additional populations representing patterns of non-representativeness that might be encountered in real-world test development research. We subsequently drew smaller normative samples from each population, and used a one-parameter logistic IRT model to generate simulated test results for each individual in the samples. Using these simulated data, we applied continuous, regression-based norming techniques, both with and without compensatory weighting. The weighting technique reduced the bias of the normed scores when the degree of non-representativeness was relatively small. The technique was less effective with larger departures from demographic representativeness, and when non-representativeness occurred at extreme person locations. However, regardless of whether weighting was applied, the observed norm errors resulting from demographic non-representativeness were small enough to be ignored in most practical applications.

Reducing the bias of norm scores in non-representative samples: Weighting as an adjunct to continuous norming methods

In the development of norm-referenced psychometric tests, demographically representative samples provide the foundation for valid norm scores. An initial task for the test developer is to identify those demographic variables that correlate most strongly with the construct to be measured by the test. These variables typically include age, gender, race/ethnicity, education level, and/or socioeconomic status. When measuring a developing cognitive ability, age is the most important variable. Age has a stronger effect on test scores than the other variables, especially when testing children and adolescents. Because of this, Wechsler (1939, Chapter 3) recommended that same-age reference populations be used when norming tests of intelligence and achievement.

> **Commented [DH1]:** Can we find a more recent reference to support this point?

Consequently, normative samples must be demographically representative not just over the entire age range of the test, but also within smaller age groups containing individuals who are at similar stages of development. Samples collected in this way create the basis for *age-stratified* norms. Age stratification tends to increase the size of the normative sample (along with its cost and the time needed to collect it). To mitigate the need for larger samples, advanced mathematical methods have been developed to model the continuous relationship between raw and normed scores across age (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).

Some norm-referenced measures require stratification on variables other than age. For example, measures of body mass index (BMI) need gender-stratified norms, because optimal BMI for women is lower than that for men (Sang-Wook et al., 2015).

In other instances, it may be counter-productive to provide separate norms for the different levels of a demographic variable. For example, some studies show that girls have higher reading skills than boys (e.g., W. Lenhard et al., 2017; Price-Mohr & Price, 2017). However, if a reading test is intended to identify those children who need additional support - for example, children at the lowest decile of reading performance - then the use of gender-stratified norms might result in biased outcomes. With gender-stratified norms, some girls might be identified as needing additional support, even though they perform better on the test than boys who are not identified as needing educational support.

Addressing demographic imbalances: Stratification and post-stratification

Besides creating stratified norms, several options exist for dealing with demographic variables that are correlated with the latent ability being measured. An obvious course is to increase the size of the normative sample. As sample size increases, the distribution of the latent ability across gender (for example) increasingly approximates the distribution in the reference population. However, cost and time constraints usually limit the size of the sample available for norming.

A second approach is stratification, in which random sampling is conducted independently within homogenous categories, or strata, defined by the demographic variable (e.g., males, females). The goal is to have the category proportions in the normative sample match, as closely as possible, the proportions in the reference population. For example, if census data indicates that the reference population is composed of 50% males and 50% females[1], then the researcher would sample males and females independently to match those proportions in the normative sample.

However, it is not always possible to replicate population distributions through stratified random sampling. One can randomly delete cases from over-represented strata, but researchers are understandably reluctant to discard data. An alternative is to apply weighting

[1] For simplicity's sake, we assume at this point that the number of individuals who identify as non-binary is negligible.

multipliers, or weights, to the data of individuals in the mis-represented strata. For example, if a sample consists of 100 males and 50 females, a weighting multiplier of 2 could be applied to the data obtained from females. Each test score from a female would then be treated as if two females had obtained such a result. A weight w_k, assigned to an observation x_i in subsample *k*, thus indicates the number of individuals that this single observation represents. The weights must therefore be calculated so that the proportion

$$p_k = \frac{w_k \cdot n_k}{\sum_k w_k \cdot n_k}$$

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

Commented [DH2]: This equation not completely clear to me. I think the *k*s in the numerator need to be distinguished from the *k*s in the denominator.

Recently, Kennedy and Gelman (2021) recommended the use of multilevel regression combined with post-stratification to correct for non-representative samples in studies of psychological intervention. The authors suggest that weights can be used to adjust the means of non-representative samples, to facilitate statistical comparisons among samples. However, in constructing test norms from non-representative samples, the application of weights is more complicated, because the norming process involves modeling both population means and percentile ranks.

It is straightforward to take weights into account when calculating percentiles. As described above, each test result is treated as if obtained by w_k individuals. But, this simple calculation runs the risk of introducing its own bias into the raw-to-norm-score relationship, especially at the tails of the raw-score distributions. This risk occurs because the weights do not change the variance of the distribution of raw scores within demographic subgroups, as would be the case if more individuals were added to the subgroups. The potential distortion of the variance of the raw score distributions increases with the magnitude of the weights themselves. The risk for bias also *increases* as the number of individuals in a subgroup *decreases* (as is expected at the tails of the raw-score distributions). Consequently, the usefulness of weighting as a corrective procedure tends to diminish as the distributions of demographic variables in the normative sample become increasingly divergent from those in the reference population.

Because the potential distortions associated with weighting are most prominent at the tails of the raw-score distributions, they can disproportionately affect the raw-to-norm-score relationships for individuals of very high and/or very low ability. Unfortunately, these extreme ability ranges are the ones where precise norm scores are most needed, because the primary clinical applications of psychometric tests are to help diagnose disabilities, or, alternatively, to identify gifted individuals.

As noted above, post-stratification is a method for dealing with normative samples that are not representative, in term of demographic distributions, of the reference populations from which they are drawn. An additional complicating factor is that the common demographic variables of gender, socio-economic status (SES), race/ethnicity, and geographic region are often inter-related, in terms of the effects they may have on test performance. For example, areas with lower household income often have higher proportions of non-white inhabitants. Because of such interactions, the most accurate approach to stratification is to consider not only the marginal distributions of the demographic variables, but also their cross-classifications, or joint distributions. In a complete crossing of the four variables mentioned above, for instance, an individual could be classified as “female, low SES, white, west region”.

There are several practical difficulties with stratification based on the joint distributions of demographic variables. For one, census data are often available only for single demographic variables considered independently from each other, not for the cross-classified categories of multiple variables. In addition, in one possible cross-classification of gender, SES, race/ethnicity, and region, 192 joint cells (2 x 4 x 6 x 4) are created, some of which require only a few individuals to meet census proportions. Collecting a sample that meets these exacting specifications becomes a costly, lengthy process. In fact, with typical sample sizes of 100 cases per age year in tests of cognitive ability (e.g., Kaufman & Kaufman, 2004; Wechsler, 2008; Wechsler, 2014), it is not possible to replicate the census proportions in every cross-classified cell, because some of the joint percentages specify less than a single individual in a cell.

Raking

The raking procedure (Ireland & Kullback, 1968; Kalton & Flores-Cervantes, 2003) is an approach to post-stratification that attempts to mitigate the practical challenges of sampling based on a complete crossing of demographic variables. Raking does not draw on the explicit joint distributions associated with all possible cross-classifications. Instead, the post-stratification weights are determined in an iterative process based on the marginal distributions of each demographic variable. That is, the weights assigned to the demographic categories are adjusted successively and, if necessary, repeatedly, until they no longer change. The procedure is termed “raking” because it is analogous to smoothing out the soil in a garden bed by repeatedly raking in different directions. Studies have shown that the raking procedure is convergent and delivers optimal asymptotically normal estimates for the joint probabilities associated with a complete crossing of demographic variables (e.g., Ireland & Kullback, 1968).

Although widely employed to correct for lack of representativeness in political polls (Kalton-Flores-Cervantes, 2003), raking apparently has not been used in the norming of psychometric tests, perhaps because it could introduce error into the raw-to-norm-score relationships. As discussed above, demographic variables may interact with one another in their effects on test scores, creating the need to consider the joint distributions of such variables in developing norms. Because raking operates only on the marginal distributions (i.e., it considers only the "main effects" of demographic variables on test scores), it may magnify sources of error that stem from the interactions of these variables. However, these potential risks remain at the level of speculation, because, to our knowledge, the effect of raking on the accuracy of norm scores has never been studied.

Effects of continuous norming on non-representativeness samples

Continuous norming methods offer the advantage of using the properties of the entire normative sample to correct local distributional anomalies in smaller subsamples (e.g., age strata). Consequently, continuous norming methods may offer at least a partial remedy to distortions caused by lack of demographic representativeness. The semi-parametric continuous norming approach (SCN), first suggested by A. Lenhard and colleagues (A. Lenhard et al., 2018; A. Lenhard et al., 2019; W. Lenhard & Lenhard, 2021), has been shown to yield accurate norm scores with several non-optimal types of normative samples. One advantage of SCN is that it does not make specific assumptions about distribution parameters, and therefore can be applied to raw score distributions that are skewed, or that show floor and/or ceiling effects. A second strength of SCN is that it performs better than parametric approaches when applied to norm samples with the typical sample size of 100 per age cohort, independent of the skewness of the raw score distributions (A. Lenhard et al., 2019).

