Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 76 (2024) 349–360

Transportation Research Procedia

www.elsevier.com/locate/procedia

12th International Conference on Transport Survey Methods

Response bias in Likert-style psychological items – an example from a large-scale travel survey in China

Miriam Magdolen[a,*], Sascha von Behren[a], Jan Vallée[a], Bastian Chlond[a], Peter Vortisch[a]

[a]*Karlsruhe Institute of Technology, Kaiserstrasse 12, 76131 Karlsruhe, Germany*

Abstract

This paper addresses the challenge of ensuring response quality when using item sets with Likert scales in travel surveys. Particularly for capturing psychology, such item sets play an important role in travel behavior research. A challenge with this kind of data is the identification of response bias. An example is straightlining, which describes selecting the same response category for each item. Since there is no universal indicator in the literature to identify unusual or strategic response patterns, we apply various indicators, compare the results and develop a new indicator based on correlations which encounters plausible straightlining.

Keywords: Response bias; Likert scale; Psychological items; Data quality; Data checking;

1. Introduction

An increasing number of travel surveys involve attitudinal item sets to examine the psychological dimension of travel behavior. The investigation of attitudes and norms allows to gain a deeper understanding of the relation between psychological factors and behavior besides sociodemographic and spatial characteristics. Psychological item sets are mostly surveyed by the application of Likert scales. The participants rate if the given statements apply, e.g. from totally agree to totally disagree. The process is simple and convenient, but also allows the participants to easily run through the given statements without careful reading and understanding of the questioned content, especially when using online surveys. Satisficing describes the phenomenon that respondents do not answer in the

* Corresponding author. Tel.: +49-721-608-47738; fax: +49-721-608-48031.
E-mail address: miriam.magdolen@kit.edu

2352-1465
10.1016/j.trpro.2023.12.060

best way, but answer with suboptimal response strategies (Kaminska et al., 2010). Krosnick (1991) highlights, that the reasons for satisficing lie in the exceeding of the motivation or the ability of the respondent. The respondents still aim to report plausible answers, but also try to avoid the cognitive work to answer the questionnaire (Krosnick et al., 2001). Different response strategies may arise, such as straightlining, which is the phenomenon of selecting the same response category for all questioned items. Response data that is highly likely to be a result of certain response strategies has to be excluded for further analyses.

In a previous study, Magdolen et al. (2020) have investigated different forms of response bias in a psychological item set in a cross-cultural travel survey. In addition to rather simple and standardized indicators to identify response bias and response styles, a special measurement has been developed. An important aspect of this *modified algorithmic measure* is the parallel consideration of the content of the items besides the identification of systematic patterns. Thus, the consistency of the participants' answers was also evaluated. This cross-cultural study revealed differences in the response behavior of people from Berlin, Shanghai, and San Francisco. In Shanghai it was found that there is a tendency towards the middle and that the participants less often chose the extreme answer categories. In general, the study by Magdolen et al. (2020) underlines that the investigation of response bias helps to better interpret participant's answers on Likert scales.

The present paper builds on this preliminary work of Magdolen et al. (2020) and examines the identification of response bias in more detail. The focus is no longer on the cross-cultural aspect but rather on survey participants of one country, in this case China. We investigate how potential response bias, which becomes evident in certain response styles or patterns, can be identified in items sets with Likert scales. This paper does not address methods in questionnaire design to prevent response bias, such as selecting the number of response categories, but provides insights on the identification of strategic response patterns in already collected data. The challenge of this identification is that even with "strange" patterns, the answers may reflect the actual psychology of the participant. This is difficult to distinguish from the response behavior of unmotivated participants who simply select categories without reading and answering the items properly as described by Krosnick (1991). The aim is to develop an algorithm that can be applied to different item sets, independent of the content or the length of the item set, the order of the items as well as the number of response categories. This universally applicable algorithm identifies biased responses and with this information observations can be removed from the data set. Thus, a higher data quality is achieved and models that use the information of the psychological item sets become more robust and reliable.

The structure of the paper is as follows: First, the literature is analyzed in terms of the occurrence of response bias in Likert scales and the identification of such response bias. This is followed by a description of the survey data and the examined item set. In the methodology, we use different indicators to identify different response styles in the item set. Based on the gained knowledge we develop a new indicator that includes the correlations between items. Finally, we interpret the results, draw conclusions from the findings and provide indications for further research.

2. Literature

In social science research, the application of Likert scales is an established and often used approach. With a Likert scale, respondents can assess whether and to what extent they agree or disagree with a given statement. The use of item sets with Likert scales in surveys serves to quantify aspects that cannot be measured with conventional techniques. The scales serve to transform the subjectivity of an individual into a more consistent quantitative measurement (Joshi et al., 2015). Surveyed items are often questioned directly among each other, which results in a grid structure. Agreeing or disagreeing on the basis of a scale makes it much easier for the participants to answer a question and this leads to less response burden during the survey. However, it can also mean that the motivation of the participants decreases and the surveyed items are answered too rashly. Straightlining describes the phenomenon when the same answer category has been chosen for all questions of an item set, resulting in a straight line. Schonlau and Toepoel (2015) analyzed the influence of participant’s experience with the LISS panel in the Netherlands on straightlining. It was found that participants who took part for at least three years in the panel show a higher tendency of straightlining as an outcome of panel conditioning. Loosveldt and Beullens (2017) examined the influence of interviewers on straightlining and non-differentiation in the European Social Survey. Evidence of an interviewer effect on the tendency for straightlining was found, which underlines that external aspects can influence

the response behavior of respondents. Further, when designing item sets, the focus should not only be on the content but also on the order of the questions and the number and wording of the answer categories, which can prevent response bias. An overview of diverse response biases is given by OECD (2017). The occurrence of certain response patterns can have other causes besides low motivation. Social desirability may be a reason for participants to fill in the questionnaire in a systematic way (Bobbio and Manganelli, 2011). The participants do not necessarily respond according to their own beliefs or thinking, but how they think it most positively presents themselves. However, this is an aspect that is extremely difficult to determine in the data and should already be considered when designing the questionnaire and choosing the survey design.

