universität wien

Diplomarbeit

Titel der Diplomarbeit

The Hybrid Rating Method:

Proposal and Assessment of a Novel Way to Measure Attitudes

Verfasser

Johann-Christoph Münscher

Angestrebter akademischer Grad

Magister der Naturwissenschaften (Mag.rer.nat)

Wien, 2015

Studienkennzahl lt. Studienblatt:	A 298
Studienrichtung lt. Studienblatt:	Diplomstudium Psychologie
Betreuer:	Univ.-Prof. i.R. Mag. Dr. Klaus D. Kubinger

Danksagung

An dieser Stelle will ich mich bei all jenen bedanken, die mich bei dem Erstellen dieser Arbeit unterstützt haben:

Herrn Univ.-Prof. Mag. Dr. Klaus D. Kubinger für die Betreuung und Unterstützung.

Claire Dahm, Martin Gödert, Tobias Griepenstroh, und Lena Sauerborn für ihre Hilfe bei der Befragung der Teilnehmer. Außerdem danke ich allen TeilnehmerInnen des Forschungsseminars für die allgemein tolle Atmosphäre und zahlreiche hilfreiche Rückmeldungen und Verbesserungen.

Meinen Eltern, für die Unterstützung während meines Studiums und die Hilfe, TeilnehmerInnen für diese Studie zu gewinnen.

Theresa Scharf, für ihre anhaltende emotionale und inhaltliche Unterstützung.

Ruben Maue, für Hilfe mit R und Empfehlungen von Literatur.

Allen TeilnehmerInnen dieser Studie, die teilweise lang und mit viel Geduld mitgemacht haben.

Abstract – English

In reaction to shortcomings with usual methods of measuring attitudes – mostly the five-point rating method – this paper presents the *hybrid rating method* as a novel approach to measuring attitudes. The hybrid rating method is characterised by its open ended design in which respondents create their own rating scale to fit their individual style. In this approach the central assumption was that respondents could express their attitudes better and the method therefore provides more and better information. The method is described in detail and procedures for scoring are presented. A comparison between the hybrid method and the five-point method, using an independent measure two-sample design ($\alpha = .05$; $\beta = .1$), highlights the aspects: Information quantity, information quality, and pragmatic value. In both conditions 40 respondents answered the NEO-PI-R with either the five-point method or the hybrid method and rated the process in an additional questionnaire. Analysis showed that the hybrid method yielded more information than the five-point method ($t = 2.823$, $p = 0.003$) while potentially allowing additional information to be gathered from the individuals rating style. Psychometric quality was found to be close to identical between the two methods (both methods show an average Cronbach's α of .88) and respondents rated the two methods similarly. The method was observed to be more demanding, both in time and effort, and therefore is restricted in its application. Criticism towards the method and this study as well as further implications are discussed.

Keywords: rating-scale, attitudes, measurement, questionnaire, psychometrics

Abstract – Deutsch

Als Reaktion auf die Mängel der üblichen Methoden zur Messung von Einstellungen – meist 5-Punkt Ratingskala – stellt dieser Artikel die *Hybride Ratingmethode* als neuen Ansatz zur Messung von Einstellungen vor. Die Hybride Ratingmethode zeichnet sich durch eine offene Gestaltung aus in der die Befragten eine eigene Ratingskala, angepasst an ihren individuellen Stil, konstruieren. Zentrale Annahme dieses Ansatzes war, dass die Befragten ihre Einstellungen besser ausdrücken können und das Verfahren so mehr und bessere Informationen liefert. Das Verfahren wird im Detail beschrieben und Verfahren zur Verrechnung werden vorgestellt. Ein Vergleich zwischen der Hybridmethode und dem 5-Punkt Verfahren anhand eines Zweigruppenplans ($\alpha = .05$; $\beta = .1$) beleuchtet die Aspekte Informationsmenge, Informationsqualität und pragmatischen Wert. In beiden Bedingungen bearbeiteten 40 Teilnehmer den NEO-PI-R, entweder mit der 5-Punkt oder der Hybridmethode und bewerteten das Verfahren mit einem zusätzlichen Fragebogen. Die Analyse zeigte, dass die Hybridmethode mehr Informationen als die 5-Punkt Methode lieferte ($t = 2.823$, $p = 0.003$), und der individuelle Antwortstil als zusätzliche Informationsquelle genutzt werden kann. Die psychometrische Qualität beider Methoden war annähernd gleich (beide Methoden zeigten ein durchschnittliches Cronbach's α von .88) und Teilnehmer bewerteten sie ebenfalls ähnlich. Die hybride Methode erwies sich als anspruchsvoller, sowohl in Zeit als auch Aufwand, und ist deshalb nur eingeschränkt einsetzbar. Das Verfahren, diese Studie sowie weitere Implikationen werden kritisch diskutiert.

Index

Introduction

Different rating methods that attempt to measure attitudes have become a hallmark of social sciences and while they are widely used, they are still heavily discussed and iterated upon. This paper proposes and assesses a new method that was developed in response to problems with existing methods. Based on the works of Thurstone (1994) and especially Likert (1932), rating-scales have been heavily researched and developed. In the past, the *five-point* method, already found to be the most useful by Likert in 1932 and therefore called the '*Likert-type scale*', was shown to be the most prevalent (Rohrmann, 1987). Some guidelines recommended five to seven points based on their overall utility (e.g., Bogner & Menold, 2015).

Naturally, rating methods exhibit shortcomings, some of which will be elaborated upon in the following – a general overview provided Podsakoff, MacKenzie, Lee, and Podsakoff (2003). On a five-point scale the granularity with which an attitude can be expressed is relatively low (Cohen, Manion, & Morrison, 2000; Döring, 2006). Complex attitudes have to be reduced into five rough groups. During this reduction, *floor and ceiling effects* (see Döring, 2006) can occur, causing attitudes that are sufficiently different to be rated equal because the available points cannot accommodate the difference, an effect explained by the *range-frequency model* (Parducci, 1965). This issue might potentially be prevented by a method in which respondents do not have to fit their attitudes to a scale but rather fit the scale to their attitudes.

A related issue, the *ambivalence-indifference problem*, describing the difficulty interpreting ratings labelled '*neutral*', causes ratings to be unclear (Kaplan, 1972). It is not possible to discern whether the respondent felt indifferent or ambivalent about the stimulus, the respondent might even have rejected the question, answered according to social acceptance, or lacked knowledge about the stimulus (Garland, 1991; Jonkisz, Moosbrugger, & Brandt, 2012; Podsakoff, et al., 2003). One might choose to remove the middle-point to try and remedy this problem, with the result that truly neutral attitudes can no longer be expressed and data distortions occur (Garland, 1991; Jonkisz, et al., 2012). The core of this problem seems to be that respondents cannot express all aspects of their attitudes and consequently deliver inconclusive data; thus a more granular and personalised method of rating seems beneficial.

Additionally, the *acquiescence bias* causes ratings to 'pile up' on the positive side, increasing ambiguity (Winkler, Kanouse, & Ware, 1982). One attempt to counteract this is to create items with reversed polarity, which has shown the side effect of introducing an additional factor loading onto the items (Jonkisz, Moosbrugger, & Brandt, 2012). Again, a more personalised rating method that allows for more differentiation might help to reduce *yea-saying* or *nay-saying* tendencies as

respondents are possibly more invested and interested in the process.

Considering both, floor-ceiling effects, and the ambivalence-indifference problem, the alarming revelation is that out of the five points on a five-point scale, only two points yield moderately concise information, whereas the remaining three produce inconclusive data. An initial response might be to increase the number of points to receive more concise data and to allow attitudes to be expressed with greater granularity. However, in typical rating-scales an increased number of points may cause diminishing returns with regards to psychometric quality – Matell and Jacoby (1971) saw no change in reliability while Ciccetti, Shoinralter, and Tyrer (1985) along with Preston and Coleman (2000) observed a limited increase – which mainly runs the risk of overburdening the respondent with too many choices (Bogner, & Menold, 2015; Döring, 2006). Preston and Coleman (2000) found that respondents preferred ten-point ratings while Hirsch (2006) found that five-point ratings were favoured. Döring (2006) and to some extent Matell and Jacoby (1971) concluded that the optimal number of points varies individually and is ideally determined by the respondent after having sampled all objects that are to be rated. This presampling of all stimuli seems not only impractical but only appears suitable to prevent floor-ceiling-effects with little potential to rectify the ambivalence-indifference problem.

In the past, alternative rating methods, such as the visual-analogue scale (see Crichton, 2001), the Q-technique (Stephenson, 1953), and the Semantic Differential (Osgood, Suci, & Tannenbaum, 1957) have been devised to address some of these shortcomings. In a new attempt the *hybrid rating method* was constructed and is presented and evaluated in this thesis.

The Hybrid Rating Method

The hybrid rating method was designed to provide the following: Reduction of floor-ceiling effects and acquiescence bias, a more accurate expression of attitudes with an individual scale, the use of more points without overburdening the respondent, and an equal or higher quality of data that yields more information. As the ambivalence-indifference problem cannot be alleviated without the introduction of further elements to differentiate respondents' reasons, the goal is not to remove the problem but to try and diminish its influence.

The heart of the method lies in an open-ended design. To prevent an accumulation of ratings at the endpoints (floor-ceiling effect), the approach was to remove the fixed endpoints to open up the scale. Respondents can use however as many points as they feel necessary to express their attitude. This concept is similar to the visual-analogue scale in that the respondent can presumably express “the underlying continuum” (Crichton, 2001) more directly and therefore more easily. In the hybrid method, there is no predetermined end so the length can be adjusted should the need arise

for a wider range – therefore eliminating floor-ceiling effects. For the respondents to easily understand the process, the instruction conveys that the task is not to choose from options but rather to rate the stimulus by assigning numerical values. The algebraic signs of these values are used to express the 'direction' of the attitude, consequently a value of 0 represents a neutral attitude. As an example: A respondent who highly agrees with a statement might rate it with + 15 points to express this opinion, while if later on an even more favourable statement might come up, the respondent can then expand the ratings and rate it with + 17 points. Obviously, it is not possible to know what a given value means on its own; therefore a reference frame has to be established. This frame is left open until the end of the questionnaire and thus remains flexible. The reference frame is determined in the end with two questions that ask respondents how they did or would have expressed their maximum of the attitude in question (these values are referred to as *minima/maxima* in the following).

