Regression-based normative data in neuropsychology: using raw scores as observed response variable outperforms transforming for normality

Javier Oltra-Cucarella[1,2], Rubén Pérez-Elvira[3,4], Beatriz Bonete-López[1,2], Clara Iñesta[2], Esther Sitges-Maciá[1,2], Rafael de Andrade Moral[5]

[1]Department of Health Psychology, Universidad Miguel Hernández de Elche, Spain

[2]SABIEX University for Seniors, Universidad Miguel Hernández de Elche, Spain

[3]Neuropsychophysiology Lab, NEPSA Rehabilitación Neurológica, Salamanca, Spain

[4]Faculty of Psychology, Universidad Pontificia de Salamanca, Spain

[5]Department of Mathematics and Statistics, Maynooth University, Ireland

Conflict of interest: the authors disclose no conflict of interest

Correspondence concerning this article should be addressed to Prof. Esther Sitges-Macià, Department of Health Psychology, Miguel Hernandez University, Avda de la Universidad s/n, Edificio Altamira, 03202 Elche, Alicante, Spain. Email: esther.sitges@umh.es

Funding

This project was partially funded by the Conselleria d'Innovació, Universitats, Ciència i Societat Digital, Generalitat Valenciana (NEUROPREVENT project, GV/2021/139)

Regression-based normative data in neuropsychology: using raw scores as observed response variable outperforms transforming for normality

Abstract

Regression-based normative data for neuropsychological variables are increasing popularity over the last years. However, some use raw data while others use transformation when the observed response variable is skewed. This work analyzes how well the linear models fit for each type of variable. We used real data from a sample of n=163 cognitively healthy individuals and compared the fit of linear regression models for raw scores and for corrected scaled scores. We then simulated a population of 1,000,000 individuals and drew 1,000 random samples of different sizes (n=100, 200, 5000, 1000, 10,000) for 7 different scenarios, analyzed the percentage of individuals scoring in the lowest 5% and analyzed the agreement between models with the Cohen's kappa statistic. Linear models for raw scores and for scaled scores were similar when the model included all the covariates, but barely identified low scores when scaled scores were corrected with covariates taken from different regressions (kappa = 0.58). Models with raw scores showed that the expected number of individuals scoring low was close to the expected 5%, whereas models with scaled scores with covariates taken from different regressions were close to 0%. The two models agreed only when the response variable was random symmetrical and uncorrelated with the covariates. When calculating normative data using linear regressions, raw scores should be the preferred choice. If residuals analysis show that the model does not fit the data well, researchers should consider using nonlinear models. Transforming data for normality of the observed response is discouraged.

Key-words: Generalized linear model; linear regression; neuropsychological assessment; normative data; residuals

Public significance statement

- If the linear model assumptions hold, linear or binomial models with skew raw scores as response variable fit the best. Models that transform raw data to scaled scores through percentiles do not fit the data well, identifying a lower-than-expected percentage of individuals scoring low. If the model assumptions do not hold, Generalized Linear Models might be a good alternative to linear models with transformed response variable.

Regression-based normative data in neuropsychology: using raw scores as observed response variable outperforms transforming for normality

Normative data play a crucial role in detecting cognitive impairments during neuropsychological assessments (Strauss et al., 2006). By utilizing data derived from population-based samples, clinicians can pinpoint cognitive impairments at an individual level for patients with confirmed or suspected brain injuries. Normative data are particularly essential for identifying memory impairments in older adults with suspected Mild Cognitive Impairment (MCI), Alzheimer's Disease (AD) or other dementias, as the diagnostic criteria necessitate objective cognitive impairment (McKhann et al., 2011; Petersen, 2004; Winblad et al., 2004).

Various methods exist for computing normative data (Strauss et al., 2006), with increasing popularity observed in the use of normative data based on linear regression equations. Unlike traditional normative data, which compute means and standard deviations for a sample or specific subgroups within a sample (e.g., age ranges or sex), regression-based normative data (RBND) use linear regression to predict a test score based on the performance from a reference sample. The individual's actual score is then compared to the expected score of people of the same age, sex, or educational level. RBND from different countries are available (delCacho-Tena et al., 2023), with some providing online calculators to facilitate use by clinicians (Calderón-Rubio et al., 2021; Iñesta et al., 2021, 2022; Shirk et al., 2011). However, some RBND apply the same methodology to different types of variables. While the predominant approach involves regressing raw scores on demographic variables such as age, sex, and level of education, alternative methodologies have been employed. In these RBND approaches, raw scores are first converted to percentiles, then further transformed into Scaled Scores (SS) using percentile ranges. Subsequently, SS are regressed on demographic variables. For instance, this methodology was employed in developing normative data at the Mayo Clinic for various tests, including the Wechsler Adult Intelligence Test – Revised (Ivnik et al., 1992b), the Free and Cued Selective Reminding Test (Ivnik et al., 1997) and the Auditory Verbal Learning Test (Ivnik et al., 1992a) among other tests (Ivnik et al., 1996) and updates thereafter

(Lucas et al., 2005; Steinberg et al., 2005; Stricker et al., 2021). In Spain, different research groups have adopted these methodologies for deriving normative data (Campos-Magdaleno et al., 2024; García-Herranz et al., 2022; Peña-Casanova, Blesa, et al., 2009), while others have utilized means and standard deviations (Campo & Morales, 2004) or applied RBND to raw scores (Alviarez-Schulze et al., 2022; Calderón-Rubio et al., 2021; Guàrdia-Olmos et al., 2015; Iñesta et al., 2021, 2022; Rivera & Arango-Lasprilla, 2017).

Nevertheless, upon closer scrutiny of the Regression-Based Normative Data (RBND) applied to Scaled Scores (SS), concerns arise regarding the usability of these normative data. Such concerns encompass both theoretical and practical implications. The objective of this study is to assess the accuracy of RBND on SS and analyze its psychometric properties.

Linear regression – Brief description

The General Linear Model (LM) refers to conventional linear regression models in which a continuous response variable is predicted using continuous or categorical predictors, as shown below:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_p + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2) \quad \text{(Eq. 1)}$$

where y_i is the score for individual i, β_0 is the intercept, $x_{1i}, x_{2i}, \ldots, x_{pi}$ are predictor variables, $\beta_1, \beta_2, \ldots, \beta_p$ are the unstandardized coefficients for each predictor, and ε_i is a normally distributed error term. In linear regression, the intercept is the mean value of the observed response variable when all the predictors are equal to zero, and the predictor's coefficient is the mean change in the observed response variable for each 1-unit increase in the predictor, while holding the remaining predictors constant. Different statistical tests and procedures under the LM can be performed, such as the ANOVA to decompose variation, or t-tests to assess whether a specific coefficient is different from zero. For instance, the value of the t-statistic and its associated p-value are the same for a t-test and for a univariate linear regression with a dummy predictor. After having summarized briefly the LM, we will now develop in detail our reasons to believe that the use of linear regression on SS is incorrect.

Normality in linear regressions

The rationale behind the use of SS as the response variable in regression analyses is that they are linear transformations of the raw scores through percentile ranks, and thus non-normal variables are accommodated to a normal distribution. For instance, Karstens et al. (2024) argued that "Standardized scores were used to minimize skewness for tests that are not normally distributed" (p. 391), which was argued by Peña-Casanova et al. (2009) and then replicated by others (Campos-Magdaleno et al., 2024; García-Herranz et al., 2022) who argue that transforming raw scores to SS through percentile ranks "…produced a normalized distribution (M = 10; SD = 3) on which linear regressions could be applied".

According to the literature cited in the work by Peña-Casanova and colleagues, the rationale behind this claim is rooted in the methodology outlined by Ivnik et al. (1992b). We contend that this assertion likely reveals an unintentional misunderstanding of the General Linear Model, primarily due to two essential reasons. First, Ivnik et al. (1992b) demonstrated that transforming raw scores on the Digit Symbol subtest from the WAIS-R to percentiles and then back to Scaled Scores (SS) resulted in an approximately normal distribution. This is likely because the nature of the Digit Symbol subtest allows for a broad range of scores, resulting in multiple scores falling within each percentile rank. Consequently, although raw scores may not follow a normal distribution, SS might exhibit an approximate normal distribution as they are derived from percentiles rather than raw scores. However, this is not applicable to skewed data such as raw scores on verbal memory tests. For instance, raw scores from the Spanish version of the Free and Cued Selective Reminding Test (FCSRT) used in the work by Peña-Casanova et al. (2009) are bounded between 0-16, with most of the normative sample scoring in the upper limit or showing ceiling effects. This leads to a negatively skewed distribution of raw scores, a common occurrence in the analysis of data from cognitively normal individuals undergoing tests of verbal memory (Girtler et al., 2015; Harrington et al., 2017; Uttl, 2005).