In travel behavior research, item sets with Likert scales are often used to query psychological constructs such as attitudes towards different transport modes or social and personal norms. Studies that use such item sets in the context of travel behavior are for example von Behren et al. (2018), Anable (2005), Steg (2005) and Hunecke et al. (2010). Models that consider psychological factors such as hybrid choice models and structural equation models require the information from Likert scales and are used in many studies, e.g. in Kroesen and Chorus (2020) and Kroesen et al. (2017). Since psychology is becoming increasingly important in explaining travel behavior, care must be taken when measuring psychological characteristics of the respondents. Only if the information on the Likert scales correspond to the person's psychology will it lead to a deeper understanding when included in models. If the participant's responses are biased, it will lead to incorrect results and misinterpretation. It is therefore of importance to identify and exclude wrong or implausible responses in studies that capture psychology.

In contrast to other information in surveys, the identification of wrong or implausible answers is not directly possible since the Likert scale does not allow the selection of answers outside the response categories. There are several indicators proposed in the literature to measure different aspects of response bias. Straightlining is one approach to measure satisficing, which describes answering questions in an easy-to-do or non-optimal way (Schonlau and Toepoel, 2015). It is possible to calculate an indicator which reflects to what extent a respondent chose the same category, but it is difficult to tell if a high value results from response bias or from the real opinion of the individual. This also applies to other indicators, such as the standard deviation of the answers given (Leiner, 2013). However, such indicators are useful to identify response bias and help to identify inattentive respondents (Morren and Paas, 2019). A combination of different indicators leads to further insights on different forms of response bias. An example for an indicator that combines different response styles is given by Leiner (2013). With the developed measurement based on an algorithm, straightlining, diagonal lines and left-right clicking are identified at once. Another measurement of random answers is speeding. The time participants need is useful to see if they take enough time to read, understand and answer the questions (OECD, 2017). However, in many cases this information is only evaluated by the survey institute or there is only the time for answering the whole questionnaire and no explicit recording of the time needed to answer the Likert scale items sets.

When answering item sets, cultural differences can also have an influence on the response. Syam (2014) highlights the relevance of the analysis of cultural influences in the context of travel behavior. A detailed overview of literature on different analyses of the cultural influences on response behavior is given in Magdolen et al. (2020). For this present study, the response behavior in the Asian culture and specifically in China is of interest, as survey data from participants from eight Chines cities is available. Both Chen et al. (1995) and Lee et al. (2002) found that East Asian respondents have a greater tendency to choose the center category and a weaker tendency to choose extreme responses compared to American respondents. Hofstede's model offers a further reference point (Hofstede, 2011). The dimension *Individualism vs. Collectivism* explains cross-cultural heterogeneity and characterizes the level of people's integration into groups. Therefore, a distinction is made between individualistic and collectivistic cultures. The latter are characterized by strong, cohesive groups that prefer harmony so that individuals tend to choose the center in rating scales (Chen et al., 1995; Hofstede, 2011). The previous study of Magdolen et al. (2020) showed that participants from Shanghai tend to choose the middle category in the Likert scale and the response behavior thus corresponds to the collectivistic culture according to Hofstede. However, it remains unclear, how response behavior differs across the country. This is addressed in the following alongside the focus on how implausible response behavior can be identified with a universally applicable algorithm.

3. Survey Data

The data collection for the present study was carried out using the survey concept of the travel skeleton. In addition to the travel behavior, which is based on typical trips and typical transport use, this concept also collects data on the participant's mobility psychology. There are already several studies with data from the travel skeleton approach, e.g. von Behren et al. (2018), Magdolen et al. (2019) and von Behren et al. (2020). In this study, we use data collected by a survey in eight Chinese cities. The surveyed cities were selected according to the categorization of cities into first-, second- and third-tier cities (The World Bank and Development Research Center of the State Council, 2014) in order to represent exemplary cities for each category in China. Cities in different tiers reflect differences in consumer behavior, income level, population size, infrastructure and business opportunity. In summary, the distinction into the three tiers serves to differentiate developed and modern cities from less developed cities. Based on the selection from different categories, the survey took place in Shanghai and Beijing as Tier-1-cities, Chongqing, Shenyang, Wuhan as Tier-2-cities and Kunming, Urumqi and Zhuhai as Tier-3-cities between May and July 2017. The study primarily focused on capturing travel behavior and psychological characteristics of people in higher income classes in urban and high dense areas. The survey of people from high-income households controls possible differences due to social status, such as the symbolic value of the car or social desirability. To generate a comparable dataset from each city, a standardized survey approach based on a computer assisted personal interview (CAPI) was used. The survey was carried out by a professional Chinese market research firm. The full sample contains 5,192 individuals with at least 550 participants from each city. The psychological item set we focus on in this study consists of 38 items on a 5-point Likert scale and is mainly based on two standardized and well-tested item sets by Hunecke et al. (2010) and Steg (2005). An additional ‘no answer’ option was given. The items are largely the same from the previous study and investigation of response bias of Magdolen et al. (2020). Table 1 shows an overview of the items questioned in this survey and their corresponding psychological constructs. In the table, the items are sorted by these psychological constructs. The order of the items used in the survey will be considered later in this paper. There is no information available on the time taken to complete the questionnaire, as the item set only accounted for a small part of the total survey. Furthermore, the survey was carried out by interviewers and only the item set was filled in by the respondents themselves to minimize the impact from social desirability.