The instruction used in this study reads as follows (translated from German):

"[...] read these statements carefully and determine how much they apply to you. State how much a statement applies to you by awarding positive or negative points. Award **however as many points as you wish**. There is no "**right**" or "**wrong**" amount and **no limit**; award the points so that you feel your opinion is adequately expressed. The goal is that you answer intuitively – answer spontaneously and without too much thought. Should you rate a statement completely neutrally award 0 points. Write the amount of points you award in the big box. Use the small box to indicate the direction with a plus or minus sign. Cross the small box out when you award 0 points."

These instructions were originally created in German based on the "NEO-Persönlichkeitsinventar nach Costa und McCrae, Revidierte Fassung (NEO-PI-R) [NEO-Personality Inventory after Costa and McCrae, Revised (NEO-PI-R)]" (Ostendorf & Angleitner, 2004). Following these instructions, an example as shown in Figure 1 helps to clarify the method.

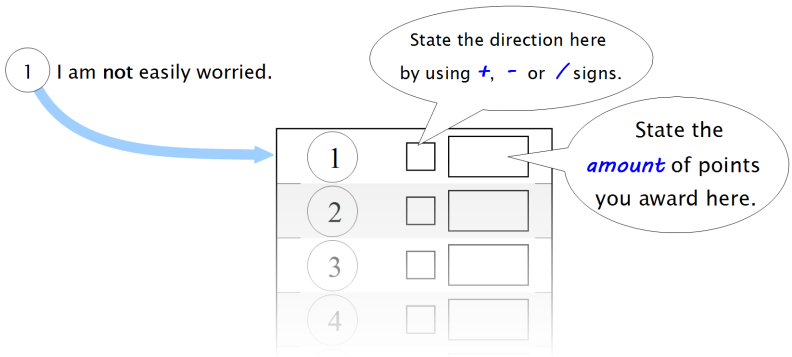


Figure 1. The example following the instructions.
The elements which respondents have to write are highlighted with a font that imitates handwriting and blue colouring. Additionally speech-bubbles were used to convey that ratings have to be entered into the boxes.

To prevent the anchor heuristic (Tversky & Kahnemann, 1974) from influencing the individual's choice of points, the example was designed without the use of explicit numbers; an alphabetic naming of the items (i.e. AA, AB, AC, …) was also tried for the same reason, but discarded in favour of visual clarity.

After all statements are rated, respondents answer the final questions (translated from German): "How many points would you award or did you award to statements which you **fully agree** with? How many points would you award or did you award to statements which you **fully disagree** with?" The resulting data describes the respondents' ratings on an individually constructed scale that can fit the respondents' need for granularity and even allows for different lengths of the two directions. Possible leanings towards a particular way of judging, such as coarse/thorough or negative/positive, might potentially be expressed this way. For example: An optimistic person might choose a wide range to express positive ratings and a narrower range for negative ones, while an introspective individual might use a high granularity (amount of different ratings) when expressing her/his attitudes. This assumption seems to be contradicted by findings by Hirsch (2006) who found no relevant connection between respondents' personality traits and their preference for, or use of rating methods. However, in the hybrid design respondents do not choose from a fixed number of options – as was done in Hirsch's (2006) study – but instead are free to define an individual scale to their liking. In addition to the number of points a person chooses, its also possible see how they use them; this possibly allows one to receive additional information about the respondents' cognitive style, akin to an objective personality test (Cattell & Warburton, 1967).

In order to allow inter-subjective comparisons, the individual scales have to be transformed into a uniform level. This is easily done by dividing the values of each direction by the respective maximum which results in a scale that represents ratings in the range [-1, 1]. Because this

transformation is linear, no information is lost or distorted. It might be argued that in the case of different maxima distortions occur because ratings are transformed differently; however, these distortions are only relative between directions with the resulting scale ranging from a minimum to a maximum with an equal distance to the middle. The resulting data can be used directly or if necessary, be transformed into a number of different formats; for example multiplying by 5 and adding 5 converts the ratings to the range [0, 10]. Figure 2 demonstrates the process with a set of imaginary data.

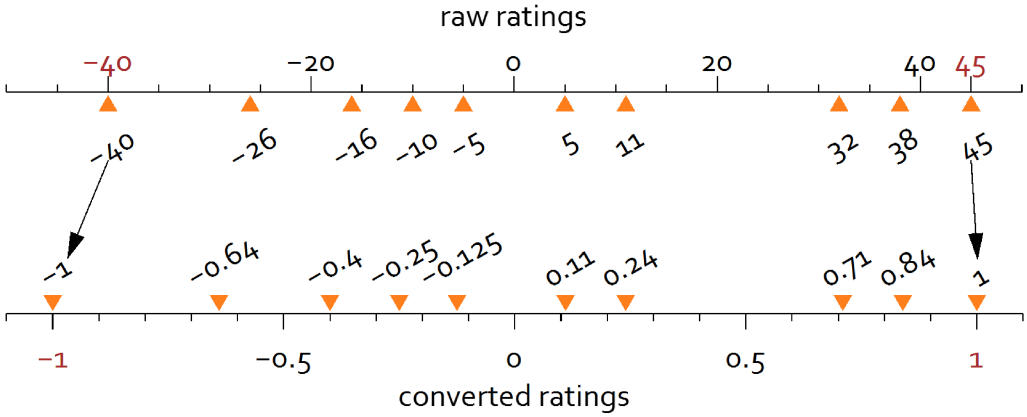


Figure 2. A visual representation of a set of data with an uneven length of the scale [-40, 45] being transformed into a uniform range from -1 to 1.

In its design, the hybrid rating method is similar to the proposal by Matell and Jacoby (1971) who speculated: "[…] desirable practical consequences might be obtained from allowing the subject to select the rating format which best suits his needs. This might result in the highly favorable consequence of increasing the subject's motivation to complete the scale. […] Indeed, it is even conceivable that the subject could record his own responses (open-ended) to each item, without a previously prepared rating format being provided."

Given the apparent complexity of the instructions and the rating process, it seems likely that the hybrid method puts a greater demand on respondents and therefore might violate the requirement for ease-of-use stated by Rice (as cited in Likert, 1932). Hence it is probably less suitable as a general-purpose method, but rather better suited as a specialized tool. Another important aspect is the number of items used. While the hybrid method should, in principle, work with only one item, it seems necessary that a reasonable number of items be presented for the respondents to become familiar to the method.

The metric appearance and potentially high granularity of hybrid ratings might suggest a false sense of accuracy and objectivity; it is important to keep in mind that these are still only ratings and therefore should be treated accordingly. Specifically, data from the hybrid method should be

assumed to be an ordinal level, assuming ratings to be an interval level is an often discussed misstep (Cohen, et al., 2000; Döring, 2006; Jamieson, 2004; Kuzon, Urbanchek, & McGabe, 1996; Rasch, Kubinger, & Yanagida, 2011; Robertson, 2012) which might easily occur when using this method.

In summary, the hybrid rating method was designed to be a versatile way of measuring attitudes and aims to combine various aspects from other methods, hence the prefix 'hybrid', while addressing issues associated with said methods.

Hypotheses and Research Questions

This study aims to assess the functionality of the hybrid method by highlighting the aspects: *Information quantity*, *information quality*, and *pragmatic value* in comparison to the five-point method. The distinction between information quantity and quality is important as it allows an independent assessment of the mathematical amount of information and semantic viability. Based on the goals and assumptions, the following hypotheses and research questions were formulated:

Information quantity.

The basic assumption was that more information can be gathered by employing the hybrid method as compared to the five-point method. On the one hand this involved the mathematical amount of information, and additional data on the individual's rating-style on the other. Therefore the hypotheses were:

- Data gathered with the hybrid method yields more mathematical information than data gathered with the five-point method.
- Data gathered with the hybrid method holds additional information about the individual rating-style.

  When using a personality questionnaire, a particular cognitive style might be echoed by the reported personality traits, therefore the subhypotheses were:
  - There is a significant correlation between the minima/maxima of an individual's ratings and their reported personality traits.
  - There is a significant correlation between the granularity of an individual's ratings and their reported personality traits.

Information quality.

This category highlights the quality of the gathered data. In short, these were psychometric data quality, prevention of floor-ceiling effects, and reduction of acquiescence bias. These aspects were explored in the form of research questions as they did not lend themselves to statistical

analysis or could not reliably be measured in the scope of this study. The research questions were:

- How does the hybrid-rating method compare to the five-point method with regards to psychometric quality?
- How successfully does the hybrid-rating method prevent floor-ceiling effects and acquiescence bias?

Pragmatic value.

This aspect was comprised of elements based on usability as well as the overall usefulness of the hybrid-rating method. This entails the difficulty of the rating process, the usefulness of the open-ended format, and ease with which an attitude can be expressed. As before, these were explored as research questions:

- What observations can be made from using both the hybrid and five-point rating method?
- How do respondents assess the hybrid and five-point rating method?
- How do respondents use the open-ended nature of the hybrid-rating method?
- How do respondents evaluate their ability to express their opinion with the two rating methods?

