Wim Van der Elst

Regression-Based Normative Data for Psychological Assessment

A Hands-On Approach Using R

MOREMEDIA

Springer

Regression-Based Normative Data for Psychological Assessment

Wim Van der Elst

Regression-Based Normative Data for Psychological Assessment

A Hands-On Approach Using R

Springer

Wim Van der Elst
Statistics and Decision Sciences
Johnson & Johnson
Leuven, Belgium

This book contains several screenshots of the Google Sheets software. Google Sheets is a trademark of Google LLC, and this book is not endorsed by or affiliated with Google in any way.

ISBN 978-3-031-50950-6 ISBN 978-3-031-50951-3 (eBook)
https://doi.org/10.1007/978-3-031-50951-3

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

Contents

Abbreviations

AIC	Akaike's information criterion
AUC	Area under the curve
CDF	Cumulative distribution function
CI	Confidence interval
DF	Degrees of freedom
EMF	Empirical mean function
EVF	Empirical variance function
GLT	General linear test
H_0	The null hypothesis
H_1	The alternative hypothesis
i.i.d.	Identically and independently distributed
LE	Level of education
OLS	Ordinary least squares
PDF	Probability density function
SD	Standard deviation
SSE	Error sum of squares
SSR	Regression sum of squares

Symbols

α	The Type I error rate
β	The Type II error rate
β_k	The k-th regression parameter of the model for the mean structure
χ^2_{BP*}	The Breusch-Pagan test-statistic
$\widehat{\delta_i}$	The standardized residuals in the normative sample
δ_i	The standardized errors in the normative sample
γ_k	The k-th regression parameter of the model for the residual structure
L^*	The Levene test-statistic
$\widehat{\pi}_0$	The estimated percentile rank for a raw test score Y_0
Φ	The cumulative distribution function of the standard normal distribution
Ψ	The cumulative distribution function of $\widehat{\delta_i}$ in the normative sample
R^2	Coefficient of determination
$\widehat{\sigma}_{\varepsilon}$	The (overall) residual standard error
$\widehat{\sigma}^2_{\varepsilon_i}$	The variance prediction function
$\widehat{\upsilon}_0$	A standardized test score
$\widehat{\varepsilon_i}$	The residuals in the normative sample
ε_i	The errors in the normative sample
W^*	The Shapiro-Wilk test-statistic
X	An independent variable
X_c	A centered independent variable
$\widehat{Y_i}$	The mean prediction function
Y_i	The raw test scores in the normative sample
Y_0	A raw test score to be normed

Chapter 1
General Introduction

1.1 Psychological Assessment and Normative Data

Psychological assessment refers to the process of collecting and interpreting information that relates to a tested person's latent characteristics (Coaley 2009). A latent characteristic cannot be directly observed but instead has to be inferred indirectly from the person's observable behavior. The observable behavior can be elicited using standardized testing procedures, rating scales, questionnaires, structured interviews, and so on. For example, a latent characteristic that cannot be directly observed is verbal memory. A large number of standardized testing procedures to assess verbal memory have been developed, of which Rey's Verbal Learning Test (VLT; Rey 1958) is one of the most commonly used procedures. In the VLT, 15 monosyllabic words are presented in 5 subsequent learning trials with a free recall procedure immediately following each trial. After a delay of approximately 20 minutes (and unexpectedly for the tested person), there is an additional free recall trial. The final part of the VLT consists of a recognition test, involving yes/no recognition of the 15 words intermixed with 15 non-target words. The observable behavior in the VLT consists of the words that are recalled or recognized by the test participants across the different trials. For example, the VLT Total Recall score corresponds to the total number of correctly recalled (non-repeated) words over the 5 learning trials. This test score is assumed to capture a tested person's latent overall verbal memory and learning abilities (Lezak 1995; Mitrushina et al. 2005; Van der Elst et al. 2005).

The Need for Normative Data
The majority of psychological testing procedures are norm-referenced (Coaley 2009; Mitrushina et al. 2005; Strauss et al. 2006; Van der Elst 2006). This means that a raw test score (e.g., the VLT Total Recall score) cannot be interpreted in a meaningful way *by itself*. For example, suppose that a patient who has memory complaints is referred to a neuropsychologist for cognitive testing. The

W. Van der Elst, *Regression-Based Normative Data for Psychological Assessment*,
https://doi.org/10.1007/978-3-031-50951-3_1

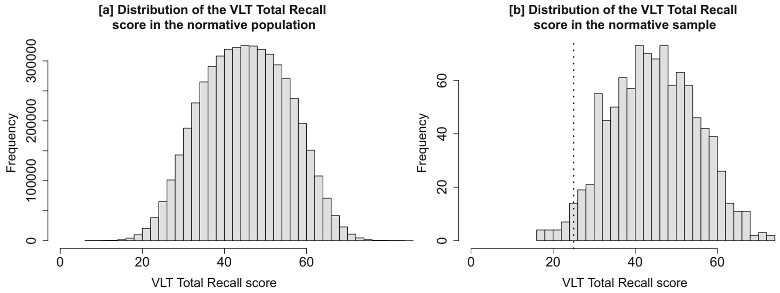

Fig. 1.1 Distributions of the VLT Total Recall scores in the normative population ($N = 5$ million; see panel (**a**)) and in the normative sample ($N = 1000$; see panel (**b**)). The vertical dotted line is the raw VLT Total Recall score = 25 that is obtained by a tested patient who has memory complaints

neuropsychologist administers the VLT to the patient, who obtains a raw VLT Total Recall score = 25. Based on this information alone, it cannot be determined whether the patient's test score is low, average, or high. To make the latter type of claim, a reference distribution of the raw VLT Total Recall score is needed such that the *relative position* of the raw test score in the population can be estimated (Capitani 2019; Coaley 2009; Mitrushina et al. 2005). A reference distribution for the test score can be obtained by administering the test at hand in a large normative sample. For example, suppose that it is of interest to establish normative data of the VLT Total Recall score for Dutch-speaking adults who live in Belgium. To this end, the VLT is administered in a large representative normative sample (e.g., $N = 1000$) of test participants who are randomly drawn from the normative population (here: all Dutch-speaking adults who live in Belgium).[1] By means of illustration, Fig. 1.1a shows a histogram of the distribution of the VLT Total Recall scores in the entire normative population (here: $N = 5$ million cognitively healthy Dutch-speaking adults who live in Belgium).[2] Figure 1.1b shows the distribution of the VLT Total Recall scores in a randomly drawn normative sample of $N = 1000$ test participants. The latter distribution provides an empirical frame of reference to determine the relative position of a VLT Total Recall score. For example, based on this reference distribution, it can be readily observed that the VLT Total Recall score = 25 that

[1] Often the normative sample is drawn from a *subset* of the normative population. For example, the normative population for a verbal memory test like the VLT will typically exclude people who have severe cognitive disorders (e.g., people who are diagnosed with Alzheimer's disease or frontotemporal dementia). The reason for this is that one is often primarily interested in obtaining reference data that reflect the population distribution of the test scores for "typical" test participants who have "normal" ability levels in a normative data context.

