**The reliability of parametric methods in the case of rating scales: a simulation study**

Harald Kindermanna\*

a University of Applied Sciences Upper Austria, Marketing and Electronic Business, Steyr, Austria

\*corresponding author: [harald.kindermann@fh-steyr.at](mailto:harald.kindermann@fh-steyr.at)

**ORCID**: <https://orcid.org/my-orcid?orcid=0000-0003-3308-2981>

**NCBI**: <https://www.ncbi.nlm.nih.gov/myncbi/harald.kindermann.2/bibliography/public/>

**Google Scholar**: <https://scholar.google.at/citations?user=ukt11NMAAAAJ&hl=en>

**Researchgate**: <https://www.researchgate.net/profile/Harald-Kindermann/research>

The reliability of parametric methods in the case of rating scales: a simulation study

**Abstract**: A recurring question is whether rating scales should be considered metrically scaled or merely ordinally scaled. This has direct implications for the permissible statistical procedures for significance testing. Based on the results of a simulation study, it is shown that the use of parametric procedures for rating scales has distinct advantages over the non-parametric alternatives. It is also shown that the parametric procedures are robust to violations of the assumption of normality, which only result in a modest loss of power compared with continuous variables. This loss should be taken into account when calculating the optimal sample size. The results suggest that sample sizes about 25% larger should be chosen for discrete rating scales than for continuous variables.

**Keywords**: Rating scales, ordinal scales, parametric methods, non-parametric methods, significance tests.

**Introduction**

Scientific experiments are conducted to test hypotheses, or to explore relationships and develop new hypotheses. The researcher manipulates one or more independent variables to see how they affect one or more dependent variables. Statistical techniques are used to carry out a test of significance. Scientific experiments are based on precise rules, and significance tests are performed using the most appropriate statistical procedures to ensure that the results are as valid and replicable as possible.

This fact applies in principle to all scientific disciplines. In many cases, the focus is on humans. It is in the nature of things that many interesting facts cannot be measured directly, but must be captured through so-called latent constructs. Various personality traits, trust, and attitudes are examples of latent constructs that are of particular interest in the social and economic sciences. Mental illness is recorded in a similar manner, using frameworks such as the American Psychiatric Association’s Diagnostic and Statistical Manual of Mental Disorders (DSM) and the World Health Organization’s International Statistical Classification of Diseases and Related Health Problems (ICD).

Typically, these constructs are measured using discrete rating scales, usually of 5, 6 or 7 items. It is common to use not just one question to measure a latent construct, but a set of validated questions for which the latent construct is responsible. In such a case we speak of a measurement model consisting of multiple indicator questions. To obtain a concrete expression of a construct, an overall mean value is usually formed over all these indicator questions. However, it is also possible to measure some latent constructs using only a single indicator question. This is referred to as the single-item measure approach (Christophersen and Konradt 2011; Fuchs and Diamantopoulos 2009; Postmes et al. 2013; Wanous et al. 1997; Hoeppner et al. 2011). Whether a single-item measure approach is sufficient or whether a validated measurement model is required is always a matter of debate—for an overview see, for example, Malhotra et al. (2012)—but the crucial question is which mathematical and statistical procedures can be used to evaluate measurement models or single-item measures. In principle, the answer to this question depends on the scale level, and higher-quality procedures can be used for a metric level than for an ordinal or nominal level. The question then arises as to whether rating scales should be treated as interval or ordinal scales.

First, there is the liberal position that rating scales can be regarded as interval scales and thus parametric tests should be used, mainly because they have higher power than non-parametric alternatives. That is, an effect, if it exists, is more likely to be detected. If rating scales are treated as ordinal, then a larger sample is needed to find the same effect with the same probability.