Method

For this study, a two-group experimental design was used to compare data gathered by the five-point and hybrid rating methods. Both groups filled in the German NEO-PI-R personality questionnaire (Ostendorf, & Angleitner, 2004); one group rated the statements with the default five-point method, in the following referred to as the '*classical condition*', whereas the second group was instructed to employ the hybrid-rating method, referred to as the '*hybrid condition*'. Participants were randomly assigned to the conditions. The study finished with a selection of written open-ended questions pertaining to the questionnaire and its rating method. Additionally, observations from the several survey administrators were collected. This design was intended to provide a way to assess the functionality of the hybrid method without the use of external criteria or retests.

In this study, the NEO-PI-R is the centrepiece and directly provides data to assess the aforementioned aspects: Information quantity and information quality. Additionally, a supplementary questionnaire provided data to assess the pragmatic value aspect. The NEO-PI-R consists of 240 items based on the Five-factor Model in which respondents have to state how much a statement applies to them using the five categories *strongly disagree* (0), *disagree* (1), *neutral* (2), *agree* (3), and *strongly agree* (4). The decision to specifically use the NEO-PI-R in this study lies in its prevalent use and good factor-analytic construction on the one hand and the fact that it suffers

from problems commonly associated with questionnaires on the other (Muck, 2004; Vassend, & Skrondal, 1995, 2011). It seems to provide a good real-world example to assess the method without any undue bias that might have been introduced by an unproven or highly artificial instrument.

The supplementary questionnaire was administered at the end and was designed to gather information on how respondents perceived the NEO-PI-R questionnaire and the rating methods used. As the goal of this questionnaire was exploratory, it was designed to use five open-ended questions that could be answered with short sentences or keywords. The questions were: (translated from German)

- How do you rate answering the questionnaire on the whole?
- How easy/hard to understand did you find the instructions?
- How easy/hard was it for you to form opinions toward the statements?
- How easy/hard was it for you to express these opinions in the questionnaire?
- How easy/hard was it for you to answer the questionnaire on the whole?

The openness of the first question was deliberate so that respondents could express their overall opinion before answering the rather specific remaining questions. Afterwards the answers were analysed and categorised. For each question the answer was subjectively categorised into one of the four groups: '*predominantly negative*', '*neutral*', '*predominantly positive*', and '*no answer*'. Additionally all points of criticism were collected and ordered into groups.

Assessment and Measures

Some of the assessments relied on comparisons between the hybrid and classical condition. The hybrid rating method's individually constructed scale cannot directly be compared to five-point ratings; to allow sensible comparisons between the two conditions the ratings from the hybrid condition were converted into five-point ratings. The range of [-1, 1] was divided into 5 segments with an equal length of .4 and the values within were converted to the corresponding labels of the five-point scale. To clarify, the ratings from the hybrid condition were used in two different formats – so called *hybrid ratings* were ratings produced by the hybrid method, whereas for *converted ratings* the same ratings were translated into five-point ratings, as described above. The ratings from the classical condition are referred to as *classical ratings*. In addition, further data was surveyed including, age, gender, education, occupation, and prior experience with personality questionnaires.

Information quantity.

Amount of mathematical Information: To determine which of the two rating methods yielded more information, the amount of mathematical information within the ratings of each

respondent was ascertained by calculating their *Entropy* (Shannon, 1948/2001); an approach similar to that outlined by Dahl and Østerås (2010). Entropy *H (Eta)* signifies how much information a signal carries and functions as a qualitative measure of variance that is determined by the relative frequencies of the symbols in the signal (Floridi, 2009; Scherer, & Brüderl, 2010; Shannon, 1948/2001). In the case of this study, a respondent's 240 ratings of the NEO-PI-R items were treated as a signal.

Based on Shannon's (1948/2001) work, Entropy is defined by:

$$\mathrm{H} = -\sum_{i=1}^{K} \hat{p}_i \ln \hat{p}_i$$

where K is the number of available symbols; in case of five-point ratings $K = 5$, and $\hat{p}_i$ is the relative frequency of the symbol i, defined by $\hat{p}_i = n_i / N$; where n_i is the absolute frequency of the symbol i and N the length of the signal, therefore $N = 240$. This procedure allows for an intra-individual Entropy value to be calculated for each respondent that expresses how much mathematical information is contained within the ratings from that person – it is important to note that the entropies were determined for the ratings exactly the way they were made by the respondents without inverting items with reverse polarity.

Entropy reaches its maximum when all elements of a signal appear with equal frequency (for $K = 5$, $H_{\max} \approx 2.32$); therefore a high variability in the signal equates to more entropy. Entropy was specifically chosen as a measure because in this scenario it holds very distinctive advantages over variance which one might use to determine said variability. The Entropy calculation treats ratings not as numerical values, but as labels and therefore as a nominal level which makes it ideal for use with ordinal ratings (Luce, 2003). In contrast, variance relies on the numerical values of these labels and their mean, therefore requiring an interval level of measurement which does not apply to ratings from rating methods (Cohen, et al., 2000; Döring, 2006; Jamieson, 2004; Kuzon, et al., 1996; Rasch, et al., 2011; Robertson, 2012). The resulting Entropy values however, are an interval level and can therefore be used for a wide range of inferential statistics (Bavaud, 2009).

From the converted ratings and the classical ratings, the intra-individual entropies for all cases were calculated using the 'entropy.empirical()' function from the Entropy Package (Hausser, Stimmer, 2013) for R (R Core Team, 2014). To compare their means, the one-sided Welch-test was chosen, based on its advantages over the *t* Test (Rasch, Kubinger, & Moder, 2011). To summarize, data from the hybrid condition was converted into five-point ratings and the intra-individual entropies for all cases were calculated and compared between the two conditions.

Additional information: To test the hypothesis that the personal rating style is linked to the construction of the individual rating scale, the correlations between the ranges of the individual

scale – derived from the minima and maxima reported by the respondents – and the personality traits were calculated. Additionally, the correlation between the granularity – the number of different values in a person's individual rating scale – and the reported personality traits was determined. The reported personality traits were derived from the scores for the five NEO-PI-R dimensions by simply summing the ratings – hybrid ratings were used for the hybrid condition – of the corresponding items. In both cases, Spearman's correlation coefficient was calculated.

Information quality.

Psychometric quality: As mentioned before, this study was designed to be mostly self-contained and did not rely on retests. Therefore psychometric quality was assessed by comparing Cronbach's α and the intercorrelations of the NEO-PI-R dimensions between the classical and hybrid condition instead. The NEO-PI-R's manual states these values as well, so they could be compared to the results from the two conditions. For ratings from the hybrid condition Cronbach's α was calculated with the converted ratings, whereas the intercorrellations were calculated with the hybrid ratings.

Floor-ceiling effects and acquiescence bias: These aspects were assessed using identical measures. The distribution and the frequencies with which the five rating categories were used in the classical condition and the converted ratings from the hybrid condition were evaluated. An increased use of the endpoints indicates floor-ceiling effects, the distribution's skewness – interpreted according to Bulmer (1979) – shows potential acquiescence bias. Additionally the distribution and density of the hybrid ratings was analysed. Inferential tests, i.e. tests for rectangular distribution, were not used because the frequencies are merely indicators for floor-ceiling effects and acquiescence bias but do not warrant the conclusion whether one or the other exists.

Pragmatic value.

Overall observations: To assess the overall use of the hybrid method observations from 6 different administrators and those made during the processing and handling of the data were gathered.

Respondents' perception, use of the open-ended scale, and expression of attitudes: These three aspects were highlighted with largely identical measures. The supplementary questionnaire contributed most of the data; for the analysis of how respondents used the open-ended scale, frequencies within the ratings and observations made by administrators were assessed.

Sample size, power, and precision

Respondents were recruited by *opportunity sampling* by 6 administrators with the goal to keep a reasonable distribution of age, gender, education, and occupation. The required sample size was determined by the one-sided Welch-test used to compare the means of the intra-individual Entropies between the conditions. For a rough estimation before starting the survey, the guidelines by Jan and Shieh (2011) were used which recommend a sample size between 25 and 54 in each condition for a power of at least .9. After 10 sets of data had been collected for each condition, the OPDOE package (Rasch, Pilz, Verdooren, & Gebhardt, 2011; Simececk, Pilz, Wan, & Gebhardt, 2014), specifically its 'size.t.test()' function, for R (R Core Team, 2014) was used to calculate the needed size more precisely. As relevant differences in the Entropy of rating scales are not readily established, it was set at two thirds of the observed entropy's standard deviation (classical SD = 0.092; hybrid SD = 0.050) which causes differences in entropy greater than 0.061 to be treated as relevant. Given a type-I-risk of .05 and type-II-risk of .1, the corresponding optimal sample size is 40 in each condition.

Results

Sample

Surveying took place from mid- 2014 to early 2015 and N = 80 respondents were surveyed. In the classical condition (n = 40, 19 females), age ranged from 23 to 76 (M = 42, SD = 18.66); in the hybrid condition (n = 40, 15 females), age ranged from 22 to 79 (M = 40, SD = 17.76). In both conditions high-school and academic degrees were the most prevalent; in terms of occupation, students were the biggest group in both conditions (hybrid: 15, classical: 17), followed by employees (classical: 8, hybrid: 13); see Table 1 for an overview. 27 respondents in the classical condition and 29 respondents in the hybrid condition reported that they had answered a personality questionnaires at least once before.

Table 1

Overview of the education and occupation within the two conditions.

	condition			condition	
Education	classical	hybrid	Occupation	classical	hybrid
Sec. education		1	Worker	1	1
Apprenticeship	6	1	Employee	13	9
General sec. education	3	2	Unemployed	1	1

University qualification	17	18	Freelance	1	
University degree	14	18	Self employed	2	2
No degree			Marginal employment		
			Housewife / man		
			Pensioned / retired	6	6
			Civil servant	1	
			Student (school)		
			Student (university)	15	21

Inferential Statistics

Information quantity

Amount of mathematical information: When comparing the intra-individual Entropies from the classical condition (M = 1.411, SD = 0.126) with those of the converted ratings from the hybrid condition (M = 1.480, SD = 0.088) the Welch-Test (t = 2.823, p = 0.003, df = 69.959) showed that the converted ratings from the hybrid condition contain significantly more mathematical information than the ratings from the classical condition, see Figure 3.

