PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Extreme response style in listening tests

Christoph Jakobs[(1)], Sebastian Böldt[(2)], Dustin Selbach[(3)], Jochen Steffens[(4)]

[(1)]Audio Communication Group TU Berlin, Germany, christoph_jakobs@campus.tu-berlin.de
[(2)]Audio Communication Group TU Berlin, Germany, sebastian.boeldt@campus.tu-berlin.de
[(3)]Audio Communication Group TU Berlin, Germany, d.selbach@campus.tu-berlin.de
[(4)]Audio Communication Group TU Berlin, Germany, jochen.steffens@tu-berlin.de

Abstract

This paper is the first to examine the influence of sociodemographic characteristics of listeners on the tendency to extreme response style (ERS) in psychoacoustical experiments. ERS can be described as the trend to exploit the edges of the rating scale in quantitative surveys leading to a systematic deviation of the respondents' reported values from the true values. According to the psychological literature, we assumed that ERS is more pronounced in women compared to men. Furthermore, we expected that ERS increases with ascending age of people and increasing duration of the survey. To test our hypotheses, we re-analyzised the data obtained by Weinzierl et al. (2018). In their study, 190 subjects rated their acoustical impression of 35 binaurally simulated rooms by means of the Room Acoustical Quality Inventory (RAQI). Results of linear mixed-effect models did not reveal any significant influences of age on ERS. In contrast, results confirmed our assumption that women show a significantly higher frequency of ERS than men, as well as increasing duration of the survey leads to a higher ERS. The results therefore highlight the role of ERS in sound evaluations and the need to consider this moderating factor when conducting and analysing psychoacoustical experiments.

Keywords: Sound evaluations; Extreme Response Style, ERS, Psychoacoustics, Room acoustics

1 INTRODUCTION

Research on the perception and evaluation of sounds often relies on data from quantitative questionnaire studies. The questionnaire method can be described as a 'targeted, systematic and rule-guided generation and recording of verbal and numerical self-disclosures of interviewees on selected aspects in written form' [1, p. 398], including both experience and behavior. Previous research from other fields suggest that such studies might be susceptible to certain response tendencies. Such response tendencies can, for example, arise from the desire to reduce cognitive effort or from attachment to social norms [2]. In addition, interviewer effects such as the presence of a third person and a perceived social or intellectual obligation to the investigator have been shown to govern certain response tendencies (ibid.). Such effects are independent of the actual content of a study, but can lead to a distortions of its results.

One phenomenon describing such response tendencies is the so-called Extreme Response Style (ERS) or, when one assumes the behavior to be stable within a person, Extreme Response Set [3]. ERS can be defined as the tendency to favour extreme over mean values in evaluation scale tests and has been observed in different fields of psychology [2]. To the best of our knowledge, ERS has not yet been connected to studies in the realm of psychoacoustics and room acoustics. Therefore, the major aim of the present paper was to fill this gap and investigate the role of ERS in acoustical research as well as to connect ERS to characteristics of the study and the participants involved.

1.1 Extreme Response Style (ERS) and person-related influences

Existing studies on ERS suggest that response behaviour is related to certain person-related variables such as age, gender as well as the educational and cultural background of study participants [3, 4]. A study by Grühn and Scheibe, for example, revealed that older people tend to give more extreme responses than younger ones [5]. Empirical findings regarding the role of gender, however, are mixed. While Borgatta et al. observed that women have a stronger tendency towards extreme answers than men [6], research by Greenleaf [3] could not confirm such a connection. Regarding the cultural influences, research suggests that test persons with an African or Hispanic cultural background tend to express their honest response by using more extreme values of a scale while Central European, for example, show a less extreme response behaviour [7]. Therefore, differences in ERS are assumed to be the result of diversities in judgment styles across cultures.

1.2 Extreme Response Style (ERS) and study-related influences

Furthermore, it is suggested that features of a study and its design can influence response behaviour. A well-known phenomenon related to ERS is the so-called central tendency bias, describing the fact that test persons show a higher tendency towards the neutral middle category in scales with an odd number of scale points [1]. In addition, a study by Couper and colleagues [8] revealed that extreme values occur less frequently when using continuous visual scales due to their higher resolution of presented response categories, relative to the restricted range of other (categorical) input types. This links to another source, namely the number of presented response categories, which has been shown to reinforce extreme response behaviour [7]. An example of a potential miss-mapping is illustrated schematically in Figure 1. In contrast to [8], Kreidler et al. [9] did not observe an effect of the length of the presented analogue scale on response behaviour.