In modeling the trajectories of percentile ranks across age groups, SCN (as implemented in the cNORM package in R, A. Lenhard et al., 2018) *does not* rely on splines. As a result, these trajectories are relatively rigid. As noted above, this feature of SCN modeling tends to reduce the influence of error variance in local age groups, including that caused by lack of demographic representativeness. As might be expected, therefore, SCN generally produces less norming error than methods that determine raw-to-norm-score mapping separately for each age group (W. Lenhard & Lenhard, 2021).

Commented [DH3]: I don't understand what "rigid" means in this context. It seems you are arguing that because of this rigidity, error in certain age groups can be reduced by the influence of age-groups that have greater representativeness. What's missing is the logical link between rigidity of trajectories and the outcome of less error. Why does the former lead to the latter?

Goals of the current simulation study

We have described two techniques (SCN, raking) that may help ameliorate bias in norm scores resulting from non-representative normative samples. To date, no research has investigated whether the combination of both methods (in the following referred to as *weighted continuous norming,* or WCN) further improves the accuracy of norm scores, compared to SCN alone. A further open question is whether the mathematical transformations wrought by SCN and raking might interact in a way that would *increase* the bias of norm scores, within certain ability ranges.

The goal of the current study, therefore, was to evaluate the benefits and risks of applying WCN[2] to non-representative normative samples. To this end, we conducted a norming procedure on measure of a simulated cognitive ability that increases with increasing age. Furthermore, we modeled the effects of three simulated demographic variables on the cognitive measure. For convenience, we labeled the simulated demographic variables as "education", "ethnicity", and "geographic region". We modeled education so that it would have a stronger effect on the cognitive measure than ethnicity or region.

To provide input for the norming procedure, we generated six simulated population-level data sets: a reference population that embodied the benchmark distributions of the three demographic variables; and five non-representative populations, in which the distribution of these demographic variables differed from the reference population. Each of these populations has six equal-sized age cohorts. Table 1 summarizes the differences among the six simulated populations.

Table 1: Simulated populations for norming input

[2] WCN incorporates SCN, as implemented in the R package cNORM (A. Lenhard et al., 2018).

No.	Label	Description	Hypothesized effects on distribution of cognitive ability variable
1	Reference	Benchmark distributions of demographic variables; the standard of comparison for describing the “representativeness” of the other simulated populations.	Not applicable (benchmark population).
2	Mild under-representation of high education	Lower proportion of high-education individuals, higher proportion of low-education individuals, than Population 1.	Both mean and variance affected.
3	Moderate under-representation of high education	The pattern of divergence of education proportions is similar to population 2, but the degree of non-representativeness is greater.	Both mean and variance affected.
4	Under-representation of both low and high education	Both tails of the education distribution have lower proportions than Population 1.	Only variance affected.
5	Biased joint distributions	Marginal distributions of demographic variables match population 1; joint distributions (cross classifications) do not	Only variance affected.

		match population 1. The pattern of non-representation alternates from over- to under-represented across the 27 (3 x 3 x 3) joint distributions.	
6	Clustered sampling	Marginal and joint distributions of demographic variables match population 1, but only when averaged across all six age cohorts. Within each age cohort, two-thirds of the joint distribution cells contain no data.	??? (original manuscript does specify)

Commented [DH4]: I would prefer the label "clustered distributions" for this population, because I want to keep a clean distinction between the generation of *population*-level data sets, and subsequent drawing of normative *samples.*

Commented [DH5]: What is the hypothesized effect on cognitive ability mean, variance for population 6?

Because raking incorporates only marginal distributions, we expect it to have little effect in populations 5 (biased joint probabilities) and 6 (clustered sampling). In these two populations, non-representativeness occurs only at the level of joint distributions (cross-classifications), not at the level of marginal distributions.

For the current study, our pre-registered hypotheses were as follows:

Commented [DH6]: I would prefer to have the information about registration in a footnote, instead of in the main narrative.

1.We expected a main effect of norming method, such that WCN would lead to less-biased estimates of the norm scores than SCN, where "bias" is quantified in terms of root mean square error (*RMSE*) and mean signed difference (*MSD*).

2. We expected an interaction between norming method and the degree of non-representativeness of the input data. Specifically, we expected that as the non-

representativeness of the normative sample increased, norm-score bias would increase for both methods, but that the increase in bias would be smaller for WCN than for SCN.

3. We expected that the simple effect of WCN in reducing norm-score bias would vary depending on person location on the cognitive variable. Specifically, we expected that WCN would be less effective at reducing bias at the tails of the cognitive ability distribution than in the central region of that distribution.

Commented [DH7]: The last part of this section (which I deleted) is too vague to be part of a hypothesis; suggest simply reviewing findings post-hoc in results or discussion.

Methods

Overview

Our study proceeded through the following steps:

1. Modeling a latent cognitive ability with a typical age-related growth curve.
2. Generating data sets for the reference population and five additional simulated populations.
3. Drawing normative samples from each simulated population.
4. Generating simulated raw scores for a test of the cognitive ability.
5. Applying WCN and SCN to the raw scores from the normative samples.
6. Generating norm scores based on the reference population, as a standard of comparison.
7. Testing the study hypotheses with ANOVAs, using RMSE and MSD as dependent variables.

Modeling cognitive ability

To provide a basis for a modeled cognitive ability that develops with age, we envisioned a reference population divided into six age cohorts, spanning one year each. We conceptualized a cognitive ability that increases in each successive age group, as is typical with cognitive development during childhood. We further specified that this cognitive ability is influenced by the three demographic variables, each of which has three categories:

- Education: low, medium, high
- Ethnicity: native, mixed, non-native
- Region: south, east, northwest

In broad terms, therefore, our model states that cognitive ability is a function of age and the three demographic variables.

We operationalized the effect of the demographic variables on cognitive ability by assigning three levels of mean cognitive ability (below average, average, above average) to the three categories of each demographic variable, according to the matrix shown in Table 1.

Table 1

Assignment of cognitive ability levels to demographic categories

Demographic variable	Below-average ability	Average ability	Above-Average ability
Education	low	medium	high
Ethnicity	native	mixed	non-native
Region	south	east	northwest

Importantly, this mapping of ability level to demographic category remains constant in all study conditions. Thus, by changing the distributions of demographic categories across simulated populations, we simultaneously manipulate the distributions of cognitive ability.[3]

We then specified benchmark distributions for the demographic variables, which would be enacted in the simulated reference population data set. The benchmark demographic distributions must be understood in terms of a complete cross-classification of the three demographic variables, which yields a 27-cell matrix with a 3 (low, medium, high education) x 3 (native, mixed, non-native ethnicity) x 3 (south, east, northwest region)

[3] To limit the complexity of the simulation, we constrained the correlations among the demographic variables to zero.

structure. Table 2 follows this structure in specifying the benchmark demographic distributions.

Table 2

Distributions of demographic variables, by category, in the reference population

		low education 40%	medium education 20%	high education 40%
ethnicity: native 30%	region: south 60%	7.2%	3.6%	7.2%
	region: east 20%	2.4%	1.2%	2.4%
	region: north-west 20%	2.4%	1.2%	2.4%
ethnicity: mixed 40%	region: south 60%	9.6%	4.8%	9.6%
	region: east 20%	3.2%	1.6%	3.2%
	region: north-west 20%	3.2%	1.6%	3.2%
ethnicity: Non-native	region: south 60%	7.2%	3.6%	7.2%
	region: east 20%	2.4%	1.2%	2.4%

30%	region: north-west 20%	2.4%	1.2%	2.4%

The benchmark marginal distributions of the demographic variables are shown in the table margins (for education and ethnicity) and in the left-most column of the nested rows (for region). The joint distributions for the complete cross-classification of the three variables are shown in the table cells.

The structure of Table 2 provides a basis for understanding the demographic manipulations that were applied to create five additional simulated populations. As described previously, each cell of the table corresponds to a certain mean cognitive ability level, which is determined by the demographic cross-classification of that cell. Thus, referring back to Table 1, the cell in Table 2 with the highest mean cognitive ability is high education, non-native ethnicity, northwest region, which appears in the lower-right corner of the table, constituting 2.4% of the reference population. Conversely, the cell with the lowest mean cognitive ability is low education, native ethnicity, south region, which appears in the upper-left corner of Table 2, constituting 7.2% of the reference population.

Within the reference population, each row of data includes age, level of cognitive ability, and classifications on education, ethnicity, and region. The values for the demographic classifications are assigned according to the percentages in Table 2. The row-wise values of cognitive ability are based on a distinct mean value for each cell[4] of Table 2. This cell-wise mean is calculated by the following polynomial equation:

[4] *SD* is constrained to 1 across all cells of Table 2.