As an example, Figure 1 shows the distribution of raw scores and the distribution of SS obtained through percentile ranks for the Total Delayed Recall variable from the FCSRT taken

from 163 cognitively healthy participants (Calderón-Rubio et al., 2021). The lower bound of scores is 9, with 48.47% of the sample scoring 16 and around 70% of the sample scoring 15 or higher. Taking the definition of percentiles as the percentage of people in the sample showing a score equal to or lower than X (Crawford et al., 2009), percentiles assigned to scores 9-16 are 0.6%, 2.5%, 4.3%, 6.1%, 12.9%, 28.8%, 51.5% and 100% respectively. Consequently, around half of the sample is assigned percentile = 100, resulting in the highest SS, clearly indicating a non-normal distribution of SS. This is consistent with tables of normative data for memory tests in the work by Peña-Casanova et al. (2009), where a score of 15 corresponds to a SS of 13 and a score of 16 corresponds to a SS of 18. Similarly, in the work by Campos-Magdaleno et al. (2024) the upper raw scores on the delayed free recall from the California Verbal Learning Test correspond to SS of 15 and the lowest scores correspond to a SS of 3, spanning more than 2SD below the mean to less than 2SD above the mean. All these data show that, for skewed data, transforming raw scores to percentile ranges and these percentiles to SS do not ensure an approximately normal distribution of SS, rendering the first assumption incorrect.

However, the assumption that the observed response variable (i.e., a factor of independent scores) must follow an approximately normal distribution for the linear regression to be applied is not correct. As shown in Eq.1, in the LM it is the residuals (i.e., the difference between observed and predicted scores), and neither the observed response variable nor the predictors which have to follow an approximately normal distribution (Tabachnick & Fidell, 2013), especially in small samples (Williams et al., 2013) for the inferential results to be trustworthy. In fact, it has been argued that normality of residuals does not significantly impact bias and outcome transformation is unnecessary and even worse than normality assumption violation (Schmidt & Finan, 2018). However, additional concerns arise when it comes to the use of SS as response variables in the regression equation.

Predictors in the regression equation with SS as response variable

Besides the distributional assumption of the observed response variable mentioned in the previous section, the most significant concerns are related to the use of SS, rather than raw

scores, as response variable in the model, and to the way previous research has dealt with the predictors in the regression equation.

As was said in the preceding sections, the intercept is the mean value of the response variable when all the predictors are zero. In neuropsychology, most of the variables used as predictors in linear regression analyses are continuous and positive, with no 0-values. For instance, raw age does not have 0-values in datasets that include people aged 50 or older. This implies that the intercept loses its meaning and no longer reflects the mean value of the response variable when all the predictors are zero, as there are no zeros in the predictor variable. To address this issue, each predictor can be centered around an arbitrary value, transforming each predictor to have a value of 0. One common method of centering the predictors is using the mean of each predictor (Arango-Lasprilla et al., 2017; Campos-Magdaleno et al., 2024; Peña-Casanova, Blesa, et al., 2009), although the lowest value in the distribution can also be used as reference (Calderón-Rubio et al., 2021; Iñesta et al., 2021, 2022). In the absence of interactions, centering the predictors does not affect either the predictor's coefficient or its associated p-value (Tabachnick & Fidell, 2013; Williams et al., 2013), but makes the intercept interpretable. When all the variables in the regression model are interpretable, the regression equation provides a predicted score that must be compared against observed scores (i.e., residuals) in order to analyze the model assumptions of normality, independency and homoskedasticity of the residuals.

In neuropsychology, in order to test the accuracy of the regression model, predicted scores must be calculated using the intercept and the predictors' coefficients as shown in Eq.1. The predicted values are then subtracted from the observed values ($y - \hat{y}$), and the difference is divided by the standard deviation of the equation. The standardized difference between observed and predicted scores provides a z-score that can be interpreted using tables of cumulative probability, assuming that the residuals follow an approximate normal distribution or the sample size is large enough. When using test scores to predict retest scores, this procedure is referred to as the regression-based Reliable Change Index (Crawford et al., 2012;

Crawford & Garthwaite, 2007), and has proven effective to identify individuals with MCI at a higher level of progressing to AD (Duff et al., 2017; Oltra-Cucarella et al., 2022). Recently, de Andrade Moral, Díaz-Orueta and Oltra-Cucarella (2022) showed that the accuracy of the regression-based Reliable Change Index to identify individuals with cognitive decline approaches 95% for samples of size 200 or larger. These data suggest that linear regression is accurate to identify cognitive impairment using a cut-off based on z-scores from standardized residuals.

The accuracy of the methodology reported in previous works to adjust SS using linear regression is unclear. Several works built the regression equation using the uncorrected SS and a linear combination of each predictor multiplied by its coefficient taken from univariate regression models, without including the intercept (Campos-Magdaleno et al., 2024; Delgado-Losada et al., 2021). This raises several concerns. First, the lack of the intercept provides no reference of the mean value of the response variable when all the predictors are zero. And second, as coefficients from separate univariate regression equations are used, there is no possibility of calculating the error of the equation. And, if residuals cannot be calculated, then the model assumptions related to independence of errors, homoskedasticity and absence of outliers or leverage cannot be tested. The rationale for using this methodology seems to be the work by Mungas et al. (1996), where adjusted scores on the Mini-Mental State Examination (Folstein et al., 1975) were calculated without the intercept from a regression equation according to the formula provided.

In summary, previous works aimed at developing normative data using regression equations with SS as response variable raise several doubts about the efficacy of their methodology: 1) they rely on the assumption of normality of the observed response variable by calculating SS for extremely skewed data, 2) they generate separate univariate regressions for each predictor, and 3) they combine coefficients from each separate univariate regression into the same equation without considering the effects of the intercepts. All these misunderstandings raise serious concerns about the utility of that methodology for the calculation of normative

data. If the variables do not behave as expected based on the statistical assumptions of the LM, the utility of the regression model is unknown. Not only does this mean that RBND have a high risk of both false negatives and false positives in the identification of cognitive impairment, but also that other situations using standard scores in this manner might be unreliable. In neuropsychology, as in other areas of psychology (e.g., intelligence or depressive symptoms), linear models can be used to test the effects of categorical variables (e.g., sex) on a continuous outcome, and some use standard scores to interpret the model (i.e., the regression-based Reliable Change Index). The aim of the present work is to analyze how well RBND with SS as response variables behave compared to RBND with raw scores as response variable for extremely skewed data, because if there are significant differences between methods many areas in Psychology and other health sciences can benefit from our results by developing more robust models. Our hypothesis is that the regression models with SS as response variable will not fit the data as expected from the LM.

Methods

Transparency and openness

This study's design and its analysis were not preregistered, but all code, scripts and data are available at the first author's website (Oltra-Cucarella, 2025) and upon reasonable request.

We begin by analyzing the Delayed Recall scores from the FCSRT observed in real data collected on 163 participants from the SABIEX (SABIduría y EXperiencia) at the Universidad Miguel Hernández de Elche (Bonete-López et al., 2021; Calderón-Rubio et al., 2021; Iñesta et al., 2021, 2022), a study on aging and cognition in highly cognitively active Spanish people aged 55 years or older. These analyses are carried out to exemplify how to analyze and interpret the residuals from the linear models, one with raw scores as the response variable and one with SS as the response variable. The SABIEX study was approved by the Ethical Committee at the Universidad Miguel Hernández de Elche.

The FCSRT is widely used to assess verbal memory through 16 items, which are presented in printed letters in four cards each with four items. Examinees are requested to read the words out loud, and a semantic cue is provided for each item for deep encoding of the learning material (e.g., which one is a tool?). Examinees are required to remember as many items as they can through 3 learning series (free recall), and the semantic cues are provided for items not recalled during learning (cued recall). Delayed recall is requested (both free and cued recall) after 30 minutes. The Spanish version of the FCSRT was used (Peña-Casanova, Gramunt-Fombuena, et al., 2009) along with an additional recognition task (Bonete-López et al., 2021). The present work focused on the Delayed Recall Total score (FCSRT-DR), which ranges from 0 to 16 and includes both free and cued recall. The FCSRT helps to differentiate between storage and retrieval impairments, and has been suggested as a reference tool for the identification of memory impairments in AD (Dubois et al., 2014).