Table 1. Psychological items in the survey

Psychological items used		
Psychological constructs	*Items*	*Questions*
Public transportation autonomy (PTA)	PTA1	I can structure my everyday life very well without a car.
	PTA2	I can take care of what I want to with public transportation.
	PTA4	If I want, it is easy for me to use public transportation instead of a car to do my things in everyday life.
Public transportation excitement (PTE)	PTE1	I appreciate public transportation because there is usually something interesting to see there.
	PTE2	I can easily use the traveling time on the bus or train for other things.
	PTE3	I like to ride buses and trains, because I don't have to concentrate on traffic while doing so.
	PTE4	I can relax well in public transportation.
Public transportation intention (PTI)	PTI1	It is my intention to use public transportation instead of a car for the things I do in everyday life.
	PTI2	I have resolved to travel the ways I need to go in everyday life using buses and trains.
Subjective norm (SN)	SN1	People who are important to me think it is good if I would use public transportation instead of a car for things I do in everyday life.
Personal norm (PN)	PN1	Due to my principles, I feel personally obligated to use eco-friendly means of transportation for the things I do in everyday life.
Car excitement (CE)	CE1	Driving a car means fun and passion for me.
	CE3	When I sit in the car I feel safe and protected.
	CE4	Being able to use my driving skill when driving a car is fun for me.
Bicycle excitement (BE)	BE1	I like to be out and about by bike.
	BE2	I can relax well when riding a bike.
	BE3	I ride a bicycle because I enjoy the exercise.

Weather resistance (WR)	WR2	I also ride my bike when the weather is bad.
Car use motive (CM)	CM1	I feel free and independent when I drive a car.
	CM2	A car can communicate status and prestige.
	CM3	The characteristics of a car can show who and what I am.
	CM7	I like to drive a car.
	CM8	There are dream cars that I would like to drive once.
	CM12	To own a car is necessary for my family, it is irreplaceable due to its convenience and flexibility.
	CM13	To own a private car is one of my life goals.
	CM14	I feel depressed to quit car ownership, because it means downward life quality.
	CM15	E-hailing is convenient, it will replace private car and become my preferred mobility mode probably. Private car will be the supplement for emergency.
On-demand mobility evaluation (ODME)	ODME1	Car sharing is not related to me, I have own car.
	ODME2	Car sharing can release the city traffic pressure, I will use it as a supplement of my daily mobility.
	ODME3	The rich and the people having high social status won't use car pooling.
	ODME4	Car pooling makes me feel unsecure/unsafe.
	ODME5	The people who like car pooling always care the price very much, it is related to the income level.
	ODME6	The people who like car pooling are young mostly, they choose it for fun and novelty, and don't concern about privacy or luxury feeling.
	ODME7	The people who like car sharing are open to new things, and concern about the public topics very much, like resource sharing, environmental protection and improvement of city traffic.
	ODME8	Only the people who don't own a private car, or the people who care about the usage cost, will accept car sharing.
	ODME9	The people who like car sharing particularly have less passion to the vehicle, they just want to complete the mobility.
	ODME10	More and more high-income people accept and use subway/light rail, it is not related to the income level or social status directly.
	ODME11	The people who insist to own a private car and drive mainly are the wealthy or elite class.

Likert scale: 1 = does not apply; 2 = rather does not apply; 3 = applies in part / does not apply in part; 4 = rather applies; 5 = applies
Number of items not consecutive, as specific items from the item sets were selected for the survey.

4. Indicators for measuring response bias

As described in the literature review, a variety of indicators are proposed to investigate different types of response styles. To comprehensively explore the response behavior in our study, we therefore apply different measurement methods. In a first, preliminary approach we calculate indicators that serve as a hint for response bias but do not exactly identify specific patterns and response styles. These indicators are the *mean*, the *standard deviation* and the *deviation previous* of the selected answers in the Likert scale for each participant. The *mean* and *standard deviation* should always be considered to draw initial conclusions about response quality. Both indicators are simple to calculate. The *deviation previous* describes the average distance between two subsequent responses and therefore considers the order of responses. The indicator *maximum sequence* counts the maximum number of subsequent items with the same response. The results of these indicators are given in Table 2. Since we aim to identify differences between the participants from the different types of cities within China, the values are summarized at the level of each of the three tier categories. The differentiation into three different categories of cities helps to better classify and understand the response behavior of the urban population in China.