Proponents of the position that rating scales can be considered interval scales include Streiner et al. (2016) and Norman (2010). Others argue that there must be at least one measurement model based on rating scales for parametric procedures to be applied. Under this view, the single-item measure approach mentioned above would not be sufficient (e.g., Carifio and Perla 2007; Boone and Boone 2012; Allen and Seaman 2007). A look at common research practice shows that rating scales are usually regarded as interval scales (Sugawara 2019; Diener 2010; Burger 2020; Krivoy 2017; Keng 2017; Mordeno 2019; González-Cabrera 2020). The diametrically opposite view is that treating inherently ordinally scaled data as interval data leads to misleading results and should therefore be rejected (e.g., Jamieson 2004; Allen and Seaman 2007).

Another important requirement for parametric methods to be applied is that the variable must be sufficiently close to normally distributed (e.g., Bühner and Ziegler 2009). However, since means of a dependent variable calculated using different samples of sufficient size are normally distributed, and this is independent of how the population is distributed, this requirement is fulfilled for sufficiently large samples (Streiner et al. 2016). With reference to the central limit theorem, we suggest n >= 30 as a sufficient sample size (Bortz and Schuster 2011; Döring and Bortz 2016). The problem is further mitigated when rating scales are used, since outliers cannot generally occur due to the restricted set of allowed responses.

These discussions about the premises to be applied to parametric scales are necessary from a theoretical point of view, but they miss the reality of the social science research community. In most of the relevant literature, neither are the premises for parametric procedures examined in the methods section, nor is there a critical discussion of whether parametric procedures can be used with rating scales. In most cases, it is the statistical competence or preference of the authors or reviewers that determines which procedures are used and accepted, rather than whether the procedures used are appropriate. The reader is urged to seek out such publications and make up their own mind. It appears to be especially difficult to find publications in the social sciences that verify that these requirements for applying parametric methods are met. For this reason, it is necessary to formulate valid and generally understandable recommendations on how to proceed statistically with rating scales.

The purpose of this article is to present the results of a simulation study, which models real experiments as closely as possible, to illustrate the differences between parametric and non-parametric procedures for rating scales and for violations of the normality assumption. In this simulation study, the parametric independent samples t-test is compared with the non-parametric Mann–Whitney U-test. False positive and false negative results are used as indicators to evaluate the procedures. Finally, specific recommendations are provided based on the results. As this issue is particularly critical in the case of a single-item measure approach, the simulation was carried out exclusively with this approach.

## Material and methods

***A typical test setup as a basis for the simulation***

The independent samples t-test and the Mann–Whitney U-test are used whenever two independent groups—an experimental group and a control group—are tested to determine whether changing a central tendency variable leads to significantly different results. In such an experiment, an intervention in the experimental group should change the selected central tendency parameter relative to the control group to such an extent that the intervention can be considered the cause and chance can be excluded except for a selected probability of error (usually 5%). The independent variable is defined here as the group membership of each subject, that is, whether the subject belongs to the experimental group or the control group. If the subject is in the control group, for example, either no intervention is performed, an intervention is simulated, a placebo is administered, or the so-called gold standard is performed (e.g., when testing a new medical treatment). However, if the subject is part of the experimental group, something new is usually tested on them. This can be a new treatment, a new drug or even new product packaging intended to improve the image of a brand.

In an experiment, the dependent variable (DV) represents the parameter that is to be influenced. For example, the DV can be the effectiveness of a new treatment method from the subjective point of view of the test subjects, or the effect of new product packaging on the image of a brand. These DVs are captured using the measurement models already mentioned, or using only a single question (the single-item measure approach). In this approach, the mean is used as the parameter for significance testing. The corresponding procedure would be the t-test for independent samples, in case the normal distribution assumption is sufficiently fulfilled. However, if the underlying scales are to be considered ordinal, then the mean would not be defined and a t-test would not be possible. Thus, one would have to resort to the parameter-free Mann-Whitney U-test. In this case, as already mentioned, one would have to accept a loss of power.

This brief description of a classical experiment serves only as a transition to the design of the simulation study that mimics such a simple experiment.

