Forschungsbericht 2021-05

Identification and Compensation of Aberrant Response Patterns: Quality of Comfort Assessments in Human Subject Trials

Hans-Jürgen Hörmann and Kevin Schudlik

German Aerospace Center (DLR)
Institute of Aerospace Medicine
Hamburg/Germany

ISSN 1434-8454

ISRN DLR-FB--2021-05

Deutsches Zentrum
DLR für Luft- und Raumfahrt

Herausgeber
Deutsches Zentrum
für Luft- und Raumfahrt e. V.
Bibliotheks- und
Informationswesen
D-51170 Köln
Porz-Wahnheide
Linder Höhe
D-51147 Köln

Telefon (0 22 03) 6 01 - 44 44
Telefax (0 22 03) 6 01 - 47 47

ISSN 1434-8454

Thermal comfort, human subject-trials, aberrant responses, outlier, data cleaning

Hans-Jürgen Hörmann and Kevin Schudlik
Institute of Aerospace Medicine, Department of Aviation and Space Psychology, DLR, Hamburg

Identification and Compensation of Aberrant Response Patterns: Quality of Comfort Assessments in Human Subject Trials

Research project Next Generation Train (NGT-BIT)
DLR-Forschungsbericht 2021-05, 2021, 81 pages, 13 figs., 28 tabs., 69 refs., 23,00 €

Whenever human samples are used to assess a set of different items by using self-reports, concerns regarding the quality of data can arise because participants potentially reply without paying sufficient attention to the contents. The present study reviews different indices of aberrant response patterns and investigates this phenomenon by applying some indicators on subjective assessments of thermal comfort of N = 160 subjects within the DLR-project Next Generation Train (NGT). Based on this approach, a full sample was compared with a cleaned sample to examine whether aberrant responses have significantly biased the results. Overall, the differences were negligible, which is a proof for the data quality and the utility of human subject trials in comfort assessments. However, the usefulness of indices of aberrant response behaviour is still demonstrated and further suggestions to prevent and reduce aberrant responses in similar studies are discussed.

Thermokomfort, Probandenversuche, abweichendes Antwortverhalten, Ausreißer, Datenbereinigung

(Published in English)

Hans-Jürgen Hörmann und Kevin Schudlik
Institut für Luft- und Raumfahrtmedizin, Abt. Luft- und Raumfahrtpsychologie, DLR, Hamburg

Identifikation und Bereinigung von abweichendem Antwortverhalten: Qualität von Komforteinschätzungen in Probandenversuchen

Forschungsprojekt Next Generation Train (NGT-BIT)
DLR-Forschungsbericht 2021-05, 2021, 81 Seiten, 13 Bilder, 28 Tabellen, 69 Literaturstellen, 23.00 € zzgl. MwSt.

Immer dann, wenn subjektive Einschätzungen verschiedener Merkmale in Probandenversuchen erhoben werden, können Bedenken hinsichtlich der Datenqualität entstehen, ob die Probanden dem Inhalt der Fragen wirklich ausreichend Beachtung geschenkt haben. In der vorliegenden Studie werden verschiedene Indikatoren abweichenden Antwortverhaltens vorgestellt. In einem Anwendungsfall aus dem DLR-Projekt Next Generation Train (NGT) werden bei N = 160 ProbandInnen mehrere Indikatoren auf subjektive Einschätzungen zum Thermokomforts angewendet. Um zu bestimmen, inwieweit abweichendes Antwortverhalten die Ergebnisse signifikant verfälscht haben könnte, wurde die Gesamtstichprobe mit der bereinigten Stichprobe verglichen. Insgesamt sind die Unterschiede vernachlässigbar, was für die Datenqualität spricht und den Nutzen von Probandenversuchen für Komfortmessungen bestätigt. Die Anwendbarkeit von Indikatoren für abweichendes Antwortverhalten wird demonstriert und präventive Vorschläge zur Vermeidung und Kompensation in ähnlichen Studien diskutiert.

Forschungsbericht 2021-05

Identification and Compensation of Aberrant Response Patterns: Quality of Comfort Assessments in Human Subject Trials

Hans-Jürgen Hörmann and Kevin Schudlik

German Aerospace Center (DLR)
Institute of Aerospace Medicine
Hamburg/Germany

81 Pages
13 Figures
28 Tables
69 References

DLR **Deutsches Zentrum für Luft- und Raumfahrt**

Thermal comfort, human subject-trials, aberrant responses, outlier, data cleaning

Hans-Jürgen Hörmann and Kevin Schudlik
Institute of Aerospace Medicine, Department of Aviation and Space Psychology, DLR, Hamburg

Identification and Compensation of Aberrant Response Patterns: Quality of Comfort Assessments in Human Subject Trials

Research project Next Generation Train (NGT-BIT)
DLR-Forschungsbericht 2021-05, 2021, 81 pages, 13 figs., 28 tabs., 69 refs., XX,XX €

Whenever human samples are used to assess a set of different items by using self-reports, concerns regarding the quality of data can arise because participants potentially reply without paying sufficient attention to the contents. The present study reviews different indices of aberrant response patterns and investigates this phenomenon by applying some indicators on subjective assessments of thermal comfort of N = 160 subjects within the DLR-project Next Generation Train (NGT). Based on this approach, a full sample was compared with a cleaned sample to examine whether aberrant responses have significantly biased the results. Overall, the differences were negligible, which is a proof for the data quality and the utility of human subject trials in comfort assessments. However, the usefulness of indices of aberrant response behaviour is still demonstrated and further suggestions to prevent and reduce aberrant responses in similar studies are discussed.

Thermokomfort, Probandenversuche, abweichendes Antwortverhalten, Ausreißer, Datenbereinigung

(Published in English)

Hans-Jürgen Hörmann und Kevin Schudlik
Institut für Luft- und Raumfahrtmedizin, Abt. Luft- und Raumfahrtpsychologie, DLR, Hamburg

Identifikation und Bereinigung von abweichendem Antwortverhalten: Qualität von Komforteinschätzungen in Probandenversuchen

Forschungsprojekt Next Generation Train (NGT-BIT)
DLR-Forschungsbericht 2021-05, 2021, 81 Seiten, 13 Bilder, 28 Tabellen, 69 Literaturstellen, XX.XX € zzgl. MwSt.

Immer dann, wenn subjektive Einschätzungen verschiedener Merkmale in Probandenversuchen erhoben werden, können Bedenken hinsichtlich der Datenqualität entstehen, ob die Probanden dem Inhalt der Fragen wirklich ausreichend Beachtung geschenkt haben. In der vorliegenden Studie werden verschiedene Indikatoren abweichenden Antwortverhaltens vorgestellt. In einem Anwendungsfall aus dem DLR-Projekt Next Generation Train (NGT) werden bei N = 160 ProbandInnen mehrere Indikatoren auf subjektive Einschätzungen zum Thermokomforts angewendet. Um zu bestimmen, inwieweit abweichendes Antwortverhalten die Ergebnisse signifikant verfälscht haben könnte, wurde die Gesamtstichprobe mit der bereinigten Stichprobe verglichen. Insgesamt sind die Unterschiede vernachlässigbar, was für die Datenqualität spricht und den Nutzen von Probandenversuchen für Komfortmessungen bestätigt. Die Anwendbarkeit von Indikatoren für abweichendes Antwortverhalten wird demonstriert und präventive Vorschläge zur Vermeidung und Kompensation in ähnlichen Studien diskutiert.

Contents

Figures

Tables

I Abstract

Whenever human samples are used to assess a set of different items by using self-reports, concerns regarding the quality of data can arise because participants potentially reply without paying sufficient attention to the contents. The present study reviews different indices of aberrant response patterns and investigates this phenomenon by applying some indicators on subjective assessments of thermal comfort of N = 160 subjects within the DLR-project Next Generation Train (NGT). Based on this approach, a full sample was compared with a cleaned sample to examine whether aberrant responses have significantly biased the results. Overall, the differences were negligible, which is a proof for the data quality and the utility of human subject trials in comfort assessments. However, the usefulness of indices of aberrant response behaviour is still demonstrated and further suggestions to prevent and reduce aberrant responses in similar studies are discussed.

II Acknowledgements

First and foremost, we want to thank our colleague Daniela Bauer for her contributions to this report. Her participation in the discussions during the preparation of this report was very helpful in determining the scope of the report. Her ideas for the data analyses with R and Excel were valuable inputs for the present version of this report. We would also like to thank Daniela Bauer for text formatting and providing the list of references.

We are grateful to Dr. Richard Yentes and his colleague F. Wilhelm for sharing and maintaining the R-package "Careless", which makes the identification of aberrant responses really convenient.

We also thank Prof.Dr. Dirk Stelling and Dr. Julia Maier for reviewing an earlier version of this report and providing helpful comments.

Lastly, we thank the DLR Division of Transport (PD-V) for the financial support of the project Next Generation Train (NGT).

1 Introduction

Thermal comfort in regard to one's surrounding is defined by the ASHRAE Standards 55 (2013) as the "condition of mind that expresses satisfaction with the thermal environment and is assessed by subjective evaluation" (p 3). According to this definition human subject trials are indispensable when modern heating, ventilation, and air conditioning systems (HVAC) are designed for example for office buildings or passenger transportation systems. Usually, in subject trials substantial variations between answers will be observed when asking individuals to assess their personal satisfaction with essentially the same environmental conditions. Different thermal preferences, physiological factors or simply clothing can account for some of these individual differences. However, the question remains how sustainable individual differences in climate satisfaction scores are for dimensioning HVAC systems and to what extend specific sources of judgment bias could impair findings or conclusions.

As part of its Next Generation Train project (NGT), the German Aerospace Center (DLR) has conducted a number of experimental studies with human subject trials to examine the effects of different ventilation concepts and air temperatures on passengers' thermal comfort in a generic train laboratory (e.g., Hörmann et al., 2017; Lange et al. 2019; Schmeling et al., 2019). To date a total number of $N = 276$ male and female subjects between 16 and 65 years of age participated in these trials. With a standardized climate comfort assessment inventory (Marggraf-Michel & Jaeger, 2007), all subjects assessed the comfort of different climate parameters related to air temperature, air draught[1] and humidity after a certain exposure time to a number of different climate scenarios. The goal of this project is to enhance general passenger satisfaction with the climate conditions created by novel HVAC systems in next generation trains.

As expected, first inspections of the data revealed some amount of variation between the opinions of the different subjects. Therefore, the question came up whether simply averaging these scores across the entire sample adequately reflects the wellbeing of all subjects or whether valuable information might get lost. Alternatively, the calculation of Fanger's comfort estimates (Predicted Mean Vote or Predicted Percentage of Dissatisfied, e.g., Fanger, 1973; ISO 7730) also does not solve this problem because predictions from these

[1] We use assessment of air draught in this document when the subjective reflection of air velocity is meant

equations only aim to satisfy "the greatest number of people" (p. 317) without considering individual differences. Of course, to some extent judgment error could account for the variation of subjective comfort assessments, but if the remaining variance reflects persisting differences between the subjects, these assessments could be a valuable source of information to further enhance thermal comfort, especially if erroneous judgments could be identified and removed from the sample.

The intention of this study is to explore several standardized methods to identify erroneously biased response patterns in subjective comfort assessments and to remove subjects' scores if flagged as aberrant or careless. The effects of this data-cleaning procedure on subjective evaluation scores for thermal comfort are reported in this document. For these analyses, a subset of $N = 160$ subjects with comparable data was retrieved from the NGT thermal-comfort database. The extent and the sources of aberrant responses are examined and discussed. In the conclusions, several means are proposed to further improve the quality of comfort assessments in future subject trials.

2 Identification of survey data with deficient quality

Concerns about the quality of survey or questionnaire data due to careless participants have been critically discussed by psychometric researchers for a long period of time (e.g., Schmitt & Stults, 1985; Groves, 1987). Besides careless response patterns (Meade & Craig, 2012; Maniaci & Rogge, 2014), other authors have labelled this phenomenon as insufficient effort responding (IER; e.g., Huang et al., 2012; McGonagle et al., 2016), or random response behaviour (e.g., Johnson, 2005; Maniaci & Rogge, 2014). The specific labels often refer to a specific form of careless response behaviour (Curran, 2016). The expression "aberrant response patterns" has been used as an overarching term (e.g., Niessen et al., 2016; Yu & Cheng, 2019).

From an applied point of view, the various concepts refer to similar problems. That is, respondents might not pay sufficient attention when completing self-report measures or deliberately alter their responses (i.e., faking) which could introduce error into a dataset (Kim & Moses, 2018; Maniaci & Rogge, 2014) and thus potentially distort the results as well as the implications for practitioners and scientists alike (Yentes, 2020). Edwards (2019) suggested that besides simple inattentive responding, respondents' personality, ability and motivation, the instrument design (e.g., instructions, length, or item construction and organization), and the method of data collection might influence the response validity as well. Goldammer et al. (2020) found careless responding to inflate item variances, to bias item means towards the scale midpoint, to increase residual variance of construct indicators, and to reduce the within-group agreement on consensus-based constructs. These influences can reduce the amount of valid variance in the variables under investigation and can therefore have detrimental effects on the credibility of research findings. Despite its significance, few scientific research articles state if or how they have addressed potential aberrant response behaviours. Studies have found varying degrees of aberrantly responding participants ranging from 3% up to over 40% of the respective samples (Berry et al., 1992; Johnson, 2005; Oppenheimer et al., 2009; Meade & Craig, 2012; Maniaci & Rogge, 2014; Nichols & Edlund, 2020). Both, the wide range of estimates as well as the ignorance in many research articles might be due to the lack of a clear indicator for aberrant responses. Below, we introduce a variety of indicators which can be used to identify aberrant response patterns in human subject trials. Huang et al. (2015) have

shown that the rigorous use of these indicators can be a worthwhile tool to partially remove error variance that would otherwise impact survey results. In applied research, this might even have consequences for the practical suggestions derived from a certain data set (Abbey & Meloy, 2017).

2.1 Approaches and indices

Different attempts to categorize and distinguish between different indices and approaches haven been proposed. For instance, Huang et al. (2012) and Niessen et al. (2016) both discriminated between a) consistency indices, b) infrequency indices, c) response patterns, and d) response time. Curran (2016) did not categorize the different indices but focused on each one individually. For the present work, we use the established distinction by DeSimone et al. (2015), which was later adjusted by DeSimone and Harms (2017), who discriminated between a) direct, b) unobtrusive (previously archival), or c) statistical techniques and indicators for data screening to detect aberrant response patterns.

2.1.1 Direct indices

Direct indices refer to the insertion of specific self-report items into a survey prior to its administration. As the label implies, these items overtly display what they aim to assess (see examples below). Thus, participants will most likely be aware of the purpose of these items. This awareness can potentially lead to distorted replies due to social desirability.

2.1.1.1 Self-reports

Self-report indices of data quality might be the most basic method to assess if a participant has invested sufficient effort on a survey. Meade and Craig (2012) labelled single items as self-reported single item (SRSI) indicators. The items (e.g., "Also, often there are several distractions present during studies (other people, TV, music, etc.). Please indicate how much attention you paid to this study. Again, you will receive credit no matter what. We appreciate your honesty!") are presented at the end of a survey and instructed as pivotal to the data quality since only such responses can be used for the analyses that have been answered with the full attention of the participants. Multiple SRSIs measuring different aspects (e.g., effort and attention) can be implemented. Meade and Craig (2012) found SRSI indicators to be moderately correlated with each other and to have mixed utility in regard to identifying aberrant response patterns. A self-report whether or not the responses should be used for data

analyses showed the best results of all SRSI indices.

2.1.1.2 Instructed items

Instructed items can be used to assess the attention of respondents during the survey (e.g., "Please leave this item blank"). This is done by using explicit instructions to certain items. The underlying assumption is that attentive respondents will comply with the given instruction (e.g., "chose answers A and C"). However, DeSimone et al. (2015) pointed out that respondents vary in regard to their attention over the course of the survey. This kind of items have also been labelled as "Screeners". Berinsky et al. (2019) advised researchers to use multi-item scales that include Screeners with high and low passage rates and to tailor a set of attention checks specific to the given research needs.

2.1.1.3 Bogus-items

Lastly, the usage of Bogus-items can be helpful to detect aberrant response patterns by containing content that is either obvious or ridiculous (DeSimone et al., 2015). For instance, an item like "I have 17 fingers on my left hand" is supposed to ensure the same answer from all respondents. It is suggested to implement multiple Bogus-items at multiple points of the survey and screen the respondents who endorse at least one of the items (Bagby et al., 1991; DeSimone et al., 2015). However, Edwards (2019) pointed out that screening items can be seen as "trick" questions and therefore annoy respondents and diminish the willingness of cooperation between the respondents and researchers. Bogus-items have also been used to detect deceptive impression management attempts (e.g., Levashina et al., 2009).

2.1.2 Unobtrusive indices

Unobtrusive indices revolve around the examination of patterns of response behaviour over the course of the survey (DeSimone et al., 2015). Although less obvious to the respondents, it might still be apparent to the respondents that certain response patterns are not desirable (e.g., always giving the same answer). Respondents might therefore attempt to avoid being overtly suspicious.

2.1.2.1 Semantic synonyms and antonyms

Semantic synonyms and antonyms are items that are theoretically

assumed to be either positively or negatively related to each other (e.g., "I prefer warm temperatures" should be positively associated with "I feel comfortable in environments with warm air"). Respondents who give dissimilar replies to similar items (in case of synonyms), can be identified as giving aberrant responses. This is conceptually related to infrequency indices that use items on which almost all non-aberrant participants will provide the same or very similar answers (Huang et al., 2012). For instance, Kam and Chan (2018) found positive correlations between respondents' replies to synonyms and instructed response items and a negative correlation with an acquiescence response style (i.e., the tendency to agree with items regardless of their content). Unsurprisingly, they also reported a significant negative correlation between instructed response items and the antonym indicator. However, it has to be noted that Kam and Chan (2018) included a larger number of semantic synonyms and antonyms in their study, but created psychometric synonyms and antonyms out of the semantic ones for their subsequent analyses (see below for further details on psychometric synonyms / antonyms). Thus, their construction can be seen as a hybrid version between semantic and psychometric synonyms and antonyms.