A linear regression of FCSRT-DR raw scores on age, sex and education (as continuous) was run and the regression assumptions were statistically tested. Normality of residuals was tested with the Shapiro-Wilk test. Independence of residuals and homoskedasticity were tested with the Durbin-Watson test and the Breusch-Pagan test respectively from the *lmtest* package. Leverage was analyzed with the *augment()* function from the *broom* package, with values ≥1 suggesting high leverage (Weisberg, 2014).

After analyzing raw scores, we replicated the Mayo procedure and transformed raw scores into Scaled Scores (SS) through the percentile range, and these SS were regressed on age, sex and education (continuous) and the same statistical tests were used to analyze the residuals.

Simulating a fictitious population

Then we simulated a population of 1 million individuals and their associated age, sex, and education levels using the same data from SABIEX. The age variable was simulated from a beta distribution with shape parameters equal to 1.57 and 2.86, multiplied by 32 (the age range in the real data) and added to 55 (the minimum age in the real data), to ensure simulated ages

are between 55 and 87 (which is the range observed in the real data). The sex variable was simulated from a Bernoulli distribution with a probability of success of 66.26%, which is the proportion of females in the real data. Finally, the education variable was simulated from a uniform distribution ranging from 3 to 22, which is the range of years of education in the real data.

We simulated the individual scores considering seven main scenarios. These included symmetric, left- and right-skewed score distributions based on a normal (scenarios 1-3) or binomial distribution (scenarios 4-6) whose means depended on the age, sex and education covariates, plus a purely random symmetric score distribution that was unrelated to the covariates, originating from a normal distribution with a mean of 10 and standard deviation of 3 (scenario 7). We selected these 7 scenarios to cover different situations that can be found in real world settings: although the first 6 scenarios simulate skewed data, the first 3 scenarios simulate data from a normal distribution, whereas scenarios 4-6 simulate data from binomial distributions for discrete bounded data, which are common in neuropsychological assessment and have proven useful in previous research (De Andrade Moral et al., 2022). Scenario 7 was proposed to understand how the two approaches would behave when the data was generated from a normal distribution in the absence of the effects of covariates.

We now present the simulation setups for each of the first six scenarios, starting with scenarios 1-3 which involve simulating from normal distributions.

Let Y_i be the random variable representing the score for individual *i*. For the scores simulated form a normal distribution (scenarios 1-3), we assume

$$Y_i \sim N(\mu_i, \sigma^2).$$

The left-skewed scores (scenario 1) were obtained by specifying

$$\mu_i = 15.1221 - 0.0325 \times \text{age}_i + 0.4699 \times \text{sex(female)}_i + 0.1364 \times \text{education}_i$$

and $\sigma^2 = 1.7824$. These parameter values were calculated by fitting a linear regression model to the real data scores. For the right-skewed scores (scenario 2), the specification was

$$\mu_i = 0.8779 + 0.0325 \times \text{age}_i - 0.4699 \times \text{sex(female)}_i - 0.1364 \times \text{education}_i$$

with the same value for the variance as before, which is simply taking the complement of the intercept from 16 (which is the maximum score in the real data), and flipping the signs of the regression coefficients. This is equivalent to fitting a linear regression model to 16 minus the real data scores. For the symmetric scores (scenario 3), firstly new scores were simulated to replace the scores in the real data, by calculating

$$p_i^* = \frac{\exp\{1 - 0.03 \times \text{age}_i + 0.5 \times \text{sex(female)}_i + 0.1 \times \text{education}_i\}}{1 + \exp\{1 - 0.03 \times \text{age}_i + 0.5 \times \text{sex(female)}_i + 0.1 \times \text{education}_i\}}$$

then simulating from a normal distribution with mean $16 \times p_i^*$ and variance $16 \times p_i^* \times (1 - p_i^*)$. After that, a linear regression was fitted using these newly simulated scores using the real data covariates, allowing us to specify

$$\mu_i = 9.5096 - 0.0802 \times \text{age}_i + 1.5911 \times \text{sex(female)}_i + 0.3501 \times \text{education}_i$$

and $\sigma^2 = 3.6856$.

We now present the simulation setups for scenarios 4-6, which involve simulating from binomial distributions. For this, we assume

$$Y_i \sim \text{Binomial}(m = 16, \pi_i).$$

The left-skewed scores (scenario 4) were obtained by specifying

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = 2.5779 - 0.0275 \times \text{age}_i + 0.4688 \times \text{sex(female)}_i + 0.1411 \times \text{education}_\text{i}.$$

These parameter values were obtained by fitting a logistic regression to the real data. For the right-skewed scores (scenario 5), we used

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = -2.5779 + 0.0275 \times \text{age}_i - 0.4688 \times \text{sex(female)}_i - 0.1411 \times \text{education}_\text{i}$$

which are obtained by simply flipping the signs in the logit scale; this is equivalent to fitting a logistic regression model to 16 minus the real data scores. For the symmetric scores (scenario 6), a logistic regression model was fitted to the simulated scores obtained in scenario 3 above, yielding

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = 0.3762 - 0.0212 \times \text{age}_i + 0.4217 \times \text{sex(female)}_i + 0.0940 \times \text{education}_\text{i}.$$

Simulation studies

After simulating this population of 1 million individuals, the simulation study consisted in drawing 1,000 samples of sizes 100, 200, 500, 1,000, 2,000, and 10,000 without replacement from this population, then calculating (i) raw regression scores based on the linear and logistic reliable change indices ("Raw Score Regression" approach; RSR), (ii) the scaled scores which use the normal distribution quantile function and covariate-based corrections ("Scaled Score Regression" approach; SSR), (iii) the percentage of the RSR scores that are equal to or less than -1.64, which would indicate reliable decline, (iv) the percentage of the SSR scores that are equal to or less than 5, which would indicate reliable decline, and (v) Cohen's kappa agreement index between the dummy variables which indicate reliable decline based on either RSR and SSR.

Calculating scores using the RSR approach

To calculate the RSR scores, first a linear regression model is fitted to the data using all available covariates (in this case, age, sex and education level). Then, a subsequent model is fitted using only the covariates that were found to be significant at a 5% level, if there was at least one non-significant covariate. After that, the linear regression-based reliable change index is calculated using the updated model. This is done by scaling the residuals by the estimated regression standard deviation. We then expect that approximately 5% of patients should have an RSR z-score equal to or lower than -1.64, which is the 5% percentile of the standard normal distribution. We selected an RSR z-score below -1.64 for comparison purposes with the SSR SS

≤ 5, as both correspond approximately to percentile 5th. We are aware that this cutoff is arbitrary, but it has been suggested as a cutoff for identifying cognitive impairment in single-case research when linear regression is used (Crawford & Garthwaite, 2012).

Calculating scores using the SSR approach

Here we are replicating what was done by the Mayo and NEURONORMA methods. To calculate the SSR scores, firstly we calculated the empirical cumulative distribution function (ecdf) of the observed scores. We then computed standard normal quantiles based on the ecdf values, multiplied them by 3, added 10, and calculated their ceiling (i.e. the next closest integer) (Peña-Casanova, Blesa, et al., 2009). Values greater than 19 were replaced with the value 19, whereas values less than 1 were replaced with the value 1. We call this variable "uncorrected SSR", or "uSSR". After that, and as reported in previous works (Campos-Magdaleno et al., 2024; Delgado-Losada et al., 2021), separate linear regressions were fitted to these scaled indices, one per covariate (in our case, three regressions, one for age, one for sex, and one for education level), with the regression for education level using a categorical variable equal to 1 if education level is between 0 and 5, equal to 2 if it is between 6 and 11, equal to 3 if it is between 12 and 15, and equal to 4 if it is greater than 16, which we call SSR. The significance of each predictor is assessed, and we calculate the following correction components:

$$\text{age}_i^* = \begin{cases} \{\text{age}_i - \text{mean(age)}\} \times \hat{\beta}_{\text{age}}, & \text{if age is significant} \\ 0, & \text{otherwise} \end{cases}$$

$$\text{sex}_i^* = \begin{cases} \text{sex}_i \times \hat{\beta}_{\text{sex}}, & \text{if sex is significant} \\ 0, & \text{otherwise} \end{cases}$$

$$\text{education}_i^* = \begin{cases} \{\text{education}_i - \text{median(education)}\} \times \hat{\beta}_{\text{ed}}, & \text{if education is significant} \\ 0, & \text{otherwise} \end{cases}$$

where $\hat{\beta}_{\text{age}}$, $\hat{\beta}_{\text{sex}}$, $\hat{\beta}_{\text{ed}}$ are the regression coefficients estimated by the three separate linear regression model fits for age, sex, and education, respectively. Finally, and following previous works (Campos-Magdaleno et al., 2024; Delgado-Losada et al., 2021), the SSR for individual *i* is obtained by calculating

$\mathrm{SSR}_i = \mathrm{uSSR}_i - (\mathrm{age}_i^* + \mathrm{sex}_i^* + \mathrm{education}_i^*)$, with low scores defined as SSR ≤ 5.