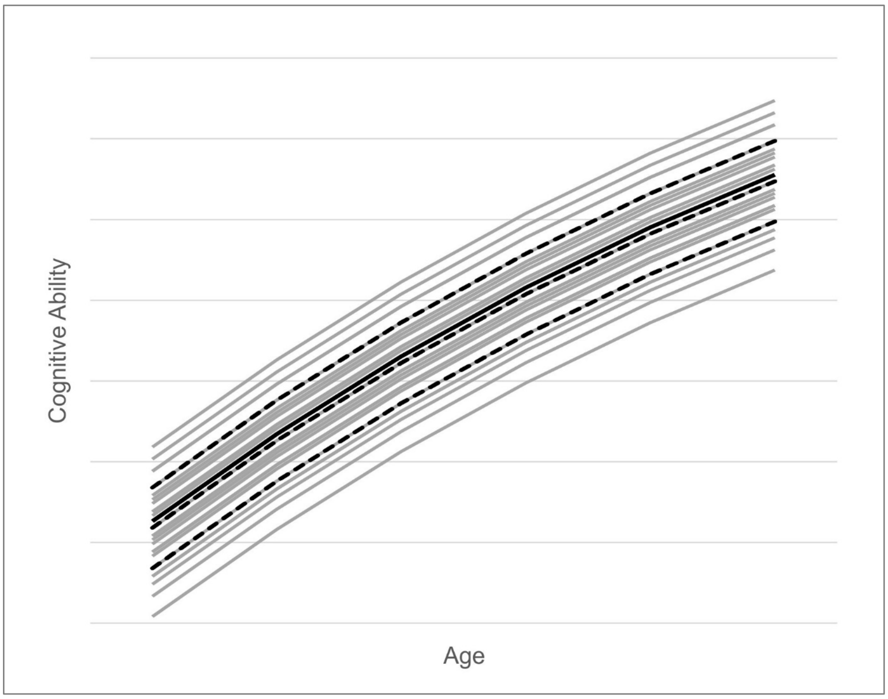
$$M(age, education, ethnicity, region) = -1.5 \cdot education - 0.25 \cdot ethnicity - 0.1 \cdot region - 0.05 \cdot ethnicity \cdot region + 1.2 \cdot age - 0.06 \cdot age^2 + 0.0001 \cdot age^4 \quad (1)$$

As a result of this equation, each demographic variable exerts a different effect on cognitive ability:

- Education: correlates at r = -.78 with cognitive ability (large effect)
- Ethnicity: *r* = -.54 (medium effect)
- Region: *r* = -.31 (small effect)

Figure 1 provides a graphic depiction of the modeled cognitive ability in the reference population.

Figure 1. Modeled cognitive ability in reference population



The figure shows mean cognitive ability increasing across the six age cohorts. The solid black line represents the reference population mean. The dashed black lines represent the marginal mean cognitive abilities for the low, medium and high categories of education, the demographic variable with the largest effect on cognitive ability. The grey lines represent the mean cognitive abilities associated with the 27 demographic cross-classifications. The highest grey line is high education, non-native ethnicity, northwest region, the cell of Table 2 with the highest mean cognitive ability. The lowest grey line is low education, native ethnicity, south region, the cell of Table 2 with the lowest mean cognitive ability.

Generation of simulated population data sets and normative samples

To generate the reference population data set, we drew 24 million[5] pairs of random numbers (4 million per age cohort), each pair representing one individual. The first random number was uniformly distributed between 0 and 6 and represented age in years. The second number 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 specific cognitive ability value for an individual by adding the mean cognitive ability for that individual's demographic cross-classification status (see Table 2 and Formula 1). Additionally, we *z*-standardized each cognitive ability value (x_1), using the reference population mean μ and standard deviation σ in formula 2,

Commented [DH8]: Suggest labelling the age cohorts (0-5?) on the figure, for clarity.

Commented [DH9]: I think we need to say explicitly that the size of the population reference sample was aligned to the US population for ages 0-6 (if that is in fact what was meant). Or, "0-6" not supposed to represent actual ages, but merely serve as a means for classifying individuals into cohorts (in other words, it's nominal)?

Commented [DH10]: Do you really mean with respect to the same stratum AND age? Or just with respect to the same age cohort? The latter makes more sense to me, because the stratum is taken into account in the next step, when the random number is added to the stratum specific mean from Formula 1.

Commented [DH11]: This needs to be clarified. Do you mean that the second random number was nested within the first, such that there were six separate random draws for the second number, one for each age cohort?

[5] The reference population size is roughly based on the number of persons in the U.S. population whose ages are within a span of six consecutive years (e.g., ages 0 to 5).

$$\theta_{pop} = \frac{x_1 - \mu}{\sigma} \quad (2)$$

where θ_{pop} represents an individual's location on the cognitive ability variable with respect to the entire reference population.

As noted previously, each individual in the reference population was assigned values on the demographic variables, such that marginal and joint distributions of these variables would match the distributions shown in Table 2. We then generated five additional simulated population data sets, using the same method described at the outset of this section. These additional simulated populations represented various violations of demographic representativeness that might be encountered in collecting normative data for the development of a psychometric tests The distributions of the demographic variables in these five additional data sets differed from the reference population as follows:

- Simulated Population 2: Mild under-representation of high education. The high education category was underrepresented (28% instead of 40%) and the low education category was overrepresented (52% instead of 40%). This manipulation affected both the mean and the variance of the cognitive ability variable.
- Simulated Population 3: Moderate under-representation of high education. The pattern of misrepresentation was the same as Population 2, but the degree of misrepresentation was greater (high education was 20% instead of 40%; low education was 60% instead of 40%). This manipulation affected both the mean and the variance of the cognitive ability variable.
- Simulated Population 4: Under-representation of both low and high education. Medium education was overrepresented (40% instead of 20%), and high and low education were underrepresented (30 % instead of 40 %). This manipulation attenuated the variance of the cognitive ability variable, but its mean was not affected.

- Simulated Population 5: Biased joint distributions. The joint distributions of the demographic variables were varied from the reference percentages shown in Table 2, such that:
    - Some of the joint cells were overrepresented, while some were underrepresented.
    - The marginal distributions were identical to those in the reference population.

  This manipulation increased the variance of the cognitive ability variable, but its mean was not affected.

Commented [DH12]: Do we need to specify the details of this manipulation in a footnote (probably too long to include in the main narrative)?

- Simulated Population 6: Clustered sampling. Within each age cohort, two-thirds of the 27 demographic cross-classification cells contained no data. In the remaining one-third of cells, the number of individuals was tripled. This manipulation was applied to different subsets of cells across age cohorts, such that when cell proportions were summed across all age cohorts, the marginal and joint distributions of the demographic variables were identical to the reference population.

Commented [DH13]: See my previous comment - I prefer the label “Clustered Distriutions”.

Commented [DH14]: For this population, how are the mean and variance of the cognitive variable affected? Also, as with Population 5, do we want to provide more details about the manipulation in a footnote?

In the five additional simulated population data sets, the cognitive variable was z-standardized using Formula 2. Importantly, the values of μ and σ were those from the reference data set (Population 1), not from the data set whose values were being standardized.

Commented [DH15]: We need a sentence explaining why it was done this way, that is, why use the reference mean and variance for *z*-standardization of the cognitive variables in the five additional simulated populations?

From each of the six simulated populations, we drew 100 random samples of 600 individuals (100 cases per age cohort). These samples served as input to the norming procedures.

Simulation of test results

Using the one-parameter logistic (1-PL) model, we simulated a 31-item test to generate test results for each individual in the normative samples. The 31 item difficulties (δ) were drawn randomly from a uniform distribution ranging from -3 and +3. The set of item difficulties covered a range of about 3.7 standard deviations (M = -0.04, SD = 1.64), therefore spanning a wide range of latent ability. The probability $p_{k,i}$ that an individual k with the z-standardized latent ability θ_{pop_k} succeeded on item i, with difficulty δ_i, was given by the following 1-PL equation:

$$p_{k,i}(x_i = 1 | \theta_{pop\_k}, \delta_i) = \frac{\exp(\theta_{pop\_k} - \delta_i)}{1+\exp(\theta_{pop\_k} - \delta_i)} \quad (3)$$

For every individual 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, each individual's scores on all 31 items were summed to yield a raw total score on the simulated test.

Application of weighted and unweighted norming procedures

For each raw score in the normative samples, we applied WCN and SCN to generate IQ-type standard scores (M = 100, SD = 15) for each norming method. These scores were labeled IQ_{WCN} and IQ_{SCN}. For both WCN and SCN, these IQ scores were calculated with *cNORM*, an R package that employs continuous norming (A. Lenhard et al., 2018). Weights were not used for SCN.

To calculate the weights for WCN, we used the *rake* function from *survey* (Lumley, 2011), an R package that implements the raking procedure described earlier in this manuscript. Additionally, we standardized the weights to make them easier to interpret. We divided each weight by the smallest weight in the respective norm sample, thereby setting the weight of the most overrepresented group in the sample to 1.[6]

Using weights required modifications to the standard cNORM functions. In WCN, weights are applied initially in the ranking procedure, where each raw score is assigned a

[6] This transformation does not affect the distribution of the weights.