Another important factor potentially influencing ERS in psychoacoustic experiments might be the duration of a study. Hui and Triandis, for example, observed that ERS increases over the course of a survey [10]. This finding could be explained by the fact that test persons become more familiar with the scale and the evaluation criteria over time and therefore develop a tendency towards more extreme evaluations. By contrast, Naemi et al. found that a shorter processing time leads to a more extreme response behaviour [11]. In their study, subjects who 'rushed' through the processing sheet showed a higher tendency towards ERS compared to participants who took their time to answer the study questions.

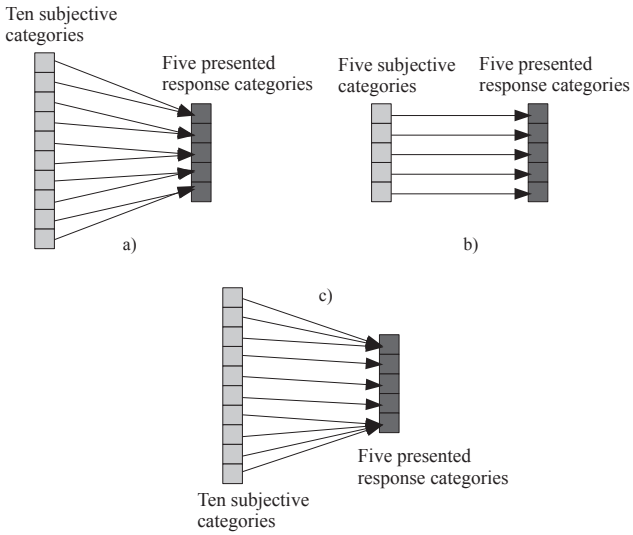


Figure 1. Matching of subjective evaluation categories to available response criteria according to [7]. a) Appropriate matching if number of subjective criteria exceeds available number of response criteria; b) Optimal matching with equal number of categories; c) Unsuitable matching of larger number of subjective categories to response criteria.

1.3 Hypotheses

As stated above, we are not aware of any investigations on ERS in the field of room acoustics or psychoacoustics. Therefore, we aimed on demonstrating ERS in these specific contexts and using on a dataset collected by Weinzierl and colleagues [12]. Based on the aforementioned literature, we formulated the following hypotheses:

H(1) Older people are more likely than younger people to show ERS (i.e. to use the edges of the scale) when evaluating their acoustic spatial perception.

H(2) Women are more inclined than men to show ERS when evaluating their spatial acoustic perception.

H(3) The longer a listening experiment (measured by the increasing number of trials), the more participants tend to show ERS.

Finally, without formulating a specific hypothesis, we explored how the likelihood of participants to show ERS might change when comparing an initial test (Session 1) with a retest (Session 2) conducted six weeks later.

2 METHOD

2.1 Computation of ERS Scores

One hundred ninety participants were presented with 35 binaurally simulated rooms from two different listening positions with symphonic orchestra, solo trumpet, and dramatic speech as audio content in the course of 14 test runs (for further details on the study see [12]). For each run, participants performed evaluations for the 46 items that were queried (i.e. room acoustical features or perceived sound characteristics) regarding their acoustical impression. In addition, a re-test exactly replicating the initial test was performed around 6 weeks later ($M = 42$ days, $SD = 37$) with $n = 88$ participants, corresponding to 46% of the original sample.

To investigate extreme response tendencies, the respective evaluations were assigned to ERS scores according to Greenleaf [3], representing the basis for our statistical analysis. With the introduction of a binary variable $l(n)$,

the evaluation of each individual item $x(n)$ was analysed for each test run with $n = [1,46]$ and in the case of an extreme response behaviour, the associated binary variable was set to 1, otherwise to 0. The allocation was made according to Equation 1:

$$l = \begin{cases} l(n) = 1, \text{if } x(n) \leq 15 \\ l(n) = 1, \text{if } x(n) \geq 85 \\ \text{else}, \quad l(n) = 0 \end{cases} \tag{1}$$

In accordance with [3], an answer in the lower or upper sixth of the rating scale was categorised as extreme response behaviour. Analogously, an answer in the remaining scale range was classified as moderate response behaviour. Based on this, an ERS score was computed for each trial of a participant. The calculation of the ERS score was determined by Equation 2:

$$P_{ERS} = \frac{1}{N} \sum_{n=1}^{N} l(n), \quad P_{ERS} \in \mathbb{Q}[0,1] \tag{2}$$

with $N = 46$ as total number of RAQI items to be evaluated and $l(n)$ as binary variable per item to be evaluated. The scores were then analysed across all trials to analyse the influence of age, gender, trial number and session on a person's tendency to show extreme response behaviour.