2.1.2.2 Response time

Multiple studies have considered the response time as an indicator for aberrant response behaviour. This approach relies on the assumption that study participants require a minimum amount of time, given that they are attentive, to read an item and answer accurately. Huang et al. (2012) suggested that it is unlikely for participants to answer items faster than with a rate of 2 seconds. While Meade and Craig (2012) found large differences in response times between participants, they still suggest to cut-off clear outliers on the low end of the distribution of response time as careless respondents. Huang et al. (2012) also considered page time on a survey to be a promising approach with sensitivity values up to 49%. Similarly, Niessen et al. (2016) reported a sensitivity of up to .51 (with a 20% proportion of careless respondents) for response time when cut-offs were derived on an empirical basis.

2.1.2.3 Longstring

A lengthy string of invariant responses (Longstring) can be indicative of aberrant response behaviour. This indicator relies on the assumption that too many consecutive identical responses may be indicative of a lack of effort (DeSimone et al., 2015). The Longstring can be further divided into the

maximum Longstring and the average Longstring. The maximum Longstring approach has been shown to empirically detect aberrant response behaviour (Meade & Craig, 2012; DeSimone et al., 2015; Niessen et al., 2016; Ward & Meade, 2018). However, it has to be noted that the Longstring indicator heavily depends on the number of items and structure of the scales and questionnaires at hand. For instance, a 30 item-scale would require a different cut-off value than a 10 item-scale. Given that no objective cut-off values are established yet (e.g., also pointed out by Niessen et al., 2016); this might be a pitfall in the application. Curran (2016) suggested to use half of the items of an instrument as a possible cut-off value but also noted that this approach might be too strict for scales with similar items and that the lack of specific cut-off values stems from the scale-specificity of the Longstring approach. Huang et al. (2012) approached possible cut-off values based on the works of Costa and McCrae (2008). Costa and McCrae (2008) suggested maxima of longest response strings for their NEO-PI-R (300 items) of 6, 9, 10, 14, and 9 times for the respective scale points (based on a 5-point scale coded from strongly agree to strongly disagree). In others words, none of their 983 cooperative participants selected the respective responses more often than the previously stated number of times. Johnson (2005) also chose these cut-offs for his study.

2.1.2.4 Intra-individual response variability

The intra-individual response variability (IRV) index has been introduced as an extension of the Longstring index (Dunn et al., 2018). Both concepts have been shown to be positively correlated with each other (DeSimone & Harms, 2017). The IRV is defined as the standard deviation of a respondent's replies to all items on a questionnaire or a selected sub set of items (Dunn et al., 2018). In comparison to the Longstring indicator, the IRV is not as apparent to the respondents (e.g., replying with 1 for 5 times in a row obviously seems unwanted but switching the reply between 1 and 2 back and forth seems to be less obvious). Dunn et al. (2018) did not provide a clear cut-off value to go with (DeSimone & Harms, 2017), but suggested that the IRV works best when it is applied to a set of 25 to 150 items. They also pointed out that researchers should be looking for rather extreme values of IRV to prevent them from excluding respondents who, for instance, simply answered the items with a central tendency response style. The cut-off value should also is dependent on the specific scale at hand. Thus, the cut-off for a uniform and homogenous scale should be higher (i.e., less deviation from one answer to another would be expected) than a comparable multidimensional scale using both positively and

negatively worded items (DeSimone et al., 2015). Dunn et al. (2018) and DeSimone and Harms (2017) both flagged the respondents with the relative lowest IRV (approximately 10%) as possible aberrant responders. The IRV has been empirically tested and has proven to be a useful indicator of aberrant response behaviour (e.g., Laconelli & Wolters, 2020). The IRV has also been utilized in applied research (e.g., Sagui-Henson et al., 2018). Marjanovic at al. (2015) proposed the inter-item standard deviation (ISD) as an intrapersonal measure of response variance calculated at the individual level which is conceptually related to the IRV. Contrary to flagging individuals with the lowest IRV (as suggested by Dunn et al., 2018), Marjanovic et al. (2015) suggested to flag individuals with the highest IRV as this might be indicative of random response behaviour. The difference is that Marjanovic et al. (2015) calculated unidimensional IRV scores separately for each construct of the NEO-FFI. On the contrary, Dunn et al. (2018) recommended to calculate the IRV across all items (including reverse coded items) representing several different constructs to ensure response variability. They also suggested to calculate a series of IRV scores for an individual across various sections of the questionnaire. If the item content of IRV is heterogenous then values at the extreme lower end of the IRV distribution should be regarded as indication of aberrant response behaviour.

2.1.3 Statistical indices

Statistical indices are based on the post-hoc calculation of indicators for data quality. As they are computed after the data is gathered, they do not require the survey to be modified, although a carefully constructed survey could be helpful to address some of the indicator-inherent limitations. DeSimone et al. (2015) pointed out that statistical indices were often developed to identify extreme responses but can also be used to detect aberrant response patterns.

2.1.3.1 Psychometric synonyms and antonyms

Psychometric synonyms (Meade & Craig, 2012) and antonyms (Goldberg, 2000, as cited in Johnson, 2005) are conceptually related to their semantic counterparts. Instead of relying on semantic relations, however, psychometric synonyms rely on item pairs based on their positive correlations. For antonyms, negative correlations are considered. Thus, whereas semantic synonyms and antonyms require content experts to identify similar items, psychometric synonyms and antonyms identify similar items by using the magnitude of the inter-item correlations. Fundamentally, both the semantic as well as the

psychometric approach assume that respondents will give similar answers to similar items. For instance, the assessment of air draught on the left and right leg in a symmetrical environment should constitute an item pair of synonyms. In comparison to the semantic approach (see 2.1.2.1), the psychometric approach has the advantage that the possibility of bias or post hoc adjustments influencing the number of pairs identified is minimized (DeSimone et al., 2015). Meade & Craig (2012) suggested a threshold of .60 (magnitude of correlation) for synonyms and -.60 for antonyms. Alternatively, one could also determine a specified number of item pairs and use either the highest or the lowest correlations (Goldberg, 2000, as cited in Johnson, 2005). However, it should be noted that the correlational cut-off still is somewhat arbitrary and should be adjusted for the given research question and questionnaire. For instance, Maniaci and Rogge (2014) used cut-offs of .64 for psychometric synonyms and -.49 for antonyms based on 5 pairs each (i.e., the specific cut-offs resulted from a relative approach). After the identification of statistical item pairs, the psychometric synonym or antonym index is computed as the within-person correlation across the item pairs that were previously identified (Meade & Craig, 2012). Respondents, with within-person correlations that are deemed too small in absolute terms are then regarded as aberrant response behaviour (Edwards, 2019). For instance, Meade & Craig (2012) flagged respondents with psychometric synonym coefficients below .22 (DeSimone et al., 2015). Concerning psychometric antonyms, Huang et al. (2012) flagged respondents with psychometric antonym coefficients greater than -.03 (DeSimone et al., 2015).

2.1.3.2 Mahalanobis Distance

The Mahalanobis Distance (Mahalanobis, 1936; De Maesschalck et al., 2000) can be used to detect aberrant response patterns as well (Yentes, 2020). It can be used for the analysis of multivariate outliers as it estimates the multivariate distance between the respondent's scores and the sample mean scores on the given items while taking the item intercorrelations into account (DeSimone et al., 2015; Edwards, 2019). Simply put, Mahalanobis *D* extends the normal outlier analysis into multivariate space (Curran, 2016). The underlying assumption when used for the detection of aberrant response patterns is that severe deviations from the normative response (in the given sample) might indicate insufficient effort or a conscious distortion. Another advantage of the Mahalanobis *D* statistic is that the square of *D* (D^2) is distributed as a chi-square variable with degrees of freedom equal to the number of items used to calculate

it (k). Therefore, a justified empirical cut-off is naturally given and doesn't need to be iteratively determined as is the case for most of the previously introduced indices. Technically, this is done by converting the Mahalanobis Distance to Chi-Square values (Zijlstra et al., 2011). Afterwards, a conventional critical *p*-value can be chosen to flag aberrant respondents. However, it has to be noted that the distribution of the D^2 chi-square variable assumes that the items are normally distributed (Curran, 2016). Conventionally, researchers might then identify the respondents in the top five percent of the chi-square distribution as having potentially shown aberrant response patterns. Curran (2016) suggested to use Mahalanobis *D* to flag individuals for deeper examination as this metric has only been tested on a limited number of occasions in the context of aberrant response patterns. Ehlers et al. (2009) have found that the Mahalanobis Distance is useful to identify inattentive responses. Furthermore, Mahalanobis *D* has been shown to correlate with other indices (Maniaci & Rogge, 2014; Meade & Craig, 2012). Meade and Craig (2012) too, found Mahalanobis *D* to be a useful metric, but also emphasized that the frequency distributions of a sample need to be inspected in order to identify a suitable cut-off before applying it.

2.2 Sensitivity of detection indices for different aberrant response patterns

Given the different characteristics of each indicator, it is understandable that they identify different kinds of aberrant response patterns. For instance, an applicant that has solely replied in a random fashion would not be detected by the Longstring index. This person would instead be flagged by either a high IRV or unfitting values on the psychometric synonyms or antonym indices. On an empirical basis, DeSimone et al. (2015) used a generated data set with four different kinds of aberrant response patterns (random, invariant, acquiescent, and extreme replies) to test the different detection indices. They have shown that each indicator fits best to a certain type of aberrant response behaviour which shows the necessity for researchers to implement more than just one kind of indictor. Meade and Craig (2012) have found similar results but also computed an exploratory factor analysis using a variety of indicators to statistically identify which indicators load on the same factors. They found psychometric synonyms, psychometric antonyms, and the Mahalanobis Distance as well as Bogus-items to load on the same factor and therefore capturing the same type of aberrant response behaviour despite the conceptual differences. On the contrary, self-report measures and the Longstring approach loaded on a

different factor. IRV was not part of this analysis but can be classified by using conceptual reasoning (Dunn et al., 2016). High IRV values might indicate haphazard and random responding without taking the item content into account. On the contrary, low IRV values might indicate that respondents also ignored the item content but decided to give the same or a very similar answer to every item instead.

Another system of different aberrant response types has been proposed by Kim and Moses (2018). They suggested that, based on a forced choice-format, six different aberrant response types exist. The first type describes random responding that can occur due to indecisiveness. Secondly, faking can occur because respondents give socially desirable statements instead of responding truthfully. Third, Kim and Moses (2018) introduced the aberrant response type mechanical responding. Mechanical responding means that systematic patterns appear in a sequence of option choices (i.e., respondents frequently chose the first response option). Fourth, the response type limited responding describes that, based on multidimensional pairs, one particular dimension is rarely chosen. Fifth, Kim and Moses (2018) suggest that peculiar groups show specific response patterns. The latter might appear because examinees with different cultural backgrounds or language skills, show atypical response patterns. Lastly, they mention omitting as a response style. Respondents might omit certain items because the items appear to be somewhat sensitive or negative (Kim & Moses, 2018). As the different response types / styles significantly differ from each other, specific detection indices for the respective aberrant response types are required.

2.3 Causes of aberrant responses and challenges in applied settings

Given above arguments, it is reasonable to assume different causes of aberrant response behaviour. Edwards (2019) listed multiple possible causes including characteristics of the respondents, the instrument, and the method of data collection. Concerning the individual respondents, their personality plays a role in the determination of what kind of reply patterns they use. Especially individuals high in conscientiousness and agreeableness have been linked to a higher response validity (Dunn et al., 2018; Maniaci & Rogge, 2014). Edwards (2019) argued that this is in line with the conceptual nature of these personality dimensions (e.g., McCrae & Costa, 2003). Compared to less conscientious persons, a highly conscientious person should thus be likely to complete a

questionnaire in a more careful, effortful, and thoughtful way to provide useful information. The second cause of aberrant responses lies within the respondent's ability. This interacts with the respective questionnaire at hand. For instance, the respondent needs to possess a sufficient verbal comprehension, vocabulary, and general language skill to produce valid responses (Curran, 2016; Edwards, 2019; Johnson, 2005). Furthermore, surveys might also require respondents to recall certain experiences and events which will produce inter-individual differences since memory capacity differs between individuals. Essentially, respondents need to fulfill demands (e.g., communicational skills and information recall) which might lead to aberrant responses if the respondent lacks a certain ability. Third, the respondents' motivation to participate plays a crucial role in the response validity. Naturally, respondents invest time and effort into deliberately answering survey items. Rationally speaking, they might intend to minimize the time and effort they invest in completing the survey (Dunn et al., 2018). This assumption is also made when the response time is used to flag aberrant responses (see above). However, Maniaci and Rogge (2014) as well as Meade and Craig (2012) noted that respondents can also be intrinsically motivated to complete a survey when the topic of the survey appeals to them. Extrinsic motivation is usually created through incentives (e.g., monetary rewards) but the payment can be linked to the requirement of completing the survey with care and attention (Abbey & Meloy, 2017). However, a high motivation can also have its downsides. For instance, if applicants try to reply in a socially desirable way (Curran, 2016; Levashina & Campion, 2009).

Fourth, Edwards (2019) mentions the characteristics of the instrument itself as a facilitating factor of response validity. This includes the construction of items as well as their organization within a survey. This specifically concerns clarity and simplicity of the wording, open versus closed response formats, and conversational norms that items invoke. Furthermore, effects of the item order and further contextual factors (e.g., blocks of uniformly positive and / or negative items) have been known for many years (Harrison & McLaughlin, 1993; Knowles, 1988). Next, every survey and set of items is accompanied by certain instructions for the respondents. Instructions can either refer to extrinsic incentives (e.g., an identification warning that careless responses lead to exclusion from payment) to emphasize that the outcome for the respondent depends on the effort they invest in the survey (Abbey & Meloy, 2017; Huang et al., 2012). Instructions can also refer to the importance of respondent's effort, attention to the study, and the overall quality of the obtained data (Meade &

Craig, 2012). In applied contexts, this might also be extended to the notion that useful conclusions can only be derived with sufficient response quality. The effects of survey length have been studied extensively. Long surveys have been linked to fatigue and a potential loss of interest (Edwards, 2019; Meade & Craig, 2012). This can either lead to a high dropout rate or a diminished response validity. Lastly, the format and method of data collection also impacts respondents' response behaviour. Especially online studies have been of some concern to researchers (Fleischer et al., 2015). Concerns regarding the possibility of distraction and the lack of control over environmental circumstances have been brought up (Johnson, 2005; Meade & Craig, 2012). However, a definitive answer as to what method of data collection is superior remains open and requires further research.

3 The present study

The literature points out a variety of causes of response validity, aberrant response behaviour, and indices that attempt to identify aberrant respondents. However, it remains unclear as to how much practical impact the application of these indices can have in different applied settings. Investigating thermal comfort in human subject trials provides an applied environment to examine aberrant response patterns. The validity of subjective measures from humans being exposed to different climate scenarios is an essential precondition for successfully customizing novel HVAC systems according to human needs. Thus, the present study is based on data from past studies on thermal comfort in next generation long-distance trains conducted by the German Aerospace Center (DLR).

As pointed out above, it remains unclear to what extent aberrant response behaviours depend on the respondent's characteristics (e.g., personality) or characteristics of the measurement (e.g., length of a questionnaire). For instance, multiple studies have shown positive effects of the personality dimensions conscientiousness and agreeableness on response validity (Dunn et al., 2018; Maniaci & Rogge, 2014; Marjanovic et al., 2014; Ward et al., 2017). On the other hand, negative effects of questionnaire characteristics (e.g., length) on response validity have been reported as well (Bowling et al., 2020; Eisele et al., 2020; Gibson & Bowling, 2020). Investigating this subject is complicated because different indicators for aberrant response behaviours reflect different types of low-quality response patterns (DeSimone et al., 2015; Meade & Craig, 2012). Thus, further research with multiple indicators is required. We approach these issues by using repeated measures data to identify if the indicators remain stable over different points in time indicating a greater trait rather than state influence. On the other hand, if respondents change significantly in their carelessness over the course of a questionnaire, a higher influence of the questionnaire itself can be assumed (Bowling et al., 2020).

Research question 1 (RQ 1): Does the extent of aberrant responses in subjective assessments of thermal comfort increase with the length of the experiment? We expect that the extent of aberrant responses increases, the longer the data gathering procedure lasts.

Research question 2 (RQ 2): Is there any evidence for a larger number of aberrant responses in subjective assessments of thermal comfort

compared to other questionnaires? With regard to the instruments, we expect a higher extent of aberrant responses for the climate assessments than for the personality ratings because the climate assessments have a more repetitive and therefore potentially more boring nature.

Research question 3 (RQ 3): Are the indicators for aberrant response behaviour in climate assessments stable over repeated measurements at consecutive points in time and across different instruments? Intercorrelations of each indicator across time for the same instrument should be significant and higher than those with other instruments.

Research question 4 (RQ 4): Are the indicators for aberrant response behaviour significantly correlated with each other? To the extent that the indicators represent similar patterns of response behaviour (e.g. LOS and IRV) there should be a positive correlation. Generally, there should not be any contradicting relations.

Research question 5 (RQ 5): Are the indicators for aberrant response behaviour in climate assessments significantly correlated with the Big Five personality traits? In line with previous research we expect the extent of aberrant responses is related to some degree to personality dimensions, especially conscientiousness.

By applying some of the aforementioned methods, the quality of subjective measurement can be enhanced (Edwards, 2019; Goldammer et al., 2020). DeSimone et al. (2015) discouraged correcting a sample by deleting the data of aberrant respondents, but analyses should be calculated for both cases instead: with and without a reduction of the sample. An improvement in data quality could thus impact the results and therefore the implications of applied research as well. However, given the relatively small portions of corrupted data (Berry et al., 1992; Johnson, 2005; Oppenheimer et.al, 2009; Meade & Craig, 2012; Maniaci & Rogge, 2014; Nichols & Edlund, 2020), it remains unclear how much of an impact cleaning up the data actually has on final conclusions. This question is examined here with respect to subjective assessments of thermal comfort in relation to a range of temperatures and ventilation techniques.

Research question 6 (RQ 6): To which degree are conclusions based on thermal comfort assessments in human subject trials biased by aberrant response behaviours of the subjects? Since this is an open question, we do not have specific expectations regarding the outcome, although the statistical null-hypothesis is that differences between subject groups

(e.g., full sample versus cleaned sample) do not differ systematically.