***The simulation procedure***

Following this simple experimental setup, we developed and ran the following simulations. First, two fictitious populations were created, each with n = 2 million objects. One population represents the AV of the control group. The other population represents the AV of the experimental group. The difference in means across all simulations was initially set to give an effect size of around 0.3. This would be a medium effect according to standard conventions (Cohen 1992). Then 1000 samples were randomly drawn from each population with sample sizes of n = 10, n = 30, n = 139 and n = 300. After each draw, a t-test and a Mann–Whitney U-test were performed (see Fig. 1).

This basic procedure was then varied with respect to the scale used (1–100, 1–7 or 1–5) and the type of distribution of the AV (normal or skewed).

[Figure 1 about here]

The selected sample size of n = 139 was chosen because it can be considered optimal for a one-sided test with an effect size of 0.3, a power of 80% and a type I error of 5%. This calculation was performed using R and the package “pwr” (see Fig. 2).

[Figure 2 about here]

Following this series of simulations with well-defined mean differences between the populations, population 1 and population 2 were chosen so that there were no mean differences. The same simulation steps as just described were then carried out. In this case, the simulation should produce a maximum of 5% significant results. This result would correspond exactly to the selected type I error. Figure 3 gives an overview of the whole simulation.

[Figure 3 about here]

**Results**

The simulation began with an ideal scenario. The continuous AV was perfectly normally distributed in populations 1 (the control group) and 2 (the experimental group). As already described, 1000 samples with different sample sizes were randomly drawn from each of these two populations (n = 10, n = 30, n = 139, n = 300). After each draw, a t-test and a U-test were performed. The true positive rate was then calculated.

Figure 4 shows the location of the two populations and the distribution of the 1000 sample means per population as a function of sample size.

[Figure 4 about here]

Table 1 contains the results of this first simulation. It can be seen that the Type I and Type II errors are quite close to the values assumed for the calculation of the optimal sample size of n = 139 (type I error = 5.1%, type II error = 19.8% and power = 79.8%). This also proves that the simulation leads to correct results.

[Table 1 about here]

The results clearly illustrate that overly small samples lose substantial power and that such “underpowered” studies are often a waste of resources because the desired evidence of an effect is unlikely to be found.

After this first simulation, the continuous variable was replaced by a discrete one and the scale range was also varied. For the scale range 1–100, the two population means were 50 (SD 16.5) and 55 (SD 16.5). For the scale range 1–7, the population means were 4.0 (SD 1.2) and 4.4 (SD 1.2). For the scale range 1–5, the population means were 3.0 (SD 1.0) and 3.3 (SD 1.0). From each of these populations, 1000 samples were drawn to determine the true and false positive rates of the two tests. Figure 5 shows the density of the distributions of the dependent variable for the fictitious populations.

[Figure 5 about here]

Tables 1–4 show the results of the simulation. The percentages correspond to the true positive rates (samples from populations 1 and 2) and the false positive rates (both samples from population 1).

[Tables 2-5 about here]

The results show that, on the one hand, the t-test has a slightly higher power than the non-parametric alternative. However, it is noticeable that this power is significantly reduced relative to a continuous variable. To compensate for this loss, when calculating the optimal sample size, one should choose a sample size that is about 25% larger. According to our simulation, this would provide the desired power for all scales.

It is now interesting to see how the results of this simulation change when it is carried out with populations for which the dependent variable deviates significantly from the assumed normal distribution. Figure 6 shows the density of arbitrarily chosen skewed distributions. They show typical deviations that occur in rating scales.

[Figure 6 about here]

The results are almost identical to the results of the simulation with a normally distributed variable. Thus, it is clear that the parametric t-test is robust to violations of the normality assumption. Again, it is recommended that the sample size chosen when calculating the optimal sample size should be approximately 25% larger if the intended power is to be achieved. Tables 5–8 show the results in detail.

[Tables 6-9 about here]

**Discussion**

Based on the results of the simulation, it can be concluded that rating scales can be treated as metrically scaled. The associated consequence of being able to use parametric procedures leads to a higher study power. A disadvantage cannot be identified on the basis of the present results.