2.2 Statistical analysis

In order to take into account the hierarchical structure of the underlying dataset (multiple repeated measures per participant), linear mixed-effects models were computed to test our hypotheses, estimating both *fixed* and *random* effects. The model used here can be described by Equation 3:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij}, \quad where \quad \beta_{0j} = \beta_0 + u_{0j} \tag{3}$$

with β_{0j} representing the Y-intercept and $u_{0j} \sim N(0, \sigma^2_{u0})$ the random intercept [13]. β_{1j} defines the slope, X_{ij} the fixed effect, and $e_{ij} \sim N(0, \sigma^2_e)$ equals the residuum. The j index constitutes the cluster variable(s) (i.e. participant ID in the present study), indicating the hierarchical structure of the data, and i is the number of observations per participant. Therefore, the regression approach chosen here corresponds to a *Random Intercept Model*. To test the respective hypotheses, separate linear mixed-effect models for each independent variable and hypothesis were computed in RStudio using restricted maximum likelihood estimates of variance components and Type III Analysis of Variance via Satterthwaite's degrees of freedom method (lmerTest package; [14]). Besides specificing a random intercept for each participant, we introduced a second random intercept for the room ID to control for the influence of the stimulus itself. For all analyses, the significance level was set to $\alpha = .05$.

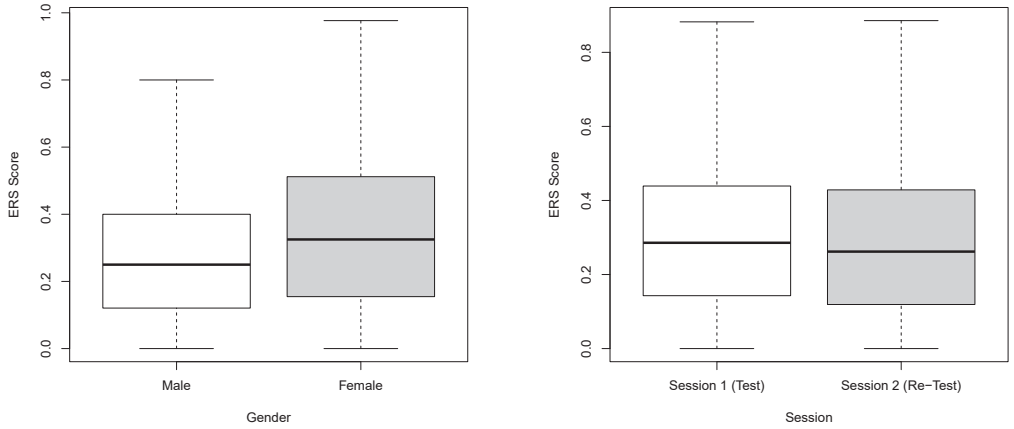
3 Results

Before testing the three hypotheses, a null model including only random intercepts for each participant and room was computed to obtain the Interclass Correlation Coefficients (ICCs). The ICCs indicate the proportion of variance explained by the respective random intercepts. Results from this analysis revealed that 49.3% of the variance in the ERS scores was explained by the random intercept for each participant ($ICC_{participants} = .493$), meaning that almost half of the variance was attributed to differences between persons. In contrast, only 4.0% of the variance was due to the different rooms evaluated by the participants ($ICC_{rooms} = .040$).

In the next step, the hypotheses and a potential effect of the session (Test vs. Retest) were tested by means of four different linear mixed-effects models. The results from the ANOVA tables as well as the variance

Table 1. Results of the four different linear mixed-effects models predicting ERS by the gender and age of the participants as well as the trial number and the study session

	SumSq.	*NumDF*	*DenDF*	*F*	*p*	R^2_{marginal}
Age	0.001	1	187.3	0.06	.80	.000
Gender	0.214	1	186.8	10.36	< .01*	.027
Trial	0.089	1	3652.3	4.36	.04*	.001
Session	0.247	1	3755.8	12.08	< .001*	.002



(a) (b)

Figure 2. ERS score dependent on the participants' gender and the test session (Test vs. Retest)

explained by the respective fixed effect (R^2_{marginal}) are displayed in Table 1. Looking at the first row of the table, it becomes evident that no significant influence of age on ERS was observed, and therefore hypothesis H(1) was not confirmed. In contrast, the results from the other three linear-mixed effects models shown in Table 1 revealed significant effects of a person's gender, the duration of the study measured by the trial number, and the session (Test vs. Retest) on ERS.