4 Methods

4.1 Design(s) and procedure

The present study is based on studies that were conducted as a part of the DLR project Next Generation Train (NGT; Winter, 2012). Specifically, we used data from studies that examined the impact of different heating, ventilation and air conditioning (HVAC) systems on thermal comfort of human subjects (e.g., Hörmann et al., 2017; Lange et al. 2019; Schmeling et al., 2019). The participants for the NGT studies were recruited by a recruiting provider and they were compensated for their participation (the specific amounts for each study are listed below). The specific requirements for the participation in the study are explained under 4.2 Sample. All experimental NGT studies were conducted in the transition periods of either September, October, March, or April. Moreover, all studies have similar and thus comparable designs. The studies were conducted in a generic train laboratory (i.e., an adjustable train compartment) located in Göttingen (Germany), which allowed to manipulate climate variables (e.g., temperature or air velocity) as well as in the installation of different HVAC systems (see Schmeling & Volkmann, 2016, 2017, 2020). The train compartment consists of six seating rows with four seats each. A comprehensive English overview of the generic train laboratory and the different HVAC systems that were used was published by Schmeling and Volkmann (2020).

The measurement concept remained the same for all NGT studies. For the present study, we focused on subjective climate assessments (instead of the objective climate values). As shown in Figure 1, each experimental study contained two variations of climate conditions (climate cases) per day. Specifically, each climate case consisted of two identical climate phases with a duration of 30 minutes each. The first half of a climate phase was a mere exposition of the participants to the respective climate conditions. During this period of time, participants were instructed not to communicate with each other and were given puzzle books as a form of entertainment. In the second 15 minutes of each climate phase, participants gave their subjective climate assessments via electronic hand-held devices (HP IPAQ 214 Enterprise Handheld). Before the climate arrived at the required conditions (e.g., the targeted room temperature), there was a lead time of 30 minutes that was used to either give the participants an instruction of the study ahead (in the beginning of a study) or the participants filled out questionnaires concerning trait variables (in the middle

of a study in between two climate cases). Overall, each study took about 3.5 hours of time. We aggregated data from the following four experimental studies that were run between 2015 and 2017 (Hörmann et al., 2018).

Figure 1: Overview chart of four experimental studies to assess different thermal

comfort conditions

4.1.1 Study 1

The first study took place on two days in October of 2015 and investigated four climate cases (with 2 climate phases each) of a displacement ventilation under the seats (CDV) with average room temperatures of 22°C, 20°C, 26°C, and 24 °C, respectively. To achieve the targeted room temperature, the inlet air was adjusted accordingly (from 13 °C to 19 °C). The standard values for the air velocity and relative air humidity for each climate phase was based on the standard EN 13129 (2016). Each day, 20 individuals participated as human subjects (altogether 21 female and 19 male individuals). The participants received a compensation of 40€. Further information can be found in Maier et al. (2015).

4.1.2 Study 2

The second study took place in March of 2016 and again investigated the climate cases of a displacement ventilation under the seats (CDV). However, higher average room temperatures of 27°C, 25°C, 29°C, and 27 °C (in the four climate cases, respectively) were examined. The respective inlet air temperatures were set to 23 °C, 21 °C, 25 °C, and 23 °C, respectively, to achieve the targeted room temperature. The standard values for the air velocity and relative air humidity for each climate phase was again based on the standard EN 13129 (2016). Equal to Study 1, Study 2 was also based on a duration of two days and 20 participants on each day (20 female and 20 male individuals). The participants received a compensation of 40€. Further information can be found in Hörmann et al. (2016).

4.1.3 Study 3

The third study took place on two days in April of 2017 and investigated the climate cases of a state-of-the-art micro-jet ventilation system (MV, see Schmeling & Volkmann, 2020) for the average room temperatures of 22°C, 20°C, 26°C, and 24 °C, respectively. To achieve the targeted room temperature, the inlet air was adjusted accordingly (from 13 °C to 21 °C). The standard values for the air velocity and relative air humidity for each climate phase was again based on the standard EN 13129 (2016). Again, 20 individuals participated on each day of the study (resulting in a total of 20 men and 20 women) and were paid 40€. Further information can be found in Hörmann et al. (2017).

4.1.4 Study 4

The fourth two-days study took place in September of 2017 and investigated a novel hatrack-integrated low-momentum ventilation (HLMV, see Schmeling & Volkmann, 2020) for the average room temperatures of 23°C, 21°C, 26°C, and 24 °C, respectively. To achieve the targeted room temperature, the inlet air was accordingly adjusted from 15 °C to 21 °C. Again, the standard values for the air velocity and relative air humidity for each climate phase were based on the standard EN 13129 (2016). Again, a total of 40 individuals (20 each day) participated in the study (22 males and 18 females). The participants received a compensation of 55€. Further information can be found in Hörmann (2017).

4.2 Sample

Based on the abovementioned studies, we included the data from a total of 160 participants in the present study. The participants were pre-selected based on requirements that were supposed to ensure comparability between individual participants. First, all participants were required to have a higher education entrance qualification ("Abitur") and possess fluent German language skills to make sure a sufficient comprehension of the survey is given. Second, the participant's body height had to be below 1.90 meters to make sure they fit into the generic train laboratory. In order to enforce a uniform extent of clothing during the experiments, all participants received information as to how they should dress in advance to the study. The participants were supposed to wear ankle free, closed shoes, a long-sleeved top, and long trousers. Wearing scarfs, turtleneck jerseys, or skirts was prohibited. The adherence to these guidelines was the third requirement and assured upon their arrival at the testing site.

85 out of the total of 160 participants identified as female (49.4%) and 87 identified as male (50.6%). On average, the participants were 33.38 years of age (SD = 12.09) with a body height of 175.34 cm (M_{male} = 182.27, SD_{male} = 5.09; M_{female} = 168.17, SD_{female} = 5.39) and an average body weight of 75.63 kg (M_{male} = 82.66, SD_{male} = 13.22; M_{female} = 68.58, SD_{female} = 16.14).

4.3 Measures

4.3.1 Subjective climate assessments

The subjective climate assessments were measured with scales originally developed by Marggraf-Micheel and Jäger (2007). The climate assessments included questions about both the room temperature as well as the air draught. The respondents were asked how they conceived the respective variable. The assessments were done in each of the four experimental phases. The temperature assessment was measured using a scale from 1 (*very cold*) to 7 (*hot*). The evaluation of the temperature was answered on a scale from 1 (*very uncomfortable*) to 5 (*very comfortable*). The air draught assessment was measured using a scale from 1 (*none*) to 7 (*very strong*) and the air draught evaluation was measured using a scale from 1 (*very uncomfortable*) to 5 (*very comfortable*). Each of the abovementioned assessments was done for the body parts face, upper body, right shoulder, left shoulder, right hand, left hand, right upper leg, left upper leg, neck, head, left foot, and right foot. Thus, the respondents replied to 48 items in each phase. Overall, 192 items were used to assess the body-part specific climate assessments. Lastly, participants were asked to rate their overall satisfaction with the climate in each phase on a scale from 1 (*very dissatisfied*) to 5 (*very satisfied*). We found Cronbach's α of .97 for the temperature assessments, .96 for the temperature evaluations, .97 for the air draught assessments, and .96 for the air draught evaluations.

4.3.2 Climate preferences

In order to assess the respondent's climate preferences, an instrument developed by Marggraf-Micheel et al. (2010) was used. The climate preferences consist of nine scales. The first scale refers to the preference of warmth vs. cold (e.g., "I am freezing quicker than most other people"). The second scale discriminates between a preference or sensitivity of heat (e.g., "I enjoy lying in the sun over an extended period of time during a hot summer day"). The third scale discriminates between a preference or sensitivity towards air draught (e.g., "I prefer closed windows when driving a car, because I dislike the air draught"). Fourth, the sensitivity vs. tolerance regarding the air quality was assessed (e.g., "I constantly need fresh air to feel comfortable"). The fifth scale concerns sensitivity vs. tolerance of dry air (e.g., "My throat or nose start to quickly feel rough when the air is dry"). The sixth scale assessed the sensitivity vs. tolerance regarding air humidity (e.g., "Stuffy air makes me have circulation problems").

Seventh, the relevance of the climate environment on the overall well-being was assessed (e.g., "The indoor climate has a great effect on my well-being"). These eight scales concerned travel-specific expectations (e.g., "I already expect an uncomfortable climate in the transportation car ahead of my travels"). Lastly, the climate preferences included a scale assessing the sense of entitlement towards comfortable thermal conditions (e.g., "I think it is very important that an appropriate air conditioning is put in place in transportation cabins"). Overall, this instrument consists of 43 items, which were answered on a scale from 1 (*not accurate*) to 5 (*completely accurate*). The following instruction was given to the respondents: "Below, you will find statements about dealing with different climate conditions and the effect those can have. Please, tick the degree to which the respective statement applies to you". For the present sample, a Cronbach's α of .65 for the climate preferences scale was found.

4.3.3 Big Five Inventory (BFI)

The Big Five personality traits (based on the Big Five personality framework; see McCrae & Costa, 1999) were assessed using the Big Five Inventory (BFI). Specifically, a German short version by Rammstedt and John (2005, 2007), called BFI-K, was used. The BFI-K uses the instruction "How well do the following statements describe your personality?" and respondents reply to items on a scale from 1 (*Disagree strongly*) to 5 (*Agree strongly*). In total, the BFI-K consists of 21 items including both reverse-worded and non-reverse-worded items. Overall, it takes respondents approximately 2 minutes to fill out the BFI-K (Rammstedt & John, 2005, 2007). The BFI-K uses four items for each factor to assess extraversion (e.g., "I see myself as someone who is reserved"), agreeableness (e.g., "I see myself as someone who tends to find fault with others"), conscientiousness (e.g., "I see myself as someone who does a thorough job"), neuroticism (e.g., "I see myself as someone who is relaxed, handles stress well"). The factor openness for new experiences is measured by five items to address its heterogeneity (e.g., "I see myself as someone who has an active imagination"). Rammstedt and John (2005) reported Cronbach's α values (extraversion: .86; agreeableness: .74; conscientiousness: .84; neuroticism: .85; openness for new experiences: .75) that indicate satisfactory reliability. A cross-validation study by Rammstedt et al. (2013) confirmed the psychometric properties of the BFI and validated the factorial structure. Therefore, we assume that this short questionnaire measures the five personality constructs reasonably well so that it can contribute to the examination of RQ5 in this study.

4.3.4 Indicators of aberrant response behaviour

The indicators of obtrusive response behaviour were computed by using the R package "Careless" by Yentes and Wilhelm (2018). The indicators were separately computed for the BFI, the climate preferences, and the climate assessments, respectively. We decided to compute one stringent and one lenient cut-off to simulate two possible approaches to identifying obtrusive respondents (for a similar approach, see McGonagle et al., 2016). We opted to compute the following four indicators.

The Longstring (LOS) indicator was computed with separate cut-offs set for each dependent variable. Based on Huang et al. (2012) and Johnson (2005), we set 10 consecutive invariant responses as the lenient threshold and 15 consecutive invariant responses as the stringent threshold for the subjective climate assessments. For the BFI, we chose a cut-off of 5 consecutive invariant responses and for climate preferences, we chose a cut-off of 6 consecutive invariant responses based on the frequencies displayed in Figure 12 and Figure 13. We decided against the implementation of two cutoffs for the BFI and the climate preferences due to the low number of items included.

The Mahalanobis-Distance (MAD) was assessed by using its Chi-Square distribution (D^2). We set the stringent cutoff to a *p*-value of .99 and the lenient cutoff to .95. However, as MAD is based on a certain distribution, the number of obtrusive respondents is relative. Additionally, we decided to also include a binary variable (flagged vs. non-flagged respondents) based on the *p*-values. The binary variable was used for subsequent statistical analyses (see Meade & Craig, 2012).

We computed the psychometric synonyms (PSY) and antonyms index based on the suggestions of Meade and Craig (2012) and Goldberg (2000, as cited in Johnson, 2005). We based our cut-off approach on the study by Meade and Craig (2012) and flagged respondents with psychometric synonym coefficients below .22 (DeSimone et al., 2015). In order to create a lenient and a stringent cut-off, we opted to use different values concerning the correlation coefficients that are included in the computation of the PSY-index. For the stringent approach we thus used the PSY based on item pairs with a minimum of .70 (as was used by Meade & Craig, 2012) and the specific cut-off value of PSY-coefficients below .22 (DeSimone et al., 2015). For the lenient approach, we instead included item pairs with a minimum correlation of .60 instead. The cut-off values of below .22 remained unchanged. Concerning psychometric

antonyms, we did not find a sufficient number of item pairs.

Lastly, we computed the IRV (Dunn et al., 2018). Dunn et al. (2018) and DeSimone and Harms (2017) both reported relative cut-off values (e.g., the lowest 10% were regarded as aberrant respondents). In accordance to these suggestions we set the stringent cut-off to the lowest 2.5% in phase 1 (.60), and the more lenient cut-off to the lowest 10% (0.77) of the IRV distribution (see results for further information on the distributions).

5 Results

5.1 Extent of aberrant responses

As a first step, we analyzed the extent to which subjects show aberrant response patterns by looking at the frequency distributions of the different indicators. In this context, aberrant response patterns were identified with the indicators "Longstring" (LOS), "Intra-individual Response Variability" (IRV), "Psychometric Synonyms" (PSY), and the "Mahalanobis Distances" (MAD). The R-package "Careless" (version 1.2.1) was used for the calculations (Yentes & Wilhelm, 2018). Subsequently, we compared the indicator distributions for the assessments of climate parameters across the four phases with the repeated analysis of variance (RQ 1). In addition, the extent of aberrant responses was examined and compared for the different instruments (48 assessments of climate parameters (CLIM), 21 personality characteristics ratings (PERS), and 43 items of climate preferences (PREF) (RQ 2).

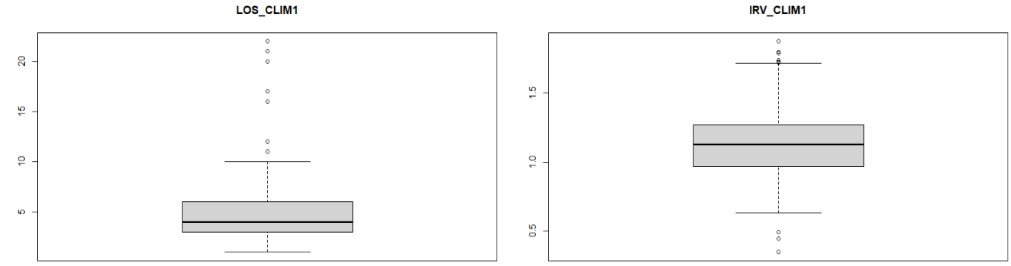

Figure 2: Boxplots for the indicators LOS and IRV within the assessments of climate parameters in phase 1 (95% intervals)

As examples, the distributions of LOS and IRV for the climate assessments in phase 1 are depicted in Figure 2 with separate boxplots. Figure 3 combines both parameters into a single scatterplot. They look very similar to those in the other three phases (see Figure 6 to Figure 13). Some subjects gave over twenty times in a row the same score for different climate assessments (Figure 2, left). Inspections of the raw scores had shown that several times the middle response categories were chosen for excessive Longstrings (e.g. 2 times "3" and 2 times "4" in phase 1). The remaining subjects checked extreme categories ("1" or

"5"). The boxplot for IRV_CLIM1 shows for some subjects a lack of response variability across the 48 items, which is close to zero. The scatterplot confirms a negative correlation of r = -.23** between the LOS-indicator and the IRV, which will be discussed below (see Table 6). The data-points in Figure 3 are labeled with the subject numbers. Most outliers are located in the lower right-hand corner with high scores for LOS and low scores for IRV.

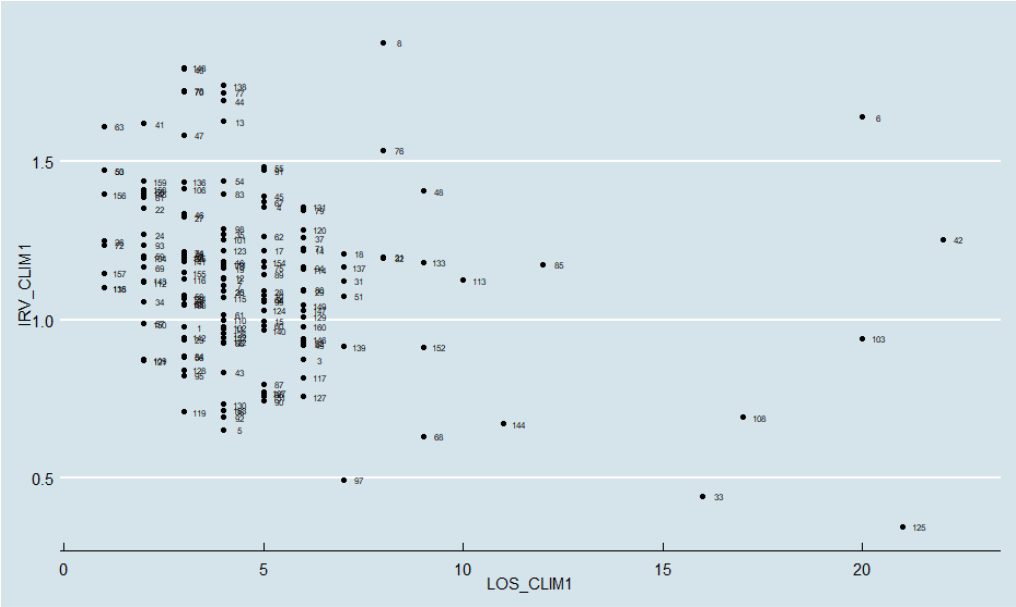

Figure 3: Scatterplot for LOS and IRV within the assessments of climate parameters in phase 1

Table 1 summarizes the results for the repeated analyses of variance between the four sequential phases of the climate assessments. The differences of the mean scores were significant only for the psychometric synonyms with a slight decrease of consistency from phase 1 to phase 4. That means the intraindividual correlations between psychometrically synonymous item pairs declined. Though, this trend was equal for all indicators, LOS, IRV, and MAD failed to show a significant change of aberrant responses over time (RQ 1). The effect sizes η^2 were also very small. Since the Mahalanobis Distances were standardized within each phase by the software, a comparison across the phases was meaningless.