Commented [DH16]: The original manuscript had "underrepresented" here, but I believe you meant overrepresented, because it is the latter groups that have the smallest weights.

percentile rank. Weighting requires that raw scores from certain demographic groups be counted more than once in this ranking process. For example, if a certain group is assigned a weight of 2, each raw score from that group must be counted twice during ranking. Because the weights, as initially calculated by the *rake* function, are floating point numbers (not integers), they were multiplied by 10^6 and then rounded to whole numbers prior to ranking. This transformation made it possible, during ranking, to count each test score the number of times indicated by its weight. Because of the high number of ties, the average rank was used for further processing, following the usual cNORM procedures.[7]

In WCN, weights are also entered in cNORM's regression-based modeling procedure. To perform the regression, cNORM draws on the *regsubsets* function of *leaps*, an R package (Lumley, 2017). *regsubsets* includes the capacity to process weights in the regression analysis.

Generating norm scores from the reference population

To test the study hypotheses, we created a measure (IQ_{best}), in the same metric as IQ_{WCN} and IQ_{SCN}, which represented the "actual" person location on the cognitive ability variable. IQ_{best} was derived from the distribution of raw scores in the entire reference population (in contrast to IQ_{WCN} and IQ_{SCN}, which were derived from the smaller normative samples).

To compute IQ_{best}, we generated raw scores on the 31-item simulated test for the 24 million individuals in the reference population, using the previously described method. We then partitioned the reference population by age, creating 365 equal-sized groups within each of the six age cohorts. Each of the resulting 2190 age-groups consisted of about 11,000 individuals with the same "birthday". The raw scores were ranked and converted into IQ

[7] For a detailed description of the *cNORM* norming process, see A. Lenhard, Lenhard, & Gary, 2019.

scores using rank-based inverse normal transformation within each age group. As a result, each row in the reference population data set included values for age, raw score and IQ_{best}.

Hypothesis testing with RMSE and MSD

As noted above, we drew 100 normative samples ($N = 600$) from each of the six simulated population data sets. We conducted ANOVAs to test the study hypotheses in each of these 600 normative samples. The ANOVAs compared IQ_{best} to IQ_{WCN} and IQ_{SCN}, respectively, with *RMSE* and *MSD* as dependent variables. Both RMSE and MSD are quantified in terms of IQ points.

RMSE is a summary measure of norming model error that includes both fixed and variable error components (Lenhard & Lenhard, 2021). It was computed using the following formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(IQ_{.} - IQ_{best})^2}, \tag{4}$$

where n is the number of cases and $IQ_{.}$ stands for either IQ_{WCN} or IQ_{SCN}.

MSD is a measure of the tendency for a norming model to overestimate ($MSD > 0$) or underestimate ($MSD < 0$) the actual person location. The formula used to calculate the *MSD* was:

$$MSD = \frac{1}{n}\sum_{i=1}^{n}(IQ_{.} - IQ_{best}). \tag{5}$$

To test Hypothesis 3, we divided the distributions of IQ_{best}, IQ_{WCN} and IQ_{SCN} into 11 intervals of 7.5 IQ points each. *RMSE* and *MSD* were calculated separately for each of these intervals.

In general, the analytic approach was to conduct 6 (simulated population) x 11 (IQ range) x 2 (norming method) mixed ANOVAs. Population was a between-groups factor, and IQ range and norming method were within-groups factors. Because of size of the simulated data sets, statistical power was high, and therefore the level of significance was set to $p = .01$. The assumption of sphericity was tested and, where indicated, degrees of freedom were corrected. Additionally, partial η^2's were computed as measures of effect size. We further specified that, in the norm score comparisons of interest, differences of less than 0.5 IQ were too small to have any practical relevance.

Commented [DH17]: How was this threshold determined?

Results

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

The 6 x 11 x 2 mixed ANOVAs yielded significant results for all main effects and interactions ($p < .001$). We focus here on the effects that are most salient for testing our hypotheses.

Hypothesis 1: Main Effect of Norming Method

The first hypothesis proposed that WCN would yield lower levels of norm-score bias than SCN. This hypothesis was supported by tests of the main effects of norming method, *RMSE*: $F(1, 594) = 94.93$, $p < .001$, $\eta^2 = .14$, *MSD*: $F(1, 594) = 3397.28$, $p < .001$, $\eta^2 = .85$. *RMSE* was smaller for WCN ($M = 2.18$, $SE = .02$) than for SCN ($M = 2.36$, $SE = .02$). The same was true for MSD (WCN: $M = 0.74$, $SE = .03$; SCN: $M = -0.24$, $SE = .03$).

Commented [DH18]: These numbers appear to contradict the claim that MSD was smaller for WCN than SCN.

However, the analysis also detected significant interactions between norming method and simulated population, indicating that the effects of weighting on norm-score bias varied among the simulated populations, *RMSE*: $F(5, 594) = 98.98$, $p < .001$, $\eta^2 = .45$, *MSD*: $F(5, 594) = 764.77$, $p < .001$, $\eta^2 = .87$. As can be seen in Figure 2 (*RMSE*) and Figure 3 (*MSD*), in two of the six simulated populations, weighting reduced bias in the normed scores. In populations 2 (mild under-representation of high education) and 3 (moderate under-representation of high education), *RMSE* was lower with WCN than with SCN, by an average

Commented [DH19]: You'll have to edit the figures to reflect the change in nomenclature from "sampling condition" to "simulated population". Please check the other figures to see if they need similar changes.

Commented [DH20]: Please add to the figure caption, or in the main narrative, that these plots show the average RMSE and MSD, average over the 600 norm draws per simulated population.

of 0.48 IQ points and 0.97 IQ points, respectively. In populations 1 (reference), 5 (biased joint probabilities) and 6 (clustered sampling), the difference in *RMSE* between WCN and SCN approximated zero. In population 4 (under-representation of both low and high education), WCN returned higher average *RMSE* than SCN, but the difference of

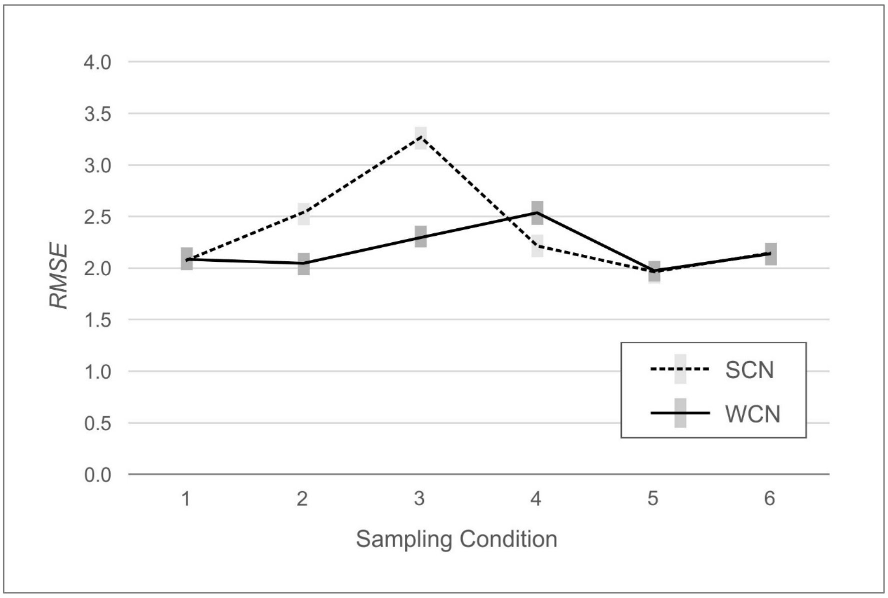


Figure 2. *RMSE* across simulated populations, with (WCN) or without (SCN) weighting. The grey rectangles represent 95% confidence intervals.

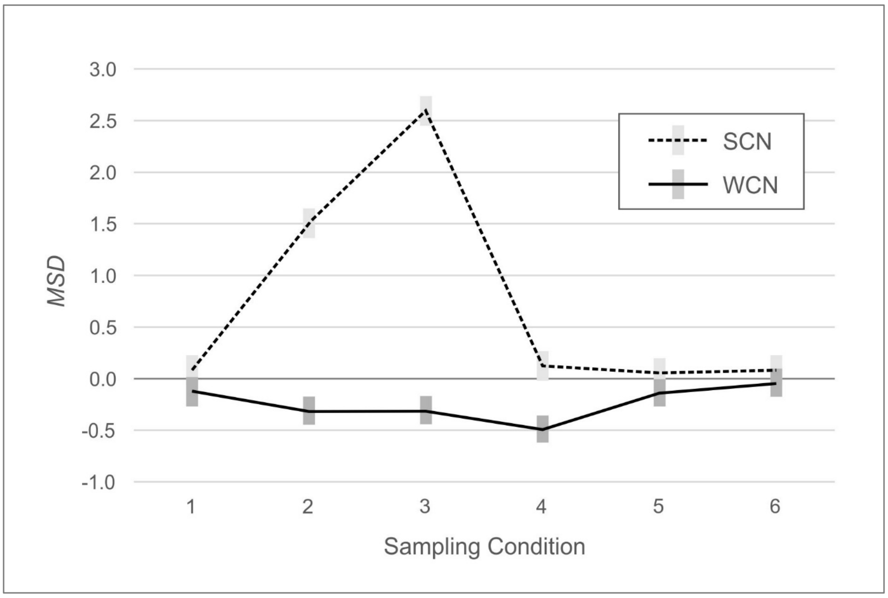


Figure 3. *MSD* across simulated populations, with (WCN) or without (SCN) weighting. The grey rectangles represent 95% confidence intervals.