However, samples that are too small significantly reduce the probability of finding an effect. Therefore, it would be advisable to always calculate the optimal sample size in advance and, in the case of rating scales, add an extra 25% or so. A sample that is too small will often not yield the desired result. This is a wasted opportunity and therefore a waste of resources. A sample that is too large is also a waste of resources because the desired effect could have been detected with less effort. Nevertheless, a sample that is too large is considered more advantageous than one that is too small.

**Disclosure** **statement**

No potential conflict of interest was reported by the author.

**References**

Allen, I. Elaine; Seaman, Christopher A. (2007): Likert scales and data analyses. In: Quality progress 40 (7), S. 64–65.

Boone, Harry N.; Boone, Deborah A. (2012): Analyzing likert data. In: Journal of extension 50 (2), S. 1–5.

Bortz, Jürgen; Schuster, Christof (2011): Statistik für Human-und Sozialwissenschaftler: Limitierte Sonderausgabe: Springer-Verlag.

Bühner, Markus; Ziegler, Matthias (2009): Statistik für Psychologen und Sozialwissenschaftler: Pearson Deutschland GmbH.

Burger, Joep; van Beuningen, Jacqueline (2020): Measuring well-being dispersion on discrete rating scales. In: Soc Indic Res. DOI: 10.1007/s11205-020-02275-1.

Carifio, James; Perla, Rocco J. (2007): Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes. In: J. of Social Sciences 3 (3), S. 106–116. DOI: 10.3844/jssp.2007.106.116.

Christophersen, Timo; Konradt, Udo (2011): Reliability, validity, and sensitivity of a single-item measure of online store usability. In: International Journal of Human-Computer Studies 69 (4), S. 269–280. DOI: 10.1016/j.ijhcs.2010.10.005.

Cohen, Yosef; Cohen, Jeremiah Y. (2008): Statistics and Data with R: an applied approach through examples: John wiley & sons.

Diener, Ed; Wirtz, Derrick; Tov, William; Kim-Prieto, Chu; Choi, Dong-won; Oishi, Shigehiro; Biswas-Diener, Robert (2010): New Well-being Measures: Short Scales to Assess Flourishing and Positive and Negative Feelings. In: Soc Indic Res 97 (2), S. 143–156. DOI: 10.1007/s11205-009-9493-y.

Döring, Nicola; Bortz, Jürgen (2016): Forschungsmethoden und evaluation. In: Wiesbaden: Springerverlag.

Fuchs, Christoph; Diamantopoulos, Adamantios (2009): Using single-item measures for construct measurement in management research: Conceptual issues and application guidelines. In: Die Betriebswirtschaft 69 (2), S. 195.

González-Cabrera, Joaquín; Machimbarrena, Juan M.; Beranuy, Marta; Pérez-Rodríguez, Priscila; Fernández-González, Liria; Calvete, Esther (2020): Design and Measurement Properties of the Online Gambling Disorder Questionnaire (OGD-Q) in Spanish Adolescents. In: Journal of clinical medicine 9 (1). DOI: 10.3390/jcm9010120.

Hoeppner, Bettina B.; Kelly, John F.; Urbanoski, Karen A.; Slaymaker, Valerie (2011): Comparative utility of a single-item versus multiple-item measure of self-efficacy in predicting relapse among young adults. In: Journal of substance abuse treatment 41 (3), S. 305–312. DOI: 10.1016/j.jsat.2011.04.005.

Jamieson, Susan (2004): Likert scales: how to (ab) use them. In: Medical education 38 (12), S. 1217–1218.

Keng, Shian-Ling; Wong, Yun Yi (2017): Association among self-compassion, childhood invalidation, and borderline personality disorder symptomatology in a Singaporean sample. In: Borderline personality disorder and emotion dysregulation 4 (1), S. 24.

Krivoy, Amir; Onn, Roy; Vilner, Yael; Hochman, Eldar; Weizman, Shira; Paz, Amir et al. (2017): Vitamin D Supplementation in Chronic Schizophrenia Patients Treated with Clozapine: A Randomized, Double-Blind, Placebo-controlled Clinical Trial. In: EBioMedicine 26, S. 138–145. DOI: 10.1016/j.ebiom.2017.11.027.