When comparing the mean values in Table 2, differences become clear both between the indicators and between the tier categories: The highest value for *mean* of the Tier-1-cities shows a higher agreement of the participants on the Likert scale. One reason for this could be the better offer of public transport and mobility on demand, which is why the participants from Beijing and Shanghai have a good experience with these services. Another interpretation would be that there is less collectivism in modern cities and the participants are more willing to choose categories at the end of the scale. The indicators *standard deviation* and *deviation previous* show similar values and the low values of 0.85 and 0.83 indicate that the participants from Tier-1-cities tend to vary less in their answers. This is a slight hint for a tendency for straightlining among the participants or less heterogeneity. The *maximum sequence*

indicates the average sequence of the same response category chosen. The average *maximum sequence* with 4.24 is by far the lowest value for the Tier-2-cities. In the other two city categories, the values are comparably high at close to eight. It can be concluded that participants in these cities show a certain degree of straightlining. The first comparison of cities based on these indicators shows that participants from Tier-2-cities tended to respond rather inconspicuously, whereas the response behavior from participants from Tier-1-cities indicate possible response bias.

Table 2. Mean values of the response bias indicators

Tier	N	Cities	Mean indicator				
			Mean	*Standard deviation*	*Deviation previous*	*Maximum sequence*	*Algorithmic measure (algM)*
1	1348	Beijing, Shanghai	3.28	0.85	0.83	7.90	0.51
2	1845	Shenyang, Wuhan, Chongqing	3.10	1.15	1.09	4.24	0.40
3	1984	Kunming, Urumqi, Zhuhai	2.99	1.08	0.93	7.71	0.54

In a second step, we apply a more sophisticated a*lgorithmic measure (algM)* which works like a penalty point system. Respondents get one point if two subsequent items have the same answer, one point if the change between subsequent items is the same as the recent change and half a point if the change is the same as the next-to-recent change. This method was developed by Leiner (2013). The *algM* combines different aspects and detects visual patterns in Likert scales such as straightlining, diagonal lines as well as left-right clicking. The higher the value, the greater the tendency that the participant's response pattern is strategic. In the present item set there are 37 items with a previous item, i.e. the maximum number of penalty points is 37. The *algM* is given as a proportion, so the maximum value is 1, if all items are answered in the same response category. If no patterns are detected, the value is 0. The combination of different aspects of response bias into one indicator is an improvement compared to the other rather simple indicators. It allows to identify different patterns in only one value. Moreover, by considering the distances to the previous item, the tendency towards response bias is identified in each row of the item set and not only overall. The results of the calculated *algM* is given in Table 2. A detailed explanation of all calculated indicators are given in Magdolen et al (2020).

The application of the *algM* demonstrates differences between the city categories. The Tier-2-cities have the lowest value. Again, the data of participants from this city category show the lowest tendency for response bias. Of interest is the comparison between Tier-1- and Tier-3-cities. The value of *algM* is higher for the Tier-3-cities, which represents a higher tendency for showing specific response styles. However, for all the previous analyzed indicators, the values for the Tier-3-cities indicate a lower tendency for response bias compared to the values for the Tier-1-cities. By investigating different response styles within the *algM*, additional aspects are considered and additional response bias is detected. Participants, in particular from Tier-3-cities, not only show straightlining but also the other patterns (left-right-clicking and diagonal lines) in their responses. The results underline the fact that one indicator on its own or simple indicators that analyze only one specific aspect of response bias are not robust enough to investigate the response quality in survey data.

5. Development of an algorithmic measure considering the correlations

The indicators presented in the previous section allow on the one hand to identify response bias with easy calculations and on the other hand to identify different response patterns with a complex algorithm. However, the indicators neglect the fact that for certain items straightlining may be a plausible and consistent way to respond. If consecutive items have similar meanings and are formulated in a similar way, it is very likely that a person will choose the same response category. However, the above described indicators assume poor data quality in those cases. This is correct if attention is paid to the order and wording when creating the item sets and if successive items are independent of each other. However, in many cases, item sets are divided in blocks with similar content.

Therefore, it is necessary to consider the possibility that straightlining is plausible. For the analysis of the response quality, the challenge arises to design an algorithm which also captures content similarities of items and quantifies them objectively for evaluation. The aim is not to find a single solution requiring a review of content and formulation of the single items by the people who work with the survey data, but to develop a general and flexible algorithm that can be used for different item sets.

5.1. Correlation analysis

The basic idea of the new algorithm is to use the captured data itself to determine which items are related to each other and are answered similarly. This is possible because one can basically assume that the majority of participants gives correct answers. The prerequisite is that the items used are tested and correctly understood by the participants, which is the case in our application. To identify related items the correlation is calculated for each two items underneath each other. At this point, as with most previously calculated indicators, it plays an important role in which order the items are listed. In our survey, the items of the first part were alternated to prevent fatigue effects. However, the items on car use motives (CM) and the on-demand mobility evaluation (ODME) were grouped together and were listed one below the other. The results of the correlation analysis between an item and the previous item are given in the original order in Table 3. The results confirm, especially in the second part of the item set, that partly high correlations between consecutive items exist.