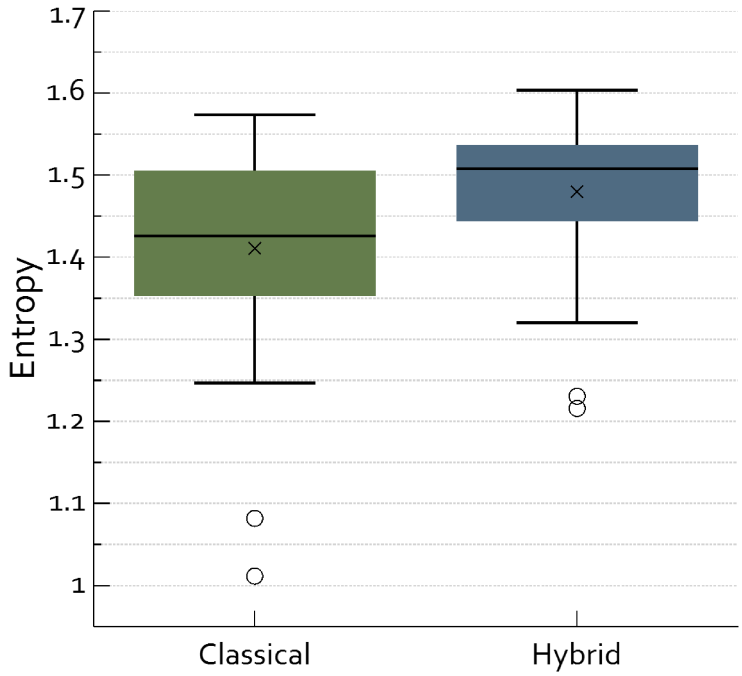


Figure 3. A boxplot displaying the intra-individual entropies within the two conditions. Crosses mark the means, circles outliers and wiskers 1.5 iqr. Entropy in the classical condition ranged from 1.01 to 1.57 and from 1.21 to 1.60 in the hybrid condition.

Ratings from the two outliers observed in the classical condition (Figure 3) were unique in that these respondents only used three of the available five rating categories; the two outliers in the hybrid condition did not seem to share a characteristic. When calculating the above results, an accidental calculation of the entropies of ratings in which the scores for items with a double negative had already been inverted – a step necessary to determine the NEO-PI-R scores – revealed an overall drop in entropy for both conditions. For the classical condition the mean dropped by 0.028 to 1.383 and for the hybrid condition by 0.015 to 1.465.

Additional information: Before the transformation into the range [-1, 1], ratings from the hybrid condition exhibited minima with an overall range of [3, 1000], M = 57.2, SD = 160.16, and maxima of [3, 1000], M = 57.82, SD = 160.25. The ratings' granularity ranged from 3 to 5 in the classical condition and from 7 to 90 in the hybrid condition, Mdn = 20. Between the minima/maxima of the individual scale and the five reported personality traits, no significant correlation was found, whereas the granularity (number of different ratings used) of the individual scales exhibited a significant correlation with the reported personality trait *openness* (r_S = .44, p = .02, r^2 = .19). See Table 2 for a summary.

Table 2

Correlations between the NEO-PI-R dimensions and the minima/maxima of the individual scales, as well as the granularity.

	N	E	O	A	C
	r_S (p)	r_S (p)	r_S (p)	r_S (p)	r_S (p)
Minimum[a]	-.05 (1)	.00 (1)	.28 (.81)	-.18 (1)	-.04 (1)
Maximum[a]	-.04 (1)	-.02 (1)	.27 (.81)	-.17 (1)	-.02 (1)
Granularity	.31 (.19)	-.02 (.91)	.43 (.03)	-.29 (.19)	-.16 (.68)

Note. NEO-PI-R dimensions are Neuroticism(N), Extraversion(O), Openness(O), Agreeableness(A), Concentiousness(C)

[a] p values for correlations of minimal and maxima were collectively holm-corrected; correlations for granularity were holm-corrected separately.

Exploratory Statistics

Information quality.

Psychometric quality: Cronbach's α ranged from .83 to .92 (M = .88) in the classical condition and from .83 to .94 (M = .88) in the hybrid condition (converted ratings). The NEO-PI-R's manual

reports values from .87 to .92 (M = .90). Intercorrelations ranged from -.554 to .217 (M = .039) in the classical condition and from -.428 to .271 (M = -.04) in the hybrid condition (hybrid ratings). The NEO-PI-R manual reports intercorrelations between -.37 and .4 (M = -.022). Cronbach's α and intercorrelations were calculated using the Psych package (Revelle, 2014) for R. See Table 3 for a summary.

Table 3

Cronbach's α and intercorrelations for the NEO-PI-R dimensions in the two conditions compared to the values from the manual.

		Dimensions[c]					
Measure		N	E	O	A	C	*M*
Cronbach's α							
	Classical	.92	.86	.89	.83	.9	.88
	Hybrid[a]	.94	.9	.83	.88	.89	.88
	Manual	.92	.89	.89	.87	.90	.90
Intercorrelations[b]							
	Classical	-.256					
E	Hybrid	-.333					
	Manual	-.27					
	Classical	.197	.277				
O	Hybrid	.120	.164				
	Manual	.05	.40				
	Classical	.020	.231	.099			
A	Hybrid	-.428	.254	-.116			
	Manual	-.04	-.05	.02			
	Classical	-.553	.205	-.034	.208		
C	Hybrid	-.350	.250	-.311	.271		
	Manual	-.37	.08	-.1	.06		

[a] Cronbach's α for the hybrid condition was determined with converted ratings.

[b] Intercorrelations for the hybrid condition were calculated with hybrid ratings.

[c] *NEO-PI-R dimensions are Neuroticism(N), Extraversion(O), Openness(O), Agreeableness(A), Concentiousness(C)*

Floor-ceiling effects and acquiescence bias: Ratings in the classical condition showed a median of overall 2 (*neutral*) compared to 3 (*agree*) in the converted ratings from the hybrid condition. Classical ratings showed relative frequencies of: *Strongly disagree* = .087, *disagree* = .267, *neutral* = .208, *agree* = .334, and *strongly agree* = .105 (SD = 0.105) with a skewness of -0.132 and kurtosis of -1.007. Converted ratings from the hybrid condition exhibited relative frequencies of: *Strongly disagree* = .143, *disagree* = .198, *neutral* = .156, *agree* = .273, and *strongly agree* = .230 (SD = 0.053) with skewness of -0.252 and a kurtosis of -1.232. Figure 4 displays detailed findings for the classical condition and the converted ratings from the hybrid condition.

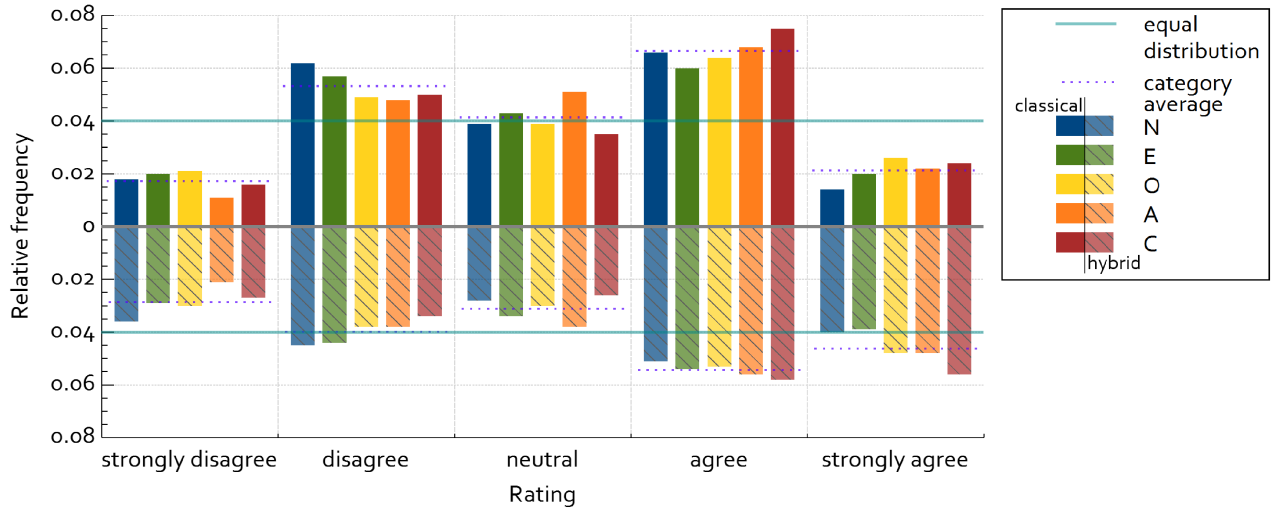


Figure 4. Relative frequencies of the five rating categories in the five NEO-PI-R dimensions. Theoretical equal distribution lies at .04. Each condition consists of 9600 ratings.

The distribution of the hybrid ratings from the hybrid condition showed a median of 0.2 (M = 0.093, SD = 0.54) with a skewness of -0.234 and a kurtosis -0.948. Figure 5 displays the relative frequencies and kernel density estimation graph.