[2] These are simulated data. The population distribution of the test scores is never known in real-life normative studies, because it is not feasible to test the entire population. For some tests, a large proportion of the normative population can be tested (e.g., for standardized school tests), but even then the entire population distribution is not known due to missing data (e.g., children who were not tested due to illness).

was obtained by the tested patient is low. Indeed, only a small fraction of the cognitively healthy test participants in the normative sample obtained a VLT Total Recall score that is equal to or below 25 (see the area on the left of the vertical dotted line in Fig. 1.1b). Under the assumption that the normative sample is a representative sample from the normative population (which is the case here, as random sampling was used), it can be inferred that the same holds for the normative population.

The approach of determining the relative position of a raw test score through visual exploration of the reference distribution is inherently subjective and imprecise. Normative data are essentially used to objectify this endeavor, i.e., they allow for estimating the relative position of a raw test score in the normative population in a more quantitative and statistically principled way. Several methods have been developed to derive normative data, of which the traditional and the regression-based methods are the most commonly used.

1.2 The Traditional Normative Method

Traditional normative data simply consist of the mean and the standard deviation (SD) of the raw test scores in the normative sample. Based on these two summary statistics, the raw score of a tested person is converted into a standardized test score in the following way:

$$\widehat{\upsilon}_0 = \frac{Y_0 - \widehat{\mu}_{Y_i}}{\widehat{\sigma}_{Y_i}}, \tag{1.1}$$

where:

- $\widehat{\upsilon}_0$ is the standardized test score.
- Y_0 is the raw test score that is being standardized.
- $\widehat{\mu}_{Y_i} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ is the mean of the raw test scores in the normative sample.
- $\widehat{\sigma}_{Y_i} = \sqrt{\frac{\sum_{i=1}^{N}\left(Y_i - \widehat{\mu}_{Y_i}\right)^2}{N-1}}$ is the SD of the raw test scores in the normative sample.
- i is the subscript that refers to the test participants in the normative sample, with $i = \{1, 2, \ldots, N\}$ and N = the total number of test participants in the normative sample.

Observe that the hat-notation is used in the above expression to distinguish estimated from true values. For example, μ_Y is the true (i.e., population-level) mean test score, whereas $\widehat{\mu}_Y$ is an estimate of the true mean test score that is based on the normative sample.

The standardized test score $\widehat{\upsilon}_0$ is a metric of relative position that quantifies how many SD units the raw test score at hand ($= Y_0$) is below the mean in the normative sample (see the $\widehat{\sigma}_Y$ and $\widehat{\mu}_Y$ in the denominator and numerator of expression (1.1), respectively). For example, in the normative sample of the VLT Total Recall score (that is shown in Fig. 1.1b), $\widehat{\mu}_Y = 45.066054$ and $\widehat{\sigma}_Y = 10.033027$. In the traditional normative approach, these summary statistics are used to standardize the raw test scores. To illustrate this, consider again the raw VLT Total Recall score $Y_0 = 25$ that was obtained by the tested patient who has memory complaints. The standardized test score $\widehat{\upsilon}_0$ for the tested patient corresponds to (see expression (1.1)):

$$\widehat{\upsilon}_0 = \frac{Y_0 - \widehat{\mu}_{Y_i}}{\widehat{\sigma}_{Y_i}} = \frac{25 - 45.066054}{10.033027} = -2.$$

As can be seen, $\widehat{\upsilon}_0 = -2$, and thus the raw VLT Total Recall score $Y_0 = 25$ of the tested patient is 2 SD units below the mean of the VLT Total Recall test score in the normative sample. To further facilitate the interpretation of the standardized test scores, the obtained $\widehat{\upsilon}_0$-values are typically converted into *percentile ranks*. A percentile rank quantifies the percentage of test scores that are equal to or lower than the raw test score at hand. Under the assumption that the standardized test scores are normally (or Gaussian) distributed, it is straightforward to convert $\widehat{\upsilon}_0$ into a percentile rank $\widehat{\pi}_0$. Indeed, this can be done by computing the Area Under the Curve (AUC) between $-\infty$ and $\widehat{\upsilon}_0$ of the standard normal distribution (i.e., a normal distribution mean $= 0$ and $SD = 1$). This is illustrated in Fig. 1.2, which shows the standard normal distribution (see the solid line) and a number of standardized test scores $\widehat{\upsilon}_0$ with their corresponding AUC values and percentile ranks $\widehat{\pi}_0$ (see the vertical dashed lines). For example, the figure shows that the standardized VLT Total Recall score $\widehat{\upsilon}_0 = -2$ that was obtained by the tested patient corresponds to a percentile rank $\widehat{\pi}_0 = 2$ (see the gray shaded area in the figure). This percentile rank is obtained by computing the AUC between $-\infty$ and $\widehat{\upsilon}_0 = -2$ (which equals 0.02; for details, see Chap. 3), and multiply the obtained value by 100 to express it as a percentage.[3] It can thus be concluded that an estimated 2% of the people in the normative population have a standardized test score $\widehat{\upsilon}_0$ that is equal to or below -2 (or equivalently, that an estimated 2% of the people in the normative population have a raw VLT Total Recall score Y_0 that is equal to or below 25). The raw VLT Total Recall score $Y_0 = 25$ that was obtained by the patient who has memory complaints is thus poor, as an estimated 98% of the people in the normative population have a raw test score that is higher than 25.

[3] In Appendix A.1, a comprehensive table of $\widehat{\upsilon}_0$-scores and their corresponding $\widehat{\pi}_0$ is provided.

1.3 Issues with the Traditional Normative Method

As was illustrated in the previous section, the traditional normative approach is straightforward and simply consists of computing the mean and the SD of the raw test scores in the normative sample. These summary statistics are then used to standardize the raw test score Y_0 of interest (see expression (1.1)), and the obtained $\widehat{\upsilon}_0$-value is subsequently converted into an easy-to-interpret percentile rank $\widehat{\pi}_0$ based on the standard normal distribution (see Fig. 1.2 and Appendix A.1).