0.32 IQ points was below the threshold of practical relevance.

The analysis of the *MSD* yielded similar results. In populations 2 (mild under-representation of high education) and 3 (moderate under-representation of high education), *MSD* was closer to the ideal value of zero for WCN (populations 2 and 3: -0.32 IQ points) than for SCN (population 2: 1.51 IQ points; population 3: 2.59 IQ points). In populations 1 (reference), 5 (biased joint probabilities) and 6 (clustered sampling), *MSD* approximated zero, regardless of the norming method. In population 4 (under-representation of both low and high education), *MSD* deviated more from zero for WCN (-0.49 IQ points) than for SCN (0.13 IQ points). As with *RSME*, these latter differences did not meet the criterion for practical significance.

Hypothesis 2: Interaction Between Norming Method and Degree of Non-Representativeness

Hypothesis 2 specified that as the non-representativeness of the normative samples increased, norm-score bias would increase for both methods, but that the increase in bias would be smaller for WCN than for SCN. To address this hypothesis, we compared populations 2 and 3. Both populations were characterized by under-representation of the high education group, but the magnitude of under-representation was greater in population 3 than in population 2. Therefore, we performed two additional ANOVAs that limited the levels of the between-groups factor to populations 2 and 3. Both analyses yielded a significant interaction between norming method and simulated population, *RMSE*: $F(1, 198) = 26.01$, $p < .001$, $\eta^2 = .12$, *MSD*: $F(1, 198) = 242.36$, $p < .001$, $\eta^2 = .55$.

The results of these analyses are visualized in Figures 2 and 3. The plots show the interaction: *RMSE* and *MSD* are greater in magnitude in population 3 (moderate under-representation) than in population 2 (mild under-representation), for both norming methods, but the magnitude of increase in the error metrics is greater for SCN than for WCN.

Considering WCN in isolation, average *MSD* was approximately equal for both populations (-0.32 IQ points). Average *RMSE*, on the other hand, did increase significantly in population 3, $F(1, 198) = 10.12$, $p = .002$, $\eta^2 = .05$.[8]

Commented [DH21]: I removed the part of this paragraph that tried to interpret the difference between the outcomes in *MSD* and *RMSE*, because it was vague and there was no additional explanation as to what caused it or why it was worth calling out.

Hypothesis 3: Effectiveness of WCN depends on person location

Hypothesis 3 proposed that WCN would be less effective at reducing bias at the tails of the cognitive ability distribution than in the central region of that distribution. We tested this hypothesis with two analytic approaches. First, we conducted 11 x 2 ANOVAs with person location and norming method (WCN vs. SCN) as within factors, and *RMSE* and *MSD* as dependent variables. We then examined how the effects of person location varied among the simulated populations. We compared the performance of WCN in populations 2, 3, 4, 5 and 6 (which yield demographically non-representative normative samples, as described earlier) to SCN in population 1 (which yields demographically representative normative samples). SCN in population 1 therefore represents a benchmark condition, against which the performance of WCN in the other non-representative populations can be measured. These latter analyses used 11 x 6 ANOVAs, with simulated population as a between-groups factor.

The results of these analyses are illustrated in Figures 4 (*RMSE*) and 5 (*MSD*). Because of the large number of comparisons, we report effects only if at least one of the differences within an analysis exceeded 0.5 IQ points.

[8]Average *RMSE* was 2.04 IQ points in sampling condition 2 and 2.29 IQ points in sampling condition 3.

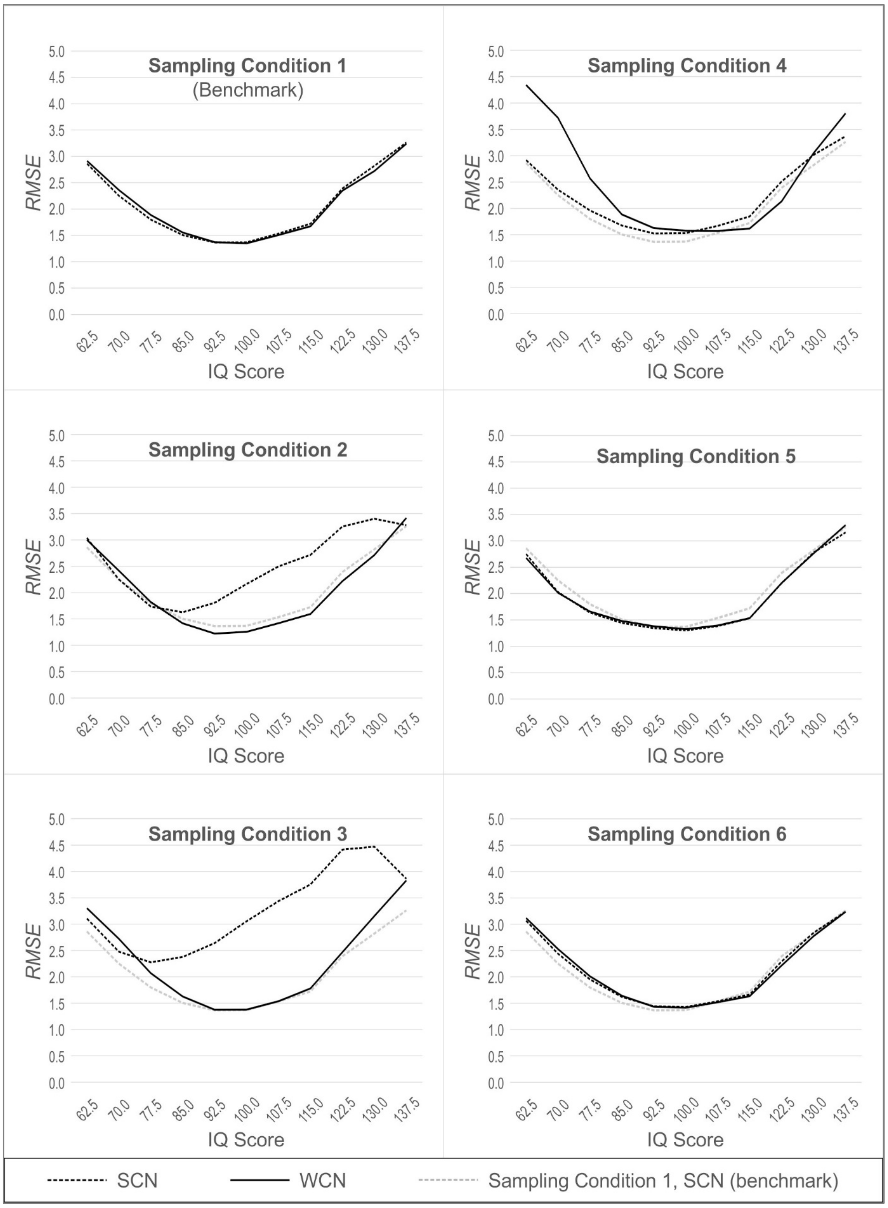
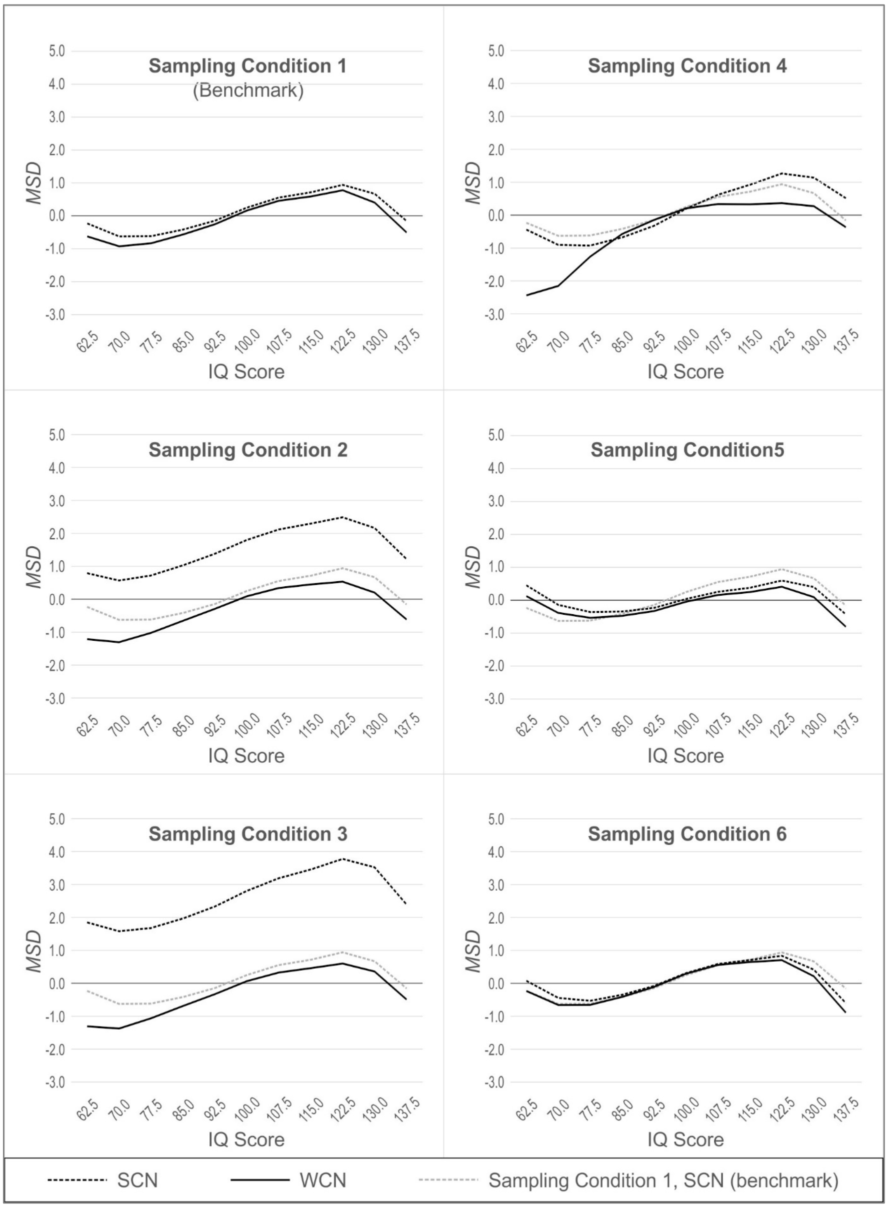