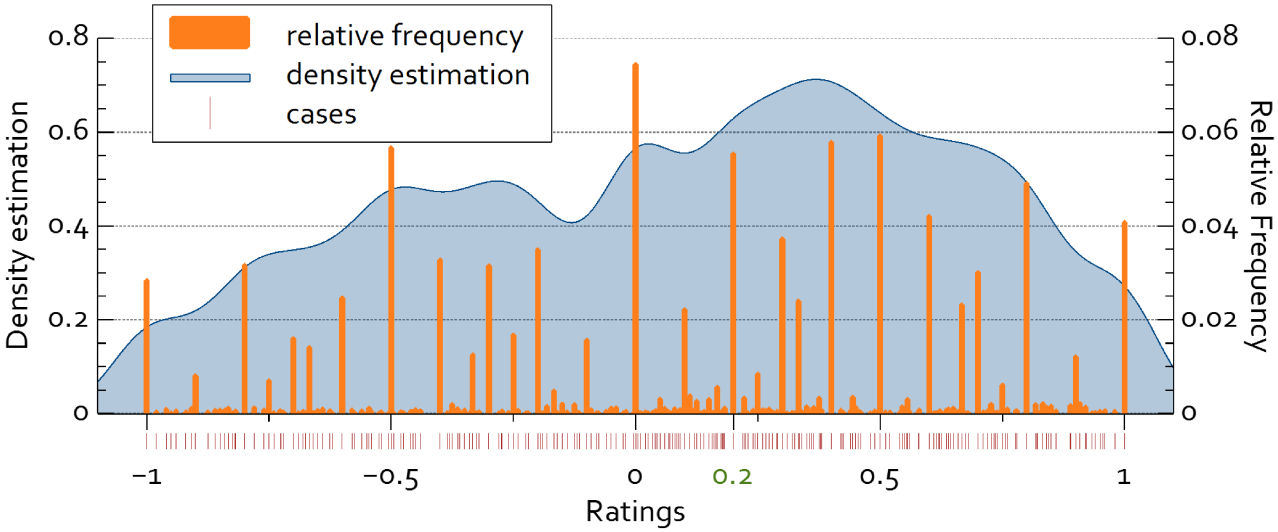


Figure 5. The relative frequencies of the hybrid ratings and the corresponding kernel density estimation graph using a Gaussian kernel with a bandwidth of 0.077.

Pragmatic value.

Overall observations: During the survey administration, the following subjective observations were made by administrators:

Most respondents needed between 30 and 90 minutes to answer all statements. Respondents in the hybrid condition needed roughly one third longer compared to those in the classical condition. Some respondents in the hybrid condition reported that they thought longer about their

answers than they otherwise would have. Reactions to the statements were usually positive; respondents regularly commented on the statements and were often amused by their wording. A majority of respondents in both conditions criticised the double negatives used in some of the statements. Many respondents commented that some statements seemed to repeat previous statements, some suspected that these were 'control-questions' to identify lying.

Two respondents, one in each condition, did not finish the questionnaire. Most respondents seemed to have no trouble using the hybrid method; some showed difficulties starting, after a reminder to read the instructions most participants continued without further problems. It often seemed as if respondents in the hybrid condition decided on a range for their ratings before or while answering the first item and kept it to the end. Respondents often appeared to be splitting the rating process in two stages, first deciding whether they agreed or not, then determining the intensity. When asked afterwards how many points they had used, respondents from the hybrid condition often named a range from zero to their maximum, seemingly unaware that they actually used a much wider range. Two respondents in the hybrid condition reported that their selected range was too big and they had trouble expressing their attitude. One respondent reported that their selected range was too narrow, the option to widen their scale had not become apparent to them. Four respondents in the hybrid condition specified ranges that did not fit their ratings (i.e. the reported range was 10 but the highest rating was 15); these ranges were adjusted accordingly. Inputting data from the hybrid condition for analysis was more time consuming and prone to errors than inputting data from the classical condition.

Respondents' perception, use of the open-ended scale, and expression of attitudes: Minima and maxima ranged from 3 to 1000; most respondents chose a minimum of 10 and a maximum of 10. See Table 4 for a summary.

Table 4

Minima and maxima of the ratings from the hybrid condition and their respective frequencies.

	Range (*f*)												
Min	- 10 (18)	- 50 (4)	- 100 (3)	- 3 (2)	- 4 (2)	- 5 (2)	- 6 (2)	- 200 (2)	- 22 (1)	- 30 (1)	- 40 (1)	- 80 (1)	- 1000 (1)
Max	10 (17)	50 (4)	100 (4)	5 (3)	3 (2)	6 (2)	200 (2)	4 (1)	9 (1)	22 (1)	30 (1)	45 (1)	1000 (1)

As described before, data from the supplementary questionnaire was processed by subjective judgements by the author. All answers were categorised into the four groups '*predominantly negative*', '*neutral*', '*predominantly positive*', and '*no answer*'. See Figure 6 for results.

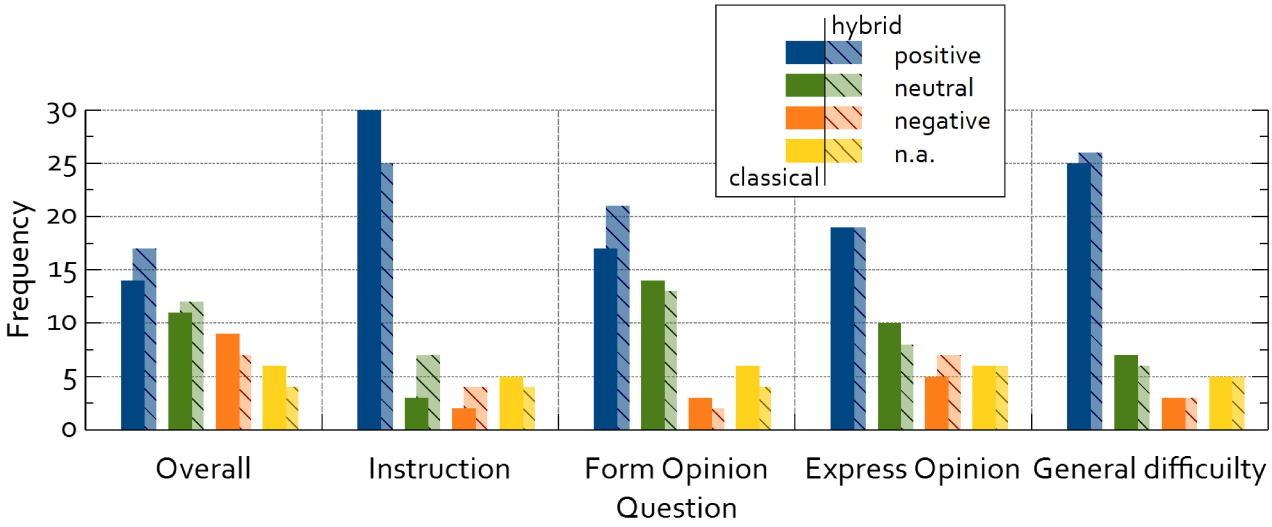


Figure 6. Overview of the answers in the supplementary questionnaire.

Additionally all critical statements made in the supplementary questionnaire were categorized into the following topics: Questions (wording and content), Rating Method, Double Negative, Motivation/Effort, Instruction, and Row Shifting. In both conditions, the two most frequent points of criticisms were directed at the questions (classical = 14, hybrid = 9), and the effort involved in answering (classical = 11, hybrid = 11). See Table 5 for results.

Table 5

Frequencies of the categories of criticism levelled towards answering the NEO-PI-R.

condition	Questions	Method	Double neg.	Effort	Instruction	Row shift
Classical	14	2	6	11	0	2
Hybrid	9	5	4	11	4	0

Some statements made in the supplementary questionnaire were particularly noteworthy and are quoted in the following (translated from German).

Classical condition:

- "The difficulty of answering was substantially lower than the effort of making crosses."
- "No problem, except for the questions in which one has to shower oneself with boundless self-praise (or the contrary). What normal person does this?"
- "I had to be very concentrated and had to take a break."
- "It was not always easy to distinguish between black and white with only two shades of grey."

Hybrid condition:

- "[…] however I had the feeling that I was trying to use every number in my range, therefore I felt the choice was rather arbitrary."
- "It was an interesting one and a half hours."
- "[…] sometimes there was a requirement to choose 'interesting' numbers; no idea why."
- "Upon the first reading, it was not completely understandable; I understood everything the second time."
- "Entertaining, relatively stimulating, differentiated ways of answering cause interesting thoughts."
- "[…] there were also lots of nonsensical questions for which I lacked the option to express this confusion."

Discussion

The results show that the hybrid method yields more mathematical information compared to five-point ratings, while providing additional information about the respondent's cognitive style. The hybrid method seems to achieve good psychometric quality. A reduction of floor-ceiling effects and acquiescence bias was not observed; however, ratings from the hybrid condition were more evenly distributed than those from the classical condition. In general the hybrid rating method has shown to be functional, with the advantages of allowing respondents to construct their own rating-scale. The overall usefulness is restricted because the rating process is demanding and time consuming; nevertheless respondents evaluated the method mostly favorably and had little to no problems using it. Considering the increased demand on respondents, the higher rate of errors and the more complicated way of scoring the ratings, a suggestion for future applications is to use a computerized version.

With regards to information quantity, the hybrid method has shown to yield more information than the five-point method. On average, the intra-individual entropies of ratings from the hybrid condition exceed those from the classical condition. Furthermore, this mean-difference shows relevance as it is greater than the minimal relevant difference defined beforehand.

Because ratings from the hybrid condition were converted into five-point ratings for this comparison, the observed difference is far from trivial. It seems obvious that the unconverted ratings from the hybrid condition would exhibit higher entropy simply because the hybrid method allows for more noise in the data. However, the fact that the ratings from the hybrid condition show a higher entropy even *after* they were converted into the five-point format, indicates that there is not

only more noise, but more information. As the calculation of entropy uses the frequencies with which symbols occur in a signal and reaches its maximum with a uniform distribution, inspecting the frequencies of the categories 'shows' where this increase in information originates: The relative frequencies of the rating categories from the classical condition have almost double the standard deviation (SD = 0.105) of those of the converted ratings from the hybrid condition (SD = 0.053), indicating that the converted ratings from the hybrid condition were more evenly distributed. It is important to keep in mind that entropy is merely a measure of mathematical information and does not allow assumptions about the semantic content; an entirely random set of ratings would produce high entropy without actual meaning.