In many real-life normative analyses, the raw test score at hand is impacted by independent variables such as Age, Gender, Level of Education, Ethnicity, and so on (Lezak 1995; Mitrushina et al. 2005; Strauss et al. 2006). For example, the "normal" VLT Total Recall score of an 80-year-old male is quite different from the "normal" test score of a 25-year-old female. Indeed, being older and being male adversely impact the VLT Total Recall scores (Mitrushina et al. 2005; Van der Elst et al. 2005). To compare apples with apples, the impact of such independent variables on the test score at hand should be properly accounted for in the normative data. In the traditional normative approach, this is done by splitting the normative sample into subgroups (Mitrushina et al. 2005; Van Breukelen & Vlaeyen 2005; Van der Elst et al. 2006). Subgroup-specific means and SDs are then used in expression (1.1) to standardize the raw test score Y_0. To illustrate this, consider again the normative sample of the VLT Total Recall score that was shown earlier in Fig. 1.1b. This normative sample included a total of $N = 1000$ test participants, of whom $N = 484$

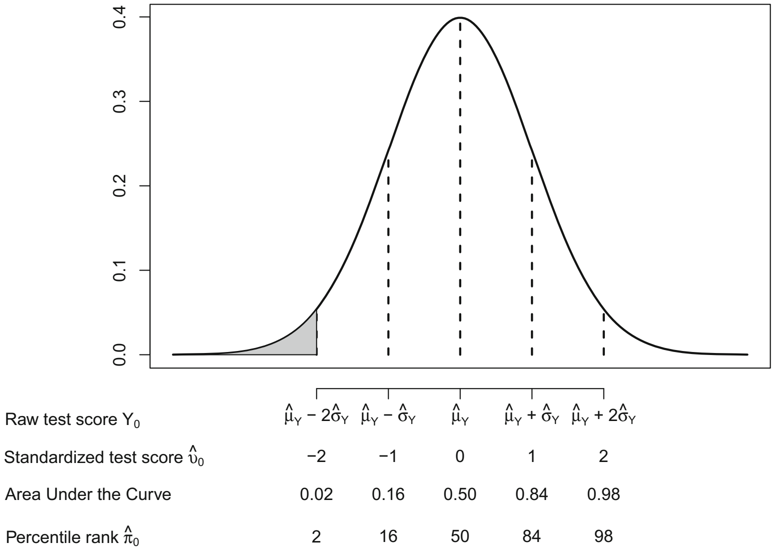

Fig. 1.2 Density of the standard normal distribution, with raw test scores Y_0 and their corresponding standardized test scores $\widehat{\upsilon}_0$, AUC values, and percentile ranks $\widehat{\pi}_0$. The gray shaded area is the AUC between $-\infty$ and $\widehat{\upsilon}_0 = -2$ that corresponds with a percentile rank $\widehat{\pi}_0 = 2$

were females and $N = 516$ were males. Figure 1.3 shows the Gender-specific distributions of the VLT Total Recall scores in the normative sample. It is well-known that females outperform males on verbal learning tests (Lezak 1995; Schmidt 1996; Van der Elst et al. 2005), and the same holds in the example VLT Total Recall normative sample. Indeed, the mean (SD) VLT Total Recall scores for females and males equal 47.247900 (9.594400) and 43.019400 (10.269400), respectively. The mean VLT Total Recall score of females is thus substantially higher than the mean score of males (i.e., 47.247900 versus 43.019400, respectively). To account for the impact of Gender on the VLT Total Recall scores, subgroup-specific summary statistics are used to conduct the normative conversions in the traditional approach.[4] For example, suppose that the tested patient who obtained a raw VLT Total Recall score $Y_0 = 25$ is a female. Her $\widehat{\upsilon}_0$-score would equal:

$$\widehat{\upsilon}_0 = \frac{25 - 47.247900}{9.594400} = -2.318842,$$

with corresponding percentile rank $\widehat{\pi}_0 = 1$ (see Appendix A.1). On the other hand, when the tested patient would have been a male, his $\widehat{\upsilon}_0$-score would equal:

$$\widehat{\upsilon}_0 = \frac{25 - 43.019400}{10.269400} = -1.754670,$$

with corresponding $\widehat{\pi}_0 = 4$ (see Appendix A.1). The same raw VLT Total Recall score $Y_0 = 25$ thus corresponds to a different percentile rank $\widehat{\pi}_0$ for females and males because a Gender-specific reference distribution is used to estimate the relative position of Y_0 in the normative population (see Fig. 1.3).

As can be seen, the traditional normative approach accounts for the impact of an independent variable on the test score at hand in a very straightforward way, i.e., by splitting the normative sample into subgroups. Unfortunately, this simple approach has some fundamental problems.

1.3.1 The Boundary Problem

A first problem with the traditional normative approach is that it cannot properly account for *quantitative* independent variables. A quantitative independent variable can take many possible outcome values that have a true numeric interpretation. For example, Age is a quantitative independent variable because it can take many possible outcome values (e.g., a tested person in the normative sample can be aged

[4] In a real-life normative analysis, a formal statistical test is typically conducted to decide whether an independent variable (like Gender) should be accounted for in the normative data (see subsequent chapters). It is assumed here that Gender has a statistically significant impact on the mean VLT Total Recall score.

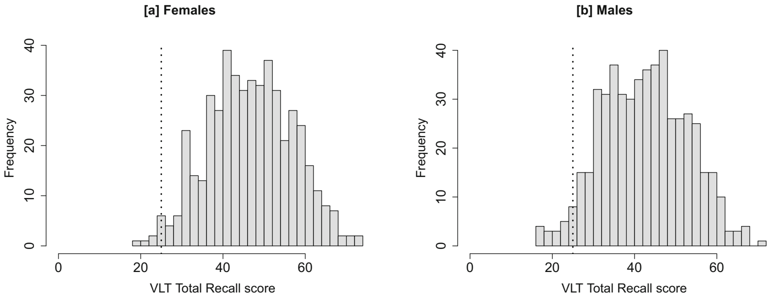

Fig. 1.3 Distributions of the raw VLT Total Recall scores for females ($N = 484$; see panel (**a**)) and males ($N = 516$; see panel (**b**)) in the VLT Total Recall normative sample

20.00, 20.01, ..., 80.00 years) that have a true numeric interpretation (e.g., we can say that a tested person who is aged 40.00 years is twice as old as a tested person who is aged 20.00 years). As was described above, the traditional normative approach accounts for the impact of an independent variable on the test score at hand by splitting the normative sample into subgroups. For example, when Gender has to be accounted for in the normative data, female and male subgroups are used. Such an approach is evidently not possible for quantitative independent variables because such variables have a large number of possible outcome values by definition. To illustrate this, consider again the example normative sample of the VLT Total Recall score that was already shown earlier in Fig. 1.1b. The test participants in this normative sample were aged between 20.00 and 80.00 years. Figure 1.4a shows a scatterplot of the VLT Total Recall test scores (on the Y-axis) against Age in years (on the X-axis) in the normative sample. As can be seen, Age clearly has a strong negative impact on the VLT Total Recall score, and thus Age-corrected normative data should be provided (it is assumed here that Age has a significant impact on the test score, see subsequent chapters). There are however a total of 927 unique Age values in the normative sample. It obviously makes no sense to split the normative sample for all possible outcome values of Age (as was done for Gender) because most of the obtained subgroups would consist of only one test participant.