We compared the frequency of each uncorrected SS with that of the SSR_i calculated as detailed above using the Two-way Random Intraclass Correlation Coefficient (ICC) as a measure of interrater agreement with the icc() function in R. According to Koo and Li (2016) values of ICC less than 0.50 are interpreted as poor, values between 0.50-0.75 as moderate, between 0.75-0.90 as good, and 0.90 or above as excellent agreement. Additionally, we calculated the agreement in the number of individuals showing a low score (i.e., SS ≤ 5) for both the uSSR_i and the SSR_i with the Cohen's Kappa statistic (Cohen, 1960) using the kappa2() function from package irr (Gamer et al., 2019), with values below 0.40, between 0.40 and 0.75, and higher than 0.75 indicating no agreement, fair to good agreement and excellent agreement respectively (Fleiss et al., 2003).

Implementation

All simulation studies, analyses and graphs were generated using R (R Core Team, 2024). The linear model for the raw scores as the outcome was calculated using package "LogisticRCI" (De Andrade Moral et al., 2022).

Results

Regressions with the real sample

For the linear model with raw scores as the response variable, the Shapiro-Wilks test showed non-normality of residuals (W = 0.87, p < .001) likely due to the relatively high sample size, the Breusch-Pagan test showed no heteroskedasticity in the residuals (BP = 5.81, p = 0.121), and the Durbin-Watson test suggested independence of errors (DW = 1.99, p = 0.97). There were no observations with high leverage. The percentage of residual z-scores ≤ -1.64 was 7.3%, close to the expected 5%. All these analyses suggest that the linear model with raw scores as the outcome holds reasonably well.

For the linear model with SS as the response variable, the Shapiro-Wilks test showed non-normality of residuals ($W = 0.94$, $p < .001$), again likely due to the relatively high sample size, the Breusch-Pagan test showed no heteroskedasticity ($BP = 1.69$, $p = 0.640$) and the Durbin-Watson test suggested independence of errors ($DW = 2.12$, $p = 0.396$). There were no observations with high leverage. The percentage of $SS \leq 6$ was 4.2%, close to the expected 5%. All these analyses suggest that the linear model might hold reasonably well.

Lastly, the agreement on the SS assigned to each individual using the Mayo procedure (one single regression with all covariates) and the SSR procedure (one regression for each covariate) was poor (ICC = 0.19, $F_{(162,163)} = 1.49$, $p = .005$, 95%CI: 0.04-0.34), with fair to good agreement in the number of individuals obtaining a low score (Mayo = 4.2%, SSR = 1.8%; Kappa = 0.59, $z = 8.25$, $p < .001$), which implies that adding the effects of covariates taken from separate regressions substantially changes the SS obtained by including all the predictors in the same model.

Data from the simulated models

The results for the 1,000,000 left-skewed, right-skewed and symmetrical populations are depicted in Figure 2. The results for the simulations of the seven different scenarios are depicted in Figure 3. The results will be presented for each method separately.

Raw Score Regressions

When regressions were run using raw scores as the response variable, the skewed models showed that the proportion of individuals showing a low score was close to the expected 5% for the logistic models (scenarios 4-6) and slightly higher than the expected 5% for the linear models (scenarios 1-3). The symmetric model (scenario 7) showed that the proportion of individuals showing a low score was close to the expected 5%.

Scaled Scores Regressions

When regressions were run using scaled scores as response variable (three separate regressions for predictors), the skewed models showed that the proportion of individuals showing a low score was close to 0% both for the linear (scenarios 1-3) and the logistic models (scenarios 4-6). Contrary to what was found for raw scores, the symmetric model (scenario 7) showed that the proportion of individuals showing a low score was slightly lower than the expected 5%.

The discrepancy between the RSR and the SSR on the proportion of individuals showing a low score is observed in the Kappa agreement statistic, with values close to 0 for models with highly skewed scores, either for the linear or the logistic models. However, when the response variable was symmetrically distributed, the agreement between the RSR and the SSR was excellent.

Discussion

The aim of the present work was to analyze whether regression models of normative data using SS as the response variables would behave similarly to regression models using raw scores as the response variable, both for highly skewed data and for symmetric data. Our hypothesis was that the regression models with SS as response variable would not fit the data as expected according to the LM. After simulating two skewed populations and one symmetric population and applying 7 different scenarios, our results showed that the proportion of individuals obtaining a z-score for the discrepancy between the obtained and the expected score was close to the expected 5% for models using raw scores as the response variable, but close to 0% for models using SS as the response variables when covariates were obtained from different regression models. Conversely, the symmetric models showed a proportion close to the 5% both for raw scores and for SS as the response variable, with the RSR performing better than the SSR.

Selecting the right model to develop normative data is important, as normative data are used to diagnose neuropsychological impairments that ultimately lead to neurological diagnoses

such as Alzheimer's disease (Strauss et al., 2006). Although normative data have been calculated traditionally using means and standard deviations for groups, linear regression is becoming the preferred technique in the last years. For example, previous works on normative data have used linear regressions, with different response variables selected for the analyses. Although using raw scores is the most common approach (Calderón-Rubio et al., 2021; Iñesta et al., 2021, 2022; Kiselica et al., 2020; Rivera & Arango-Lasprilla, 2017; Shirk et al., 2011), others have opted to transform raw scores to percentiles, then percentiles to scaled scores and finally use these SS as response variable (Campos-Magdaleno et al., 2024; Karstens et al., 2024; Peña-Casanova, Blesa, et al., 2009). The rationale for this methodology is that converting raw scores to scaled scores helps to normalize variables that do not follow a normal distribution (Campos-Magdaleno et al., 2024; Karstens et al., 2024; Peña-Casanova, Blesa, et al., 2009). However, as we have shown, this rationale might be true for variables that deviate only slightly from normality or have a wide range of scores (e.g., the Symbol Digit Modalities Test), but is not true for variables that are highly skewed such as delayed recall scores from verbal memory tests such as the Free and Cued Selective Reminding Test (Campo & Morales, 2004; Ehrenreich, 1995; Grau-Guinea et al., 2020; Larrabee et al., 2000).

The first misconception is testing normality for the observed response variable in the linear model. Several works have emphasized that it is the residuals which must follow an approximate normal distribution in linear regression (Kéry & Hatfield, 2003; Schmidt & Finan, 2018; Williams et al., 2013), and even normality of residuals is the least important for the linear model to fit if the sample size is large enough (Schmidt & Finan, 2018). How large is large is a matter of debate, but some authors argue that even samples with two subjects per variable are large enough to estimate unbiased coefficients and unbiased standard errors and confidence intervals (Austin & Steyerberg, 2015).

The second concern was related to the use of SS as the outcome variable in the linear regression model. Our results showed that transforming highly skewed scores does not guarantee an approximate normal distribution, and using SS as the outcome was associated with

a very low proportion of individuals scoring SS ≤ 5 except when the outcome followed an approximate normal distribution. These results are in line with the example provided by Ivnik et al. (1992b) for the Digit Symbol test, a test whose score distribution follows an approximate normal distribution (Morlett Paredes et al., 2024; Williamson et al., 2022). The assumption of normality in linear regression must be checked on the residuals, and as Pek et al. (2017) showed non-normality of residuals is not a concern for large samples and transformation are unnecessary as they might produce more damage than maintaining raw scores. Our results are in line with this suggestion, because although residuals did not follow an approximate normal distribution, the model fitted the data well and provided the expected number of low scores.