Figure 4. *RMSE* across simulated populations, with (WCN) or without (SCN) weighting, as a function of person location. The dotted grey line represents SCN with norm samples drawn from Population 1 (benchmark).



515 ***Figure 5.*** *MSD* across simulated populations, with (WCN) or without (SCN) weighting, as a function of person
516 location. The dotted grey line represents SCN with norm samples drawn from Population 1 (benchmark).

Population 1: Reference

In normative samples drawn from the reference population, both ANOVAs yielded a significant main effect of person location, *RMSE*: $F(2.68, 264.87) = 70.68$, $p < .001$, $\eta^2 = .42$, *MSD*: $F(2.34, 231.65) = 35.54$, $p < .001$, $\eta^2 = .26$. In general, *RMSE* increased as person location moved towards either tail of the distribution, away from the average IQ of 100. This effect, also seen in the other simulated populations, is visualized as a parabolic shape in Figure 4. By contrast, in the analysis with *MSD*, the main effect of person location is visualized as a sinusoidal pattern (see Figure 5 and discussion section below). This effect of person location on norming bias is a previously reported feature of continuous norming procedures (cf. A. Lenhard et al., 2019). As such, this effect is not directly relevant to the question of whether weighting, per se, reduces norm bias due to non-representative sampling. What is important to note (and is readily seen in Figures 4 and 5) is that WCN and SCN perform equally well, in terms of error measures, when processing normative samples drawn from a demographically representative, reference population. This makes intuitive sense, because with representative samples, there are no cell-wise departures from expected demographic proportions, to which weights could be applied to correct for bias in the norming process.

Population 2: Mild under-representation of high education

In samples drawn from Population 2, we found a main effect of norming method, *RMSE*: $F(1, 99) = 74.94$, $p < .001$, $\eta^2 = .43$, *MSD*: $F(1, 99) = 1924.64$, $p < .001$, $\eta^2 = .95$. WCN was superior to SCN in reducing norm bias resulting from non-representative samples. With *RMSE*, we also observed an interaction between person location and norming method, $F(2.72, 268.76) = 41.72$, $p < .001$, $\eta^2 = .30$. As shown in Figure 4, WCN reduced the error measure to a greater degree in the upper range of person location than in the lower range. In Population 2, individuals of higher education (and consequently, higher cognitive ability) are under-represented. Thus, the interaction shows that WCN is correcting for norm bias in the region of person location that is under-represented in the normative samples.

In the comparison of WCN in Population 2 to the benchmark of SCN in Population 1, there was no main effect of population on *RMSE*. That is, even under the conditions of non-representativeness in Population 2, WCN did not differ from the benchmark on the error measure. This suggests that weighting successfully compensated for any norm bias due to demographic non-representativeness in Population 2, when that bias was measured by *RMSE*. The results differed for *MSD,* where we observed a main effect of population, $F(1, 198) = 15{,}37$, $p < .001$, $\eta^2 = .07$, and an interaction between population and person location, $F(2.31, 456.93) = 2.91$, $p = .048$, $\eta^2 = .01$. These findings indicated that, with respect to *MSD*, WCN did not fully correct norm bias in samples from Population 2. In addition, this difference in MSD between WCN and the benchmark condition[9] exceeded the threshold of practical relevance in the lower region of person location, with absolute differences of 0.68 IQ points at IQ 70 and 0.99 IQ points at IQ 62.5.

Population 3: Moderate under-representation of high education

For both error measures, the ANOVAs with normative samples drawn from Population 3 yielded significant main effects of norming method, *RMSE*: $F(1, 99) = 155.76$, $p < .001$, $\eta^2 = .61$, *MSD*: $F(1, 99) = 2707.76$, $p < .001$, $\eta^2 = .97$, and significant interactions between norming method and person location, *RMSE*: $F(2.79, 276.13) = 71.89$, $p < .001$, $\eta^2 = .42$, *MSD*: $F(2.63, 260.05) = 6.58$, $p = .001$, $\eta^2 = .06$. These analyses produced larger effect sizes than those for Population 2, mirroring the difference in representativeness between the two populations. This suggests that WCN exerts a larger corrective effect on norm-score bias with normative samples that display greater deviations from demographic representativeness.

[9] In the remainder of this section, the absolute difference in error measure between norming methods will be labeled |ΔRMSE| or |ΔMSD|, as appropriate.

In comparing WCN in Population 3 to SCN in Population 1, we observed significant main effects of population, *RMSE*: $F(1, 198) = 9.52$, $p = .002$, $\eta^2 = .05$, *MSD*: $F(1, 198) = 12.52$, $p = .001$, $\eta^2 = .06$, and significant interactions between population and person location, *RMSE*: $F(2.81, 557.01) = 3.06$, $p = .031$, $\eta^2 = .02$, *MSD*: $F(2.26, 447.76) = 3.28$, $p = .033$, $\eta^2 = .02$. In the normative samples drawn from Population 3, WCN yielded greater norming error than the benchmark within the low and high regions of person location. However, a practically significant difference in IQ scores occurred at only one location for *RMSE* (IQ 137.5, $|\Delta RMSE| = 0.57$ IQ points), and one location for *MSD* (IQ 62.5, $|\Delta MSD| = 0.68$ IQ points).

Population 4: Under-representation of both low and high education

As with Populations 2 and 3, the analyses of normative samples drawn from Population 4 produced significant main effects of norming method, *RMSE*: $F(1, 99) = 39.02$, $p < .001$, $\eta^2 = .28$, *MSD*: $F(1, 99) = 164.63$, $p < .001$, $\eta^2 = .62$, and significant interactions between norming method and person location, *RMSE*: $F(3.02, 299.01) = 42.43$, $p < .001$, $\eta^2 = .30$, *MSD*: $F(2.53, 250.47) = 42.44$, $p = .001$, $\eta^2 = .30$. However, with Population 4, where both tails of the education distribution were under-represented, WCN did not provide greater reduction of norm-score bias than SCN. The interactions revealed that in the low region of person location *RMSE* was greater for WCN than SCN (IQ 77.5 $|\Delta RMSE| = 0.61$ IQ points, IQ 70.0 $|\Delta RMSE| = 1.37$ IQ points, IQ 62.5 ($|\Delta RMSE| = 1.43$ IQ points). For *MSD,* the interaction between norming method and person location was more complex, with SCN providing greater reduction of norm bias than WCN in the low region of person location (IQ 70.0: $|\Delta MSD| = 1.25$ IQ points; IQ 62.5: $|\Delta MSD| = 2.00$ IQ points), and WCN outperforming SCN in the high region of person location (IQ 115.0: $|\Delta MSD| = 0.60$ IQ points; IQ 122.5: $|\Delta MSD| = 0.90$ IQ points; IQ 130.0: $|\Delta MSD| = 0.87$ IQ points; IQ 137.5: $|\Delta MSD| = 0.88$ IQ points).