Table 3. Correlation between item and previous item

No.	*Items*	*Correlation with the item above*	*No.*	*Items*	*Correlation with the item above*
1	PTA1	-	19	CM1	0.121
2	SN1	0.569	20	CM2	0.573
3	PTE2	0.538	21	CM3	0.688
4	BE1	0.375	22	CM7	0.607
5	PTE3	0.425	23	CM8	0.601
6	CE1	-0.002	24	CM12	0.565
7	PTI1	0.087	25	CM13	0.592
8	BE2	0.385	26	CM14	0.548
9	PTA2	0.410	27	CM15	0.176
10	PTE1	0.727	28	ODME1	0.265
11	PTI2	0.593	29	ODME2	0.289
12	PTE4	0.444	30	ODME3	0.249
13	CE3	0.107	31	ODME4	0.340
14	PN1	0.029	32	ODME5	0.238
15	CE4	0.022	33	ODME6	0.447
16	PTA4	0.139	34	ODME7	0.402
17	BE3	0.324	35	ODME8	0.278
18	WR2	0.460	36	ODME9	0.517
	Item set continues →		37	ODME10	0.270
			38	ODME11	0.239

Items in original order of the survey

In Fig. 1 we classify the correlation (C) between two subsequent items into three categories of relation. No relation is identified when C is below 0.3. Out of 37 subsequent items there are 15 item pairs in this category. A moderate relation is identified for C between 0.3 and 0.7. There are 21 correlations between subsequent items assigned in this category. When the correlation between two items is higher than 0.7 we see a high relation between these items. In our item set we identify one item pair (PTA2 and PTE1) as high related. Especially the items 20 to 26 are all related to each other. Awarding full penalty points for people who answered these questions the same way

would mean that these people would be ranked as suspicious although they might have answered all questions correct and plausible. Hence, we adjust the penalty points for moderate and high related items.

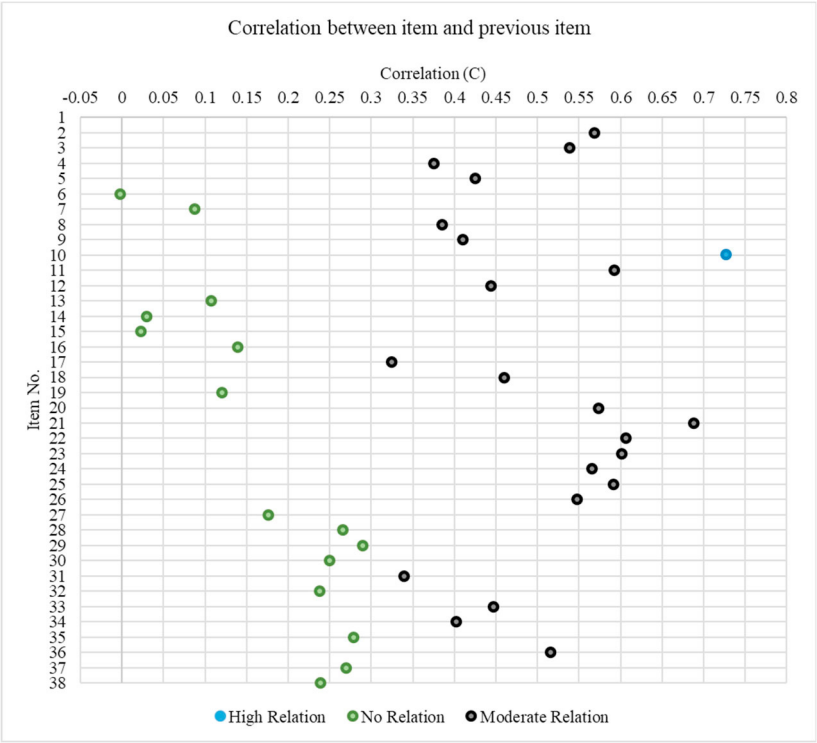

Fig. 1. Correlation between subsequent answers

5.2. Inclusion of correlation into the algorithmic measure

In our new developed algorithm non-related items are treated the same way regarding the penalty points as in the *algM* presented above (case 1). However, when there is a high relation between two items no penalty point (P) is given for the same answer. In fact, we assign a penalty point, if different answer categories are selected for items with a high relation (case 3). When there is a moderate relation between two items, P given for that answer depends on the degree of correlation between these items. We reduce P by the value of the correlation (C) (case 2). The differentiation of P depending on the correlation of the subsequent answers is shown in Table 4. The final score for each participant is the average of the points awarded.

Table 4. Penalty points depending on the correlation

Case	*Correlation (C) with previous item*	*Penalty points (P) for same answer*	*Penalty points (P) for different answer*	*Penalty Points (P) for diagonal lines*	*Penalty points (P) for left right clicking*
1	$C < 0.3$	$P = 1.0$	P=0	P=1.0	P = 0.5
2	$0.3 \leq C \leq 0.7$	$P = 1.0 - C$	P=0	P=1.0	P = 0.5
3	$C > 0.7$	P = 0	P=1	P=1.0	P = 0.5

Table 5 gives an example to illustrate how P is awarded for various answer schemes differentiated by the three cases introduced in Table 4. The new algorithm prevents participants who give the same answers to related items from being classified as suspicious by considering the correlation between two subsequent items.