The third concern reported on in the present work was the methodology used to correct the SS according to a set of predictors obtained in separate linear regression models. The model where SS were regressed on age, sex and education as predictors in one regression model proved to be almost as reliable as the model using raw scores as the outcome, but computationally more complicated. However, the model used to correct SS using the coefficients of age, sex and education from separate regressions was found to identify a lower number of individuals as showing a low score, with only fair to good agreement with the model including all the predictors in one regression. These results suggest that converting raw scores of highly skewed scores to SS through percentile ranges and with coefficients for predictors taken from different regressions is likely to bias the normative data towards higher SS, rendering the ability of the normative data to identify impairment more difficult. To avoid these issues, we recommend checking carefully for the regression assumptions and also use the most suited method for analysis. For example, if linearity does not hold, predictors can be transformed in order to find an association between predictors and outcome other than linear, as curvilinear relationships have been reported in previous studies providing normative data for the FCSRT when the quadratic effect of age and education has been added into the model (Calderón-Rubio et al., 2021; Iñesta et al., 2021, 2022). Additionally, a logistic model for discrete outcome variables could be applied as it has been shown to perform well compared to

the linear model even in small samples, since the scores are discrete and bounded, and such type of response variable can be modeled using a binomial distribution (De Andrade Moral et al., 2022).

Finally, our simulation studies are not without limitations. Firstly, it is very common for discrete proportion data (discrete scores bounded between 0 and a maximum) to present variability that is either lower or greater than the expected by the binomial model (phenomena referred to as under- and over-dispersion, respectively). There are extended models that can accommodate such features (Demétrio et al., 2014), and future studies would benefit from taking these into account to better understand how under- or over-dispersion affects the results. Second, although in the present study this did not seem to be an issue, there are ways to accommodate heterogeneity of variances within the LM framework. One such way is to allow the variance to be modeled with predictors, within a distributional regression framework (Klein, 2024). Third, the focus of this work was to analyze the goodness of fit of linear regression models for different response variables, and thus we used a group of cognitively healthy individuals to simulate data. This implies that there is no external validator variable (e.g., disability) that might provide evidence of the superiority of one model over the other, which warrants further research in future studies.

Constraints on Generality

We used education, sex, and age to establish the "true" relationships in the simulated population. However, there could be many other, omitted predictors that can bear influence in the observed scores. Future studies considering the inclusion of other predictors, or even a latent variable approach, would be useful to understand how the methodologies explored here perform when used to identify individuals that show reliable decline. Relatedly, although out of the scope of this work, we are aware that modifying the number of covariates might have an impact on the results reported here. We included in our models the most commonly demographic variables used to predict scores on neuropsychological tests, but their specific influence on different cognitive variables differs. For example, Peña-Casanova et al. (2009) showed that sex

had almost no effect on the FCSRT-DR score, whereas age had a very low effect on the Boston Naming Test (Peña-Casanova, Quinones-Ubeda, et al., 2009). Future works will unravel whether modifying the number of covariates has any effect on the regression model used to develop normative data.

In summary, the present work has shown that when RBND are to be calculated, the best approach seems to be using raw scores as observed response variables and check the regression assumptions to make sure that the model fits the data well. Conversely, using transformed scores corrected using coefficients from different regressions is associated with an upward bias and a marked decrease in the number of individuals scoring in the expected lower tail of the distribution, rendering that methodology prone to diagnostic errors. In case that the model does not fit the data well, there are several methods in the Generalized Linear Model that allow researchers to analyze the association between a set of predictors and an observed response variable that has a non-normal distribution (Akram et al., 2023).

References

Akram, M., Cerin, E., Lamb, K. E., & White, S. R. (2023). Modelling count, bounded and skewed continuous outcomes in physical activity research: Beyond linear regression models. *International Journal of Behavioral Nutrition and Physical Activity*, *20*(1), 57. https://doi.org/10.1186/s12966-023-01460-y

Alviarez-Schulze, V., Cattaneo, G., Pachón-García, C., Solana-Sánchez, J., Tormos, J. M., Pascual-Leone, A., & Bartrés-Faz, D. (2022). Validation and normative data of the Spanish version of the Rey Auditory Verbal Learning Test and associated long-term forgetting measures in middle-aged adults. *Frontiers in Aging Neuroscience*, *14*. https://www.frontiersin.org/articles/10.3389/fnagi.2022.809019

Arango-Lasprilla, J. C., Rivera, D., Ramos-Usuga, D., Vergara-Moragues, E., Montero-López, E., Adana Díaz, L. A., Aguayo Arelis, A., García-Guerrero, C. E., García de la Cadena, C., Llerena Espezúa, X., Lara, L., Padilla-López, A., Rodriguez-Irizarry, W., Alcazar Tebar, C., Irías Escher, M. J., Llibre Guerra, J. J., Torales Cabrera, N., Rodríguez-Agudelo, Y., & Ferrer-Cascales, R. (2017). Trail Making Test: Normative data for the Latin American Spanish-speaking pediatric population. *NeuroRehabilitation*, *41*(3), 627-637. https://doi.org/10.3233/NRE-172247

Austin, P. C., & Steyerberg, E. W. (2015). The number of subjects per variable required in linear regression analyses. *Journal of Clinical Epidemiology*, *68*(6), 627-636. https://doi.org/10.1016/j.jclinepi.2014.12.014

Bonete-López, B., Oltra-Cucarella, J., Marín, M., Antón, C., Balao, N., López, E., & Macià, E. S. (2021). Validation and norms for a recognition task for the spanish version of the free and cued selective reminding test. *Archives of Clinical Neuropsychology*, *36*(6), 954-964. https://doi.org/10.1093/arclin/acaa117

Calderón-Rubio, E., Oltra-Cucarella, J., Bonete-López, B., Iñesta, C., & Sitges-Maciá, E. (2021). Regression-based normative data for independent and cognitively active Spanish older

adults: Free and Cued Selective Reminding Test, Rey-Osterrieth Complex Figure test and Judgement of Line Orientation. *International Journal of Environmental Research and Public Health*, *18*(24), Article 24. https://doi.org/10.3390/ijerph182412977

Campo, P., & Morales, M. (2004). Normative data and reliability for a Spanish version of the verbal Selective Reminding Test. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *19*(3), 421-435. https://doi.org/10.1016/S0887-6177(03)00075-1

Campos-Magdaleno, M., Nieto-Vieites, A., Frades-Payo, B., Montenegro-Peña, M., Facal, D., Lojo-Seoane, C., & Delgado-Losada, M. L. (2024). Normative data for the Spanish versions of the CVLT, WMS-Logical Memory, and RBMT from a sample of middle-aged and old participants. *Psychological Assessment*, *36*(2), 114-123. https://doi.org/10.1037/pas0001292

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Crawford, J. R., & Garthwaite, P. H. (2007). Using regression equations built from summary data in the neuropsychological assessment of the individual case. *Neuropsychology*, *21*(5), 611-620. https://doi.org/10.1037/0894-4105.21.5.611

Crawford, J. R., & Garthwaite, P. H. (2012). Single-case research in neuropsychology: A comparison of five forms of t-test for comparing a case to controls. *Cortex*, *48*(8), 1009-1016. https://doi.org/10.1016/j.cortex.2011.06.021

Crawford, J. R., Garthwaite, P. H., Denham, A. K., & Chelune, G. J. (2012). Using regression equations built from summary data in the psychological assessment of the individual case: Extension to multiple regression. *Psychological Assessment*, *24*(4), 801-814. https://doi.org/10.1037/a0027699

Crawford, J. R., Garthwaite, P. H., & Slick, D. J. (2009). On percentile norms in neuropsychology: Proposed reporting standards and methods for quantifying the

uncertainty over the percentile ranks of test scores. *The Clinical Neuropsychologist*, *23*(7), 1173-1195. https://doi.org/10.1080/13854040902795018

De Andrade Moral, R., Díaz-Orueta, U., & Oltra-Cucarella, J. (2022). Logistic versus linear regression-based reliable change index: A simulation study with implications for clinical studies with different sample sizes. *Psychological Assessment*, *34*(8), 731-741. https://doi.org/10.1037/pas0001138

delCacho-Tena, A., Christ, B. R., Arango-Lasprilla, J. C., Perrin, P. B., Rivera, D., & Olabarrieta-Landa, L. (2023). Normative Data Estimation in Neuropsychological Tests: A Systematic Review. *Archives of Clinical Neuropsychology*, acad084. https://doi.org/10.1093/arclin/acad084