In the traditional normative approach, this issue is dealt with by *discretizing* the quantitative independent variable. For example, in the VLT Total Recall normative sample, the quantitative independent variable Age could be discretized into 6 subgroups that each have a span of approximately 10 years, i.e., as (20.00, 30.00], (30.00, 40.00], ... and (70.00, 80.00] years.[5] The Age subgroup-specific means and SDs are then used in expression (1.1) to perform the normative conversions. There are however several problems with this method:

[5] In the notation that is used for the Age subgroup intervals, a round bracket (i.e., the "("-symbol) indicates that the specified value is not included in the interval, whereas a square bracket (i.e., the "]"-symbol) indicates that the specified value is included in the interval. For example, the Age subgroup (20.00, 30.00] contains test participants who are aged >20.00 years and ≤30.00 years.

- It is implicitly assumed that the Age subgroup-specific means and SDs follow a step-function, i.e., that the means and SDs within an Age subgroup are identical. This is visually illustrated in Fig. 1.4b, which shows the subgroup-specific means for the VLT Total Recall scores (see the horizontal lines) in the different Age subgroups (indicated by the vertical dashed lines). Such an assumption is clearly unrealistic, i.e., the overall pattern in the normative sample suggests that the relation between Age and the (mean) VLT Total Recall score is of a more smooth

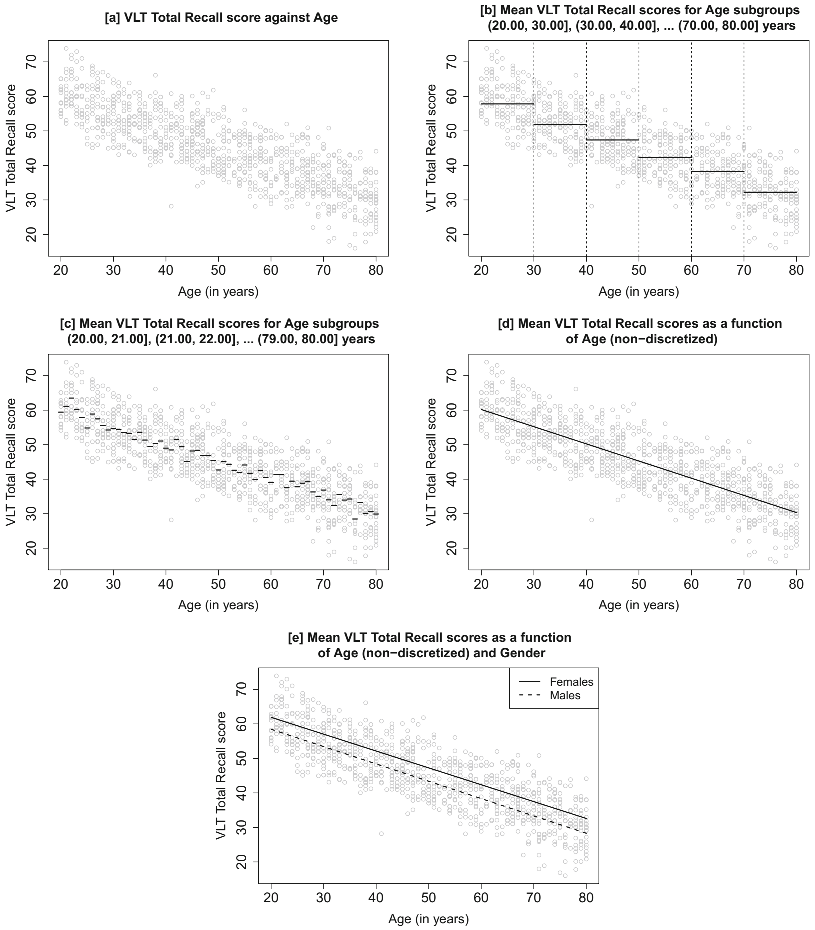

Fig. 1.4 Scatterplot of the VLT Total Recall scores against Age (see panel (**a**)) that is supplemented with the mean VLT Total Recall scores for Age subgroups (20.00, 30.00], (30.00, 40.00], …and (70.00, 80.00] years (see panel (**b**)), with the mean VLT Total Recall scores for Age subgroups (20.00, 21.00], (21.00, 22.00], …and (79.00, 80.00] years (see panel (**c**)), with the mean VLT Total Recall scores that are modeled as a function of Age (not-discretized in Age subgroups; see panel (**d**)), and with the mean VLT Total Recall scores that are modeled as a function of both Age (non-discretized) and Gender (see panel (**e**))

and gradual nature (see Fig. 1.4a) – as opposed to a step-function with abrupt changes in the means for different Age subgroups.

- When a quantitative independent variable is discretized, the so-called *boundary problem* is encountered (Capitani 2019). To illustrate this phenomenon, consider a scenario where a patient who has memory complaints is administered the VLT at his or her 70th birthday versus a few days later. Suppose that the patient obtained a raw VLT Total Recall score $Y_0 = 25$. In the first scenario (where the patient is aged exactly 70.00 years), the raw test score is standardized using the mean and the SD of the VLT Total Recall score in the (60.00, 70.00] years Age subgroup, yielding $\widehat{\upsilon}_0 = \frac{25-38.225610}{5.769952} = -2.292153$ with corresponding percentile rank $\widehat{\pi}_0 = 1$ (see Appendix A.1). In the second scenario (where the patient is aged 70.01 years at the time of test administration), the mean and SD in the (70.00, 80.00] years Age subgroup are used to perform the normative conversion, yielding $\widehat{\upsilon}_0 = \frac{25-32.242424}{6.357396} = -1.139212$ with corresponding percentile rank $\widehat{\pi}_0 = 13$. The small difference in the timing of the test administration thus has a dramatic impact on the obtained percentile ranks (i.e., $\widehat{\pi}_0 = 1$ versus 13), which is obviously not acceptable (Capitani 2019).
- The number of subgroups and the specific bounds that are used in the discretization of a quantitative independent variable are determined arbitrarily in the traditional normative approach (Parmenter et al. 2010). For example, instead of using 6 Age subgroups with intervals (20.00, 30.00], (30.00, 40.00], ... and (70.00, 80.00] years, we could alternatively have discretized Age into 3 Age subgroups with intervals (20.00, 40.00], (40.00, 60.00], and (60.00, 80.00] years. No proper justification can be given as to which of these discretization schemes is "the best," but the choice for one discretization scheme or the other can nonetheless have a major impact on the norms that are obtained in the traditional normative method. To illustrate this, consider a 61.00-year-old patient who obtained a raw VLT Total Recall score $Y_0 = 25$. When the first discretization scheme of Age is used, the raw test score will be standardized using the mean and the SD of the VLT Total Recall score in the (60.00, 70.00] Age subgroup, yielding $\widehat{\upsilon}_0 = \frac{25-38.225610}{5.769952} = -2.292153$ with corresponding percentile rank $\widehat{\pi}_0 = 1$. When the second discretization scheme is used, the raw test score would be standardized using the mean and the SD of the VLT Total Recall score in the (60.00, 80.00] Age subgroup, yielding $\widehat{\upsilon}_0 = \frac{25-35.224924}{6.762377} = -1.512031$ with corresponding percentile rank $\widehat{\pi}_0 = 7$. As can be seen, the essentially unjustifiable choice for one discretization scheme of Age or the other has again a major impact on the norms that are obtained in the traditional normative approach. So even though *exactly the same* normative sample is used in both analyses, a substantially different set of norms would be obtained.