The cut-off scores, which were applied to identify the outliers with

aberrant responses, were determined according to the indicator distributions and recommendations from the literature as described in the section 4.3.4. We introduced two thresholds ("high" and "low") for each indicator with a different degree of strictness. According to the average scores across all phases, between 3% (for IRV) and 12% (for MAD) of the subjects in our sample seemed to have given low quality ratings when assessing the climate parameters (see Table 1, column "Overall"), if we applied the threshold "high" (stringent). Between 6% (for PSY) and 21% (for MAD) were flagged for the threshold "low" (lenient).

When we compared the amount of aberrant responses between the different questionnaires usually the climate assessments showed a higher amount than the personality questionnaire or the climate preferences (RQ 2). In both of these other questionnaires neither IRV nor LOS flagged any subject for obtrusive response patterns. Only the psychometric synonyms for the climate preferences identified a higher number of noticeable response patterns compared to the other instruments. PSY could not be calculated for the BFI because only an insufficient number of psychometric synonyms was identified. Therefore, this cell remained empty. However, it should be kept in mind that these quantities depend on the settings for the cutoff scores. Because the applied instruments used different scales and item numbers, the comparisons between these methods remain descriptive.

Table 1: Distribution scores of response characteristics for the different phases and instruments of data gathering

	Phase 1	Phase 2	Phase 3	Phase 4	Overall	Sig.
	Min M Max SD *N Outliers high/low*	*Min M Max SD* *N Outliers high/low*	*Min M Max SD* *N Outliers high/low*	*Min M Max SD* *N Outliers high/low*	*Min M Max SD* *Ø Outliers high/low*	*p η^2*
LOS_CLIM	1 4.80 22 3.50 6/9	1 4.80 25 4.10 6/14	1 5.40 24 4.10 6/18	1 5.40 24 4.00 7/18	1 5.10 25 4.00 6.3/14.8	0.20 0.01
IRV_CLIM	0.35 1.14 1.88 0.28 3/16	0.49 1.13 1.90 0.30 4/21	0.48 1.09 2.17 0.31 7/21	0.51 1.11 2.13 0.32 7/24	0.35 1.12 2.17 0.30 5.3/20.5	0.09 0.01
PSY_CLIM	-0.70 0.77 1.00 0.27 9/7	-0.25 0.76 1.00 0.24 4/7	-0.03 0.71 1.00 0.24 9/10	-0.16 0.69 1.00 0.24 8/14	-0.70 0.77 1.00 0.27 7.5/9.5	<0.01 0.04
MAD_CLIM	5.13 47.70 109.70 20.34 17/36	6.55 47.70 94.53 20.27 18/29	4.82 47.70 100.65 21.72 23/32	3.60 47.70 100.28 22.36 20/37	3.60 47.70 109.70 21.14 19.5/33.5	1.0 0.00
LOS_PREF					2 3.55 9 1.25 9	
IRV_ PREF					0.63 1.19 1.90 0.23 0	
PSY_PREF					-0.54 0.54 1.00 0.38 31	
MAD_ PREF					21.66 42.70 87.02 12.95 6	
LOS_PERS					1 2.94 7 0.97 8	
IRV_PERS					0.59 1.10 1.69 0.22 1	
PSY_PERS					- - - - -	
MAD_PERS					4.21 20.87 55.30 7.98 3	

5.2 Stability of indicators

In this section we examined whether the same indicator for aberrant responses showed some stability over time (RQ 3). To analyze this question, we correlated all scores of each indicator across the repeated measurements in the four different phases of climate assessments. If the source for aberrance was linked to characteristics of the respective person, the respective indicator should have shown reasonable stability between different points in time. Therefore, we expected that the response profiles in the entire sample for each indicator in climate assessment phase n would correlate significantly with the response profile in phase n+1. In contrast, correlations between the same indicators across different sets of questions should be substantially lower. For this latter comparison we included the corresponding indicators for the personality questionnaire and for the climate preferences in the following tables.

Table 2: Correlations of Longstrings across the climate assessments, personality and climate preferences

Pearson's Correlations						
Variable	**LOS_CLIM1**	**LOS_CLIM2**	**LOS_CLIM3**	**LOS_CLIM4**	**LOS_PERS**	**LOS_PREF**
LOS_CLIM1	—					
LOS_CLIM2	0.460 ***	—				
LOS_CLIM3	0.183 *	0.296 ***	—			
LOS_CLIM4	0.191 *	0.182 *	0.435 ***	—		
LOS_PERS	0.079	0.021	0.035	0.000	—	
LOS_PREF	0.051	-0.057	0.100	0.070	0.039	—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 2 shows the situation for the Longstring indicator. All intercorrelations of LOS for the four climate assessments were significant. Correlations between adjacent climate phases were the highest. Correlations of the same indicator for different sets of questions (LOS_CLIM with LOS_PERS and LOS_PREF) were not significant. For LOS an average stability coefficient of $\bar{r} = 0.42$ was calculated (ranging from $r = 0.30$ to $r = 0.46$).

Table 3: Correlations of Intra-Individual Response Variability across the climate assessments, personality and climate preferences

Pearson's Correlations						
Variable	**IRV_CLIM1**	**IRV_CLIM2**	**IRV_CLIM3**	**IRV_CLIM4**	**IRV_PERS**	**IRV_PREF**
IRV_CLIM1	—					
IRV_CLIM2	0.715 ***	—				
IRV_CLIM3	0.648 ***	0.664 ***	—			
IRV_CLIM4	0.535 ***	0.631 ***	0.802 ***	—		
IRV_PERS	0.044	0.044	0.003	0.101	—	
IRV_PREF	0.236 **	0.273 ***	0.249 **	0.153	0.320 ***	—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

For the intra-individual response variability, the stability correlations were even higher than for the long-string indicator (Table 3). The average stability coefficient was $\bar{r} = 0.93$, ranging from $r = 0.66$ to $r = 0.80$. In addition, some of the correlations between the IRV-indicator for different sets of questions were also significant, but as expected substantially lower (e.g. correlations between IRV_CLIM and IRV_PREF). According to these results, IRV seemed to be more closely related to person characteristics than LOS.

In Table 4 the correlation matrix for the psychometric synonyms is shown. As mentioned before, PSY could not be calculated for the personality questionnaire. Therefore, these cells remain empty. The stability coefficients for PSY varied across the different phases. They were obviously, lower for the first two phases of climate assessments than for the phases three and four. The average stability was $\bar{r} = 0.47$ (ranging from $r = 0.28$ to $r = 0.66$).

Table 4: Correlations of psychometric synonyms across the climate assessments, personality and climate preferences

Pearson's Correlations

Variable	PSY_CLIM1	PSY_CLIM2	PSY_CLIM3	PSY_CLIM4	PSY_PERS	PSY_PREF
PSY_CLIM1	—					
PSY_CLIM2	0.283 ***	—				
PSY_CLIM3	0.300 ***	0.323 ***	—			
PSY_CLIM4	0.188 *	0.385 ***	0.658 ***	—		
PSY_PERS[a]					—	
PSY_PREF	0.095	0.168 *	0.073	0.095		—

Note. [a] Number of observations is < 3; * $p < .05$, ** $p < .01$, *** $p < .001$

A strong temporal stability could again be observed for the Mahalanobis Distances. The average was $\bar{r} = 0.83$ (ranging from $r = 0.66$ to $r = 0.69$). The picture appeared similar to IRV. According to these findings MAD and IRV indicators of aberrant responses seemed to reflect more stable person characteristics than LOS and PSY.

Table 5: Correlations of Mahalanobis Distances across the climate assessments, personality and climate preferences

Pearson's Correlations

Variable	MAD_CLIM1	MAD_CLIM2	MAD_CLIM3	MAD_CLIM4	MAD_PERS	MAD_PREF
MAD_CLIM1	—					
MAD_CLIM2	0.684 ***	—				
MAD_CLIM3	0.652 ***	0.689 ***	—			
MAD_CLIM4	0.519 ***	0.613 ***	0.662 ***	—		
MAD_PERS	0.152	0.149	0.089	0.179 *	—	
MAD_PREF	0.338 ***	0.245 **	0.290 ***	0.220 **	0.363 ***	—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

5.3 Consistency of indicators

With each of the indices, aberrant responses were identified for some subjects in the sample. The next research question referred to the consistency of the patterns of identified individuals when comparing between the different indices (RQ 4). For this purpose, correlation coefficients are reported between the raw values of each of the four indicators in the entire sample. Therefore, the cutoff scores did not influence the results.

Table 6: Correlations between Longstrings and Intra-Individual Response Variabilities for the climate assessments in phase 1 to 4

	Pearson's Correlations		
			Pearson's r
Phase 1	LOS_CLIM1	- IRV_CLIM1	**-0.232 ****
	LOS_CLIM1	- IRV_CLIM2	-0.207 **
	LOS_CLIM1	- IRV_CLIM3	-0.173 *
	LOS_CLIM1	- IRV_CLIM4	-0.156 *
Phase 2	LOS_CLIM2	- IRV_CLIM1	-0.214 **
	LOS_CLIM2	- IRV_CLIM2	**-0.247** **
	LOS_CLIM2	- IRV_CLIM3	-0.175 *
	LOS_CLIM2	- IRV_CLIM4	-0.111
Phase 3	LOS_CLIM3	- IRV_CLIM1	-0.040
	LOS_CLIM3	- IRV_CLIM2	0.014
	LOS_CLIM3	- IRV_CLIM3	**0.057**
	LOS_CLIM3	- IRV_CLIM4	0.123
Phase 4	LOS_CLIM4	- IRV_CLIM1	-0.053
	LOS_CLIM4	- IRV_CLIM2	-0.057
	LOS_CLIM4	- IRV_CLIM3	-0.020
	LOS_CLIM4	- IRV_CLIM4	**0.026**

Note. * $p < .05$, ** $p < .01$, *** $p < .001$; bold are the correlations between corresponding phases

As Table 6 shows, the LOS-indicator and the IRV-scores were significantly correlated in phase 1 and phase 2. The coefficients were negative, indicating

that higher LOS-values corresponded to lower IRV-values as illustrated above in Figure 3. In phase 3 and 4 these correlations disappeared, which means that some consistency between these two indicators existed, but it was limited to the first two phases of repeated measurements.

Table 7: Correlations between Longstrings and psychometric synonyms for the climate assessments in phase 1 to 4

	Pearson's Correlations			
				Pearson's r
Phase 1	LOS_CLIM1	-	PSY_CLIM1	**-0.090**
	LOS_CLIM1	-	PSY_CLIM2	0.043
	LOS_CLIM1	-	PSY_CLIM3	0.026
	LOS_CLIM1	-	PSY_CLIM4	-0.058
Phase 2	LOS_CLIM2	-	PSY_CLIM1	-0.089
	LOS_CLIM2	-	PSY_CLIM2	**-0.067**
	LOS_CLIM2	-	PSY_CLIM3	-0.010
	LOS_CLIM2	-	PSY_CLIM4	-0.056
Phase 3	LOS_CLIM3	-	PSY_CLIM1	-0.115
	LOS_CLIM3	-	PSY_CLIM2	-0.183 *
	LOS_CLIM3	-	PSY_CLIM3	**-0.032**
	LOS_CLIM3	-	PSY_CLIM4	0.046
Phase 4	LOS_CLIM4	-	PSY_CLIM1	-0.164 *
	LOS_CLIM4	-	PSY_CLIM2	-0.047
	LOS_CLIM4	-	PSY_CLIM3	0.012
	LOS_CLIM4	-	PSY_CLIM4	**0.031**

Note. * $p < .05$, ** $p < .01$, *** $p < .001$; bold are the correlations between corresponding phases

In Table 7 the intercorrelations between the LOS-indicator and the PSY-scores are shown. None of these correlations were significant for corresponding phases. This means, that these two indicators were sensitive to different aspects of aberrant responses.

The intercorrelations between the LOS-indicator and the MAD-scores are shown in Table 8. The correlation pattern was similar as in Table 6. The two

indicators showed only for phases 1 and 2 some correspondence. However, unexpectedly the correlations had a negative sign, which meant that subjects with longer strings of identical responses tended to have slightly smaller average distances to the other subjects. The correlation should be positive if the two indicators would reflect similar underlying conditions. The results seem to confirm that in our data LOS reflected more a neutral answer style towards the middle of the scale.

Table 8: Correlations between Longstrings and Mahalanobis Distances for the climate assessments in phase 1 to 4

	Pearson's Correlations			
				Pearson's r
Phase 1	LOS_CLIM1	-	MAD_CLIM1	**-0.161** *
	LOS_CLIM1	-	MAD_CLIM2	-0.169 *
	LOS_CLIM1	-	MAD_CLIM3	-0.182 *
	LOS_CLIM1	-	MAD_CLIM4	-0.150
Phase 2	LOS_CLIM2	-	MAD_CLIM1	-0.044
	LOS_CLIM2	-	MAD_CLIM2	**-0.159** *
	LOS_CLIM2	-	MAD_CLIM3	-0.127
	LOS_CLIM2	-	MAD_CLIM4	-0.119
Phase 3	LOS_CLIM3	-	MAD_CLIM1	0.076
	LOS_CLIM3	-	MAD_CLIM2	0.019
	LOS_CLIM3	-	MAD_CLIM3	**-0.045**
	LOS_CLIM3	-	MAD_CLIM4	-0.014
Phase 4	LOS_CLIM4	-	MAD_CLIM1	0.078
	LOS_CLIM4	-	MAD_CLIM2	-0.033
	LOS_CLIM4	-	MAD_CLIM3	-0.092
	LOS_CLIM4	-	MAD_CLIM4	**-0.148**

Note. * $p < .05$, ** $p < .01$, *** $p < .001$; bold are the correlations between corresponding phases

By far clearer was the picture for the correlations between the IRV-scores, the PSY-scores and the MAD-scores as shown in Table 9 and Table 10. The correlations between the PSY-scores and the MAD-scores were also conclusive. Higher scores for the psychometric synonyms were related to smaller

average distances of the response profiles to the rest of the subjects (Table 11). Therefore, significant correlations for the corresponding phases were negative, which was what we expected.

This pattern was fully in line with our expectations that different indicators of aberrant responses correspond significantly if they were calculated for the same set of variables during the same climate scenario. All correlations were positive and statistically significant. The highest correlations were between the corresponding phases. The only exception was in the third phase where IRV correlated almost equally high with MAD in phase 2 and 3.

Table 9: Correlations between intra-individual response variabilities and psychometric synonyms for the climate assessments in phase 1 to 4

	Pearson's Correlations		
			Pearson's r
Phase 1	IRV_CLIM1	- PSY_CLIM1	**0.397** ***
	IRV_CLIM1	- PSY_CLIM2	0.198 *
	IRV_CLIM1	- PSY_CLIM3	0.178 *
	IRV_CLIM1	- PSY_CLIM4	0.211 **
Phase 2	IRV_CLIM2	- PSY_CLIM1	0.235 **
	IRV_CLIM2	- PSY_CLIM2	**0.376** ***
	IRV_CLIM2	- PSY_CLIM3	0.191 *
	IRV_CLIM2	- PSY_CLIM4	0.315 ***
Phase 3	IRV_CLIM3	- PSY_CLIM1	0.286 ***
	IRV_CLIM3	- PSY_CLIM2	0.171 *
	IRV_CLIM3	- PSY_CLIM3	**0.426** ***
	IRV_CLIM3	- PSY_CLIM4	0.402 ***
Phase 4	IRV_CLIM4	- PSY_CLIM1	0.192 *
	IRV_CLIM4	- PSY_CLIM2	0.188 *
	IRV_CLIM4	- PSY_CLIM3	0.316 ***
	IRV_CLIM4	- PSY_CLIM4	**0.441** ***

Note. * $p < .05$, ** $p < .01$, *** $p < .001$; bold are the correlations between corresponding phases

The correlations between the PSY-scores and the MAD-scores were also

conclusive. Higher scores for the psychometric synonyms were related to smaller average distances of the response profiles to the rest of the subjects (Table 11). Therefore, significant correlations for the corresponding phases were negative, which was what we expected.

With respect to IRV one could argue that not only low but also high IRV values represent obtrusive responses. Therefore, non-linear correlations of IRV with MAD and with PSY should be carried out. This was accomplished by inserting a quadratic and a cubic IRV-term into polynomial regressions with MAD and PSY as the dependent variables and IRV-linear, IRV-quadratic, and IRV-cubic as the independent variables. In fact, a significant increase for the respective relationships could be found for *R*(IRV_CLIM1;MAD_CLIM1) of .42 (instead of *r* = .39 for the linear correlation). For the relationship between IRV and PSY even three correlations increased significantly by the quadratic IRV-term: *R*(IRV_CLIM1;PSY_CLIM1) of .42 (instead of *r* = .40 for the linear correlation), *R*(IRV_CLIM2;PSY_CLIM2) of .40 (instead of *r* = .38), *R*(IRV_CLIM3;PSY_CLIM3) of .47 (instead of *r* = .43).

Table 10: Correlations between Individual Response Variabilities and Mahalanobis Distances for the climate assessments in phase 1 to 4

	Pearson's Correlations			
				Pearson's r
Phase 1	IRV_CLIM1	-	MAD_CLIM1	**0.390 *****
	IRV_CLIM1	-	MAD_CLIM2	0.304 ***
	IRV_CLIM1	-	MAD_CLIM3	0.189 *
	IRV_CLIM1	-	MAD_CLIM4	0.137
Phase 2	IRV_CLIM2	-	MAD_CLIM1	0.299 ***
	IRV_CLIM2	-	MAD_CLIM2	**0.355 *****
	IRV_CLIM2	-	MAD_CLIM3	0.235 **
	IRV_CLIM2	-	MAD_CLIM4	0.137
Phase 3	IRV_CLIM3	-	MAD_CLIM1	0.262 ***
	IRV_CLIM3	-	MAD_CLIM2	0.309 ***
	IRV_CLIM3	-	MAD_CLIM3	**0.305 *****
	IRV_CLIM3	-	MAD_CLIM4	0.184 *
Phase 4	IRV_CLIM4	-	MAD_CLIM1	0.261 ***
	IRV_CLIM4	-	MAD_CLIM2	0.260 ***
	IRV_CLIM4	-	MAD_CLIM3	0.210 **
	IRV_CLIM4	-	MAD_CLIM4	**0.302** ***

Note. * $p < .05$, ** $p < .01$, *** $p < .001$; bold are the correlations between corresponding phases

In conclusion, these analyses have demonstrated some consistency between several of the indicators for aberrant response behaviour. The results for IRV, PSY, and MAD are more in line with each other than with LOS. If IRV is analyzed for non-linear relationships with MAD and PSY, the correlations increased slightly, but did not change the entire picture. According to our findings, LOS seems to be related to a middle-tendency response style with respect to the climate assessments, which is not necessarily reflecting a careless or inattentive answer style.