Malhotra, Naresh K.; Mukhopadhyay, Soumya; Liu, Xiaoyan; Dash, Satyabhusan (2012): One, Few or Many?: An Integrated Framework for Identifying the Items in Measurement Scales. In: International Journal of Market Research 54 (6), S. 835–862. DOI: 10.2501/IJMR-54-6-835-862.

Mordeno, Imelu G.; Nalipay, Ma. Jenina N.; Luzano, Jelli Grace C.; Galela, Debi S.; Ferolino, Michelle Anne L. (2019): Development and validation of a DSM-5-based generalized anxiety disorder self-report Scale: Investigating frequency and intensity rating differences. In: Curr Psychol. DOI: 10.1007/s12144-019-00475-8.

Postmes, Tom; Haslam, S. Alexander; Jans, Lise (2013): A single-item measure of social identification: reliability, validity, and utility. In: The British journal of social psychology 52 (4), S. 597–617. DOI: 10.1111/bjso.12006.

Streiner, David L.; Norman, Geoffrey R.; Cairney, John (2016): Health measurement scales: a practical guide to their development and use. In: Care Management Journals 17 (3).

Sugawara, Hiroko; Tsutsumi, Takahiro; Inada, Ken; Ishigooka, Jun; Hashimoto, Mamoru; Takebayashi, Minoru; Nishimura, Katsuji (2019): Association between anxious distress in a major depressive episode and bipolarity. In: Neuropsychiatric disease and treatment 15, S. 267–270. DOI: 10.2147/NDT.S188947.

Wanous, J. P.; Reichers, A. E.; Hudy, M. J. (1997): Overall job satisfaction: how good are single-item measures? In: The Journal of applied psychology 82 (2), S. 247–252. DOI: 10.1037/0021-9010.82.2.247.

**Tables**

Table 1: Results of simulations 1a and 1b

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Reference: continuous scale | | | | | |
|  | *n* | t-test |  | U-test |  |
| Samples out of **two** populations | 10 | 17.0 |  | 9.5 | True positive rate  (%) |
| 30 | 31.9 |  | 20.0 |
| 139 | 80.2 |  | 68.7 |
| 300 | 98.4 |  | 95.1 |
| Samples out of **one** population | 10 | 5.4 |  | 5.3 | False positive rate  (%) |
| 30 | 5.9 |  | 3.9 |
| 139 | 5.1 |  | 4.9 |
| 300 | 4.4 |  | 4.1 |

Table 2: Sample size n = 10 / approximately normally distributed / effect size (ES)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| n = 10 | Population 1 | Population 2 |  | Samples out of populations 1 and 2 | | Both samples out of population 1 | |
|  | mean 1 (SD 1) | mean 2 (SD 2) | ES | t-test | U-Test | t-test | U-Test |
| scale 1–100 | 50.0 (16.5) | 55.0 (16.5) | 0.3 | 10.2% | 9.4% | 4.4% | 4.1% |
| scale 1–7 | 4.00 (1.2) | 4.4 (1.2) | 0.3 | 10.0% | 9.2% | 4.2% | 3.9% |
| scale 1–5 | 3.0 (1.0) | 3.3 (1.0) | 0.3 | 10.1% | 9.6% | 3.6% | 3.5% |

Table 3: Sample size n = 30 / approximately normally distributed / effect size (ES)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| n = 30 | Population 1 | Population 2 |  | Samples out of populations 1 and 2 | | Both samples out of population 1 | |
|  | mean 1 (SD 1) | mean 2 (SD 2) | ES | t-test | U-Test | t-test | U-Test |
| scale 1–100 | 50.0 (16.5) | 55.0 (16.5) | 0.3 | 21.4% | 19.5% | 4.8% | 5.3% |
| scale 1–7 | 4.0 (1.2) | 4.4 (1.2) | 0.3 | 20.5% | 19.3% | 5.1% | 5.0% |
| scale 1–5 | 3.0 (1.0) | 3.3 (1.0) | 0.3 | 21.4% | 20.6% | 4.9% | 4.6% |