1.3.2 The Splitting Problem

A potential solution for the boundary problem is to use narrower subgroups in the discretization of the quantitative independent variable at hand. For example, if we would discretize Age into very narrow subgroups with a span of only 1 year, the boundary problem would be substantially reduced. Unfortunately, at the same time the so-called *splitting problem* would become more pronounced. The splitting problem refers to the phenomenon that dividing the normative sample into subgroups reduces the precision by which the subgroup-specific summary statistics can be estimated. Indeed, it is well-known that the precision by which summary statistics such as the mean and the SD can be estimated is strongly impacted by the sample size. To illustrate this, consider the standard errors of the means and SDs, respectively:[6]

$$\widehat{\sigma}_{\mu_Y} = \frac{\widehat{\sigma}_{Y_i}}{\sqrt{N}}, \tag{1.2}$$

$$\widehat{\sigma}_{\sigma_Y} = \frac{\widehat{\sigma}_{Y_i}}{\sqrt{2(N-1)}}. \tag{1.3}$$

The above expressions immediately show that the use of more narrow subgroups in the discretization of a quantitative independent variable (to try to deal with the boundary problem) will substantially reduce the precision of the estimated means and SDs. Indeed, this will result in smaller subgroup-specific sample sizes N and thus in larger standard errors for the summary statistics (because the $\sqrt{N}$ and $\sqrt{2(N-1)}$ components in the denominators of expressions (1.2) and (1.3) will decrease). When the means and the SDs of the test scores in the normative sample are estimated with a larger standard error (or equivalently, with a lower precision), the same will obviously hold for the $\widehat{\upsilon}_0$- and $\widehat{\pi}_0$-values because the latter metrics of relative position are functions of the estimated means and SDs (see Sect. 1.2).

To illustrate the splitting problem, consider again the VLT Total Recall normative sample and suppose that we would discretize Age into narrow subgroups that each have a span of approximately 1 year (i.e., as (20.00, 21.00], (21.00, 22.00], …and (79.00, 80.00] years). As noted above, this would substantially reduce the boundary problem – but at the same time the splitting problem would become very pronounced. Indeed, the normative sample would now have to be split into 61 Age subgroups, which each contain (on average) only approximately 16 test participants (i.e., $N = 1000/61 = 16.393440$). As a result, the summary statistics would be estimated with poor precision (see expressions (1.2) and (1.3)). To illustrate this, Fig. 1.4c shows the mean VLT Total Recall scores for the 61 Age subgroups

[6] The standard error of an estimated parameter like the mean or the SD reflects the uncertainty in the estimated values (i.e., the sample-to-sample variability, see Chap. 3). When the standard error increases, the precision of the estimated summary statistic decreases.

in the normative sample (see the small horizontal black lines). It can be readily observed that the means jump up and down from one Age subgroup to another in a rather inconsistent way. For example, the estimated mean VLT Total Recall score for 50-year-old test participants equals 42.666667, whereas the estimated mean VLT Total Recall score for 55-year-old test participants equals 44.111111. It is very unlikely that the *true* (i.e., population-level) mean VLT Total Recall score of 55-year-old people is higher than the true mean score of 50-year-old people, because the vast majority of the cognitive ageing studies have shown that there is a consistent negative impact of Age on the verbal memory abilities of adults (Hedden & Gabrieli 2004; Van der Elst & Jolles 2012). Instead, the inconsistent pattern in the sample means of the VLT Total Recall scores across the different Age subgroups is attributable to the splitting problem, i.e., the phenomenon that the subgroup-specific means and SDs are estimated with poor precision after splitting the normative sample into many subgroups.

Observe that the splitting problem and the boundary problem are inversely related to each other. Indeed, the boundary problem can be ameliorated by using narrower subgroups in the discretization of a quantitative independent variable – but this results in an exacerbation of the splitting problem and vice versa.

Fine-Grained Norms and Sample Size Requirements

The splitting problem is not only an issue in the discretization of quantitative independent variables, but it also makes it difficult to derive fine-grained normative data that account for multiple independent variables. The reason for this is that the consideration of each additional independent variable will further reduce the sample size per subgroup. For example, suppose that we would like to account for the impact of *both* Gender and Age on the VLT Total Recall score. Even if we would totally ignore the boundary problem and it would be acceptable to use, e.g., Age subgroups with a span of 5 years, the normative sample would have to be split into a total of 26 subgroups (i.e., $2 \cdot 13$ Gender by Age subgroup combinations). This would result in subgroup-specific sample sizes of (on average) only approximately 39 test participants (i.e., $N/26 = 38.461540$). It obviously makes no sense to derive normative data based on such small samples.

In fact, recommendations regarding the minimum subgroup-specific sample sizes in the traditional normative approach vary substantially and range between $N = 75$ and 300 (Bridges & Holler 2007; Charter 1999; Evers et al. 2009; Piovesana & Senior 2018). Even in the most optimistic scenario where $N = 75$ per subgroup would be considered sufficient to achieve the required estimation precision for the summary statistics, it can be readily observed that very large sample sizes are typically needed in the traditional normative approach. Indeed, in the above example where it is of interest to derive normative data for the VLT Total Recall score that accounts for both Gender and Age (using Age subgroups with a span of 5 years), the total required sample size would be as large as $N = 1950$ ($= 2 \cdot 13 \cdot 75$). Moreover, each additional independent variable that is considered in the norms would increase the total sample size even further. For example, suppose that Level of Education (with 3 levels: low, average and high) would also have to be accounted for in the normative data. The total sample size would now correspond to $N = 5850$ test participants ($= 2 \cdot 13 \cdot 3 \cdot 75$). So even in the most optimistic scenario (i) where $N = 75$ per subgroup is considered to be sufficient to achieve a good estimation precision for the summary statistics, and (ii) where the boundary problem is ignored, very large sample sizes are typically needed in the traditional normative approach when multiple independent variables have to be accounted for. Conducting normative studies with such large sample sizes is often not feasible from a practical perspective because this would be a very time-intensive and costly endeavor.