Table 11: Correlations between psychometric synonyms and Mahalanobis Distances for the climate assessments in phase 1 to 4

Pearson's Correlations			
			Pearson's r
PSY_CLIM1	-	MAD_CLIM1	**-0.205** **
PSY_CLIM1	-	MAD_CLIM2	-0.119
PSY_CLIM1	-	MAD_CLIM3	-0.079
PSY_CLIM1	-	MAD_CLIM4	-0.094
PSY_CLIM2	-	MAD_CLIM1	-0.288 ***
PSY_CLIM2	-	MAD_CLIM2	**-0.343** ***
PSY_CLIM2	-	MAD_CLIM3	-0.214 **
PSY_CLIM2	-	MAD_CLIM4	-0.279 ***
PSY_CLIM3	-	MAD_CLIM1	-0.365 ***
PSY_CLIM3	-	MAD_CLIM2	-0.280 ***
PSY_CLIM3	-	MAD_CLIM3	**-0.415** ***
PSY_CLIM3	-	MAD_CLIM4	-0.386 ***
PSY_CLIM4	-	MAD_CLIM1	-0.317 ***
PSY_CLIM4	-	MAD_CLIM2	-0.325 ***
PSY_CLIM4	-	MAD_CLIM3	-0.372 ***
PSY_CLIM4	-	MAD_CLIM4	**-0.512** ***

Note. * $p < .05$, ** $p < .01$, *** $p < .001$; bold are the correlations between corresponding phases

Of practical value could be information about the overlap of subsets of subjects identified by the different indicators as showing aberrant response behaviour across the different indicators. Do they flag the same or different individuals? On the basis of the four analyzed indicators, binary variables were computed, which flagged the subjects as being an outlier or not. With crosstabulation these binary variables were combined for the corresponding phases of climate assessments and the respective level of the threshold (high vs low). Because of the low observed frequencies of outliers for most indicators, we did not perform tests of significance for the Chi-square statistic. The percentages of overlapping subjects identified as outliers are shown in Table 12.

The figures in Table 12 showed that the majority of analyzed indicators

of aberrant response patterns reflected independent characteristics. Especially, if a more stringent threshold was applied to flag subjects, the overlap was often zero. For the lower threshold, some overlap existed between LOS and IRV. However, it is decreasing along the four phases of consecutive assessments. Rather high was also the overlap between PSY and MAD (with the exception of phase 1). However, given the rather low frequency of subjects flagged by the indicator for psychometric synonyms (see Table 1) these figures should be interpreted with caution.

Based on the correlations and crosstabulation we can assert a moderate degree of consistency between the indicators. Since the overlap between the subsets of subjects flagged by different indicators is in most cases below 50%, a multi-hurdle approach with all four indicators involved seemed to be the most appropriate approach to clean the sample from aberrant responses.

Table 12: Crosstabulation results about outliers flagged by the different indicators

			% overlap	
		Thresholds	Low	High
Phase 1	LOS_CLIM1	- IRV_CLIM1	44.4	33.3
	LOS_CLIM1	- PSY_CLIM1	33.3	0.0
	LOS_CLIM1	- MAD_CLIM1	0.0	0.0
Phase 2	LOS_CLIM2	- IRV_CLIM2	42.9	50.0
	LOS_CLIM2	- PSY_CLIM2	0.0	0.0
	LOS_CLIM2	- MAD_CLIM2	0.0	0.0
Phase 3	LOS_CLIM3	- IRV_CLIM3	33.3	0.0
	LOS_CLIM3	- PSY_CLIM3	5.6	0.0
	LOS_CLIM3	- MAD_CLIM3	11.1	0.0
Phase 4	LOS_CLIM4	- IRV_CLIM4	16.7	0.0
	LOS_CLIM4	- PSY_CLIM4	5.6	0.0
	LOS_CLIM4	- MAD_CLIM4	11.1	0.0
Phase 1	IRV_CLIM1	PSY_CLIM1	37.5	0.0
	IRV_CLIM1	MAD_CLIM1	0.0	0.0
Phase 2	IRV_CLIM2	PSY_CLIM2	14.3	0.0
	IRV_CLIM2	MAD_CLIM2	0.0	0.0
Phase 3	IRV_CLIM3	PSY_CLIM3	9.5	14.3
	IRV_CLIM3	MAD_CLIM3	0.0	0.0
Phase 4	IRV_CLIM4	PSY_CLIM4	25.0	14.3
	IRV_CLIM4	MAD_CLIM4	4.2	0.0
Phase 1	PSY_CLIM1	MAD_CLIM1	0.0	33.3
Phase 2	PSY_CLIM2	MAD_CLIM2	42.9	25.0
Phase 3	PSY_CLIM3	MAD_CLIM3	50.0	44.4
Phase 4	PSY_CLIM4	MAD_CLIM4	64.3	62.5

Note. Low threshold is based on the lenient cutoff; high threshold on the stringent cutoff

5.4 Relationship to personality scales

In line with earlier research, correlations of indices for aberrant responses were correlated with the personality scales of the Big5 model (RQ 5). The coefficients are shown in with Table 13 to Table 16. Most of the correlations were insignificant for our dataset. Three coefficients (for IRV and MAD) were negative for the scale Openness. Only for phase 1 Conscientiousness is negatively related to the Longstring-indicator and positively to the Mahalanobis Distances. However, the latter coefficient was poled opposite to our expectations.

Table 13: Correlations of the Longstring indicator of climate assessments with personality scales

Variable	LOS_CLIM1	LOS_CLIM2	LOS_CLIM3	LOS_CLIM4
Extraversion	-0.075	-0.088	0.062	-0.032
Agreeableness	0.018	0.023	0.092	0.087
Conscientiousness	-0.181 *	-0.072	0.081	0.024
Neuroticism	0.153	0.051	0.036	-0.094
Openness	0.065	0.115	0.074	0.007

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 14: Correlations of the Longstring indicator of climate assessments with personality scales

Variable	IRV_CLIM1	IRV_CLIM2	IRV_CLIM3	IRV_CLIM4
Extraversion	0.014	0.009	-0.016	-0.036
Agreeableness	-0.087	-0.089	-0.122	-0.121
Conscientiousness	0.078	0.105	0.058	0.066
Neuroticism	-0.009	-0.018	-0.077	0.006
Openness	-0.179 *	-0.217 **	-0.108	-0.111

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

It could be argued that linear correlations were not the best measures to examine the relationship of IRV to personality scales. If not only low IRV values but also high scores were regarded as an indication of obtrusive answers, a non-

linear term in the regression equation could increase the correlation. Therefore, we used polynomial regression to analyze the significance of the quadratic and the cubic term of IRV for the prediction of the personality scales. The only significant result was found for the scale Openness. In the phases P1, P2, and P3 the multiple correlation coefficient *R* increased significantly if the quadratic IRV-term entered the equation. The correlation coefficients increased for P1 from *R* = -.18 to *R* = -.24, for P2 from *R* = -.22 to *R* = -.29, and for P3 from *R* = -.11 to *R* = -.20. The cubic term did not further improve to the prediction. In summary, the non-linear analysis did not change the entire picture of the relations between IRV and personality because the pattern of the correlations did not change completely in our data-set. The personality scale Openness showed already significant linear correlations with IRV, at least in phase 1 and 2. The existing correlations increased just a bit further.

Table 15: Correlations of the Longstring indicator of climate assessments with personality scales

Variable	PSY_CLIM1	PSY_CLIM2	PSY_CLIM3	PSY_CLIM4
Extraversion	-0.052	-0.065	-0.086	-0.171 *
Agreeableness	-0.087	-0.100	0.033	-0.044
Conscientiousness	0.002	-0.054	-0.016	0.003
Neuroticism	0.004	0.064	-0.040	0.022
Openness	-0.094	0.032	-0.063	-0.025

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 16: Correlations of the Longstring indicator of climate assessments with personality scales

Variable	MAD_CLIM1	MAD_CLIM2	MAD_CLIM3	MAD_CLIM4
Extraversion	0.146	0.107	0.086	0.066
Agreeableness	-0.010	0.012	0.008	-0.011
Conscientiousness	0.217 **	0.101	0.078	0.044
Neuroticism	-0.101	-0.051	-0.118	0.031
Openness	-0.128	-0.190 *	-0.131	-0.135

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

In summary, the relationship of aberrant response indicators to personality scales is less clear than in previous research (e.g., Marjanovic et al.

2015, Dunn et al., 2018; Maniaci & Rogge, 2014). Aberrant responses in the climate assessments seem to be rooted primarily in other factors not fully covered by the Big 5 personality scales.

Not exactly personality variables, but as a side note we also analyzed the relationship of the various indicators of aberrant responses with age and gender. Out of 23 *T*-tests with gender as the group variable only one significant difference resulted. Male subjects had somewhat lower IRV scores for the climate preferences compared to female subjects (*mean* IRV-PREF(female) = 1.23; *mean* IRV-PREF(male) = 1.14, p = .017). Three of the 23 correlations with age were also significant: r(IRV-P1: AGE) = -0.20*; r(MAD-P3:AGE) = -.21**; r(MAD-P4:AGE) = -.18* . Since these relations are very small, we assumed that in our data set neither age nor gender played any important role for aberrant response behaviours.

5.5 Thermal comfort full sample vs cleaned sample

After examining four possible indicators of aberrant response patterns in the previous sections of this report, the effects of data-cleaning on conclusions about comfort assessments were finally analyzed. The question remained, whether conclusions about thermal comfort assessments in different climate scenarios would change significantly, if the sample of subjects was cleaned to account for aberrant response patterns (RQ 6).

A multi-hurdle approach was applied to eliminate subjects from the sample who showed obtrusive response patterns as reflected by any of the four applied indicators (Table 17). That means subjects had been filtered out, if one or more of the aberrant response indicators displayed a red flag. Each phase of the climate comfort experiment was treated separately.

Table 17: Number of subjects eliminated in a multi-hurdle approach based on all applied indicators (N=160 subjects from study 1 to study 4)

	Stringent threshold	Lenient threshold
Phase 1	30	58
Phase 2	28	59
Phase 3	40	66
Phase 4	36	73

As shown in Table 17 the numbers of eliminated subjects ranged between 17.5% and 25.0% for the high threshold and 36.3% and 45,6% for the low threshold. Since it did not really make sense to eliminate 30% or 40% of the sample if reasonable data quality is to be assumed, we only investigated whether the climate assessments of subjects flagged in the high-threshold approach would systematically differ from those of the remaining subjects.

For each of the four phases with a different climate situation, a multivariate analysis of variance was conducted with the respective threshold as the independent factor and 12 climate assessments as dependent variables. The climate assessments were distinguished into sensations and evaluations of air draught and local temperatures as perceived in different regions of the body (see section 4.3.1). Of altogether 16 analyses only four resulted in significant differences according to the multivariate *F*-Tests (Table 18).

Table 18: MANOVA results of the climate assessments for subjects flagged vs. unflagged (N=160 subjects from study 1 to study 4)

Dependent variables		*F*	*df*	*p*	η_p^2
Sensations of air draught					
	Phase 1	0.945	12/147	0.505	0.072
	Phase 2	2.891	12/147	0.001	0.191
	Phase 3	1.529	12/147	0.120	0.111
	Phase 4	1.716	12/147	0.069	0.123
Evaluations of air draught					
	Phase 1	0.560	12/147	0.871	0.044
	Phase 2	1.556	12/147	0.111	0.113
	Phase 3	2.117	12/147	0.019	0.147
	Phase 4	1.296	12/147	0.227	0.096
Sensations of air temperatures					
	Phase 1	1.186	12/147	0.298	0.088
	Phase 2	1.299	12/147	0.225	0.096
	Phase 3	1.005	12/147	0.447	0.076
	Phase 4	2.323	12/147	0.009	0.159
Evaluations of air temperatures					
	Phase 1	0.707	12/147	0.742	0.055
	Phase 2	1.248	12/147	0.256	0.092
	Phase 3	1.510	12/147	0.126	0.110
	Phase 4	1.868	12/147	0.043	0.132

By inspection of these analyses results, hardly any systematic pattern can be determined. Two of the four MANOVAs were significant for phase 4. None of the four different sets of dependent variables collected more significant results than the others. Additionally, we calculated independent samples *T*-tests to compare the overall climate satisfaction in the four phases for the group of unflagged and flagged subjects. Only in one out of four *T*-tests the climate satisfaction was significantly lower for the flagged subjects (Phase 1: $T(158) = 2.286$, $p = .024$, $M1 = 3.02$, $M2 = 2.60$). The remaining differences were insignificant. In total, we concluded that the effects of aberrant responses did not significantly affect the quality of the comfort assessments. The climate assessments and climate satisfaction scores are more or less the same regardless of whether the response pattern showed symptoms of an aberrant response style or not.

As described in section 4.1, the climate satisfaction trials with human subjects were conducted with three different ventilation techniques. From the findings reported so far, we could not exclude that the variance induced by the ventilation techniques had outshined the error variance of aberrant responses. Therefore, we conducted further comparisons in the largest sub-sample with $N = 80$, which was exposed to only one HVAC system, the displacement ventilation (CDV), with different degrees of indoor temperatures. The frequencies of flagged subjects in this smaller sample are listed in Table 19.

Table 19: Number of subjects eliminated in a multi-hurdle approach based on all applied indicators (N=80 subjects from study 1 and 2)

	High threshold	Low threshold
Phase 1	15	24
Phase 2	9	24
Phase 3	18	25
Phase 4	14	30

The MANOVA analyses for the reduced sample showed even less significant differences between the subjects' groups with and without symptoms of aberrant responses. None of the air draught assessments was significant and only two effects of temperature assessments reached the $\alpha < 5\%$ threshold (temperature sensations in phase 2 and temperature evaluations in phase 3). The

mean scores of the dependent variables in these two significant findings are displayed in Figure 4 and in Figure 5.

Table 20: MANOVA results of the climate assessments for subjects flagged vs. unflagged (N=80 subjects from study 1 and 2)

Dependent variables		*F*	*df*	*p*	η_p^2
Sensations of air draught					
	Phase 1	0.709	12/67	0.738	0.113
	Phase 2	1.774	12/67	0.071	0.241
	Phase 3	1.750	12/67	0.076	0.239
	Phase 4	0.786	12/67	0.663	0.123
Evaluations of air draught					
	Phase 1	0.858	12/67	0.592	0.133
	Phase 2	1.544	12/67	0.131	0.217
	Phase 3	1.779	12/67	0.070	0.242
	Phase 4	1.473	12/67	0.157	0.209
Sensations of air temperatures					
	Phase 1	1.237	12/67	0.278	0.181
	Phase 2	2.022	12/67	0.035	0.266
	Phase 3	1.119	12/67	0.360	0.167
	Phase 4	1.204	12/67	0.299	0.177
Evaluations of air temperatures					
	Phase 1	0.605	12/67	0.830	0.098
	Phase 2	1.200	12/67	0.301	0.177
	Phase 3	2.948	12/67	0.002	0.346
	Phase 4	1.060	12/67	0.407	0.160

The expectations from climate comfort studies with displacement ventilation techniques is usually that in correspondence with the vertical temperature gradient (low temperatures at the floor level and increasingly higher temperature towards the ceiling) the perceived temperatures should decrease from head down to the feet (Hörmann et al., 2017). While the temperature sensation data of the unflagged subjects seemed to follow this expectation of the experiment, the data of the flagged subjects do not as well. In phase 2 these differences became significant in the MANOVA analysis (Figure 4; $F(12/67) = 2.022$, $p = .035$)). Subjects flagged for aberrant responses were also more negative with evaluations of the temperatures. However, these differences are

only significant in phase 3 as shown in Figure 5 ($F(12/67) = 2.948$, $p = .002$).