Delgado-Losada, M. L., López-Higes, R., Rubio-Valdehita, S., Facal, D., Lojo-Seoane, C., Montenegro-Peña, M., Frades-Payo, B., & Fernández-Blázquez, M. A. (2021). Spanish Consortium for Ageing Normative Data (SCAND): Screening tests (MMSE, GDS-15 and MFE). *Psicothema*, *33*(1), 70-76. https://doi.org/10.7334/psicothema2020.304

Demétrio, C. G. B., Hinde, J., & Moral, R. A. (2014). Models for overdispersed data in entomology. En C. P. Ferreira & W. A. C. Godoy (Eds.), *Ecological Modelling Applied to Entomology* (pp. 219-259). Springer International Publishing. https://doi.org/10.1007/978-3-319-06877-0_9

Dubois, B., Feldman, H. H., Jacova, C., Hampel, H., Molinuevo, J. L., Blennow, K., DeKosky, S. T., Gauthier, S., Selkoe, D., Bateman, R., Cappa, S., Crutch, S., Engelborghs, S., Frisoni, G. B., Fox, N. C., Galasko, D., Habert, M.-O., Jicha, G. A., Nordberg, A., … Cummings, J. L. (2014). Advancing research diagnostic criteria for Alzheimer's disease: The IWG-2 criteria. *The Lancet Neurology*, *13*(6), 614-629. https://doi.org/10.1016/S1474-4422(14)70090-0

Duff, K., Hammers, D. B., Dalley, B. C. A., Suhrie, K. R., Atkinson, T. J., Rasmussen, K. M., Horn, K. P., Beardmore, B. E., Burrell, L. D., Foster, N. L., & Hoffman, J. M. (2017). Short-term

practice effects and amyloid deposition: Providing information above and beyond baseline cognition. *The Journal of Prevention of Alzheimer's Disease*, *4*(2), 87-92. https://doi.org/10.14283/jpad.2017.9

Ehrenreich, J. H. (1995). Normative data for adults on a short form of the Selective Reminding Test. *Psychological Reports*, *76*(2), 387-390. https://doi.org/10.2466/pr0.1995.76.2.387

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed). J. Wiley.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). «Mini-mental state». A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, *12*(3), 189-198. https://doi.org/10.1016/0022-3956(75)90026-6

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.84.1. https://CRAN.R-project.org/package=irr

García-Herranz, S., Díaz-Mardomingo, M. D. C., Suárez-Falcón, J. C., Rodríguez-Fernández, R., Peraita, H., & Venero, C. (2022). Normative data for the Spanish version of the California Verbal Learning Test (TAVEC) from older adults. *Psychological Assessment*, *34*(1), 91-97. https://doi.org/10.1037/pas0001070

Girtler, N., De Carli, F., Amore, M., Arnaldi, D., Bosia, L. E., Bruzzaniti, C., Cappa, S. F., Cocito, L., Colazzo, G., Ghio, L., Magi, E., Mancardi, G. L., Nobili, F., Pardini, M., Picco, A., Rissotto, R., Serrati, C., & Brugnolo, A. (2015). A normative study of the Italian printed word version of the free and cued selective reminding test. *Neurological Sciences*, *36*(7), 1127-1134. https://doi.org/10.1007/s10072-015-2237-7

Grau-Guinea, L., Pérez Enríquez, C., García-Escobar, G., Arrondo Elizarán, C., Pereira Cutiño, B., Florido Santiago, M., Piqué Candini, J., Planas, A., Paez, M., Peña Casanova, J., & Sánchez-Benavides, G. (2020). Development, equivalence study, and normative data of

version B of the Spanish-language Free and Cued Selective Reminding Test. *Neurología (English Edition)*, S2173580820300043. https://doi.org/10.1016/j.nrleng.2018.02.001

Guàrdia-Olmos, J., Peró-Cebollero, M., Rivera, D., & Arango-Lasprilla, J. C. (2015). Methodology for the development of normative data for ten Spanish-language neuropsychological tests in eleven Latin American countries. *NeuroRehabilitation*, *37*(4), 493-499. https://doi.org/10.3233/NRE-151277

Harrington, K. D., Lim, Y. Y., Ames, D., Hassenstab, J., Rainey-Smith, S., Robertson, J., Salvado, O., Masters, C. L., & Maruff, P. (2017). Using robust normative data to investigate the neuropsychology of cognitive aging. *Archives of Clinical Neuropsychology*, *32*(2), 142-154. https://doi.org/10.1093/arclin/acw106

Iñesta, C., Oltra-Cucarella, J., Bonete-López, B., Calderón-Rubio, E., & Sitges-Maciá, E. (2021). Regression-based normative data for independent and cognitively active Spanish older adults: Digit Span, Letters and Numbers, Trail Making Test and Symbol Digit Modalities Test. *International Journal of Environmental Research and Public Health*, *18*(19), 9958. https://doi.org/10.3390/ijerph18199958

Iñesta, C., Oltra-Cucarella, J., & Sitges-Maciá, E. (2022). Regression-based normative data for independent and cognitively active Spanish older adults: Verbal fluency tests and Boston Naming Test. *International Journal of Environmental Research and Public Health*, *19*(18), 11445. https://doi.org/10.3390/ijerph191811445

Ivnik, R. J., Malec, J. F., Smith, G. E., Tangalos, E. G., & Petersen, R. C. (1996). Neuropsychological tests' norms above age 55: COWAT, BNT, MAE token, WRAT-R reading, AMNART, STROOP, TMT, and JLO. *The Clinical Neuropsychologist*, *10*(3), 262-278. https://doi.org/10.1080/13854049608406689

Ivnik, R. J., Malec, J. F., Smith, G. E., Tangalos, E. G., Petersen, R. C., Kokmen, E., & Kurland, L. T. (1992a). Mayo's older americans normative studies: Updated AVLT norms for ages 56

to 97. *Clinical Neuropsychologist*, *6*(sup001), 83-104. https://doi.org/10.1080/13854049208401880

Ivnik, R. J., Malec, J. F., Smith, G. E., Tangalos, E. G., Petersen, R. C., Kokmen, E., & Kurland, L. T. (1992b). Mayo's older americans normative studies: WAIS-R norms for ages 56 to 97. *Clinical Neuropsychologist*, *6*(sup001), 1-30. https://doi.org/10.1080/13854049208401877

Ivnik, R. J., Smith, G. E., Lucas, J. A., Tangalos, E. G., Kokmen, E., & Petersen, R. C. (1997). Free and cued selective reminding test: MOANS norms. *Journal of Clinical and Experimental Neuropsychology*, *19*(5), 676-691. https://doi.org/10.1080/01688639708403753

Karstens, A. J., Christianson, T. J., Lundt, E. S., Machulda, M. M., Mielke, M. M., Fields, J. A., Kremers, W. K., Graff-Radford, J., Vemuri, P., Jack, C. R., Knopman, D. S., Petersen, R. C., & Stricker, N. H. (2024). Mayo normative studies: Regression-based normative data for ages 30-91 years with a focus on the Boston Naming Test, Trail Making Test and Category Fluency. *Journal of the International Neuropsychological Society: JINS*, *30*(4), 389-401. https://doi.org/10.1017/S1355617723000760

Kéry, M., & Hatfield, J. S. (2003). Normality of Raw Data in General Linear Models: The Most Widespread Myth in Statistics. *Bulletin of the Ecological Society of America*, *84*(2), 92-94. https://doi.org/10.1890/0012-9623(2003)84[92:NORDIG]2.0.CO;2

Kiselica, A. M., Kaser, A. N., Webber, T. A., Small, B. J., & Benge, J. F. (2020). Development and preliminary validation of standardized regression-based change scores as measures of transitional cognitive decline. *Archives of Clinical Neuropsychology*, *35*(7), 1168-1181. https://doi.org/10.1093/arclin/acaa042

Klein, N. (2024). Distributional regression for data analysis. *Annual Review of Statistics and Its Application*, *11*(1), 321-346. https://doi.org/10.1146/annurev-statistics-040722-053607

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Larrabee, G. J., Trahan, D. E., & Levin, H. S. (2000). Normative data for a six-trial administration of the Verbal Selective Reminding Test. *The Clinical Neuropsychologist*, *14*(1), 110-118. https://doi.org/10.1076/1385-4046(200002)14:1;1-8;FT110