Table 5. Example for algorithmic measure with correlations (algMC) in comparison to algorithmic measure (algM)

Item No.	*Likert scale*	*Correlation (C) with previous item*	*Case*	*Penalty Points (P) with algMC*	*Penalty Points (P) with algM*
5	☐ ☐ ☐ ☐ ☒ ☐				
		0.00	1 (same answer)	1.0	1.0
6	☐ ☐ ☐ ☐ ☒ ☐				
		0.09	1 (different answer)	0	0
7	☐ ☐ ☒ ☐ ☐ ☐				
		0.39	2 (same answer)	P = 1.0 - C = 0.61	1.0
8	☐ ☐ ☒ ☐ ☐ ☐				
		0.41	2 (diagonal)	1.0	1.0
9	☐ ☐ ☐ ☒ ☐ ☐				
		0.72	3 (same answer)	0	1.0
10	☐ ☐ ☐ ☒ ☐ ☐				
		0.82	3 (different answer)	1.0	0
11	☐ ☒ ☐ ☐ ☐ ☐				

Overall, the developed algorithm differs to commonly used indicators in two regards: First, compared to most other indicators such as *standard deviation* and *maximum sequence*, the algorithm allows to identify different patterns at once (straightlining, diagonal lines and left-right-clicking). Second, the algorithm considers plausible straightlining in item sets which is the case when subsequent items are highly correlated.

For the comprehensibility of all indicators and algorithms developed and calculated, the used code is publicly available (see Appendix A).

6. Results of the algorithmic measure based on correlations

By applying the developed algorithm to the given data, we can see a clear difference between the new *algorithmic measure with correlations (algMC*) and the *algM*. Table 6 shows the results. Again, the indicators are given as a score, which equals the average P over the item set, differentiated by tier category. The values of the indicators decrease, depending on the tier category. In Tier-1- and Tier-2-cities we observe an adjustment by 25%, whereas in Tier-3-cities the value decreases by 30%. From this, two conclusions can be drawn. First, the overall number of P in the new *algMC* is lower because a respondent gets less points for subsequent answers in the same response category if the items are related. In our data, most correlations belong to case 1 (no relation) or case 2 (moderate relation). Hence, the overall amount of P is reduced. The second conclusion is that people from Tier-3-cities have a higher decrease in the score than participants from other cities. That indicates that straightlining is more common in these cities although straightlining occurred on related questions. Hence, a decrease in P benefits the plausibility of the responses.

Table 6. Comparison of the algorithmic measure and the algorithm with correlations

Tier	**N**	**Cities**	**Mean Indicator**	
			Algorithmic measure (algM)	*Algorithmic measure with correlations (algMC)*
Tier 1	1348	Beijing, Shanghai	0.51	0.38
Tier 2	1845	Shenyang, Wuhan, Chongqing	0.40	0.30
Tier 3	1984	Kunming, Urumqi, Zhuhai	0.54	0.38

In Fig. 2, the distribution of the scores of *algM* and the *algMC* in the data is shown. This comparison again shows the reduction in the proportion of identified probable response bias by taking the correlations into account. We assess this as an improvement of the algorithm, since we expect plausible straightlining as a response pattern in the item set. Especially in the second part of the item set several items with similar content are listed below each other (see Table 3). Further, the distribution of the *algMC* in the three tier categories is shown. Especially in the Tier-1-cities there are some participants who answered remarkably. Differences between the tier categories become again clear and confirm the previous results. It underlines that the data quality of the individuals from the different tier categories differs and that it is worthwhile to examine the survey data more closely with regard to response bias.

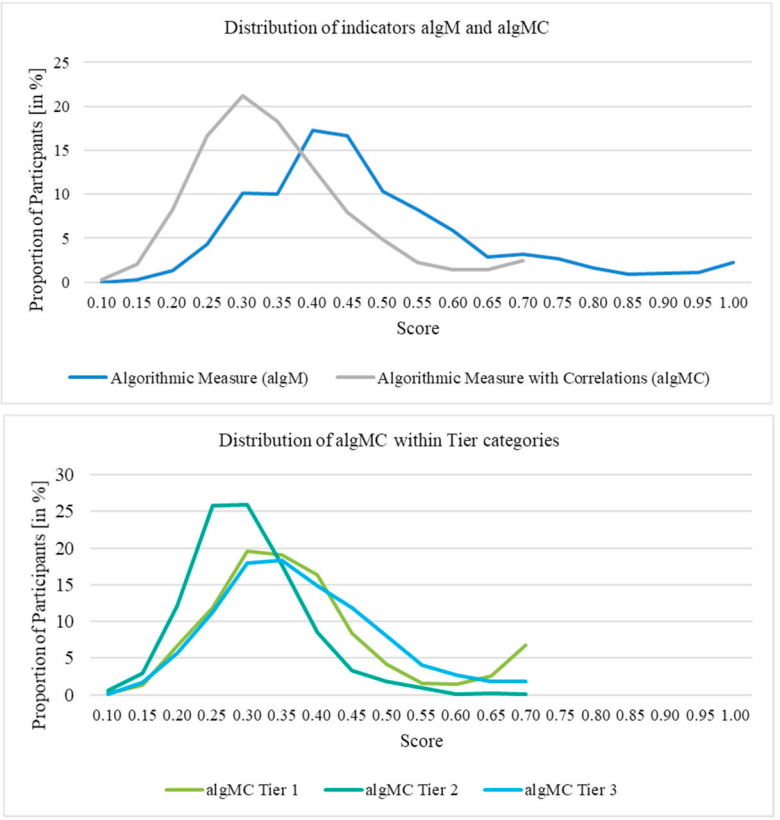

Fig. 2. Frequency distribution of the algorithmic measure (algM) and the algorithmic measure with correlations (algMC)