Analyzing the correlations between the minima/maxima of the unconverted ratings and the reported personality traits showed no significant link, whereas the granularity of the individual ratings exhibited a significant correlation with the score in the NEO-PI-R dimension *openness.* Again, this result is far from trivial. It not only provides evidence that the hybrid method can supply additional information about the cognitive style of the respondent, but also reinforces the notion that the information gathered by the hybrid method is meaningful. The clear-cut correlation with specifically one dimension in the NEO-PI-R could not have been observed with overly noisy data. This correlation supports the hypothesis that the individual's cognitive style influences the rating process. It seems reasonable to assume that more open-minded individuals approach the method accordingly and more readily engage with it.

Calculations of psychometric quality showed that the reliability and intercorrelations (Table 3) are similar between the two conditions and also are very close to the metrics provided by the NEO-PI-R manual. They appear to be sufficiently similar to conclude that the difference in rating method did not influence the psychometric quality of the instrument. This conclusion not only echoes the findings by Matell and Jacoby (1971) – who found no influence of the number of rating points on internal consistency – but more importantly, provides a strong indication that the hybrid method functions just as well as other methods with regards to reliability while carrying the instrument's semantic content.

Based on the distribution of ratings within the two conditions, the conclusion is that no recognizable floor-ceiling effects occurred. Visual inspection of the distributions (Figures 4 & 5), reveals no signs of an accumulation on the ends; furthermore their skewness can be interpreted as “fairly symmetrical” (Bulmer, 1979). Therefore inferring that one or the other method was more adept at preventing floor-ceiling effects has no merit. Consequently there did not seem to have been a reason for respondents to extend their range. Therefore the assumption is that most respondents stuck to the range they decided upon in the beginning. This notion is supported by observations

made during administration and the fact that most minima and maxima were multiples of five.

The relative frequencies of the five rating categories in the classical and hybrid conditions (converted ratings) show that the converted ratings from the hybrid condition were more evenly distributed, indicated by the lower standard deviation of their relative frequencies and the differences in entropy. The skewness of classical ratings was lower than for hybrid ratings, indicating that acquiescence bias was to some extent present with both methods, but stronger for ratings from the hybrid condition. However, using Bulmer's (1979) guide for interpretation suggests that, as mentioned above, both distributions were close to symmetrical and the difference between them is negligible.

Reviewing the distribution of hybrid ratings (Figure 5) shows that there were several 'hot spots' in the distribution that were used particularly often, for example 1, 0.5, 0.2 or 0. The compacting of ratings at these points was a result of different levels of granularity used by respondents and is inherent with the method's design. The result however is that differences between ratings, specifically from different respondents are not explicitly interpretable, which reinforces the argument that ratings from the hybrid method should not be falsely treated as being an interval level. This conclusion brings up the question: Can respondents' ratings be compared at all, especially when different levels of granularity were used? The answer is: Yes, as long as the limitations of ordinal data are respected. For some comparisons it might be beneficial to round the ratings.

The general observations support the assumption that the hybrid method is more demanding, for respondents, as well as administrators. Although the usually greater demand on time can be interpreted as a downside, especially when quick ratings are needed, it seems to be acceptable considering that respondents reported that they thought more about their answers.

The reported splitting of the rating process was a surprising observation and might be the reason why the hybrid method worked as well as it did. Presumably, many respondents did not actively rate statements using their full range [min, max], but instead formulated one range [0, max] which they used to express their agreement and disagreement. They seem to have unwillingly used a range double the size than what was apparent to them. The fact that 10 was the most prevalent choice for minima and maxima is in accordance with the findings by Preston and Colman (2000) – who found respondents favoring the 10-point scale – while at the same time actually providing a 21-point scale, a length considered close to optimal by Matell and Jacoby (1971). Visually inspecting Figure 5 gives the impression that the splitting of the rating process caused two distinct bell-shaped distributions of the hybrid ratings, one for positive and one for negative ones. They seem to be divided by the neutral point (0) which shows a frequency that seemingly does not fit within the two

distributions. Therefore there might have actually been three distributions at play – positive ratings, negative ratings and ambivalent/indifferent ratings. One might reason that this division is artificially distorting the ratings. It can however be argued that attitudes are not one-dimensional attributes and the hybrid rating method allows for this characteristic to show.

When assessing the results from the supplementary questionnaire, keeping in mind that they are influenced by subjective judgments, the hybrid method and five-point method seemed very close, if not equal in the way they were perceived by respondents. Comments on all five questions were predominantly positive with similar numbers for both conditions; equivalent to the points of criticism which were brought up in both conditions with comparable frequencies.

As mentioned before, an accidental calculation of the intra-individual entropies for datasets with some items inverted revealed a drop in entropy for both conditions. This drop might be coincidental and without much meaning, but it coincides with the findings by Jonkisz et al. (2012) who found that items with reversed polarity were influenced by an additional factor. Whether or not this caused the drop in entropy remains unclear; it nevertheless brings up the question: Is the calculation of entropy a useful measure in the application of questionnaires? Given how the two respondents who only used three of five rating categories were easily identifiable by their low intra-individual entropies, it seems plausible that entropy values below a certain threshold might be used as a marker for response sets. Considering how well the calculation of entropy fits the data gathered by questionnaires and surveys, it is surprising that it is not used more frequently.

Given the observation that most respondents chose a range of [-10, 10] and presumably did not change it over the course of the questionnaire, a suggestion for cases in which quicker and less personalised ratings are needed, is to remove the open-ended aspect and set a fixed range of 10 for both directions. For this *closed hybrid rating method***,** a possible and much more simplified instruction could be: “[...] state whether you agree or not with + or – and specify how much you agree or disagree on a scale of 1 to 10. Assign 0 when your opinion is neutral”. This way the presumed benefit of splitting the rating process into two stages is an explicit part of the method, while complexity is greatly reduced. The ratings' granularity might still hold additional information as respondents can choose how many of the 21 points they wish to use. Obviously this way of rating would need its own evaluation and test for functionality.

Reflecting on this study and its design, reveals that the decision to avoid the use of external criteria, except for metrics from the NEO-PI-R manual, while necessary was not optimal. A design utilizing additional instruments such as retests would have better highlighted the method and its use. Furthermore, the pragmatic value aspect of this study heavily relied on the supplementary questionnaire for which the categorization solely relied on subjective judgments; this aspect would

have benefited from a more thorough, objective, and reliable design. However, this study also succeeded in highlighting the core aspects of attitude measurement with the hybrid rating method. The results provide a broad overview of the method and allowed for a good first analysis along with interesting observations.

In conclusion, this study has shown that the hybrid rating method is a functional method for measuring attitudes. It naturally exhibits downsides but has also shown characteristic advantages over related methods.

References

Bavaud, F. (2009). Information theory, relative entropy and statistics. In Sommaruga, G. (Ed.), *Formal theories of information: From Shannon to semantic information theory and general concepts of information* (pp. 54-78). Springer Berlin Heidelberg.

Bulmer, M. G. (1979). *Principles of statistics*. New York: Courier Dover Publications.

Cattell, R. B. & Warburton, F. W. (1967). *Objective Personality and Motivation Tests*. Uraban, IL.: University of Illinois Press.

Cicchetti, D. V., Shoinralter, D., & Tyrer, P. J. (1985). The effect of number of rating scale categories on levels of interrater reliability: A Monte Carlo investigation. *Applied Psychological Measurement*, 9(1), 31-36. doi:10.1177/014662168500900103

Cohen, L., Manion, L. & Morrison, K. (2000) *Research Methods in Education (5th edition)* London: Routledge Falmer.

Crichton, N. (2001). Visual analogue scale (VAS). *Journal Of Clinical Nursing*, 10(5), 697-706.

Dahl, F. A., & Østerås, N. (2010). Quantifying information content in survey data by entropy. *Entropy*, 12(2), 161-163. doi:10.3390/e12020161

Döring, N. (2006). Quantitative Methoden der Datenerhebung [Quanttitative Methods of Data Acquisition]. In Bortz, J., & Döring, N. (Eds.), *Forschungsmethoden und Evaluation [Research Methods and Evaluation]* (pp. 137-293). Springer Berlin Heidelberg.

Floridi, L. (2009). Philosophical conceptions of information. In Sommaruga, G. (Ed.), *Formal theories of information: From Shannon to semantic information theory and general concepts of information* (pp. 13-53). Springer Berlin Heidelberg.

Garland, R. (1991). The mid-point on a rating scale: Is it desirable. *Marketing bulletin*, 2(1), 66-70.

Hirsch, C. (2006). Inwiefern hängt die Präferenz eines multiple-choice-Antwortformats, mit wenigen (bis dichotom) oder mehr (bis Analogskala bei Persönlichkeitsfragebogen) Antwortmöglichkeiten mit der Persönlichkeit der Testperson zusammen? [How is the preference for a multiple-choice method with few (up to dichotomous) or many (up to analogue-scale) options linked to the respondent's personality?](Unpublished master's thesis). University of Vienna, Austria.

Jamieson, S. (2004). Likert scales: how to (ab) use them. *Medical education*, 38(12), 1217-1218. doi:10.1111/j.1365-2929.2004.02012.x

Jan, S. L., & Shieh, G. (2011). Optimal sample sizes for Welch's test under various allocation and cost considerations. *Behavior research methods*, 43(4), 1014-1022. doi:10.3758/s13428-011-0095-7

Jean Hausser and Korbinian Strimmer (2013). entropy: Estimation of Entropy, Mutual Information and Related Quantities. R package version 1.2.0. http://CRAN.R-project.org/package=entropy

Jonkisz, E., Moosbrugger, H., & Brandt, D. P. H. (2012). Planung und Entwicklung von Tests und Fragebögen [Planning and Development of Tests and Surveys]. In Moosbrugger, H., & Kelava, A. (Eds.), *Testtheorie und Fragebogenkonstruktion [Testtheory and Questionnaire-construction]* (pp. 27-74). Springer Berlin Heidelberg.