Lucas, J. A., Ivnik, R. J., Smith, G. E., Ferman, T. J., Willis, F. B., Petersen, R. C., & Graff-Radford, N. R. (2005). Mayo's Older African Americans Normative Studies: Norms for Boston Naming Test, Controlled Oral Word Association, category fluency, animal naming, Token Test, WRAT-3 Reading, Trail Making Test, Stroop Test, and Judgment of Line Orientation. *The Clinical Neuropsychologist*, *19*(2), 243-269. https://doi.org/10.1080/13854040590945337

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., Klunk, W. E., Koroshetz, W. J., Manly, J. J., Mayeux, R., Mohs, R. C., Morris, J. C., Rossor, M. N., Scheltens, P., Carrillo, M. C., Thies, B., Weintraub, S., & Phelps, C. H. (2011). The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 263-269. https://doi.org/10.1016/j.jalz.2011.03.005

Morlett Paredes, A., Tarraf, W., Gonzalez, K., Stickel, A. M., Graves, L. V., Salmon, D. P., Kaur, S. S., Gallo, L. C., Isasi, C. R., Lipton, R. B., Lamar, M., Goodman, Z. T., & González, H. M. (2024). Normative data for the Digit Symbol Substitution for diverse Hispanic/Latino adults: Results from the Study of Latinos-Investigation of Neurocognitive Aging (SOL-INCA). *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *16*(2), e12573. https://doi.org/10.1002/dad2.12573

Mungas, D., Marshall, S. C., Weldon, M., Haan, M., & Reed, B. R. (1996). Age and education correction of Mini-Mental State Examination for English and Spanish-speaking elderly. *Neurology*, *46*(3), 700-706. https://doi.org/10.1212/wnl.46.3.700

Oltra-Cucarella, J. (2025). Research files. *Universidad Miguel Hernández de Elche*. https://sabiex.umh.es/lineas-de-investigacion/neuropsicologia-y-envejecimiento-files/

Oltra-Cucarella, J., Sánchez-SanSegundo, M., & Ferrer-Cascales, R. (2022). Predicting Alzheimer's disease with practice effects, APOE genotype and brain metabolism. *Neurobiology of Aging*, *112*, 111-121. https://doi.org/10.1016/j.neurobiolaging.2021.12.011

Pek, J., Wong, O., & Wong, C. M. (2017). Data transformations for inference with linear regression: Clarifications and recommendations. *Practical Assessment, Research & Evaluation*, *22*(1), 1-11. https://doi.org/doi.org/10.7275/2w3n-0f07

Peña-Casanova, J., Blesa, R., Aguilar, M., Gramunt-Fombuena, N., Gomez-Anson, B., Oliva, R., Molinuevo, J. L., Robles, A., Barquero, M. S., Antunez, C., Martinez-Parra, C., Frank-Garcia, A., Fernandez, M., Alfonso, V., Sol, J. M., & for the NEURONORMA Study Team. (2009). Spanish Multicenter Normative Studies (NEURONORMA Project): Methods and sample characteristics. *Archives of Clinical Neuropsychology*, *24*(4), 307-319. https://doi.org/10.1093/arclin/acp027

Peña-Casanova, J., Gramunt-Fombuena, N., Quiñones-Úbeda, S., Sánchez-Benavides, G., Aguilar, M., Badenes, D., Molinuevo, J. L., Robles, A., Barquero, M. S., Payno, M., Antúnez, C., Martínez-Parra, C., Frank-García, A., Fernández, M., Alfonso, V., Sol, J. M., & Blesa, R. (2009). Spanish multicenter normative studies (NEURONORMA project): Norms for the Rey-Osterrieth complex figure (copy and memory), and Free and Cued Selective Reminding Test. *Archives of Clinical Neuropsychology*, *24*(4), 371-393. https://doi.org/10.1093/arclin/acp041

Peña-Casanova, J., Quinones-Ubeda, S., Gramunt-Fombuena, N., Aguilar, M., Casas, L., Molinuevo, J. L., Robles, A., Rodriguez, D., Barquero, M. S., Antunez, C., Martinez-Parra, C., Frank-Garcia, A., Fernandez, M., Molano, A., Alfonso, V., Sol, J. M., Blesa, R., & for the NEURONORMA Study Team. (2009). Spanish Multicenter Normative Studies (NEURONORMA Project): Norms for Boston Naming Test and Token Test. *Archives of Clinical Neuropsychology*, *24*(4), 343-354. https://doi.org/10.1093/arclin/acp039

Petersen, R. C. (2004). Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine*, *256*(3), 183-194. https://doi.org/10.1111/j.1365-2796.2004.01388.x

Rivera, D., & Arango-Lasprilla, J. C. (2017). Methodology for the development of normative data for Spanish-speaking pediatric populations. *NeuroRehabilitation*, *41*(3), 581-592. https://doi.org/10.3233/NRE-172275

Schmidt, A. F., & Finan, C. (2018). Linear regression and the normality assumption. *Journal of Clinical Epidemiology*, *98*, 146-151. https://doi.org/10.1016/j.jclinepi.2017.12.006

Shirk, S. D., Mitchell, M. B., Shaughnessy, L. W., Sherman, J. C., Locascio, J. J., Weintraub, S., & Atri, A. (2011). A web-based normative calculator for the uniform data set (UDS) neuropsychological test battery. *Alzheimer's Research & Therapy*, *3*(6), 32. https://doi.org/10.1186/alzrt94

Steinberg, B. A., Bieliauskas, L. A., Smith, G. E., Langellotti, C., & Ivnik, R. J. (2005). Mayo's Older Americans Normative Studies: Age- and IQ-adjusted norms for the Boston Naming Test, the MAE Token test, and the Judgment of Line Orientation test. *The Clinical Neuropsychologist*, *19*(3-4), 280-328. https://doi.org/10.1080/13854040590945229

Strauss, E., Sherman, E. M. S., Spreen, O., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary* (3rd ed). Oxford University Press.

Stricker, N. H., Christianson, T. J., Lundt, E. S., Alden, E. C., Machulda, M. M., Fields, J. A., Kremers, W. K., Jack, C. R., Knopman, D. S., Mielke, M. M., & Petersen, R. C. (2021). Mayo Normative Studies: Regression-based normative data for the Auditory Verbal Learning Test for ages 30–91 years and the importance of adjusting for sex. *Journal of the International Neuropsychological Society*, *27*(3), 211-226. https://doi.org/10.1017/S1355617720000752

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics.* (Sixth Ed.). Pearson Education Inc.

Uttl, B. (2005). Measurement of individual differences: Lessons from memory assessment in research and clinical practice. *Psychological Science*, *16*(6), 460-467. https://doi.org/10.1111/j.0956-7976.2005.01557.x

Weisberg, S. (2014). *Applied Linear Regression* (4th ed.). Wiley.

Williams, M. N., Gómez Grajales, C. A., & Kurkiewicz, D. (2013). Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research & Evaluation*, *18*(11), 1-14. https://doi.org/10.7275/55hn-wk47

Williamson, M., Maruff, P., Schembri, A., Cummins, H., Bird, L., Rosenich, E., & Lim, Y. Y. (2022). Validation of a digit symbol substitution test for use in supervised and unsupervised assessment in mild Alzheimer's disease. *Journal of Clinical and Experimental Neuropsychology*, *44*(10), 768-779. https://doi.org/10.1080/13803395.2023.2179977

Winblad, B., Palmer, K., Kivipelto, M., Jelic, V., Fratiglioni, L., Wahlund, L.-O., Nordberg, A., Backman, L., Albert, M., Almkvist, O., Arai, H., Basun, H., Blennow, K., de Leon, M., DeCarli, C., Erkinjuntti, T., Giacobini, E., Graff, C., Hardy, J., … Petersen, R. C. (2004). Mild cognitive impairment - beyond controversies, towards a consensus: Report of the International Working Group on Mild Cognitive Impairment. *Journal of Internal Medicine*, *256*(3), 240-246. https://doi.org/10.1111/j.1365-2796.2004.01380.x