In comparing WCN in Population 4 to SCN in Population 1, we found significant main effects of population, *RMSE*: $F(1, 198) = 38.46$, $p < .001$, $\eta^2 = .16$, *MSD*: $F(1, 198) = 31.10$, $p < .001$, $\eta^2 = .14$, and significant interactions between population and person location, *RMSE*: $F(2.83, 560.75) = 22.19$, $p < .001$, $\eta^2 = .10$, *MSD*: $F(2.55, 505.02) = 19.44$, $p < .001$, $\eta^2 = .09$. For *MSD*, we examined the simple effects underlying the interaction and found that WCN reduced norm bias more than the benchmark only at the highest person locations (IQ >122.5). At IQ 122.5, the difference was 0.58 IQ points.

Populations 5 (Biased joint distributions) and 6 (Clustered distributions)

For populations 5 and 6, the ANOVAs revealed no main effects of norming method. With respect to *RMSE*, the differences between WCN and SCN did not exceed 0.5 IQ points at any point in the range of person location. When we compared WCN in Population 5 to the benchmark (SCN in Population 1), the ANOVA for *MSD* returned a significant main effect of population, $F(1, 198) = 5.02$, $p = .026$, $\eta^2 = .03$, and a significant interaction between population and person location, $F(2.29, 453.98) = 6.25$, $p = .001$, $\eta^2 = .03$. At some person locations, the absolute difference between WCN and the benchmark exceeded 0.5 IQ points. However, the valence of these differences varied over the range of person location. At IQ 122.5 and IQ 130.0, MSD was closer to zero for WCN than for the benchmark, but at IQ 137.5 this pattern was reversed ($\Delta MSD| = 0.73$). The results suggest that weighting offers no clearcut advantage in reducing the error associated with norming, when normative samples are drawn from a population with biased joint distributions of the demographic variables.

Discussion

Summary of results

The present study examined whether compensatory weighting at the raw score level, when combined with continuous norming procedures, would reduce bias in norm scores derived from demographically non-representative norm samples. To pursue this aim, we simulated six populations in which the distributions of demographic variables departed to various degrees from expected proportions. We modeled a latent cognitive ability, which we used as the input for a one-parameter logistic IRT model to create raw test scores. We drew normative samples from the six populations, and generated IQ-type norm scores by applying weighted continuous norming (WCN) and semi-parametric continuous norming without weighting (SCN). We used mean square error (*RMSE*) and mean signed difference (*MSD*) as measures of norm-score bias.

Our first hypothesis proposed that when processing non-representative normative samples, WCN would produce less-biased norm scores than SCN. The predicted advantage of WCN was most apparent in samples drawn from populations 2 and 3, in which individuals with high levels of education were under-represented. In samples drawn from populations 4 (Under-representation of both low and high education) and 6 (Clustered sampling), WCN showed no benefit over SCN, but neither did it degrade the quality of norm scores, relative to continuous norming without compensatory weighting. In population 5 (Biased joint probabilities), we found that WCN led to a small increase in norm score error, but only at certain points in the range of cognitive ability.

In normative samples drawn from Population 1, which served as the standard of representativeness for the demographic variables, WCN demonstrated no advantage over SCN. This result is not surprising: WCN creates weights to compensate for departures from representativeness. Because Population 1 was the benchmark, in terms of demographic composition, random samples drawn from it were expected to be demographically representative.

Population 2 introduced deviations from the benchmark distribution of education, the demographic variable with the strongest effect on cognitive ability. Specifically, level 1 (high education/high ability) was mildly under-represented, and level 3 (low education/low ability) was proportionately over-represented. In samples drawn from Population 2, WCN yielded greater reduction in norm-score bias than SCN, for both error measures, across the entire range of cognitive ability.

Population 3 presented a pattern of non-representativeness on education that was similar to that in Population 2, but greater in magnitude. The comparison of normative samples drawn from Populations 2 and 3 was relevant to testing our second hypothesis, which specified that as the non-representativeness of the normative sample increased, norm-score bias would increase for both methods, but that the increase in bias would be smaller for WCN than for SCN. Our findings provided support for this hypothesis: with samples drawn from Population 3, the magnitude of norming error for WCN was larger than it was in the Population 2 analyses, although WCN retained its superiority to SCN in terms of reducing norm score bias. With population 3, the increase in norming error associated with WCN depended on person location – it occurred at either extreme of the range of the cognitive ability variable, but not in the middle region. In no instances, however, did these increases in the error measures exceed 1 IQ point.

Population 4 embodied a further scenario of demographic non-representativeness, in which both tails of the education distribution were under-represented, and the central region of the distribution was proportionately over-represented. In terms of the average degree of misrepresentation across the three levels of education, Population 4 did not differ from Population 3. Where the effect of the demographic manipulation differs is on the raw score distributions. In Population 4, the manipulation attenuates the variance of the raw score distributions, because under-sampling both tails of the education distribution results in an under-sampling of the very high and low raw scores that reside in those regions. In addition, whereas in Population 3 the pattern of misrepresentation affects the mean of the raw score distribution, in Population 4 the mean is not affected, because there is equal under-representation of both tails of the raw score distribution.

In normative samples drawn from Population 4, we observed that in certain regions of person location, WCN was less effective than SCN in reducing norm-score bias, which is consistent with our third hypothesis. Specifically, we found that the disparity between WCN and SCN increased at both tails of the cognitive ability distribution, with WCN showing the greatest magnitude of norming error in the lowest region of person location.

To put this finding into context, consider how the raw score distribution of a demographic subgroup is affected differentially by adding additional individuals, as opposed to weighting the existing raw scores without increasing sample size. Adding more individuals increases the variance of the raw score distribution, whereas weighting existing raw scores does not affect the variance. In Population 4, furthermore, the variance of the low and high ability groups was reduced by the pattern of under-representation, which results in fewer individuals in each of these groups. Therefore, weighting the raw scores of the under-represented groups increases the influence of any sampling error that exists in the raw score distributions. This phenomenon may explain our finding that WCN resulted in greater norming error in the under-represented, low region of person location. By contrast, WCN did not yield increased norming error in the central region of person location, where there are more observations present and a consequent reduction in sampling error. Our findings suggest that researchers should employ WCN with caution when processing normative samples where the non-representative subgroups are also those containing few individuals.

In Population 5, the joint distributions of the demographic variables (resulting from a complete cross-classification of the three variables) were manipulated in a pattern of alternating over- and under-representation. This was accomplished so that the marginal distributions of the variables closely approximated those of the reference population. Thus, Population 5 simulates a sampling scenario wherein demographic misrepresentation occurs at a level that is “beyond the reach” of cNORM’s raking and weighting methods, which operate only on marginal distributions.

Under these conditions, WCN did not provide any improvement in the reduction of norm-score bias over SCN. However, our manipulation of the joint probabilities, as it turned out, did not strongly affect the means and variances of the raw score distributions. Thus, this analysis leaves unanswered the question of how WCN might perform when misrepresentation at the level of joint probabilities does bias the parameters of the raw score distributions.

It is important to keep in mind that in our simulation study, the three demographic variables were modeled so that education had the strongest relationship with cognitive ability, and thus had more impact on norm score accuracy than ethnicity or region. Thus, our findings with Population 5 do not reflect the range of possible relationships between demographic factors and cognitive ability (e.g., other variables that are highly correlated with ability, or that interact with each other). In these alternate scenarios, it is unknown how misrepresentation in the joint distributions might affect raw score means and variances. Later in this section, we provide guidance on how to address these scenarios in practice.

Commented [DH22]: I found the original line of argument about the Population 5 findings to be convoluted and difficult to follow. I tried to simplify the argument here, but please check against the original to make sure that the new narrative accurately reflects the points you were trying to make, and that it omits no important details.

In Population 6 (clustered distributions), the distributions of the demographic variables were manipulated *within* each of the six age cohorts. This manipulation is best understood in comparison to Population 1, in which the marginal and joint probabilities of the entire population are replicated within each age cohort. In Population 6, by contrast, two-thirds of the joint distribution cells contained no data, meaning that the overall demographic distributions were *not* replicated *within* the age cohorts. However, the pattern of data deletion was such that the marginal and joint probabilities of Population 6, averaged over the entirety of the population (across all age cohorts), matched those of Population 1.

In normative samples drawn from Population 6, the age-specific patterns of demographic non-representativeness affected the parameters of the raw score distributions within each age cohort. Raking per se cannot compensate for discrepancies of this nature, because raking operates on marginal probabilities of the entire normative sample, not those within each age cohort. It is therefore counter-intuitive to find, as we did, that neither WCN nor SCN yielded increases in norm-score bias, when compared to the benchmark condition. We attribute this finding to the influence of the semi-parametric continuous norming method that underlies both WCN and SCN. As noted previously, this method models a developing cognitive ability as a monotonic function of age and person location. It thereby uses the stable variance of the entire normative sample to smooth the parameters of the raw score distributions across age cohorts, even when those age-specific distributions are affected by varying levels of demographic non-representativeness.