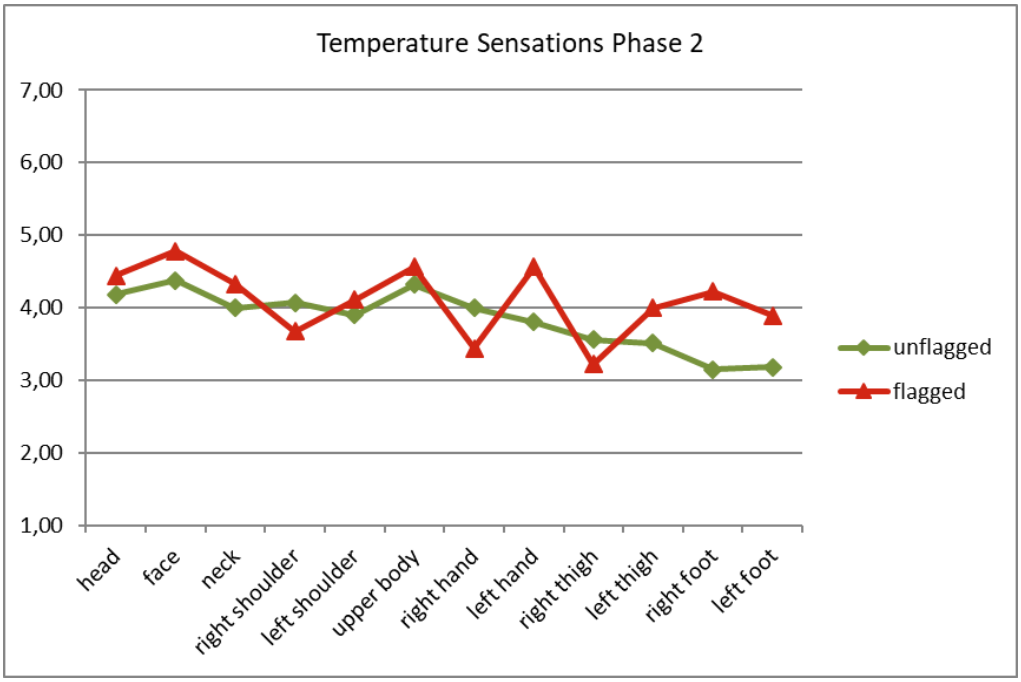

Figure 4: Temperature sensations for different body parts. N = 80 flagged vs unflagged subjects

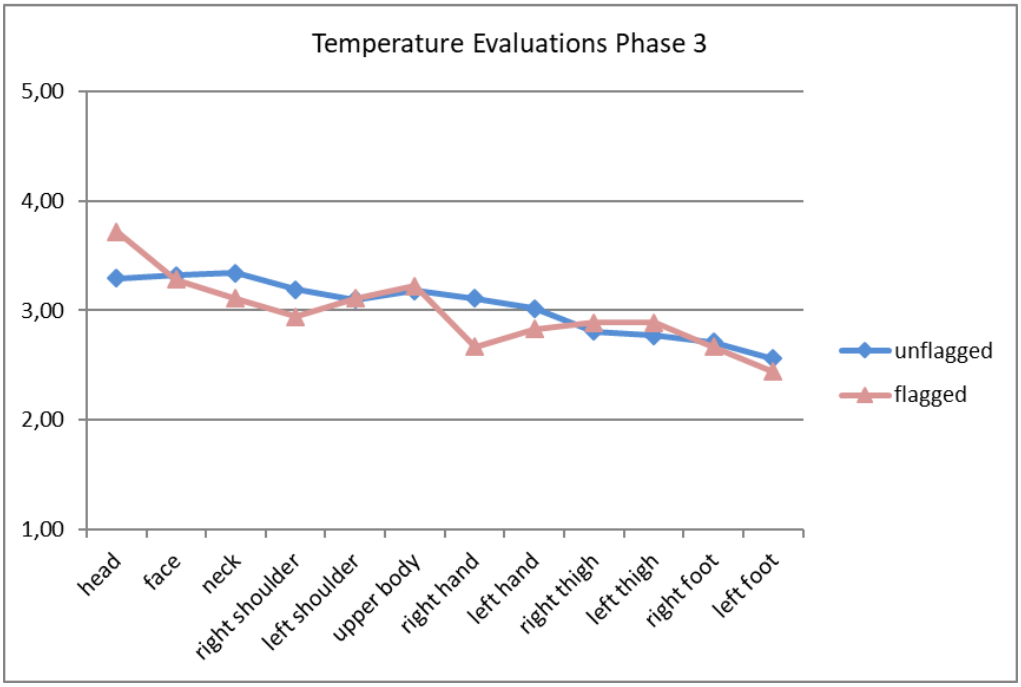

Figure 5: Temperature evaluations for different body parts. N = 80 flagged vs unflagged subjects

Finally, it would be interesting to see, whether the correlations between objectively measured temperatures and subjectively assessed temperatures would change substantially, if subjects were removed for indications of aberrant response behaviours. These correlation analyses were conducted with the full sample (N = 160) in comparison to the sample in which subjects were eliminated for signs of aberrant responses. Both samples were therefore overlapping and tests for significant differences not possible. As can be seen in Table 21 as expected all correlations were significant but the differences of the coefficients for the two samples were rather small. If at all, a consistent tendency could only be stated for the measurements at the foot level. Therefore, we concluded that the data quality of the subjective measures was not substantially degraded by certain response styles of the subjects.

Table 21: Correlations between objective and subjective temperature measures for the full and the cleaned sample

	Head Full /clean sample	Upper body Full /clean sample	Feet Full /clean sample
Phase 1	.39** / .42**	.36** / .39**	.37** / .39**
Phase 2	.27** / .25*	.43** / .46**	.46** / .53**
Phase 3	.32** / .38**	.44** / .49**	.57** / .64**
Phase 4	.42** / .36**	.46** / .45**	.59** / .63**

Note. N = 160 for the full sample. In the clean sample, N_H = 114-124 for head measures, N_B = 89-95 for body measures, N_F = 118-134 for feet measures

6 Discussion

Human subject trials are an indispensable step when determining the most suitable thermal comfort conditions for occupants of indoor environments such as homes, offices, or transportation means. However, if different members in a sample are being asked to assess several climate parameters under controlled conditions in the same situation, usually considerable variation in their assessments will be observed. This variation could be regarded as error variance due to careless response patterns, casting doubt on the credibility of such assessments (Kim & Moses, 2018; Maniaci & Rogge, 2014; Yentes, 2020). Or it can be seen as a reflection of true individual differences, which is valuable information to better adjust the climatization system to the individual needs.

Naturally, subjective judgements about thermal conditions are not free of errors. For example, some subjects might be unable to perceive and quantify different climate characteristics accurately or they might be simply unmotivated to fill in endlessly tiring survey questions with persistent care (Dunn et al., 2018). Besides person attributes also the measurement instrument itself, the instruction text, or the method of data collection could influence the extent of odd-looking responses. Edwards (2019) provided a comprehensive overview of such error sources.

In this report we analyzed the response profiles of N = 160 participants of a thermal comfort study of new train compartments for aberrant response patterns, which might be related to careless response styles or insufficient effort. The overall aim of this research is to determine the quantity of flawed climate assessments and to provide an empirical estimate for the effects of that bias on the general conclusions about climate comfort.

In a broader literature review, several techniques to detect aberrant responses were compared. Based on studies by DeSimone et al. (2015) and DeSimone and Harms (2017), these techniques can be categorized into *direct indices* (e.g. self-reports about attentive responding that are obvious to the respondent), *unobtrusive indices* (e.g. less obvious indicators derived from the pattern of responses), and *statistical indices* (e.g. calculations based on certain items). The following four indices were applied to the comfort assessments of the subject trials:

- Longstrings (LOS)

- Intra-individual response variabilities (IRV)
- Psychometric synonyms (PSY)
- Mahalanobis Distances (MAD)

For LOS and MAD, higher values indicate suspicious response behaviours (i.e., aberrant response patterns), for IRV and PSY lower scores would raise doubts about the quality of the respective assessments.

The computation of theses indices can conveniently be done by using the R-Package Careless from Yentes and Wilhelm (2018). However, a clear drawback of these indices is the lack of undisputable threshold scores for response patterns being classified as aberrant or not aberrant (McGonagle et al., 2016; Niessen et al., 2016). To a significant degree, the cut-off definition relies on the survey items (number, polarity, format, wording etc.), the homogeneity of the sample, and on the assumptions of the researcher as to how strict the cutoff scores for the different indices should be defined. Since these parameters usually change from study to study it is often not possible to determine whether aberrant responses are more frequent in study A than in study B. For this reason, we have decided to define two cutoff scores for each indicator, one more lenient and one more stringent (see McGonagle et al., 2016). Under these conditions six different research questions (RQ) were examined.

In RQ 1, we were examining whether the *length of the experiment* accounted for an increasing amount of aberrant responses in subjective assessments of climate parameters. In fact, across the four times of repeated measurements an increasing trend for obtrusive assessments could be observed. However, only for the index of psychometric synonyms this trend was statistically significant. Though, we cannot clearly distinguish whether the repetitive nature of the climate assessments or the plain time length itself was the main contributing factor, this result confirms that specific modalities of the data gathering can increase or diminish the extent of aberrant responses by the participants. Our finding is in line with results of Bowling et al. (2020) who reported an increasing level of careless responding in a 500 items survey with 358 university students as the participants progressed further through the questionnaire. Findings by Eisele et al. (2020) and Gibson and Bowling (2020) also indicate that a greater survey length is positively associated with careless responses.

In RQ2, the intended comparison of aberrant responses between

different measurement instruments (climate assessments, climate preferences, personality questionnaire) was not conclusive. Because of different item numbers and specific item wordings, not all indicators could be calculated for all instruments. For example, the algorithm for the psychometric synonyms did not identify any synonymous item-pair for the short form of the Big-5. Furthermore, there is no universal rule available for the definition of the cutoff scores for different measurement instruments. Therefore, a comparison between completely different instruments is not possible. This displays that using indicators to detect aberrant response patterns is highly dependent on the questionnaire and data at hand (Bowling et al., 2020; Eisele et al., 2020; Gibson & Bowling, 2020). Thus, researches should carefully decide on which indicators they want to use for their respective study.

In our climate comfort study, the same climate assessments were gathered from the same subjects in four consecutive phases within a time-span of roughly 3 to 3.5 hours. As part of the experimental conditions some climate parameters were gradually changing during this time period. In RQ 3 we analyzed whether *interindividual differences in the amount of aberrant responses remained stable* over this time. Pearson correlations resulted in inter-phase stability coefficients of $\bar{r}$ = .42 for LOS, $\bar{r}$ = .93 for IRV, $\bar{r}$ = .47 for PSY, and $\bar{r}$ = .83 for MAD. The stabilities were generally highest between adjacent phases. According to these findings, the interindividual differences of aberrant responses seemed to be quite stable across different points of time. Especially, the response behaviours represented by IRV and MAD seem to be related to stable person characteristics, which do not change randomly between occasions. This is in line with findings of positive associations between personality traits and aberrant response behaviour (Dunn et al., 2018; Edwards, 2019; Maniaci & Rogge, 2014).

The differences in the stabilities let us look into the *consistency between the four analyzed indicators* (RQ4). We used Pearson correlations between the values of the different indicators for the same phase of the climate assessments. This resulted altogether in six averaged consistency correlations ((4 * 3) / 2 = 6), which showed an average consistency of $\bar{r}$ = .44 between IRV and PSY and of $\bar{r}$ = -.39 between PSY and MAD. Consistencies between LOS and the other indicators were low (between -.04 and -.14). MAD showed consistencies in the opposite direction than expected. Higher Mahalanonis distances corresponded to shorter Longstrings ($\bar{r}$ = -.13) and to higher intra-individual response variabilities. McGonagle (2019) reported similar results. In her study, careless

responding metrics were generally correlated with the exception of the Mahalanobis Distances. The correlation between MAD and IRV was also positive (r = .49) and therefore opposite to the expectations. According to our data, the only indicator which was consistent with all other indicators was that of psychometric synonyms. The usefulness of the psychometric synonyms was also previously demonstrated by Maniaci & Rogge (2014) and Meade and Craig (2012). The advantage of our correlation analyses was that any cutoff settings for the indicators were irrelevant.

However, if we complement these consistency coefficients with the number of subjects flagged simultaneously by different indicators as showing aberrant responses, we found that this overlap remained rather small. For the more lenient cutoffs, an average of 34% of the subjects were flagged simultaneously by LOS and IRV (21 % for the more stringent cutoff). The largest overlap was found between PSY and MAD. On average 39% subjects were flagged for the lenient and 41% for the stringent cutoff.

In conclusion, the consistency analyses demonstrated that the practical commonalities of these indicators are only moderate. Subjects flagged for aberrant responses by one indicator were not necessarily flagged by the other indicators as well. It rather seemed that each indicator has a specific sensitivity for different aspects of response behaviours. Prior analyses of intercorrelations between different indicators have also revealed mediocre associations between the indicators and thus shown similar patterns (Huang et al., 2012; Maniaci & Rogge, 2014; Meade & Craig, 2012). However, we found that the psychometric synonyms showed repeatedly the highest degree of consistency with the other indicators, especially with MAD.

If aberrant response behaviour is rooted in stable person characteristics it can be expected to find *correlations between the indicators and personality* measures (RQ5). However, according to our findings, we can summarize that the relationship of aberrant response indicators and personality scales is less clear than in previous research (e.g., Marjanovic et al. 2015, Dunn et al., 2018; Maniaci & Rogge, 2014). In our study, the extent of aberrant responses in the climate assessments cannot be explained with personality factors such as the Big 5 personality scales. The size of the correlations we found is not even worth to mention and does not confirm our expectations. Also, relations to age and gender were negligible. Given the rather extensive data assessment we relied on (i.e., experiments that took over 3 hours in total), other variables such as fatigue (Gibson & Bowling, 2019) or boredom proneness (e.g., Harris, 2000) might play

a more important role in explaining aberrant response variance.

A large part of this report was dedicated to the analysis of the nature of indicators of aberrant response behaviours in subjective climate assessments. Four commonly used indicators could be examined in our data set. In repeated measurements they demonstrated a reasonable degree of stability. Each indicator seemed to represent distinct aspects of subjects' response behaviour. With the final research question (RQ6) we tried to determine the extent to which *practical conclusions about thermal comfort* in indoor environments could become misleading by subjects displaying insufficient effort when responding to the climate assessments. A multi-hurdle approach was applied by which subjects were eliminated from the sample if they had been flagged by any of the indicators. This approach resulted in 28 (17.5%) to 40 (25%) eliminated subjects for the stringent threshold and 58 (36.3%) to 73 (45.6%) eliminated subjects for the lenient threshold. Since we did not see any reasons to reduce the sample by more than 30%, we decided to pursue only the more stringent approach. With multivariate analyses of variance, we examined the effects of the independent factor "aberrant responses" (flagged vs. not flagged) on the climate assessment variables (sensations and evaluations of air draught and ambient air temperature) in each of the four phases of measurement. In 16 analyses only two effects were found as being significant. In all other cases the two groups did not differ from each other systematically. The two significant effects were found for the temperature sensations in phase 2 and the temperature evaluations in phase 3. The HVAC system under investigation was the displacement ventilation which usually has a stronger vertical air temperature gradient with colder temperature close to the floor mounted outlets and warmer temperatures close to the extraction slots near the ceiling where the contaminated air is being removed. This vertical gradient was somewhat clearer recognizable in the assessments of subjects not being flagged for aberrant responses. While this can be seen as a weak indication of higher response quality, further research is needed to consolidate and expand this finding.

In addition to this comparison of mean climate assessments for the flagged and the unflagged subjects, we analyzed whether the correlation between objective and subjective temperatures differed significantly in relation to aberrant response patterns. These correlations are practically identical for the full sample and the cleaned sample, from which the flagged subjects had been removed. The average correlations for the full sample versus the clean sample were for the head level $\bar{r}$ = .37 vs $\bar{r}$ = .37, for the upper body level $\bar{r}$ = .45 vs. $\bar{r}$ =

.48 and for the feet level $\bar{r}$ = .55 vs $\bar{r}$ = .63. Basically, the differences are close to zero although they tend to go into the desired direction.

In total, we concluded that the data quality of the subjective climate assessments was not substantially degraded by certain response styles of the subjects. The potential shift is rather small and not distinctively biasing the results. We regard these findings as providing vital evidence that subjective climate assessments in human subject trials with a reasonably sound measurement instrument are not considerably flawed by judgment error. On the contrary, this report demonstrates that human subject trials as part of a study of thermal comfort in indoor environments contribute indeed true variance, which is essential to better adjust the climate parameters to the needs of the customers.

Some specific conditions of this study might limit the generalizability of our findings. First and unlike Goldammer et al. (2020) or Huang et al. (2012), we did not follow an experimental approach were the amount of aberrant responses is controlled by the experimental setting (e.g., specific instruction texts, variations of measurement instruments). Therefore, we cannot determine with absolute certainty whether the response behaviours, which we identified as being aberrant were really related to error sources or simply to a lack of individual sensitivity or even absence of measurable differences in the environment. This might have led to the disappearance of correlations to personality traits found elsewhere. A second restriction is the short form of the personality questionnaire, which we used. It contains only 21 items which limits the applicability of some of the indicators. Even though, the short form has proven reasonable psychometric properties, its construct validity may be lower than that of the full questionnaire version. Finally, our sample size is not very large. Of course, the advantage of this study is that we used survey data from a real practical application, but tests of significance are losing statistical power if the sample is getting smaller. Therefore, smaller effects could have been overlooked. Lastly, the participants (i,e., the samples) were recruited by a professional agency and selected based on certain requirements. Whereas some variables were evenly distributed (e.g., gender), others have been standardized (e.g., education level). Moreover, a monetary incentive was given and a certain degree of self-selection could have taken place so that only interested individuals participated. Overall, this might have resulted in a restriction of variance compared to a representative sample. In other words, we recommend further experimental studies of the validity of the applied indicators of "unusual"

response behaviours of survey respondents.

7 Conclusions

Based on the preceding discussion of our results on the amount and consequences of aberrant responses in human subject trials, we provide a number of recommendations to prevent and to mitigate this potential source of errors in subjective assessments of thermal comfort. Our recommendations are regarded as complementary to already existing compilations of countermeasures and specifically geared to applied research settings. Detailed examples can be found in DeSimone et al. (2015), McGonagle et al. (2016), or Bowling et al. (2020).

- Probably, the best preventive measure is to limit the reasons for the survey participants to become inattentive or careless. Shorter surveys with a less repetitive nature are more motivating to respond in compliance to the instructions (Eisele et al., 2020; Gibson & Bowling, 2020). For instance, Bowling et al. (2020) showed that a more careless response style develops after 79 items in a row. Breaks could be scheduled for longer surveys and less important questions could be postponed to the end.
- Motivation to comply with the questionnaire instruction can be enhanced by a transparent briefing, which emphasizes and explains why accurate and careful responses of each individual participant are important for the results and purposes of the study. Whenever possible time-length and survey contents should be transparent to the participants in advance so they can decide themselves whether to participate or not. Relatedly, research has shown that a warning instruction can be correlated with less careless response behaviour (Huang et al., 2012; Meade & Craig, 2012; Ward & Pond, 2015). In applied settings such as ours, a sufficient attentiveness and diligence might be framed as requirement for the payment.
- If possible, a contact person should be available while participants are working on a questionnaire. Bowling et al. (2020) call this an "in-person proctor". However, besides supervisory control the contact person should have an encouraging or assisting role for example if issues with the question contents or entry formats might arise. This is even more important the more time is needed for the data gathering. If breaks are scheduled, the contact person could be open to some entertaining small talk as well.