Kaplan, K. J. (1972). On the ambivalence-indifference problem in attitude theory and measurement: A suggested modification of the semantic differential technique. *Psychological Bulletin*, 77(5), 361. doi:10.1037/h0032590

Kuzon Jr, W. M., Urbanchek, M. G., & McCabe, S. (1996). The seven deadly sins of statistical analysis. *Annals of plastic surgery*, 37(3), 265-272.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology.*

Luce, R. D. (2003). Whatever happened to information theory in psychology?. *Review of general psychology*, 7(2), 183.

Matell, M. S., & Jacoby, J. (1971). Is There an Optimal Number of Alternatives for Likert Scale Items? Study I. *Educational and psychological measurement*, 31, 657-674. doi:10.1177/001316447103100307

Menold, N., & Bogner, K., (2015). Gestaltung von Ratingskalen in Fragebögen [Design of Rating-scales in Questionnaires]. *Mannheim, GESIS – Leibniz-Institut für Sozialwissenschaften (SDM Survey Guidelines)*. doi:10.15465/sdm-sg_015

Muck, P. M. (2004). Rezension des "NEO-Persönlichkeitsinventar nach Costa und McCrae (NEO-PI-R)"von F. Ostendorf und A. Angleitner [Review of the "NEO-Personalityinventory by Costa und McCrae (NEO-PI-R)" by F. Ostendorf und A. Angleitner]. *Zeitschrift für Arbeits- und Organisationspsychologie A&O*, 48(4), 203-210. doi:10.1026/0932-4089.48.4.203

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological bulletin*, 49(3), 197. doi:10.1037/h0055737

Ostendorf, F., & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae, Revidierte Fassung (NEO-PI-R)[Personality Inventory after Costa and McCrae, Revised (NEO-PI-R)]*. Göttingen: Hogrefe.

Parducci, A. (1965). Category judgment: a range-frequency model. *Psychological review*, 72(6), 407.

Simecek, P., Pilz, J., Wang, M., & Gebhardt, A. (2014). OPDOE: OPtimal Design Of Experiments. R package version 1.0-9. http://CRAN.R-project.org/package=OPDOE

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases

in behavioral research: a critical review of the literature and recommended remedies. *Journal of applied psychology*, 88(5), 879-903. doi:10.1037/0021-9010.88.5.879

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta psychologica*, 104(1), 1-15. doi:10.1016/s0001-6918(99)00050-5

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rasch, D., Kubinger, K. D., & Moder, K. (2011). The two-sample t test: pre-testing its assumptions does not pay off. *Statistical papers*, 52(1), 219-231. doi:10.1007/s00362-009-0224-x

Rasch, D., Kubinger, K., & Yanagida, T. (2011). *Statistics in Psychology using R and SPSS*. John Wiley & Sons.

Rasch, D., Pilz, J., Verdooren, L. R., & Gebhardt, A. (2011). *Optimal experimental design with R*. CRC Press.

Revelle, W. (2014) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, http://CRAN.R-project.org/package=psych Version = 1.4.8.

Robertson, J. (2012). Likert-type scales, statistical methods, and effect sizes. *Communications of the ACM*, 55(5), 6-7. doi:10.1145/2160718.2160721

Rohrmann, B. (1978). Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung [Empirical Studies on the Development of Rating-scales for Social-Scientific Research]. *Zeitschrift für Sozialpsychologie*, 9(3), 222-245.

Scherer, S., & Brüderl, J. (2010). Sequenzdatenanalyse [Analysing Sequential Data]. In Wolf, C., & Best, H. (Eds.), *Handbuch der sozialwissenschaftlichen Datenanalyse [Handbook of Social-Scientific Dataanalysis]* (pp. 1031-1051). VS Verlag für Sozialwissenschaften. doi:10.1007/978-3-531-92038-2_39

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1), 3-55.(Reprinted from The Bell System Technical Journal, 27, 379–423, 623–656. by C. E., Shannon, 1948). doi:10.1145/584091.584093

Stephenson, W. (1953). *The study of behavior: Q-technique and its methodology*. Chicago: University of Chicago Press.

Thurstone, L. L. (1994). A law of comparative judgment. *Psychological Review*,101(2), 266. doi:10.1037/h0070288

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*,

185(4157), 1124-1131. doi:10.1126/science.185.4157.1124

Vassend, O., & Skrondal, A. (1995). Factor analytic studies of the NEO Personality Inventory and the five-factor model: The problem of high structural complexity and conceptual indeterminacy. *Personality and Individual Differences*, 19(2), 135-147. doi:10.1016/0191-8869(95)00041-4

Vassend, O., & Skrondal, A. (2011). The NEO personality inventory revised (NEO-PI-R): Exploring the measurement structure and variants of the five-factor model. *Personality and Individual Differences*, 50(8), 1300-1304. doi:10.1016/j.paid.2011.03.002

Winkler, J. D., Kanouse, D. E., & Ware, J. E. (1982). Controlling for acquiescence response set in scale development. *Journal of Applied Psychology*, 67(5), 555.

Anhang

Abbildungsverzeichnis

Tabellenverzeichnis

Verfahrensmaterialien

An dieser Stelle sind nur die Materialien, die in der hybriden Bedingung verwendet wurden aufgeführt. Die 240 Aussagen des NEO-PI-R sind nicht aufgeführt.

Hybride Bedingung

Fragenheft

Vielen Dank für Ihre Teilnahme an dieser Studie. Diese Untersuchung erfolgt im Rahmen einer Diplomarbeit. Es soll durch die Bearbeitung eines Persönlichkeitsfragebogens eine neue Methode der Befragung untersucht werden. Es handelt sich nicht um einen Leistungstest und es gibt keine richtigen oder falschen Antworten.

- Die Teilnahme ist vollständig freiwillig und kann von Ihnen jederzeit ohne die Angabe von Gründen unterbrochen werden. Unvollständige Daten liefern leider keine auswertbaren Informationen, für die Untersuchung ist es daher vorteilhaft wenn Sie die Durchführung nicht abbrechen.

- Die von Ihnen gemachten Angaben erfolgen anonym und können nicht auf ihre Person zurück geführt werden. Die gesammelten Daten werden ausschließlich für diese Studie verwendet und werden nicht an Dritte weiter gegeben.

- Am Ende erhalten Sie die Möglichkeit sich für eine Benachrichtigung über die Ergebnisse der Untersuchung zu registrieren. Sollten Sie dies wünschen, beachten Sie bitte, dass dadurch die Anonymität nicht mehr gesichert ist.

Nehmen Sie zunächst den **Antwortbogen** zur Hand und füllen Sie die Felder auf der Vorderseite aus. Blättern Sie dann diese Seite um und lesen Sie die Anleitung auf der folgenden Seite

Bitte lesen Sie diese Anleitung sorgfältig durch. Schreiben Sie bitte alle Ihre Angaben auf den **Antwortbogen** in die dafür vorgesehenen Felder. **SCHREIBEN SIE BITTE NICHT** in dieses Heft.

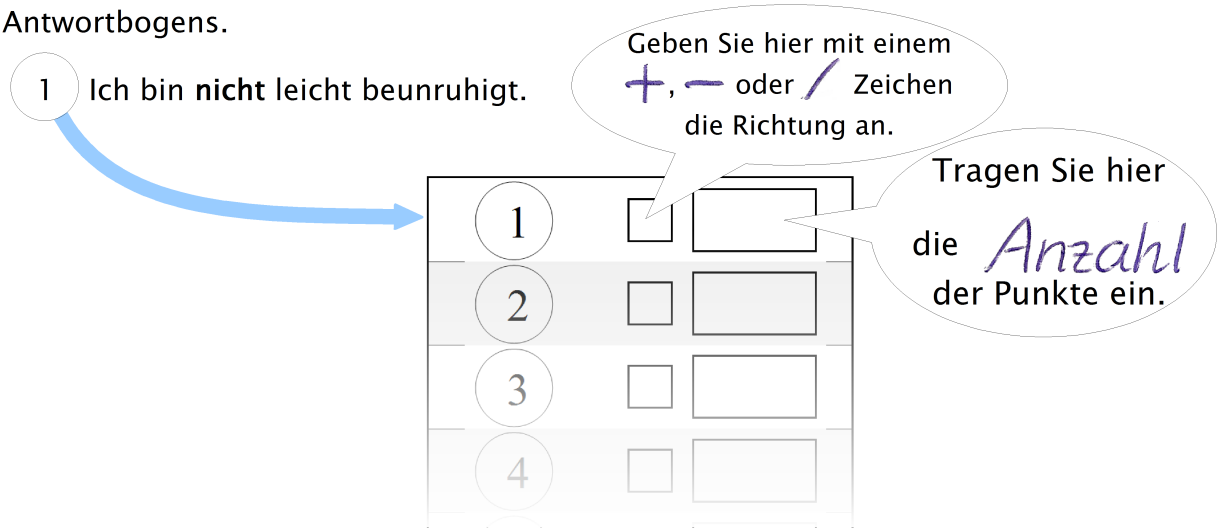
Auf den folgenden Seiten finden Sie insgesamt 240 Aussagen, die sich zur Beschreibung Ihrer eigenen Person eignen könnten. Lesen Sie bitte jede Aussage aufmerksam durch und überlegen Sie, wie sehr diese auf Sie zutrifft.

Wie sehr eine Aussage zutrifft sollen Sie durch das Vergeben von **Plus- oder Minuspunkten** ausdrücken.