Besides the identification of differences between the two indicators and between the tier categories, the figure can be used to determine possible cut-offs, to remove respondents with implausible reports from the data. If the overall *algM* is considered, a cut-off of 0.65 is reasonable. There is an “elbow” in the curve indicating that the measured response bias increases. As a result, all participants with a higher value than 0.65 should be examined more closely. The curve of the *algMC* is smoothed. This is due to the fact that in some cases straightlining is not assessed so strictly. A cut-off at 0.65 would be also conceivable as the proportion increases above this score. The discussion at which point a cut-off should be made is very difficult. A general cut-off based on e.g. quartiles is not effective, since the same proportion of people would always be excluded regardless of the questions in the survey. If the distributions of the *algMC* for the individual tier categories are considered, it can be seen that the data quality varies strongly. More respondents should be excluded in Tier-1-cities than from the other two categories. Our study indicates that the cut-off can be made at a local minimum. We expect the rising proportions at the end of the scale identify those people who have not answered the questions fairly. A cut-off at 0.65 excludes 12.7% of the participants for *algM* and only 2.5% for the overall *algMC*. With the new developed algorithm, a larger sample is

retained and the chance of excluding participants who answered truthfully but randomly in specific response patterns is reduced.

In general, the curves show that a certain amount of response patterns is present in the entire sample and can therefore partly be regarded as a plausible response behavior. Even in the case of participants with high scores, the answers may reflect their actual assessment. However, they are declared as suspicious by the algorithms. We recommend not to remove persons with high scores directly from the data but to take a closer look at them. If, for example, other important information is missing, e.g. socio-demographic characteristics, it is likely that the person has not answered accurately. However, if all other information is complete and plausible, a reason may be given to leave the person in the data. It should also be checked whether the respondents use other response strategies due to respondent burden. If this is the case, a systematic problem with the responses can be assumed.

For all indicators presented, existing and developed, it has to be considered that their values depend on the questionnaire. Factors that influence the indicators are, for example, content and wording of the items, order of the items and number of response categories. Nevertheless, the indicators allow to compare groups of people within a survey, as shown by the comparison of the answer quality of people from different cities in this study.

7. Conclusion

With this paper we address the issue of data quality in item sets with Likert scales. Different response styles and patterns can occur for different reasons, e.g. lack of motivation of the respondent. This research is not addressing strategies in questionnaire design to prevent response bias. Instead, we focus on the identification of response bias in collected data. Indicators are used to identify such biased data and exclude them for subsequent analysis. We applied six indicators to a large dataset of a survey conducted in eight different cities in China. It was found that response bias in the item set with 38 items occurs to different degrees. The differentiation of the cities regarding their level of modernity through the classification into tier categories showed that people from the modern Tier-1-cities Beijing and Shanghai have a similar level of response bias as people from the cities Kunming, Urumqi, Zhuhai (Tier-3 cities). By contrast, participants from the Tier-2-cities Shenyang, Wuhan and Chongqing show the lowest levels of response bias in the indicators. Thus, no relation between the modernity of a city and the tendency of biased answers in the Likert scale item set is identified. Errors in translation or wording are expected to occur in all analyzed cities to the same extent. Further, we controlled for social status by surveying only people from high income households.

With the improvement of an existing algorithmic measure of response bias based on a penalty point system, we account for plausible straightlining in item sets. Such is the case if subsequent items have a similar meaning that overlaps. This is a relevant aspect that has not been considered in other indicators. A correlation analysis showed moderate and high correlations in our item set. This information helps us in deciding how to score the selection of the same response category in subsequent items. If there is a moderate correlation between an item and its previous item, we reduce the amount of penalty points compared to subsequent items that do not correlate. The new indicator resulted in fewer respondents being noticed with a biased response behavior and the distribution of the values of this indicator became more balanced. However, even the new algorithm does not directly indicate whether a respondent answered with response bias. Under the given circumstance, that systematic response patterns could really represent the respondent's attitude, e.g. choosing the central response category with a neutral opinion, the identification with measurement indicators serves only as a hint for response bias. We assume that a straightlining pattern does not automatically imply a response bias but might also reveal the true respondent's attitude. Therefore, an individual examination of the respondent and the given information is necessary to decide whether the data should be kept or excluded. The presented approach based on correlations requires that a large majority of the participants answers the items truthfully. Otherwise the identification of related items based on the correlation is biased. In addition, the sample size must be large enough to compensate for variations in the responses. This also argues for using established and well-tested item sets instead of developing new items for similar issues.

The focus of this study was on the identification and the development of an algorithm to identify response bias that is universally applicable. However, it should be emphasized that the investigation of the influences on response behavior is relevant and further research is needed in this respect. Further research should be done on the choice of cut-offs to exclude participants who report poorly. Our study suggests that the cut-off can be drawn at a local

minimum in the distribution of the indicator. However, the transferability to other surveys and item sets, other sample sizes and other distributions of the indicators should be checked. In addition, future research should focus on the refinement of both the correlation thresholds and the penalty point assignment. Overall, this paper provides a flexible approach to validate the quality of data supplied by survey institutions, especially when using online surveys and access panels.

Appendix A.

The code for the algorithm developed in this study is available on https://github.com/kit-ifv/likert_response_bias.

References

Anable, J., 2005. 'Complacent Car Addicts' or 'Aspiring Environmentalists'? Identifying travel behaviour segments using attitude theory. Transport Policy 12, 65–78. https://doi.org/10.1016/j.tranpol.2004.11.004.

Bobbio, A., Manganelli, A.M., 2011. Measuring social desirability responding - A short version of Paulhus' BIDR 6. Testing, Psychpometrics, Methodology in Applied Psychology, 117–135.