Table 4: Sample size n=139 / approximately normally distributed / effect size (ES)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| n = 139 | Population 1 | Population 2 |  | | Samples out of populations 1 and 2 | | Both samples out of population 1 | | |
|  | mean 1 (SD 1) | mean 2 (SD 2) | | ES | t-test | U-Test | | t-test | U-Test | |
| scale 1–100 | 50.0 (16.5) | 55.0 (16.5) | | 0.3 | 71.9% | 69.8% | | 5.3% | 5.5% | |
| scale 1–7 | 4.0 (1.2) | 4.4 (1.2) | | 0.3 | 70.5% | 69.6% | | 5.1% | 5.6% | |
| scale 1–5 | 3.0 (1.0) | 3.3 (1.0) | | 0.3 | 72.2% | 72.4% | | 5.3% | 5.1% | |

Table 5: Sample size n = 300 / approximately normally distributed / effect size (ES)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **n** **=** **300** | Population 1 | Population 2 |  | Samples out of populations 1 and 2 | | Both samples out of population 1 | | |
|  | mean 1 (SD 1) | mean 2 (SD 2) | ES | t-test | U-Test | | t-test | U-Test |
| scale 1–100 | 50.0 (16.5) | 55.0 (16.5) | 0.3 | 96.4% | 95.3% | | 6.1% | 6.7% |
| scale 1–7 | 4.0 (1.2) | 4.4 (1.2) | 0.3 | 96.0% | 95.1% | | 6.5% | 6.2% |
| scale 1–5 | 3.0 (1.0) | 3.3 (1.0) | 0.3 | 96.3% | 95.8% | | 5.9% | 5.8% |

Table 6: Sample size n = 10 / skewed / effect size (ES)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **n** **=** **10** | Population 1 | Population 2 |  | | Samples out of populations 1 and 2 | | | Both samples out of population 1 | | |
|  | mean 1 (SD 1) | mean 2 (SD 2) | | ES | | t-test | U-Test | | t-test | U-Test |
| scale 1–100 | 39.2 (20.4) | 45.6 (20.5) | | 0.3 | | 9.9% | 9.0% | | 4.4% | 4.1% |
| scale 1–7 | 3.1 (1.7) | 3.6 (1.9) | | 0.3 | | 9.3% | 7.9% | | 5.0% | 4.5% |
| scale 1–5 | 1.7 (0.8) | 2.0 (0.9) | | 0.3 | | 8.6% | 7.3% | | 5.2% | 4.2% |

Table 7: Sample size n = 30 / skewed / effect size (ES)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **n** **=** **30** | Population 1 | Population 2 |  | | Samples out of populations 1 and 2 | | | Both samples out of population 1 | | |
|  | mean 1 (SD 1) | mean 2 (SD 2) | | ES | | t-test | U-Test | | t-test | U-Test | |
| scale 1–100 | 39.2 (20.4) | 45.6 (20.5) | | 0.3 | | 22.2% | 21.5% | | 5.1% | 5.0% | |
| scale 1–7 | 3.1 (1.7) | 3.6 (1.9) | | 0.3 | | 21.9% | 19.8% | | 5.1% | 4.9% | |
| scale 1–5 | 1.7 (0.8) | 2.0 (0.9) | | 0.3 | | 20.3% | 17.2% | | 5.3% | 5.5% | |

Table 8: Sample size n = 139 / skewed / effect size (ES)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **n** **=** **139** | Population 1 | Population 2 |  | Samples out of populations 1 and 2 | | Both samples out of population 1 | |
|  | mean 1 (SD 1) | mean 2 (SD 2) | ES | t-test | U-Test | t-test | U-Test |
| scale 1–100 | 39.2 (20.4) | 45.6 (20.5) | 0.3 | 73.2% | 72.1% | 5.5% | 5.3% |
| scale 1–7 | 3.1 (1.7) | 3.6 (1.9) | 0.3 | 71.0% | 67.8% | 5.2% | 5.2% |
| scale 1–5 | 1.7 (0.8) | 2.0 (0.9) | 0.3 | 71.3% | 60.2% | 5.2% | 5.1% |