1.4 The Regression-Based Normative Approach

This book focuses on regression-based methods to derive normative data (Zachary & Gorsuch 1985). In this approach, regression models that capture the mean and the residual structures in the normative dataset are used to derive norms (see later chapters). An important advantage of the regression-based normative method is that it can handle quantitative independent variables in a straightforward way *without* the need to discretize these. Indeed, in the regression-based normative approach the mean test scores are directly modeled as a function of the quantitative independent variable at hand. To illustrate this, consider again the example VLT Total Recall normative sample that was shown earlier in Fig. 1.4a. Figure 1.4d shows the model-predicted mean test scores that are obtained in the regression-based approach (see the solid black line). As there is no need to discretize the quantitative independent variable Age, the issues that were discussed in Sect. 1.3.1 also do not occur:

- In the traditional normative method, the unrealistic assumption had to be made that the subgroup-specific means follow a step-function (see the horizontal lines in Fig. 1.4b).
- In the traditional normative approach, the boundary problem was encountered. For example, it was illustrated that a small difference in the timing of the test administration can have a dramatic impact on the obtained percentile ranks. Similarly, it was illustrated that the use of a different discretization scheme for the quantitative independent variable Age can have a major impact on the obtained normative data.

Moreover, in the regression-based normative approach, all information from the entire normative sample is used when the mean and the residual structures are modeled. The issues that were detailed in Sect. 1.3.2 thus also do not occur:

- In the traditional normative approach, the normative sample had to be split into subgroups to account for the impact of independent variables. This adversely affects the precision of the estimated summary statistics. In contrast, the regression-based normative approach uses all available information in the normative sample to model the mean and the residual variance structures. This typically leads to a higher estimation precision (Van Breukelen & Vlaeyen 2005).

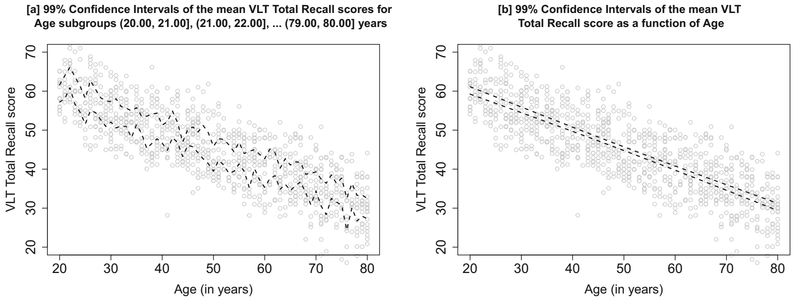

Fig. 1.5 99% Confidence intervals of the mean VLT Total Recall scores for the Age subgroups (20.00, 21.00], (21.00, 22.00], …and (79.00, 80.00] years (panel (**a**)), and for the mean VLT Total Recall scores that are predicted based on a regression model (panel (**b**))

To illustrate this, Fig. 1.5a shows the 99% Confidence Intervals (CIs) of the mean VLT Total Recall scores for the Age subgroups (20.00, 21.00], (21.00, 22.00], …and (79.00, 80.00] years that are obtained when the traditional normative method is used. Figure 1.5b shows the 99% CIs for the mean VLT Total Recall scores that are obtained in the regression-based normative approach. It can be readily observed that the 99% CIs are substantially narrower when the regression-based normative approach is used. The reason for this is that the summary statistics in the traditional normative approach are estimated for each subgroup separately. It is thus essentially assumed that the summary statistics of the VLT Total Recall score in one subgroup (e.g., in the (20.00, 21.00] Age subgroup) *tell us nothing* about the summary statistics in another subgroup (e.g., in the (21.00, 22.00] Age subgroup), and thus a lot of (potentially) useful information is ignored. For example, Fig. 1.1a clearly indicates that there is an approximately linear relation between Age and the (mean) VLT Total Recall scores in the normative sample. In the regression-based approach, this information is explicitly used in the normative analysis to achieve a higher estimation precision (see later chapters).

Notice that even in the simplest normative analysis where one independent variable with two possible outcome values (a binary variable such as Gender) has to be accounted for, the traditional normative approach is often suboptimal. Indeed, in such a scenario the summary statistics are computed for each subgroup separately, but this is suboptimal for the SDs when certain assumptions hold (in particular, when the so-called homoscedasticity assumption is valid; for details, see subsequent chapters).

- The traditional normative approach strongly limits the number of independent variables that can be accounted for in the normative data. In contrast, regression-based normative methods allow for deriving more personalized and fine-grained norms where the impact of multiple independent variables can be jointly taken into account. For example, suppose that it is of interest to derive normative data for the VLT Total Recall score that account for both Age and Gender. In the regression-based normative approach, all information in the normative sample is again used to model the mean and the residual variance structures as a function of Age and Gender in a straightforward way. This is visually illustrated in Fig. 1.4e, which shows the predicted mean VLT Total Recall scores as a function of Age for females (see the solid black line) and males (see the dashed black line). As all information is used in the analysis, more precise estimates of the means and the residual variances can be obtained in the regression-based normative approach (Van Breukelen & Vlaeyen 2005). This in turn results in lower sample size requirements. For example, Oosterhuis et al. (2016) found that regression-based normative methods require a sample size that is 2.5 to 5.5 times smaller than what is the case for traditional normative methods to achieve the same levels of precision. This substantially reduces the time investment and cost of the normative study.

As illustrated above, the regression-based normative approach has some substantial advantages over the traditional method (in particular when quantitative independent variables have to be accounted for, and/or when the normative analysis involves multiple independent variables), but at the same time it poses some additional challenges. Indeed, the use of the regression-based normative approach is more complex than the traditional method, because it involves the fitting of statistical models for the mean and the residual variance structures (in contrast to the traditional normative approach, which simply consists of computing subgroup-specific means and SDs). If the models that are used to derive the normative data are not appropriate (e.g., a linear association between Age and the mean test score is assumed, whereas the true association is non-linear), the obtained normative data will be incorrect as well. Moreover, it is important to ensure that the distributional assumptions that are made in the regression-based normative method are valid because violations of one or more of these assumptions can lead to incorrect normative data (see subsequent chapters).

1.5 Outline of the Book

The remainder of this book is organized in 7 chapters. Chapter 2 provides a brief introduction to the R statistical programming language. R will be used throughout this book to illustrate the application of the regression-based normative methods in case studies. Readers who are familiar with R can skip this chapter. Chapter 3 focuses on regression-based normative methods that account for one binary independent variable. A binary independent variable has only two possible outcome values, such as Gender. Chapter 4 discusses the assumptions that are made in a regression-based normative data setting and the remedial actions that can be taken when these assumptions are violated. Chapter 5 details regression-based normative methods that account for a (non-binary) qualitative independent variable. A qualitative independent variable has a limited number of possible outcome values that do not have a true numeric meaning (such as Level of Education). Chapter 6 focuses on regression-based normative methods that account for a quantitative independent variable (such as Age). Chapter 7 details regression-based normative methods that account for multiple independent variables simultaneously (i.e., any combination of binary, non-binary qualitative, and/or quantitative variables). Finally, in Chap. 8, a general approach is proposed to quantify the uncertainty in the obtained norms (i.e., the estimated percentile ranks) based on a bootstrap procedure.