Vergeben Sie so **viele oder wenig Punkte wie Sie wünschen**. Es gibt **keine „richtige“ oder „falsche“ Anzahl** an Punkten und auch **keine Begrenzung**; vergeben Sie die Punkte so wie Sie fühlen, dass Ihr Urteil am besten wiedergegeben wird. Ziel ist es, dass Sie intuitiv urteilen - antworten Sie also möglichst spontan und ohne viel nachzudenken.

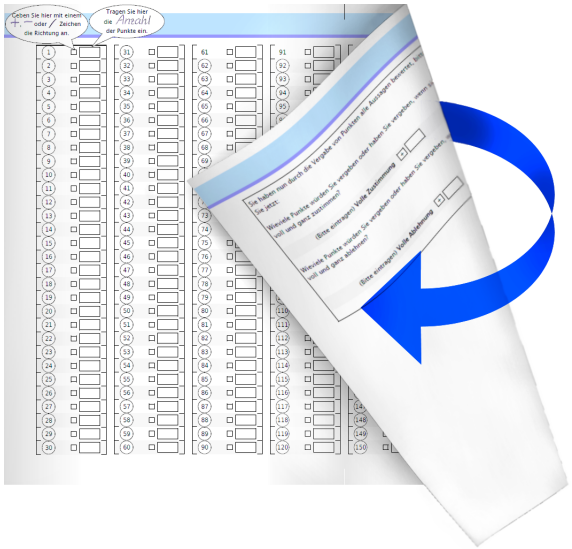
Wenn Sie eine Aussage vollkommen neutral bewerten, vergeben Sie weder Plus- noch Minuspunkte, also Null. Tragen Sie die Anzahl der Punkte, die Sie vergeben, auf dem Antwortbogen in das große Kästchen ein. Geben Sie im kleinen Kästchen mit einem Plus- oder Minuszeichen an ob die Punkte positiv oder negativ sind, streichen Sie das kleine Kästchen diagonal durch, wenn Sie null Punkte vergeben.

Beispiel: Hier sehen Sie eine Aussage aus dem Fragenheft und einen Ausschnitt des Antwortbogens.

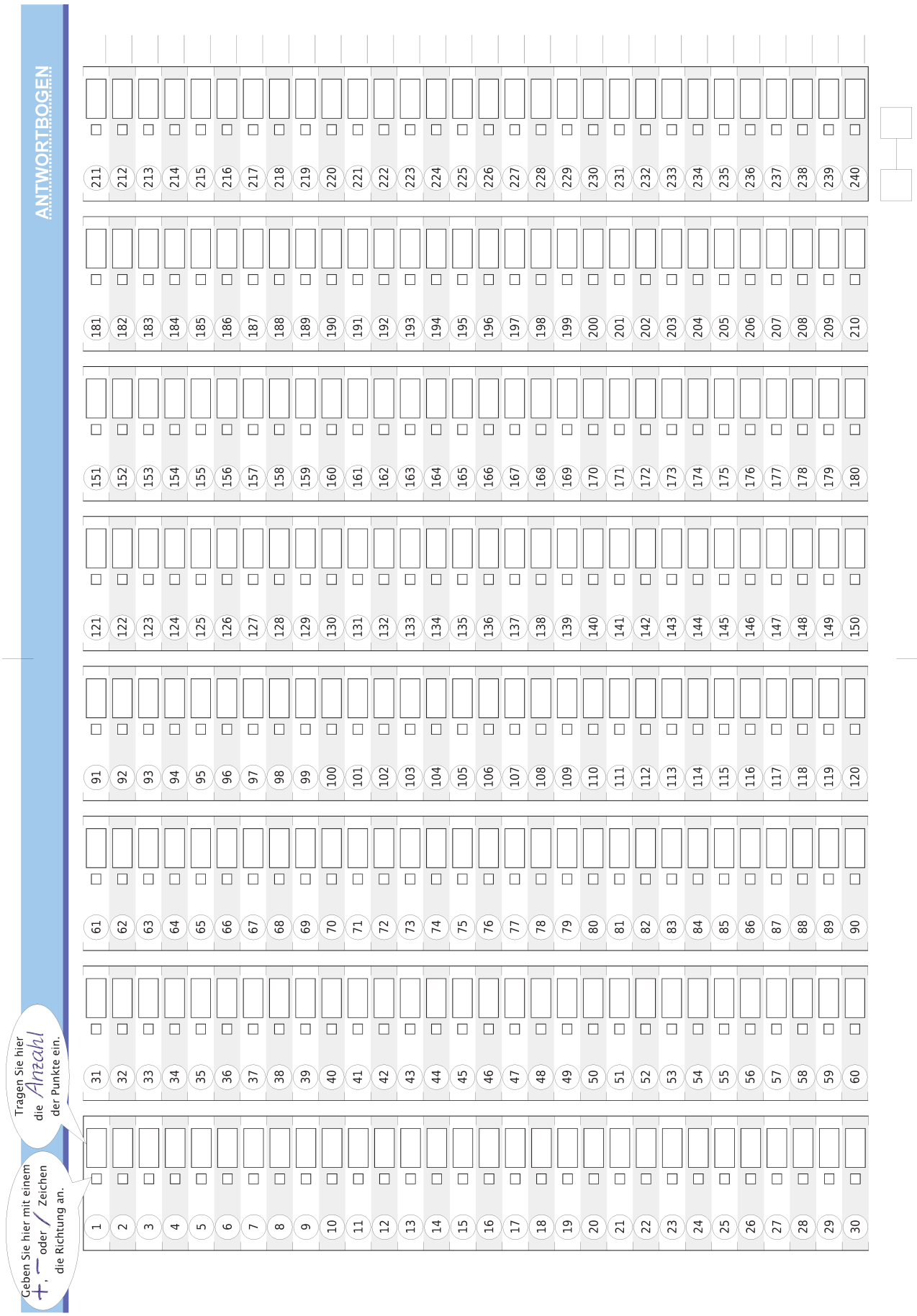


Bitte blättern Sie nun um und beginnen Sie mit der Bewertung der Aussagen. Schlagen Sie dazu den Antwortbogen auf und tragen Sie die Punktwerte in die vorgesehenen Kästchen.

Sie haben nun alle Aussagen beantwortet. Bitte schlagen Sie nun die Rückseite des **Antwortbogens** auf bearbeiten Sie die Abschlussfragen.



Antwortbogen



ANTWORTBOGEN

Sie haben nun durch die Vergabe von Punkten alle Aussagen bewertet, bitte überlegen Sie jetzt:

Wie viele Punkte würden Sie vergeben oder haben Sie vergeben, wenn Sie einer Aussage **voll und ganz zustimmen**?

(Bitte eintragen) **Volle Zustimmung** [+] ____

Wie viele Punkte würden Sie vergeben oder haben Sie vergeben, wenn Sie eine Aussage **voll und ganz ablehnen**?

(Bitte eintragen) **Volle Ablehnung** [-] ____

Wenn Sie über die Ergebnisse dieser Studie informiert werden wollen tragen Sie bitte ihre E-Mail Adresse hier ein: _____

(Bitte beachten Sie, dass die Anonymität dann nicht mehr gegeben ist.)

Vielen Dank für die Teilnahme an dieser Studie. Füllen Sie die folgenden Felder aus und blättern Sie dann im **Fragenheft** auf die nächste Seite.

Datum: _____

Geburtsjahr: _____ Geburtsmonat: _____

Geschlecht: ☐ Weiblich ☐ Männlich

Höchster abgeschlossener Bildungsabschluss:
- ☐ Hauptschulabschluss
- ☐ Lehrabschluss/Ausbildung
- ☐ Berufsbildende mittlere Schule/Realschule
- ☐ Matura/Abitur
- ☐ Hochschulabschluss/Studium
- ☐ Kein Bildungsabschluss
- ☐ Sonstiges: _____

Derzeitige Berufstätigkeit:
- ☐ Arbeiter/in
- ☐ Angestellte/r
- ☐ Arbeitslos
- ☐ Freiberuflich
- ☐ Selbstständig
- ☐ Geringfügig beschäftigt
- ☐ Hausfrau/-mann
- ☐ In Rente/Pension
- ☐ Beamte/r
- ☐ Schüler/in
- ☐ Student/in
- ☐ Sonstiges: _____

Haben Sie schon einmal einen Persönlichkeitsfragebogen beantwortet?
- ☐ Ja
- ☐ Nein

Blättern Sie nun auf die nächste Seite im **Fragenheft** und lesen Sie die Anleitung.

Zusatzfragebogen

Zusatzfragen

Bitte beantworten Sie noch kurz folgende Fragen, kurze Sätze oder Stichworte reichen aus:

- Wie beurteilen Sie die Bearbeitung des Fragebogens allgemein?

Im Fragebogen haben Sie Ihre Einstellungen mit einer besonderen Methode, die anfangs erklärt wurde, ausgedrückt.

- Wie leicht/schwer verständlich empfanden Sie die Anleitung?

- Wie leicht/schwer fiel es Ihnen, Sich eine Meinung zu den Aussagen zu bilden?

- Wie sehr/wenig hatten Sie das Gefühl, Ihre Meinung angemessen ausdrücken zu können?

- Wie leicht/schwer fiel Ihnen die Beantwortung des Fragebogens allgemein?

Lebenslauf

Johann-Christoph Münscher

*1987 in Hamburg, Deutschland

	Schule:
1993 – 1997	Grundschule Schulkamp in Hamburg
1997 – 2007	Bilinguales Abitur am Gymnasium Hochrad in Hamburg
	Studium:
Ab 2007	Diplomstudiengang Psychologie an der Universität Wien
	Inskription: 2007
	Beginn: 2008
	Vordiplom: 2011
2014	6-Wochenpraktikum in der Justizvollzugsanstalt Fuhlsbüttel, Hamburg
2014 – 2015	Diplomarbeit am Arbeitsbereich psychologische Diagnostik: begonnen SS 2014
	Sprachkenntnisse:
	Englisch (sehr gut in Wort und Schrift)