Table 9: Sample size n = 300 / skewed / effect size (ES)

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **n** **=** **300** | Population 1 | Population 2 |  | | Samples out of populations 1 and 2 | | | Both samples out of population 1 | | | |
|  | mean 1 (SD 1) | mean 2 (SD 2) | | ES | | t-test | U-Test | | t-test | U-Test |
| scale 1–100 | 39.2 (20.4) | 45.6 (20.5) | | 0.3 | | 95.5% | 94.8% | | 6.4% | 6.6% |
| scale 1–7 | 3.1 (1.7) | 3.6 (1.9) | | 0.3 | | 94.4% | 93.0% | | 6.0% | 5.8% |
| scale 1–5 | 1.7 (0.8) | 2.0 (0.9) | | 0.3 | | 94.6% | 87.1% | | 5.5% | 5.9% |

**Figures**

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

Figure 1: The basic test procedure

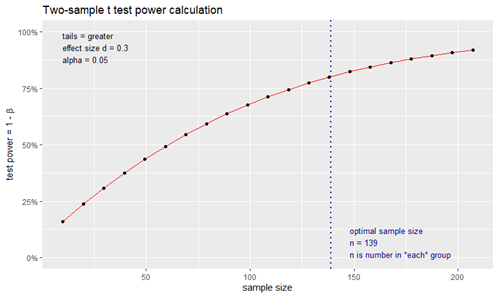


Figure 2: Results of the sample size calculation

Ein Bild, das Text, Quittung enthält.

Automatisch generierte Beschreibung

Figure 3: Overview of the simulations

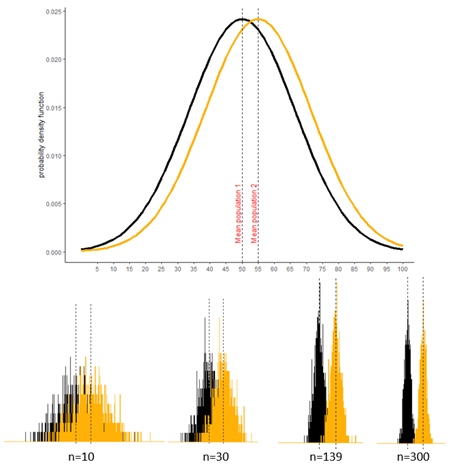


Figure 4: Populations 1 and 2 and the distribution of the sample means.

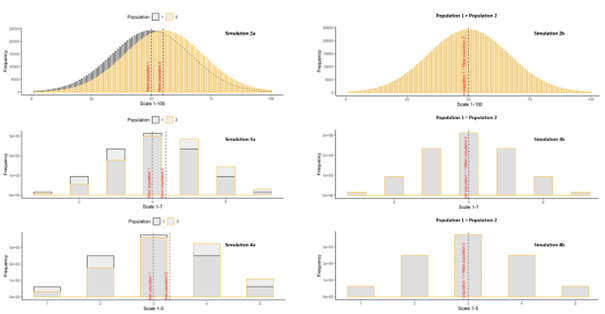


Figure 5: Density of populations—normally distributed (scales: 1–100, 1–7, 1–5)

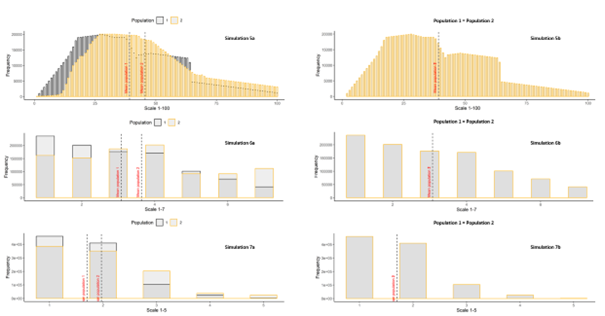


Figure 6: Density of populations—skewed (scales: 1–100, 1–7, 1–5)