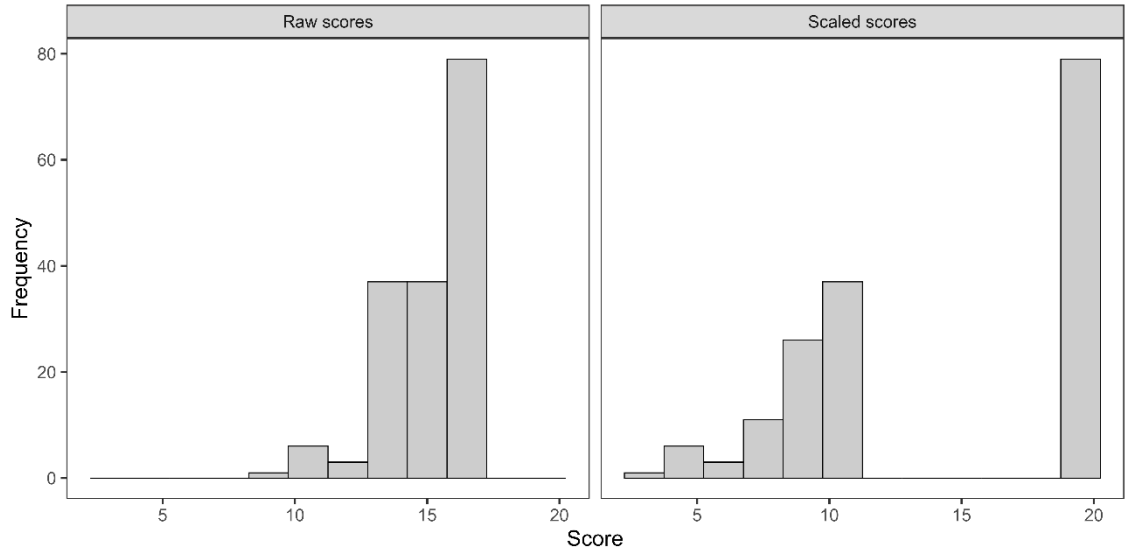

Figure 1. Histograms for FCSRT – Delayed recall raw scores and scaled scores

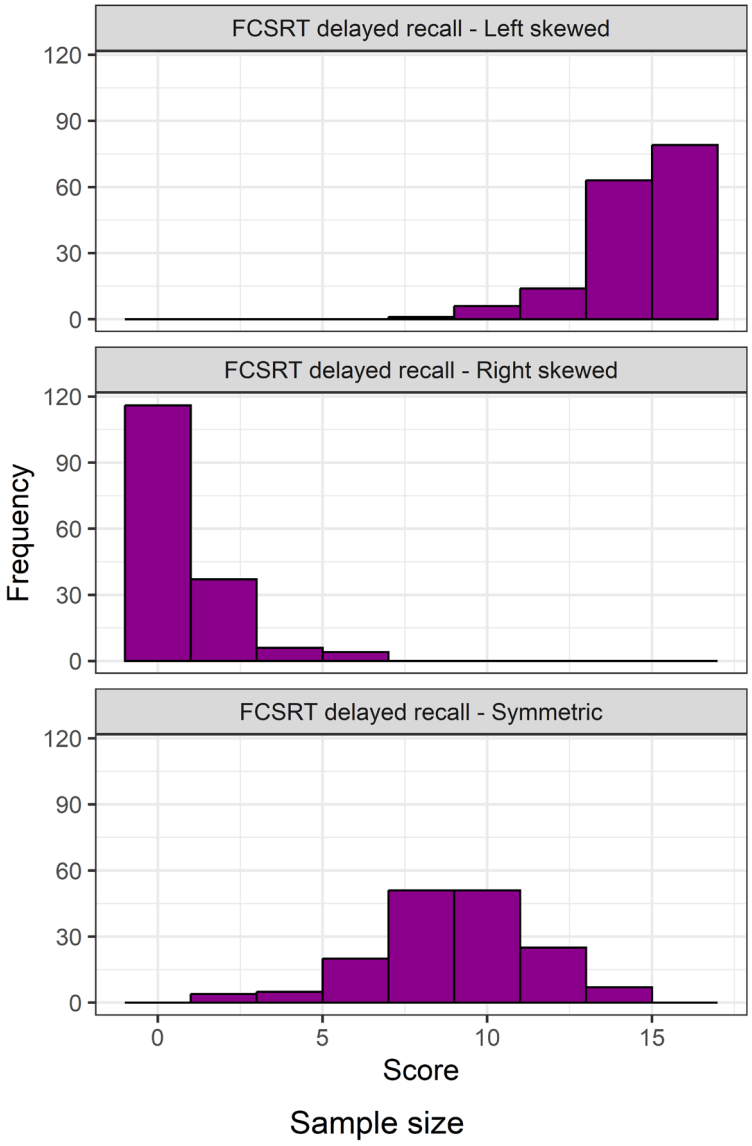

Figure 2. Histograms for simulated FCSRT – Delayed recall raw scores

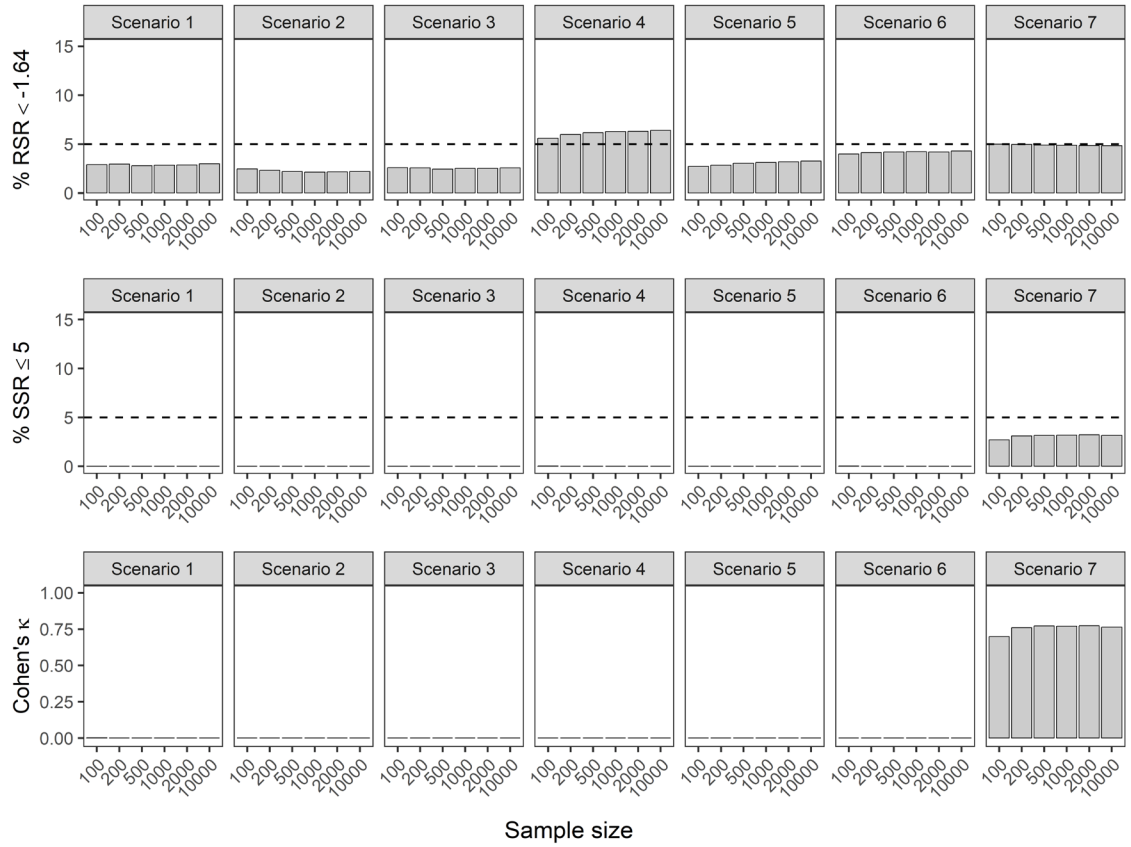

Figure 3. Percentage of low scores for each of the seven scenarios by type of regression.

RSR: raw score regression. SSR: scaled score regression. Scenarios 1-3: symmetric, left- and right-skewed score distributions based on a normal distribution. Scenarios 4-6: symmetric, left- and right-skewed score distributions based on a binomial distribution. Scenario 7: a purely random symmetric score distribution unrelated to the covariates, originating from a normal distribution with a mean of 10 and standard deviation of 3