The gender effect assumed by H(2) and confirmed by our analyses shows that, on average, higher ERS scores were observed for women ($M = 0.346$, $SE = 0.016$) than for men ($M = 0.275$, $SE = 0.016$). This mean difference is illustrated by Figure 2a. The R^2_{marginal} reported in the second row of Table 1 shows that 2.7% of the variance in the ERS scores can be attributed to gender differences.

Furthermore, concerning the influence of the test duration (H(3)), the regression estimate b indicates that, on average, the ERS score increased by 0.0013 ($SE = 0.0006$) per trial with increasing test duration (see Figure 3). The trial number, however, only accounts for 0.1% of the variance in the data (see third row of Table 1).

Finally, the exploratory analysis also revealed a significant influence of the test session (Test vs. Retest). Figure 2b illustrates that the ERS decreased when comparing the initial test ($M = 0.306$, $SE = 0.013$) with the retest ($M = 0.287$, $SE = 0.014$) conducted six weeks later. The effect of the session, however, is small and only explains 0.2% of the overall variance. Here, post-hoc analyses suggest that the decrease of ERS is slightly stronger for male ($M_{\text{Test}} = 0.282$, $SE = 0.016$, $M_{\text{Retest}} = 0.252$, $SE = 0.017$) than for female participants ($M_{\text{Test}} = 0.347$,

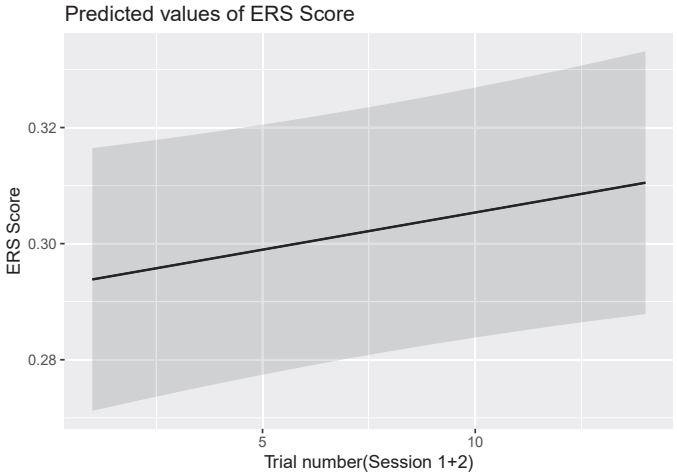


Figure 3. ERS score as a function of the trial number predicted by the linear mixed-effect model

$SE = 0.019$, $M_{\text{Retest}} = 0.342$, $SE = 0.020$). This *gender* x *session* interaction effect (added to both main effects) is significant, $F(1, 3718.4) = 4.71$, $p = .03$.

4 Discussion

The present study investigated extreme response behaviour (ERS) and influencing person- and study-related factors in the course of room acoustical and psychoacoustical experiments. Based on an analysis of the dataset collected by Weinzierl and colleagues [12], we found that the gender of the participants and the duration of the listening test measured by the number of trials significantly influenced ERS. In addition, we observed a decrease of ERS when comparing initial test results with those from a retest conducted about six weeks later. This decrease was particularly pronounced in male participants. Contrary to our assumptions, however, we did not find an effect of the participants' age on extreme response behaviour.

The results related to an influence of gender and test duration are in line with the presented literature, in particular with the studies by Borgatta et al. [6] and Hui and Triandis [10]. Therefore, they support the assumption that the different genders indeed differ in the way they use rating scales such as those used in psychoacoustical experiments. However, more research is needed to proof that the observed differences in ERS across genders are solely due to different uses of the rating scales and not due to differences with regard to the actual perception of the acoustical phenomena under investigation (meaning that women could generally perceive auditory stimuli in a more extreme way than men). Such a proof could be realised by comparing data from quantitative questionnaires with non-reactive, for example (neuro-) physiological, measures which are not affected by behavioural tendencies.

Regarding the null effect of age, the results contradict findings by Grühn et al. [5] who found an influence of age on ERS. The discrepancies in the results could be related to the specific age sample in our study which predominantly involved participants of younger age (83.7% were younger than 40 years). Therefore, in order to generalize these findings, the analysis should be repeated with a sample of broader age range in the course of future studies.