- The item inventory should contain a number of bogus-items and allow to measure response times item-wise or per page time. These measures have proven effectiveness for the detection of aberrant response styles in other studies (Bagby et al., 1991; DeSimone et al., 2015). In dependence of content, Bogus items can also help to screen for responses that are distorted due to deceptive impression management (e.g., Levashina et al., 2009). Example lists of such items, which can be distributed across the entire survey are provided for example by Bowling et al. (2020).
- Berinsky et al. (2019) suggested to use "Screeners" to identify inattentive respondents. Screeners instruct participants to select a specific and otherwise atypical response to demonstrate their attention and diligence. Berinsky et al. (2014, 2019) and Meade et al. (2012) have shown the practical usefulness of Screeners. Specifically, Bersinky et al. (2019) suggest to use four Screeners of which two should be based on a grid format while the two remaining should be multiple-choice based.
- Before the data are being processed for hypothesis testing, the regular data cleaning phase should be augmented by a carelessness analysis (DeSimone et al., 2015). With the R-package by Yentes and Wilhelm (2018), this step is quite straightforward and does not require extended efforts (the code we have used can be found in the supplementary material). According to our results the indicators IRV and PSY appeared to be most conclusive in identifying subjects with obtrusive response behaviours. Cut-offs have to be defined based on the inspection of the respective indicator distributions. If subjects are being flagged, a multi-hurdle approach is recommended for their elimination. Overall, using statistical and unobtrusive indices to identify aberrant responses is a useful way to ensure data quality even after the data has been collected (DeSimone & Harms, 2017).

To some extent, aberrant responses could be identified in the subjective comfort assessments in our study. However, the overall data quality was not significantly degraded.

8 References

Abbey, J. D., & Meloy, M. G. (2017). Attention by design: Using attention checks to detect inattentive respondents and improve data quality. *Journal of Operations Management*, *53-56*, 63-70. https://doi.org/10.1016/j.jom.2017.06.001

Ashrae (2013). *Thermal environmental conditions for human occupancy*. ANSI/ASHRAE Standard 55-2013. ASHRAE, Atlanta/GA.

Bagby, R. M., Gillis, J. R., & Rogers, R. (1991). Effectiveness of the millon clinical multiaxial inventory validity index in the detection of random responding. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *3*(2), 285-287.

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, *58*(3), 739-753.

Berinsky, A. J., Margolis, M. F., Sances, M. W., & Warshaw, C. (2019). Using screeners to measure respondent attention on self-administered surveys: Which items and how many? *Political Science Research and Methods*, 1-8. https://doi.org/10.1017/psrm.2019.53

Berry, D. T. R., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment, 4*(3), 340–345. https://doi.org/10.1037/1040-3590.4.3.340

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2020). Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organizational Research Methods*, *00*(0), 1-21.https://doi.org/10.1177/1094428120947794

Costa, P. T., Jr., & McCrae, R. R. (2008). *The revised NEO personality inventory (NEO-PI-R).* In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE handbook of personality theory and assessment, Vol. 2. Personality measurement and testing* (p. 179–198). Sage Publications, Inc. https://doi.org/10.4135/9781849200479.n9

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in

survey data. *Journal of Experimental Social Psychology*, *66*, 4-19. https://doi.org/10.1016/j.jesp.2015.07.006

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The Mahalanobis distance. *Chemometrics and intelligent laboratory systems*, *50*(1), 1-18.

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, *36*(2), 171-181.https://doi.org/10.1002/job.1962

DeSimone, J. A., & Harms, P. D. (2017). Dirty Data: The effects of screening respondents who provide low-quality data in survey research. *Journal of Business and Psychology, 33(5)*, 559-577. https://doi.org/10.1007/s10869-017-9514-9

Dunn, A. M., Heggestad, E. D., Shanock, L. R., & Theilgard, N. (2018). Intra-individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationships with individual differences. *Journal of Business and Psychology*, *33*(1), 105-121.https://doi.org/10.1007/s10869-016-9479-0

Edwards, J. R. (2019). Response invalidity in empirical research: Causes, detection, and remedies. *Journal of Operations Management*, *65*(1), 62-76.https://doi.org/10.1016/j.jom.2018.12.002

Ehlers, C., Greene-Shortridge, T. M., Weekley, J. A., & Zajack, M. D. (2009). *The exploration of statistical methods in detecting random responding.* Paper presented at the annual meeting of the Society for Industrial/Organizational Psychology, Atlanta, GA

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2020). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*, 00(0), 1-16. https://doi.org/10.1177/1073191120957102

European Committee for Standardization (2016). *EN 13129:2016 Railway applications – Air conditioning for mainline rolling stock – Comfort parameters and type tests*. CEN-CENELEC Management Centre, Brussels.

Fanger, P.O. (1973). Assessment of man's thermal comfort in practice. *British Journal of Industrial Medicine*, *30*, 313-324.

https://doi.org/10.1136/oem.30.4.313

Fleischer, A., Mead, A. D., & Huang, J. (2015). Inattentive responding in MTurk and other online samples. *Industrial and Organizational Psychology*, *8*(2), 196-202. https://doi.org/10.1017/iop.2015.25

Gibson, A. M., & Bowling, N. A. (2020). The effects of questionnaire length and behavioral consequences on careless responding. *European Journal of Psychological Assessment, 36*(2), 410–420

Goldammer, P., Annen, H., Stöckli, P. L., & Jonas, K. (2020). Careless responding in questionnaire measures: Detection, impact, and remedies. *The Leadership Quarterly*, *31*(4). https://doi.org/10.1016/j.leaqua.2020.101384

Groves, R. M. (1987). Research on survey data quality. *The Public Opinion Quarterly*, *51*, 156-172.

Harris, M. B. (2000). Correlates and characteristics of boredom proneness and boredom. *Journal of Applied Social Psychology*, *30*(3), 576-598.

Harrison, D. A., & McLaughlin, M. E. (1993). Cognitive processes in self-report responses: tests of item context effects in work attitude measures. *Journal of Applied Psychology*, *78*(1), 129.

Hörmann, H.-J., Maier, J. & Zierke, O. (2016). *NGT III-AP0240: Ergebnisse NGT-Versuche 2.1+2.2. Objektive Daten.* Ergebnisbericht DLR-ME-PSY, Hamburg.

Hörmann, H.-J., Maier, J. & Zierke, O. (2017). Passengers' assessments of thermal comfort in a railway car mock-up with displacement ventilation. Paper presented at the *European Transport Conference*, 04 – 06 Oct 2017, Barcelona, Spain. https://elib.dlr.de/114384/

Hörmann, H.-J. (2017). *NGT III-AP0240: Ergebnisse NGT-Versuche 4.1+4.2. Objektive Daten.* Ergebnisbericht DLR-ME-PSY, Hamburg.

Hörmann, H.-J., Goerke, P., Koelzer, A., Maier, J. & Zierke, O. (2018). *NGT III: Dritter und vierter Probandenversuch zu Belüftungskonzepten*. Meilensteinbericht 02401712. DLR-ME-PSY, Hamburg.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99-114. https://doi.org/10.1007/s10869-011-9231-8

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*(3), 828.https://doi.org/doi: 10.1037/a0038510

Iaconelli, R., & Wolters, C. A. (2020). Insufficient effort responding in surveys assessing self-regulated learning: Nuisance or fatal flaw? *Frontline Learning Research*, *8*(3), 104-125. https://doi.org/10.14786/flr.v8i3.521

International Organization for Standardization (ISO) (2005). *ISO-7730:2005(E) - Ergonomics of the thermal environment - Analytical determination and interpretation of thermal comfort using calculation of the PMV and PPD indices and local thermal comfort criteria*. ISO, Geneva.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality, 39*, 103–129.https://doi.org/10.1016/j.jrp.2004.09.009

Kam, C. C. S., & Chan, G. H. H. (2018). Examination of the validity of instructed response items in identifying careless respondents. *Personality and Individual Differences*, *129*, 83-87.

Kim, S., & Moses, T. (2018). The impact of aberrant responses and detection in forced-choice noncognitive assessment. *ETS Research Report Series*, *2018*(1), 1-15. https://doi.org/10.1002/ets2.12222

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of personality and social psychology*, *55*(2), 312-320.

Lange, P., Schmeling, D., Hörmann, H.-J. & Volkmann, A. (2019). Comparison of local equivalent temperatures and subjective thermal comfort rating in a train compartment. Proceedings of the *10th International Conference on Indoor Air Quality, Ventilation and Energy Conservation in Buildings*. Bari/Italy, September 5-7, 2019. https://www.iaqvec2019.org/

Levashina, J., Morgeson, F. P., & Campion, M. A. (2009). They don't do it often, but they do it well: Exploring the relationship between applicant mental abilities and faking. *International Journal of Selection and Assessment*, *17*(3), 271-281.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India, 2*, 49–55.

Maier, J., Hörmann, H.-J. & Zinn, F. (2014). *NGT III-AP0240: Bericht zur Entwicklung von Versuchsdesign, Messkonzept und Erhebungsinstrument für Probandenversuche zum Klimakomfort*. Meilensteinbericht MS 02401412, DLR-ME-PSY, Hamburg.

Maier, J., Hörmann, H.-J., Kölzer, A. & Zinn, F. (2015). *NGT III-AP0240: Bericht zu den ersten Probandenversuchen - Klima*. Meilensteinbericht 02401512. DLR-ME-PSY, Hamburg.

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61-83.

Marggraf-Micheel, C., & Jaeger, Sabine (2007) *Erfassung des subjektiven Wohlbefindens in der Flugzeugkabine*. DLR-Research report. DLR-FB-2007-07. DLR, Cologne.

Marggraf-Micheel, Claudia und Piewald, Carina und Winzen, Julia und Berg, Jana (2010) *Thermischer Komfort in der Flugzeugkabine - Forschung im Do 728 Kabinen-Mock-Up.* DLR-Research report. DLR-FB 2010-07. DLR, Cologne.

Marjanovic, Z., Struthers, C. W., Cribbie, R., & Greenglass, E. R. (2014). The conscientious responders scale: A new tool for discriminating between conscientious and random responders. *Sage Open*, *4*(3), 1-10.

Marjanovic, Z., Holden, R., Struthers, W., Cribbie, R. & Greenglass, E. (2015). The inter-item standard deviation (ISD): An index that discriminates between conscientious and random responders. *Personality and Individual Differences, 84*, 79-83.

McCrae, R. R., & Costa, P. T., Jr. (1999). *A Five-Factor theory of personality.* In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (p. 139–153). Guilford Press.

McGonagle, A. D. (2019). *Caring more about careless responding: Applying the theory of planned behavior to reduce careless responding on online surveys*. Doctoral thesis. Colorado State University, Fort Collins/CO.

McGonagle, A. K., Huang, J. L., & Walsh, B. M. (2016). Insufficient effort survey responding: An under-appreciated problem in work and organisational health psychology research. *Applied Psychology*, *65*(2), 287-321.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455. https://doi.org/10.1037/a0028085

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, *63*, 1-11.

Nichols, A. L., & Edlund, J. E. (2020). Why don't we care more about carelessness? Understanding the causes and consequences of careless participants. *International Journal of Social Research Methodology*, 1-14.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology*, *45*(4), 867-872.

Rammstedt, B., & John, O. (2005). Kurzversion des Big Five Inventory (BFI-K) [Short version of the Big Five Inventory (BFI-K)]. *Diagnostica, 51*, 195-206. https://doi.org/10.1026/0012-1924.51.4.195

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five inventory in english and german. *Journal of Research in Personality, 41*(1), 203-212. https://doi.org/10.1016/j.jrp.2006.02.001

Rammstedt, B., Kemper, C.J., Klein, M.C., Beierlein, C., Kovaleva, A. (2013). A short scale for assessing the Big Five dimensions of personality: 10 item Big Five inventory (BFI-10) *Methoden, Daten, Analysen, 7*(2), 233-249. DOI: 10.12758/mda.2013.013

Sagui-Henson, S. J., Levens, S. M., & Blevins, C. L. (2018). Examining the psychological and emotional mechanisms of mindfulness that reduce stress to enhance healthy behaviours. *Stress and Health*, *34*(3), 379-390.

Schmeling, D. & Volkmann, A. (2016). *NGT III-AP0250: Messtechnik- und Belüftungskonzept für GZG und ZVG erarbeitet*. Zwischenbericht DLR-AS-FLY, Göttingen.

Schmeling, D. & Volkmann, A. (2017). *NGT III-AP0250: Kabinenmessanlagen ZVG und GZG einsatzbereit*. Zwischenbericht DLR-AS-FLY, Göttingen.

Schmeling, D., Hörmann, H.-J., Volkmann, A. & Goerke, P. (2019). Impact of local comfort zones in long-distance rolling stock on objective and

subjective thermal comfort rating. Proceedings of the *12th World Congress on Railway Research*. Tokio, Oct. 28th – Nov 3rd.

Schmeling, D. & Volkmann, A. (2020). On the experimental investigation of novel low-momentum ventilation concepts for cooling operation in a train compartment. *Building and Environment, 182*, 1-12. https://doi.org/10.1016/j.buildenv.2020.107116

Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, *9*, 367–373.

Ward, M. K., Meade, A. W., Allred, C. M., Pappalardo, G., & Stoughton, J. W. (2017). Careless response and attrition as sources of bias in online survey assessments of personality traits and performance. *Computers in Human Behavior*, *76*, 417-430. https://doi.org/10.1016/j.chb.2017.06.032

Ward, M. K., & Meade, A. W. (2018). Applying social psychology to prevent careless responding during online surveys. *Applied Psychology*, *67*(2), 231-263.

Winter, J. (2012). Next Generation Train. *Der Eisenbahningenieur*, 4, 32-26.

Yentes, R.D., & Wilhelm, F. (2018) *Careless: Procedures for computing indices of careless responding*. R packages version 1.2.1 url: https://github.com/ryentes/careless

Yentes, R. D. (2020). *In search of best practices for the identification and removal of careless responders*. Doctoral thesis, North Carolina State University.

Yu, X., & Cheng, Y. (2019). A change-point analysis procedure based on weighted residuals to detect back random responding. *Psychological Methods, 24*(5), 658–674. https://doi.org/10.1037/met0000212

Zijlstra, W. P., van der Ark, L. A., & Sijtsma, K. (2011). Outliers in questionnaire data. *Journal of Educational and Behavioral Statistics, 36*(2), 186-212. https://doi.org/10.3102/1076998610366263

9 Supplementary material

9.1 Extent of aberrant responses

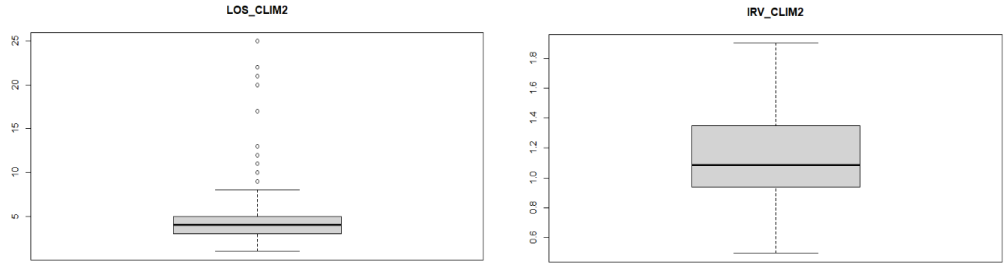

Figure 6: Boxplots for the indicators LOS and IRV within the assessments of climate parameters in phase 2 (95% intervals)

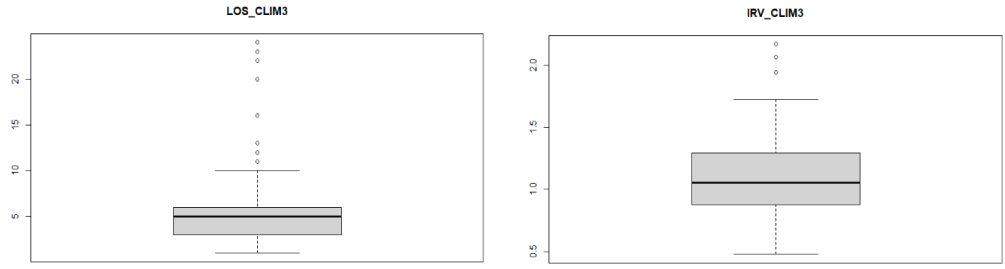

Figure 7: Boxplots for the indicators LOS and IRV within the assessments of climate parameters in phase 3 (95% intervals)

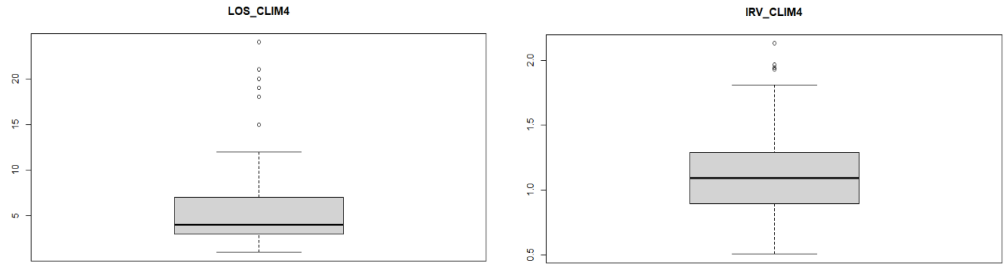

Figure 8: Boxplots for the indicators LOS and IRV within the assessments of climate parameters in phase 4 (95% intervals)