Chen, C., Lee, S., Stevenson, H.W., 1995. Response Style and Cross-Cultural Comparisons of Rating Scales Among East Asian and North American Students. Psychol Sci 6, 170–175. https://doi.org/10.1111/j.1467-9280.1995.tb00327.x.

Hofstede, G., 2011. Dimensionalizing Cultures - The Hofstede Model in Context. Online Readings in Psychology and Culture 2. https://doi.org/10.9707/2307-0919.1014.

Hunecke, M., Haustein, S., Böhler, S., Grischkat, S., 2010. Attitude-Based Target Groups to Reduce the Ecological Impact of Daily Mobility Behavior. Environment and Behavior 42, 3–43. https://doi.org/10.1177/0013916508319587.

Joshi, A., Kale, S., Chandel, S., Pal, D., 2015. Likert Scale: Explored and Explained. BJAST 7, 396–403. https://doi.org/10.9734/BJAST/2015/14975.

Kaminska, O., McCutcheon, A.L., Billiet, J., 2010. Satisficing Among Reluctant Respondents in a Cross-National Context. Public Opinion Quarterly 74, 956–984. https://doi.org/10.1093/poq/nfq062.

Kroesen, M., Chorus, C., 2020. A new perspective on the role of attitudes in explaining travel behavior: A psychological network model. Transportation Research Part A: Policy and Practice 133, 82–94. https://doi.org/10.1016/j.tra.2020.01.014.

Kroesen, M., Handy, S., Chorus, C., 2017. Do attitudes cause behavior or vice versa? An alternative conceptualization of the attitude-behavior relationship in travel behavior modeling. Transportation Research Part A: Policy and Practice, 190–202. https://doi.org/10.1016/j.tra.2017.05.013.

Krosnick, J.A., 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. Appl. Cognit. Psychol. 5, 213–236. https://doi.org/10.1002/acp.2350050305.

Krosnick, J.A., Holbrook, A.L., Berent, M.K., Carson, R.T., Hanemann, W.M., Kopp, R.J., Mitchell, R.C., Presser, S., Ruud, P.A., Smith, V.K., Moody, W.R., Green, M.C., Conaway, M., 2001. The Impact of "No Opinion" Response Options on Data Quality. Public Opinion Quarterly 66, 371–403. https://doi.org/10.1086/341394.

Lee, J.W., Jones, P.S., Mineyama, Y., Zhang, X.E., 2002. Cultural differences in responses to a Likert scale. Research in nursing & health 25, 295–306. https://doi.org/10.1002/nur.10041.

Leiner, D.J., 2013. Too Fast, Too Straight, Too Weird: Post Hoc Identification of Meaningless Data in Internet Surveys. SSRN Journal. https://doi.org/10.2139/ssrn.2361661.

Loosveldt, G., Beullens, K., 2017. Interviewer Effects on Non-Differentiation and Straightlining in the European Social Survey. Journal of Official Statistics 33, 409–426. https://doi.org/10.1515/jos-2017-0020.

Magdolen, M., von Behren, S., Chlond, B., Hunecke, M., Vortisch, P., 2019. Combining attitudes and travel behavior - A comparison of urban mobility types identified in Shanghai, Berlin and San Francisco, in: TRB 98th Annual Meeting Compendium of Papers. TRB 98th Annual Meeting Compendium of Papers, Washington, D.C.

Magdolen, M., von Behren, S., Hobusch, J., Chlond, B., Vortisch, P., 2020. Comparison of Response Bias in an Intercultural Context – Evaluation of Psychological Items in Travel Behavior Research. Transportation Research Procedia: 15th World Conference on Transport Research - WCTR 2019 in Mumbai, India.

Morren, M., Paas, L.J., 2019. Short and Long Instructional Manipulation Checks: What Do They Measure? International Journal of Public Opinion Research. https://doi.org/10.1093/ijpor/edz046.

OECD, 2017. OECD guidelines on measuring trust. OECD Publishing, Paris, 211 pp.

Schonlau, M., Toepoel, V., 2015. Straightlining in Web survey panels over time. Survey Research Methods 9, 125–137. https://doi.org/10.18148/srm/2015.v9i2.6128.

Steg, L., 2005. Car use: Lust and must. Instrumental, symbolic and affective motives for car use. Transportation Research Part A: Policy and Practice 39, 147–162. https://doi.org/10.1016/j.tra.2004.07.001.

Syam, A.A., 2014. Cultural Values: A New Apporach to Explain People's Travel Behaviour and Attitudes toward Transport Mode. Auckland, New Zealand.

The World Bank and Development Research Center of the State Council, 2014. Urban China: Toward Efficient, Inclusive, and Sustainable Urbanization. The World Bank, Washington, D.C, 624 pp.

von Behren, S., Kirn, M., Heilig, M., Bönisch, L., Chlond, B., Vortisch, P., 2020. The role of attitudes in on-demand mobility usage - an example from Shanghai, in: Goulias, K.G., Davis, A.W. (Eds.), Mapping the Travel Behavior Genome. Elsevier, pp. 103–124.

von Behren, S., Minster, C., Magdolen, M., Chlond, B., Hunecke, M., Vortisch, P., 2018. Bringing travel behavior and attitudes together: An integrated survey approach for clustering urban mobility types, in: TRB 97th Annual Meeting Compendium of Papers, Washington, D.C.