> **Commented [DH23]:** Please compare this paragraph to the original manuscript. The original used jargon such as "wavy iso-percentiles" and "stiffness of the method" which was unfamiliar to me. I tried to rewrite it in a simpler fashion that relied on basic concepts of continuous norming that were introduced earlier in the manuscript. Please confirm that I'm capturing your intended meaning here.

Implications for the use of WCN in test norming

Our study showed that WCN reduces norm-score bias under certain patterns of non-representativeness of a demographic variable, where that variable is strongly correlated with the test score being normed. The pattern of results across the six simulated populations, however, suggested that even when a demographic variable has a strong effect on raw scores, it produces relatively small distortions in resulting norm scores. Even under conditions representing large departures from demographic representativeness, the differences in *RMSE* between norm scores derived using WCN and those from representative samples did not exceed 2 IQ points.

With norming methods that generate norms independently for each age group, we would expect departures from demographic representativeness to cause greater levels of norm-score bias. These conventional norming methods lack the previously noted advantage of continuous norming, which can smooth out local effects of non-representativeness.

Consistent with this view, we have demonstrated previously that with conventional norming per age group, *RMSE* is about twice as high, on average, as with semi-parametric continuous norming, even with representative random samples (W. Lenhard & Lenhard, 2021). The selection of an appropriate norming method is therefore a critical prerequisite for accurate test norms, regardless of whether this procedure is used with or without weighting. By contrast, the size of the normative sample is less critical, if continuous norming is used. For example, we found that increasing sample size from 100 to 250 per age group did not yield significant reduction in *RMSE*, when continuous norming methods were used (A. Lenhard et al., 2019).

Clearly, the best practice is to prevent problems associated with non-representativeness in the first place, by collecting an adequately sized, demographically representative sample for norming. Post-hoc weighting procedures are no substitute for a well-planned data collection effort that draws randomly from the general population. Care must also be taken to avoid over-sampling from clinical settings, as this will bias the sample towards individuals of lower ability.

The current study demonstrates the utility of weighting procedures in reducing norm-score error under conditions of mild-to-moderate non-representativeness of a demographic variable. Nevertheless, we also found that the ability of WCN to reduce norm-score bias was degraded, when we reduced the marginal probability of the high level of education to 20% from the reference value of 40% (that is, when the size of that subgroup was half that needed for a representative sample). Our work further shows that the effectiveness of weighting depends on the location of under-represented demographic groups on the spectrum of person ability. With a typical cognitive ability that is normally distributed in the general population, random sampling will yield relatively small subgroups at either tail of the ability distribution. If these extreme subgroups are under-sampled to begin with, any sampling error embodied in the raw score distributions will only be multiplied by the application of compensatory weights. This can lead to increased norm-score bias, as illustrated in our results. The remedy, of course, is to ensure that these low- and high-ability groups are represented in adequate numbers.

As described previously, the raking procedures used in this study operate only on the marginal distributions of the demographic variables. Census information on the joint distributions of the demographic variables (e.g., the expected probability for the joint category of low education/non-white ethnicity) is not always available. However, when that information is available, it can be incorporated in the raking procedures through a recoding process. For example, the crossing of two demographic variables, each of which has three categories, results in nine cross-classification cells. These classifications can be recoded into nine levels of single dummy variable. The expected joint probabilities of the cross-classified cells thereby become the expected marginal probabilities of the dummy variable. The risk in this approach comes from increasing the number of categories, which also increases the likelihood that one or more category would have a very low expected probability. Under these circumstances, of course, even adequately sampled categories may hold only a few individuals, thus increasing the influence of sampling error when weights are applied.

To counter this tendency, we often recommend reducing the number of demographic categories by combining groups that are not expected to differ significantly in mean location on the ability variable. This practice can be applied to either marginal categories or joint cross-classifications, when the latter are subject to the recoding procedure described in the previous paragraph.

Limitations of the study

This study evaluated only one method of post-stratification: raking with marginal probabilities as the input. We did not examine fully cross-classified post-stratification (i.e., a method that takes joint probabilities into account). Instead, we analyzed norm samples drawn from Population 5 (biased joint distributions), to determine the performance of raking under conditions where the marginal probabilities are representative, but the joint probabilities are not. In Population 5, we did not find that WCN, which includes raking, yielded increased norm-score bias compared to the benchmark condition. This may have been due to the magnitude of non-representativeness in the cross-classification cells. The demographic deficiencies in these cells may not have been great enough to expose the inability of raking to compensate for such deficiencies.

In our study, we simulated three demographic variables (education, ethnicity, region), with varying levels of correlation with the latent cognitive ability (strong, moderate, weak, respectively). We did not model any interactions among these three variables. Demographic variables that interact in their effects on cognitive ability might yield larger disturbances in the raw score distributions of the cross-classification cells. Under these conditions, as we have demonstrated, weighting carries the risk of increasing norm-score bias. However, the main effect of education on test scores in our study probably represents the upper limit of analogous effects that could occur in real-world normative samples. With demographic variables that have smaller effect sizes, of course, we can expect resulting norm-score biases to also diminish in magnitude.

A second limitation was that our study modeled only one latent psychological variable: a cognitive ability that increases monotonically with increasing age. Other variables measured by psychometric tests (e.g., the “big five” personality traits, Donnellan & Lucas, 2008) may not manifest the same dependency on age, and they may be affected by demographic variables with different characteristics than the ones simulated in our study. When norming tests of personality traits, therefore, it may be appropriate to apply a weighting method that is not combined with continuous norming procedures.

A third caution relates to the mathematical underpinnings of the cNORM norming process. cNORM uses a semi-parametic continuous norming method that requires the expansion of a Taylor polynomial (for more details, see [INSERT CITATION]). The modeling process calls for specification of a parameter (*k)*, that sets an upper bound on the exponents of person location and age. In the current study, we used a default value of $k = 4$ for both location and age. It is possible that more precise models of the latent cognitive ability could have been obtained with different values of *k*. Simulation studies published elsewhere ([INSERT CITATION]) have compared norm-score bias across a range of values of *k*. These findings suggest that $k = 5$ for location and $k = 3$ for age provide an optimal balance between norm score accuracy and processing load. As a result, we have selected these values as the defaults for the current version of cNORM.

Finally, our study examined weighting only as applied to the semi-parametric continuous norming method implemented in the cNORM package. We did not combine weighting with other continuous norming approaches, such as parametric continuous norming (e.g., Stasinopoulos et al., 2018). Earlier in this paper, we pointed out that the semi-parametric continuous method, because it does not rely on splines to model age-related changes in ability, may be better suited for certain conditions of non-representativeness in normative samples (e.g., the clustered distributions modeled in Population 6). Moreover, we have demonstrated elsewhere (see A. Lenhard, 2019) that the cNORM approach yields less norm-score bias than parametric continuous norming with skewed raw score distributions, and with sample sizes of 150 or less per age group. Yet, the efficiency of post-stratification techniques combined with parametric continuous norming remains to be investigated.

Concluding Remarks and Outlook

The application of weighting techniques to the norming of psychometric tests is a relatively new area of study. Unsurprisingly, therefore, several additional research questions emerged from the current simulation protocol. For example, we implemented raking weights twice within cNORM: Once during ranking of raw scores, and then again during regression modeling. But we did not evaluate the relative value, in terms of reducing norm-score error, of the second step. It is therefore possible that applying weights to the regression analysis was of little benefit, or that it may have even increased norm score bias. The latter might occur because, as noted previously, weighting can multiply the effects of sampling error in under-represented groups.

This last caution serves as a final reminder that weighting techniques are no substitute for the painstaking process of assembling a demographically representative normative sample. Our study has shown, however, that if such samples still exhibit reasonably small departures from representativeness, the weighting methods implemented in cNORM offer a useful way of mitigating any resulting norm-score bias.

Acknowledgements

We extend our gratitude to WPS (Torrance, CA) for funding the simulation studies. We acknowledge that the fourth author (DSH) is employed by WPS.

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. Modeled cognitive ability in reference population.

Figure 2. *RMSE* across simulated populations, with (WCN) or without (SCN) weighting. The grey rectangles represent 95% confidence intervals.

Figure 3. *MSD* across simulated populations with (WCN) or without (SCN) weighting. The grey rectangles represent 95% confidence intervals.

Figure 4. *RMSE* across simulated populations, with (WCN) or without (SCN) weighting, as a function of person location. The dotted grey line represents SCN with norm samples drawn from Population 1 (benchmark).

Figure 5. *MSD* across simulated populations, with (WCN) or without (SCN) weighting, as a function of person location. The dotted grey line represents SCN with norm samples drawn from Population 1 (benchmark).