1.6 Setting the Scene

Before delving into the regression-based normative approach in the subsequent chapters, it is important to establish some general comments and assumptions.

Psychometric Properties of Test Scores

As noted above, psychological assessment refers to the process of collecting and interpreting information that relates to a tested person's latent characteristics (Coaley 2009). In psychological assessment, there is never a perfect agreement between a raw test score and the latent characteristic that is being assessed. The extent to which a raw test score (e.g., the VLT Total Recall score) adequately captures the latent characteristic at hand (e.g., the overall verbal memory and learning abilities of a person) is referred to as the validity of a test score. A necessary condition to have a high level of validity is that the test score should have a high level of reliability. Reliability essentially refers to the repeatability of a test score (e.g., over time or when using different raters/test administrators). A test score can only have high levels of reliability and validity when other peripheral conditions are fulfilled. For example, the procedure to administer the test should be properly standardized, there should be clear and objective scoring rules, and so on. This book focuses on regression-based methods to derive normative data. A prerequisite to derive normative data for any test score is that it should have sufficiently high levels of reliability and validity (and thus peripheral conditions such as the proper standardization of the test administration should be fulfilled as well). In the current book, it will be assumed that all raw test scores to be normed have good psychometric properties, without actually testing these.[7]

Uncertainty in the Normative Data

Normative data allow for converting a raw test score Y_0 into a metric of relative position such as a percentile rank $\widehat{\pi}_0$. It is important to keep in mind that the obtained percentile ranks are *estimates* of the true percentile ranks in the normative population. Indeed, in a real-life normative analysis there is always uncertainty in

[7] A wide range of methods to estimate the reliability and the validity of test scores have been developed. These methods are beyond the scope of this book, but the interested reader is referred to, e.g., Coaley (2009), Furr (2021), or Van der Elst et al. (2016).

the estimated percentile ranks $\widehat{\pi}_0$ because the population distribution of the raw test scores is never known. This uncertainty is often not explicitly acknowledged in test manuals or other publications that provide normative data. For example, normative tables typically show the estimated percentile ranks without providing a metric of uncertainty (such as a CI or standard error for $\widehat{\pi}_0$). The percentile ranks $\widehat{\pi}_0$ (that are estimated based on a normative sample) are thus treated as if they are the true percentile ranks (i.e., as if they are estimated without any error), which is evidently not the case. The extent to which an estimated percentile rank corresponds to the true (i.e., population-level) percentile rank depends on several factors, such as the sample size of the normative study, how the normative sample was collected (e.g., whether random sampling was used), and the validity of the assumptions that were made in the derivation of the normative data.

Details are provided in subsequent chapters, but it is good to always keep in mind that the percentile ranks (or other metrics of relative position) that are obtained in a normative procedure are always estimated with uncertainty. Chapter 8 will cover a general method for quantifying the uncertainty in the estimated percentile ranks.

Item Response Theory

As mentioned above, the majority of psychological testing procedures are norm-referenced. This means that a raw test score cannot be interpreted in a meaningful way by itself (i.e., without normative data). This is particularly the case for tests that are rooted in the so-called Classical Test Theory (CTT). In CTT, the focus of the analyses is at the level of the total raw test score (e.g., the VLT Total Recall score), and score meaning is determined based on the relative position of the test score in a reference distribution (see above).

In contrast, in Item Response Theory (IRT), the focus of the analyses is on the individual items of the test (e.g., the items of a questionnaire or a multiple-choice test), and score meaning is determined by relating a tested person's estimated ability level to the item properties (Embretson & Reise 2000). In IRT, the difference between a tested person's ability level and the item characteristics has direct meaning *by itself* because both the ability levels and the item characteristics are calibrated on a common latent scale. For example, suppose that an IRT analysis is conducted on a rating scale for activities of everyday living that are challenging for older people (e.g., walking independently, cooking, taking the stairs, and so on). In IRT, a tested person's estimated ability level is meaningful by itself in the sense that it can be directly related to the probabilities that this person can still conduct these activities (i.e., the probabilities that the tested person can walk independently, cook, take the stairs, and so on). Normative data are thus not strictly necessary to interpret test performance in a meaningful way in the IRT framework, although they can be useful in this setting as well (for an example, see Van der Elst et al. 2013a).

Regression Models

This book focuses on regression-based methods to derive normative data. Many different types of regression models have been developed. Throughout this book, the classical linear regression model that is based on ordinary least squares estimation will be used (for details on this model, see subsequent chapters). This model has the advantages (i) that most psychologists have at least a basic familiarity with it, (ii) that the model is straightforward to fit because closed-form expressions are available to estimate the relevant model parameters (in contrast to more complex models that use iterative estimation procedures, which may or may not converge in a real-life normative analysis), (iii) that the underlying assumptions are straightforward to check, (iv) that model violations can be relatively easily remedied should they occur, (v) that the model parameters have a clear substantive interpretation (in terms of, e.g., the predicted mean test scores), and (vi) that the model is implemented in all standard statistical software packages (including R, SPSS, JMP, and SAS).

Many other regression types of models have been developed in the statistical literature, such as quantile regression (Koenker 2005), Generalized Additive Models for Location, Scale and Shape (GAMLSS; Stasinopoulos & Rigby 2007), linear mixed-effects models (Verbeke & Molenberghs 2000), and multivariate regressions models (i.e., regression models that consider multiple test scores at the same time; Johnson & Wichern 2007). These models are more flexible than the classical linear regression model (i.e., they relax some of the assumptions that are made by the classical model), but they are also substantially more complex to fit. In the current book, the focus is on the classical linear regression model, and these more complex models are not considered. The interested reader can find examples of how normative data can be derived using quantile regression, GAMLSS, linear mixed-effects models, and multivariate regression models in Crompvoets et al. (2021), Timmerman et al. (2021), Van der Elst et al. (2013b), and Van der Elst et al. (2017), respectively.

Continuous Test Scores

The classical linear regression model that is used in this book assumes that the dependent variable (i.e., the raw test score) is a continuous variable. A continuous variable is essentially a special case of a quantitative variable that can take an infinite number of real values between the lowest and highest values. In a psychological assessment context, the raw test scores are never *truly* continuous. For example, the VLT Total Recall score ranges between 0 and 75 and thus can take only 76 possible outcome values (instead of an infinite number of outcome values). A time score could in principle be a continuous variable (e.g., the time that is needed to complete a cognitive test), but in practice this is also not the case because the actually measured value is rounded to, e.g., a whole second or to one decimal place.

The fact that the raw test scores are not truly continuous is not problematic as long as (i) the number of possible outcome values for the test score at hand is sufficiently high (say, at least 10 possible outcome values), (ii) there is sufficient variability in the test scores, and (iii) the relevant assumptions are fulfilled for the fitted model (Fox 2016).