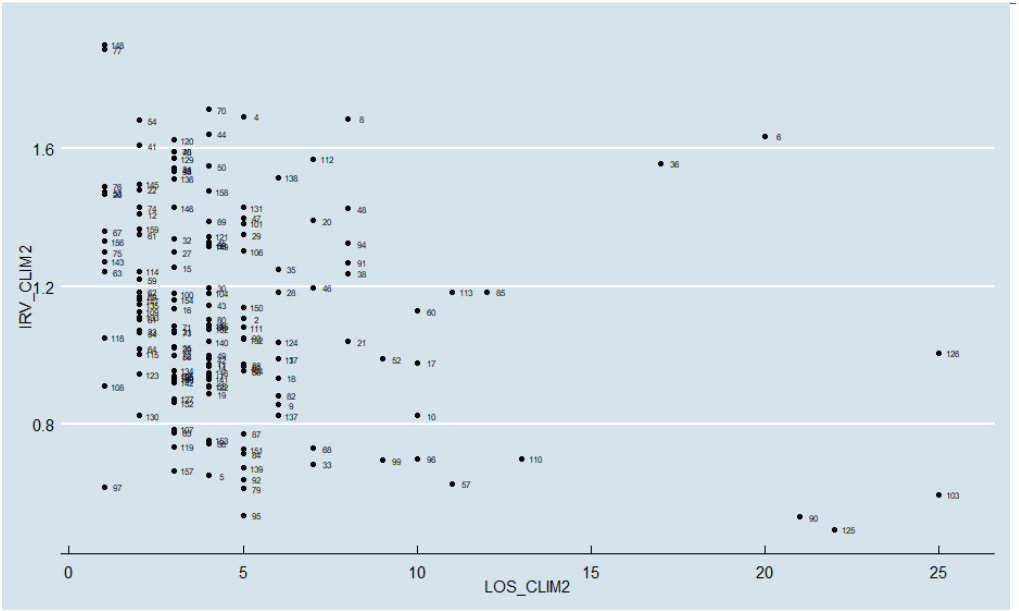

Figure 9: Scatterplot for LOS and IRV within the assessments of climate parameters in phase 2

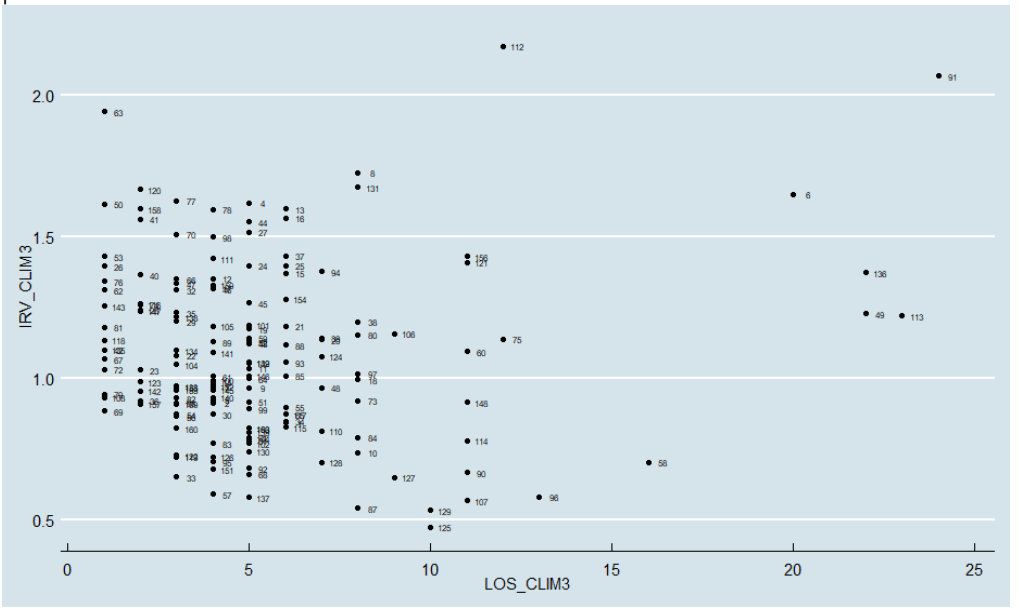

Figure 10: Scatterplot for LOS and IRV within the assessments of climate parameters in phase 3

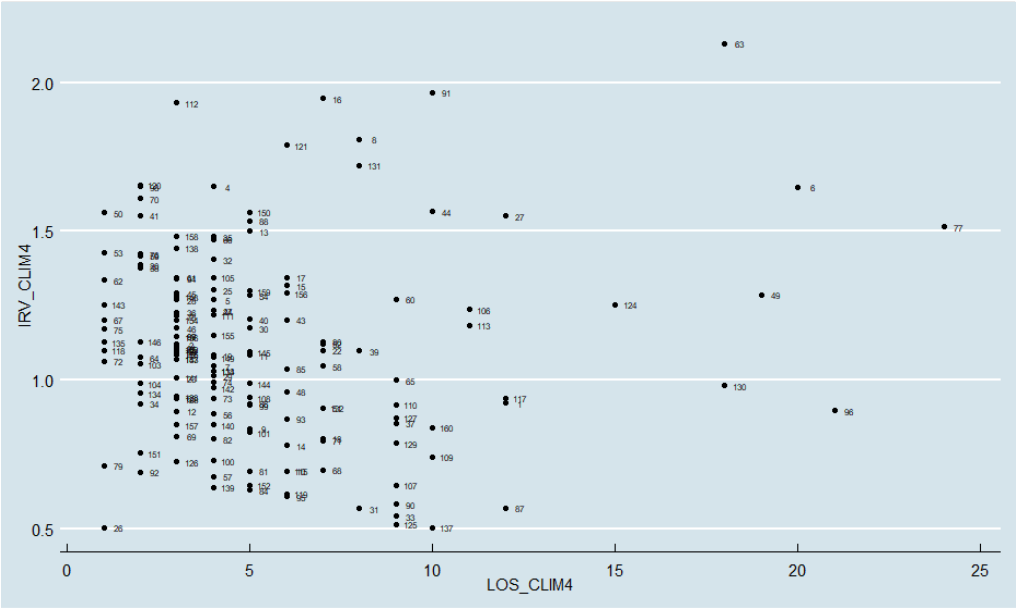

Figure 11: Scatterplot for LOS and IRV within the assessments of climate parameters in phase 4

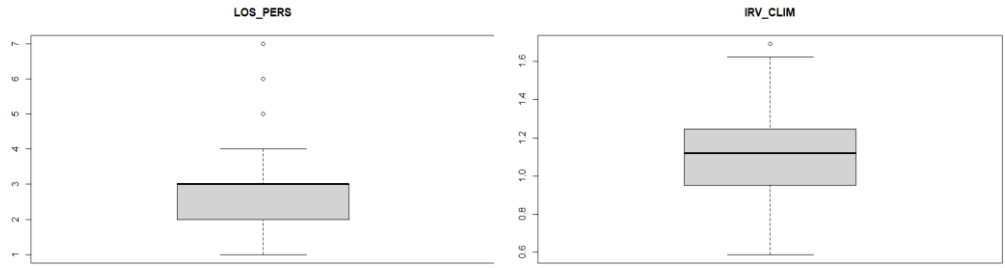

Figure 12: Boxplots for the indicators LOS and IRV within the personality questionnaire

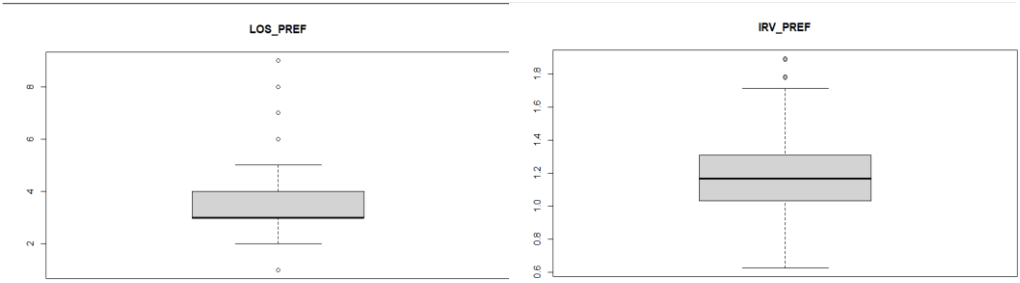

Figure 13: Boxplots for the indicators LOS and IRV within the climate preferences

9.2 Consistency of indicators

Table 22: Correlation matrix of Long-strings and Individual Response Variabilities for the climate assessments in phase 1 to 4

Pearson's Correlations

Figure 14: Boxplots for the indicators LOS and IRV within the climate preferences (95% intervals)

LOS_CLIM4	0.191	0.182	0.435	—				
IRV_CLIM1	-0.232	-0.214	-0.040	-0.053	—			
IRV_CLIM2	-0.207	-0.247	0.014	-0.057	0.715	—		
IRV_CLIM3	-0.173	-0.175	0.057	-0.020	0.648	0.664	—	
IRV_CLIM4	-0.156	-0.111	0.123	0.026	0.535	0.631	0.802	—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 23: Correlation matrix of Long-strings and Psychometric Synonyms for the climate assessments in phase 1 to 4

Pearson's Correlations

Variable	LOS_CLIM1	LOS_CLIM2	LOS_CLIM3	LOS_CLIM4	PSY_CLIM1	PSY_CLIM2	PSY_CLIM3	PSY_CLIM4
LOS_CLIM1	—							
LOS_CLIM2	0.460 ***	—						
LOS_CLIM3	0.183 *	0.296 ***	—					
LOS_CLIM4	0.191 *	0.182 *	0.435 ***	—				
PSY_CLIM1	-0.090	-0.089	-0.115	-0.164 *	—			
PSY_CLIM2	0.043	-0.067	-0.183 *	-0.047	0.283 ***	—		
PSY_CLIM3	0.026	-0.010	-0.032	0.012	0.300 ***	0.323 ***	—	
PSY_CLIM4	-0.058	-0.056	0.046	0.031	0.188 *	0.385 ***	0.658 ***	—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 24: Correlation matrix of Long-strings and Mahalanobis Distances for the climate assessments in phase 1 to 4

Pearson's Correlations

Variable	LOS_CLIM1	LOS_CLIM2	LOS_CLIM3	LOS_CLIM4	MAD_CLIM1	MAD_CLIM2	MAD_CLIM3	MAD_CLIM4
LOS_CLIM1	—							
LOS_CLIM2	0.460 ***	—						
LOS_CLIM3	0.183 *	0.296 ***	—					
LOS_CLIM4	0.191 *	0.182 *	0.435 ***	—				
MAD_CLIM1	-0.161 *	-0.044	0.076	0.078	—			
MAD_CLIM2	-0.169 *	-0.159 *	0.019	-0.033	0.684 ***	—		
MAD_CLIM3	-0.182 *	-0.127	-0.045	-0.092	0.652 ***	0.689 ***	—	
MAD_CLIM4	-0.150	-0.119	-0.014	-0.148	0.519 ***	0.613 ***	0.662 ***	—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 25: Correlation matrix of Individual Response Variabilities and Psychometric Synonyms for the climate assessments in phase 1 to 4

Pearson's Correlations

Variable	IRV_CLIM1	IRV_CLIM2	IRV_CLIM3	IRV_CLIM4	PSY_CLIM1	PSY_CLIM2	PSY_CLIM3	PSY_CLIM4
IRV_CLIM1	—							
IRV_CLIM2	0.715 ***	—						
IRV_CLIM3	0.648 ***	0.664 ***	—					
IRV_CLIM4	0.535 ***	0.631 ***	0.802 ***	—				
PSY_CLIM1	0.397 ***	0.235 **	0.286 ***	0.192 *	—			
PSY_CLIM2	0.198 *	0.376 ***	0.171 *	0.188 *	0.283 ***	—		
PSY_CLIM3	0.178 *	0.191 *	0.426 ***	0.316 ***	0.300 ***	0.323 ***	—	
PSY_CLIM4	0.211 **	0.315 ***	0.402 ***	0.441 ***	0.188 *	0.385 ***	0.658 ***	—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 26: Correlation matrix of Individual Response Variabilities and Mahalanobis Distances for the climate assessments in phase 1 to 4

Variable	IRV_CLIM1	IRV_CLIM2	IRV_CLIM3	IRV_CLIM4	MAD_CLIM1	MAD_CLIM2	MAD_CLIM3	MAD_CLIM4
IRV_CLIM1	—							
IRV_CLIM2	0.715 ***	—						
IRV_CLIM3	0.648 ***	0.664 ***	—					
IRV_CLIM4	0.535 ***	0.631 ***	0.802 ***	—				
MAD_CLIM1	0.390 ***	0.299 ***	0.262 ***	0.261 ***	—			
MAD_CLIM2	0.304 ***	0.355 ***	0.309 ***	0.260 ***	0.684 ***	—		
MAD_CLIM3	0.189 *	0.235 **	0.305 ***	0.210 **	0.652 ***	0.689 ***	—	
MAD_CLIM4	0.137	0.137	0.184 *	0.302 ***	0.519 ***	0.613 ***	0.662 ***	—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Table 27: Correlation matrix of Psychometric Synonyms and Mahalanobis Distances for the climate assessments in phase 1 to 4

Pearson's Correlations

Variable	PSY_CLIM1	PSY_CLIM2	PSY_CLIM3	PSY_CLIM4	MAD_CLIM1	MAD_CLIM2	MAD_CLIM3	MAD_CLIM4
PSY_CLIM1	—							
PSY_CLIM2	0.283 ***	—						
PSY_CLIM3	0.300 ***	0.323 ***	—					
PSY_CLIM4	0.188 *	0.385 ***	0.658 ***	—				
MAD_CLIM1	-0.205 **	-0.288 ***	-0.365 ***	-0.317 ***	—			
MAD_CLIM2	-0.119	-0.343 ***	-0.280 ***	-0.325 ***	0.684 ***	—		
MAD_CLIM3	-0.079	-0.214 **	-0.415 ***	-0.372 ***	0.652 ***	0.689 ***	—	
MAD_CLIM4	-0.094	-0.279 ***	-0.386 ***	-0.512 ***	0.519 ***	0.613 ***	0.662 ***	—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

9.3 R-Code to determine aberrant response indicators

```
########################################################################
# R-Code to calculate different indicators of aberrant reponses in survey data
#
# cleaning workspace
rm(list = ls())

# set working directory
setwd("E:/R-WD-Files/NGT")

# load packages
library(ggplot2)
library(careless)
library(readxl)
library(ggthemes)
library(dplyr)
library(xlsx)

# Read data matrix from Excel-spreadsheet
climdata_p1 <- read_excel("E:/R-WD-Files/NGT/datasets/surveydata.xlsx", range =
"BL1:DG161", na = "#NULL!")
climdata_p2 <- read_excel("E:/R-WD-Files/NGT/datasets/surveydata.xlsx", range =
"GE1:HZ161", na = "#NULL!")
climdata_p3 <- read_excel("E:/R-WD-Files/NGT/datasets/surveydata.xlsx", range =
"OQ1:QL161", na = "#NULL!")
climdata_p4 <- read_excel("E:/R-WD-Files/NGT/datasets/surveydata.xlsx", range =
```

```
"TJ1:VE161", na = "#NULL!")
persdata <- read_excel("E:/R-WD-Files/NGT/datasets/surveydata.xlsx", range =
"KX1:LR161", na = "#NULL!")
prefdata_p1 <- read_excel("E:/R-WD-Files/NGT/datasets/surveydata.xlsx", range =
"ME1:NU161", na = "#NULL!")

# Calculation of indicators  ****************************************
# Climate assessment data phase 1
# LongString
LOS_CLIM1 <- longstring(climdata_p1, avg = F)
boxplot(LOS_CLIM1, main = "LOS_CLIM1")
LOS_OVER_15 <-matrix(LOS_CLIM1)
# IRV
IRV_CLIM1 <- irv(climdata_p1)
IRV_CLIM1_obstrusive <- matrix(IRV_CLIM1 <= 0.60)
IRV_CLIM1_Data <- matrix(IRV_CLIM1)
boxplot(IRV_CLIM1, main = "IRV_CLIM1")
# 2D Graphic LongString WITH IRV
ggplot(climdata_p1, aes(x=LOS_CLIM1, y=IRV_CLIM1)) + geom_point
()+theme_economist()+theme(legend.position = "bottom") + scale_fill_economist() +
geom_text(label=rownames (climdata_p1), nudge_x=0.4,size=2)
# Mahalanobis Distances for high (99%) and low (95%) thresholds
MAD_CLIM1_99 <- mahad_raw <- mahad(climdata_p1, flag = TRUE, confidence = 0.99,
na.rm = TRUE)
MAD_CLIM1_95 <- mahad_raw <- mahad(climdata_p1, flag = TRUE, confidence = 0.95,
na.rm = TRUE)
# Psychometric Synonyms for high (r=.70) and low (r=.60) thresholds
PSY_CLIM1_cor <- psychsyn_critval(climdata_p1)
PSY_CLIM1_70 <- psychsyn(climdata_p1, .70)
PSY_CLIM1_70_Values <- matrix(PSY_CLIM1_70)
PSY_CLIM1_60 <- psychsyn(climdata_p1, .60)
PSY_CLIM1_60_Values <- matrix(PSY_CLIM1_60)
# same for phase 2, 3, and 4
# Climate preferences data
# LongString
LOS_PREF <- longstring(prefdata, avg = T)
boxplot(LOS_PREF, main = "Boxplot Longstring Climate Preferences")
boxplot(LOS_PREF$avg, main = "Boxplot Longstring Climate Preferences")
LOS_PREF <- longstring(prefdata)
#IRV
IRV_PREF <- irv(prefdata)
IRV_PREF_obstrusive <- matrix(IRV_PREF <= 0.60)
```

```
IRV_PREF_Data <- matrix(IRV_PREF)
boxplot(IRV_PREF, main = "Boxplot IRV Climate Preferences")
# 2D Graphic LongString WITH IRV
longs_prefdata <- longstring(prefdata)
irv_prefdata <- irv(prefdata)
ggplot(prefdata, aes(x=longs_prefdata, y=irv_prefdata)) +
geom_point()+theme_economist()+theme(legend.position = "bottom") +
scale_fill_economist() + geom_text(label=rownames(l_and_t_Values_p4),
nudge_x=0.4,size=2)
# Mahalanobis Distances for (99%) threshold
MAD_PREF <- mahad_raw <- mahad(prefdata, flag = TRUE, confidence = 0.99)
MAD_PREF_Data <- mahad_raw <- mahad(prefdata, na.rm = TRUE)
# Psychometric Synonyms for r=.60 threshold
PSY_PREF_cor <- psychsyn_critval(prefdata)
PSY_PREF <- psychsyn(prefdata, .60)
PSY_PREF_Values <- matrix(PSY_PREF)
# same for the personality data

# save working directory
save.image("E:/R-WD-Files/NGT")
#########################################################################
```