Furthermore, our results showed that, with an increasing number of trials, participants were more likely to show extreme response behaviour. These results are particularly interesting, as it could be assumed that the continuous scales used in the experiment counteracted the ERS compared to categorical scales [8]. Also, the question arises

whether ERS might change in a stimulus-differentiated way, meaning whether the increase of ERS is dependent or independent of the stimuli presented over the course of the study. In this context, it must be stated that we cannot estimate a time threshold in a listening experiment when ERS increases to a critical value, as our dataset did not include the absolute time participants spent on each trial. This also should be subject of future studies. As significant differences already become apparent for the relative duration measured by the trial number, it is reasonable to expect that the effect size increases when the absolute duration of the experiment and the single trials is taken into account [10].

Finally, the results regarding the effect of a retest (compared to an initial test) on extreme response behaviour are striking as this is (to the best of our knowledge) the first study suggesting that ERS slightly decreases when repeating the same experiment with the same participants six weeks later. This decrease of ERS might be explained by the fact that participants were able to better differentiate between the stimuli in the retest after already listening to each of them in the initial test. The gender differences regarding this retest effect could be explained by the theory that ERS in men is more due to relative comparisons across the stimuli presented in an experiment ("relative frame of reference") whereas ERS in women is assumed to represent more the expression of their absolute judgment regardless other stimuli in the experiment ("absolute frame of difference"). However, this theory is quite speculative, thus more empirical evidence is needed to further support it.

5 Conclusion

Our study demonstrated that extreme response behaviour is also present in listening tests dealing with room acoustical and psychoacoustical phenomena. In particular, we showed that the gender of the participants and an increasing number of test trials are associated with extreme response tendencies. This should be considered when conducting listening experiments. One practical conclusion drawn from this paper could be that listening tests should be kept as short as possible or limited to a certain time in order to avoid extreme response behaviour caused by the test duration. Furthermore, it is conceivable to develop mathematical formulas, which quantify and correct for the distortion of the response caused by ERS. Finally, it is advisable to track the time participants spend on single trials and the whole experiment to potentially correct for the response behaviour caused by an extensive duration of an experiment.

REFERENCES

[1] Jürgen Bortz and Nicola Döring. *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Springer-Lehrbuch. Springer, Berlin and Heidelberg, 5. vollständig überarbeitete, aktualisierte und erweiterte auflage edition, 2016.

[2] Kathrin Bogner and Uta Landrock. Antworttendenzen in standardisierten umfragen. *Mannheim, GESIS–Leibniz Institut für Sozialwissenschaften (SDM Survey Guidelines)*, 2015.

[3] Eric A. Greenleaf. Measuring extreme response style. *Public Opinion Quarterly*, 56(3), 1992.

[4] Gerhard Meisenberg and Amandy Williams. Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences*, 44(7):1539–1550, 2008.

[5] Daniel Grühn and Susanne Scheibe. Age-related differences in valence and arousal ratings of pictures from the international affective picture system (iaps): Do ratings become more extreme with age? *Behavior Research Methods*, 40(2):512–521, 2008.

[6] Edgar F. Borgatta and David C. Glass. Personality concomitants of extreme response set (ERS). *The Journal of Social Psychology*, 55:213–221, 1961.

[7] C Harry Hui and Harry C Triandis. Effects of culture and response format on extreme response style. *Journal of cross-cultural psychology*, 20(3):296–309, 1989.

[8] Mick P. Couper, Roger Tourangeau, Frederick G. Conrad, and Eleanor Singer. Evaluating the effectiveness of visual analog scales. *Social Science Computer Review*, 24(2):227–245, 2006.

[9] David Kreindler, Anthony Levitt, Nicholas Woolridge, and Charles J. Lumsden. Portable mood mapping: the validity and reliability of analog scale displays for mood assessment via hand-held computer. *Psychiatry Research*, 120(2):165–177, 2003.

[10] C Harry Hui and Harry C Triandis. The instability of response sets. *Public Opinion Quarterly*, 49(2):253–260, 1985.

[11] Bobby D. Naemi, Daniel J. Beal, and Stephanie C. Payne. Personality predictors of extreme response style. *Journal of Personality*, 77(1):261–286, February 2009.

[12] Stefan Weinzierl, Steffen Lepa, and David Ackermann. A measuring instrument for the auditory perception of rooms: The room acoustical quality inventory (raqi). *The Journal of the Acoustical Society of America*, 144(3):1245, 2018.

[13] Nicholas W Galwey. *Introduction to mixed modelling: beyond regression and analysis of variance*. John Wiley & Sons, 2014.

[14] Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–27, 2017.