Notation

Throughout this book, Roman letters will be used to refer to the observed (or measured) independent and dependent variables. The letter Y refers to dependent variables (e.g., the VLT Total Recall score), and the letter X refers to independent variables (e.g., Age or Gender). Greek letters will be used to refer to true population parameters, and the hat-notation is used to distinguish true population parameters from their sample estimates. For example, μ_Y is the true mean test score in the normative population, and $\widehat{\mu}_Y$ is the estimated population mean based on the data in the normative sample.

Furthermore, the subscript $i = \{1, 2, \ldots N\}$ will be used to index the test participants in the normative sample. The index $i = 0$ is used to refer to a new person (not included into the normative study, e.g., a tested patient who has memory complaints).

References

Bridges, A. J., & Holler, K. A. (2007). How many is enough? Determining optimal sample sizes for normative studies in pediatric neuropsychology. *Child Neuropsychology, 13,* 528–538.

Capitani, E. (2019). Normative data and neuropsychological assessment. Common problems in clinical practice and research. *Neuropsychological Rehabilitation, 7,* 295–309.

Charter, R. A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology, 21,* 559–566.

Coaley, K. (2009). *An introduction to psychological assessment and psychometrics*. SAGE Publications Ltd.

Crompvoets, E. A. V., Keuning, J., & Emons, W. H. M. (2021). Bias and precision of continuous norms obtained using quantile regression. *Assessment, 28,* 1735–1750.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Lawrence Erlbaum Associates.

Evers, A., Lucassen, W., Meijer, R. R., & Sijtsma, K. (2009). *COTAN beoordelingssysteem voor de kwaliteit van tests* [COTAN assessment system for the quality of tests]. Nederlands Instituut van Psychologen.

Fox, J. (2016). *Applied regression analysis & generalized linear models* (3^{rd} edition). Sage.

Furr, R. M. (2021). *Psychometrics: an introduction*. SAGE Publications Ltd.

Hedden, T., & Gabrieli, J. D. E. (2004). Insights into the ageing mind: A view from cognitive neuroscience. *Nature Reviews, 5*, 87–97.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis*. Pearson Prentice-Hall.

Koenker, R. (2005). *Quantile regression* (Econometric Society Monographs). Cambridge University Press.

Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). Oxford University Press.

Mitrushina, M. N., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed.). Oxford University Press.

Oosterhuis, H. E. M., Van der Ark, L. A., & Sijtsma, K. (2016). Sample size requirements for traditional and regression-based norms. *Assessment, 23*, 191–202.

Parmenter, B. A., Testa, S. M., Schretlen, D. J., Weinstock-Guttman, B., & Benedict, R. H. (2010). The utility of regression-based norms in interpreting the minimal assessment of cognitive function in multiple sclerosis (MACFIMS). *Journal of the International Neuropsychological Society, 16*, 6–16.

Piovesana, A., & Senior, G. (2018). How big is big: Sample size and skewness. *Assessment, 25*, 793–800.

Rey, A. (1958). *L'examin clinique en psychologie* [The clinical examination in psychology]. Presses Universitaires de France.

Schmidt, M. (1996). *Rey auditory verbal learning test: A handbook*. Western Psychological Services.

Stasinopoulos, D. M., & Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software, 7*, 1–47.

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

Timmerman, M. E., Voncken, L., & Albers, C. J. (2021). A tutorial on regression-based norming of psychological tests with GAMLSS. *Psychological Methods, 26*, 357–373.

Van Breukelen, G. J., & Vlaeyen, J. W. (2005). Norming clinical questionnaires with multiple regression: The Pain Cognition List. *Psychological Assessment, 17*, 336–344.

Van der Elst, W. (2006). *The neuropsychometrics of Aging. Normative studies in the Maastricht Aging Study*. Neuropsy Publishers.

Van der Elst, W., & Jolles, J. (2012). *Verbal learning and aging*. In N. M. Steel (Ed.), *Encyclopedia of the sciences of learning* (pp. 3397–3400). Springer-Verlag.

Van der Elst, W., Ouwehand, C., van Rijn, P., Lee, B., Van Boxtel, M., & Jolles, J. (2013a). The shortened Raven standard progressive matrices: Item response theory-based psychometric analyses and normative data. *Assessment, 20*, 48–59.

Van der Elst, W., Molenberghs, G., Hilgers, R., Verbeke, G., & Heussen, N. (2016). Estimating the reliability of repeatedly measured endpoints based on linear mixed-effects models. A tutorial. *Pharmaceutical Statistics, 15*, 486–493.

Van der Elst, W., Molenberghs, G., Van Boxtel, M. P. J., & Jolles, J. (2013b). Establishing normative data for repeated cognitive assessment: a comparison of different statistical methods. *Behavior Research Methods, 45*, 1073–1086.

Van der Elst, W., Molenberghs, G., van Tetering, M., & Jolles, J. (2017). Establishing normative data for multi-trial memory tests: The multivariate regression-based approach. *The Clinical Neuropsychologist, 31*, 1173–1187.

Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2005). Rey's Verbal Learning Test: Normative data for 1855 healthy participants aged 24–81 years and the influence of age, sex, education, and mode of presentation. *Journal of the International Neuropsychological Society, 11*, 290–302.

Van der Elst, W., Van Boxtel, M. P. J., Van Breukelen, G. J. P., & Jolles, J. (2006). The Stroop Color-Word Test: Influence of age, sex, and education; and normative data for a large sample across the adult age range. *Assessment, 13*, 62–79.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer-Verlag.

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

Chapter 2
The R Programming Language

2.1 What Is R?

Chapters 3 to 8 of this book focus on regression-based methods to derive normative data. Each of these chapters starts with a theoretical part, which is followed by a practical part in which the methodology is exemplified in two case studies. The R software will be used to conduct the normative analyses of the case studies. In recent years, R has become one of the most popular statistical software tools to analyze data. There are several reasons for this:

- R is open-source software. In contrast to other major statistical software packages (such as SAS, JMP, or SPSS), no expensive software licenses are needed.
- R is platform-independent and runs on all major operating systems, including Windows, MacOS, and Linux.
- R is very capable "out of the box" and allows for conducting a wide range of classical statistical analyses (such as t- or χ^2-tests, ANOVA, linear regression analysis, and factor analysis). In addition, the capabilities of R can be extended by thousands of freely available add-on packages (or libraries). These packages can be downloaded in a straightforward way (see Sect. 2.2), and they allow for conducting a wide variety of more specific statistical and graphical analyses. For example, the package `NormData` (that accompanies this book) allows for deriving regression-based normative data in a straightforward way (see subsequent chapters).
- R has extensive graphical capabilities and can produce high-resolution plots of publication quality.
- R is compatible with other programming languages, such as C++ and Python.

The main disadvantage of R is that it has a relatively steep learning curve, particularly for people who have no prior programming experience. Indeed, R is a command-based software environment that has no point-and-click graphical user interface.

W. Van der Elst, *Regression-Based Normative Data for Psychological Assessment*,
https://doi.org/10.1007/978-3-031-50951-3_2