



Agree to agree: correcting acquiescence bias in the case of fully unbalanced scales with application to UK measurements of political beliefs

Phil Swatton¹ 

Accepted: 16 April 2024
© The Author(s) 2024

Abstract

A methodologically important area in political science is measuring the ideology of voters. This task can be difficult, and researchers often rely on ‘off the shelf’ datasets. Many of these datasets contain unbalanced Likert scales, which risk acquiescence bias. This paper proposes a strategy for dealing with this issue. I first demonstrate using two comparable datasets from the UK how unbalanced scales produce distorted distributions and can affect regression results. Then, building on past research that utilises factor analysis to eliminate the influence of acquiescence bias, I demonstrate how researchers can utilise a person intercept confirmatory factor analysis model to obtain factor scores corrected for acquiescence in the case of fully unbalanced scales. I conclude with practical recommendations for researchers and survey designers moving forward.

Keywords Survey · Acquiescence · Public opinion · Measurement · Factor analysis · Confirmatory factor analysis · Latent variable model

1 Introduction

In political science, a substantive area of interest is the ideology of voters. It follows that a methodologically important area of research is the measurement of voter ideology. A popular approach to measuring individual beliefs in surveys is the Likert scale. To construct a Likert scale, survey respondents are given several statements called Likert items¹ to respond to and a range of ordinal responses to choose from; typically ranging from ‘strongly disagree’ to ‘strongly agree’. The responses to the statements are then additively aggregated to produce the Likert scale.

¹ Some scholars may prefer to refer to the individual item as the Likert scale. For the purposes of this paper however, ‘Likert scale’ refers to the aggregate scale.

Earlier version of this paper presented at APSA Conference, 1–3 Oct 2021

✉ Phil Swatton
philswatton@gmail.com

¹ University of Essex, Colchester, England

Three types of Likert scale exist. Balanced scales which are built from an equal number of indicators representing both sides of the ideological dimension, partially unbalanced scales which have more indicators from one of the sides of the ideological spectrum than the other, and fully unbalanced scales which have indicators from only one of the sides of the ideological spectrum.

The indicators which are used to build Likert scales are typically subject to acquiescence bias and so where unbalanced scales are used, so too are the final scales. If researchers utilising these scales fail to consider the survey design and response stage of the data generating process (DGP), this can lead to incorrect research conclusions being drawn. In practice, due to the time and expense involved in collecting survey data, most political scientists (and indeed social scientists more broadly) are reliant on 'off the shelf' survey data produced by other researchers. Where unbalanced Likert scales are used in survey design, this typically exacerbates the problem as many researchers are forced knowingly or unknowingly to use the biased scales.

While past research has dealt with modelling acquiescence in balanced scales (see Mirowsky and Ross 1991; Billiet and McClendon 2000; Savalei and Falk 2014; Primi et al. 2019; Primi et al. 2019), it has not dealt with the more difficult problem of fully unbalanced scales. All methods designed to correct acquiescence bias rely on contradiction in survey responses. Without contradiction, it is logically impossible to identify which respondents are expressing agreement due to acquiescence, and which respondents are expressing agreement due to genuine agreement. In fully unbalanced scales, this contradiction does not exist, rendering acquiescence impossible to identify. In this paper I therefore discuss the problem of fully unbalanced scales, propose a model-based solution predicated on the introduction of additional indicators, and conclude by providing recommendations to both survey designers and users. The model-based solution I propose is an adaptation of a model previously proposed for balanced and partially unbalanced scales, which I label the person-intercept confirmatory factor analysis (CFA) approach. This approach leverages the common factor model to capture the acquiescence component latent in survey responses.

I begin with a substantive discussion of acquiescence bias in terms of the common factor model. This model divides latent variation between content factors, measurement factors, and unique variation. I use this model to discuss the assumptions underlying Likert scales, and how unbalanced scales introduce acquiescence bias. I proceed with a demonstration of acquiescence bias between two comparable datasets in the form of the British Election Study and the British Social Attitudes survey. These results serve as a baseline for the correction methods I apply. I discuss past work on person intercept CFA, and develop four variations of the model. I discuss the need for empirical identification in the case of fully unbalanced Likert scales. I then apply these to a third dataset in the form of the 14th wave of the British Election Study internet panel. The results broadly show that the correction methods applied succeed in producing results more akin to fully balanced scales. I conclude with recommendations for researchers and survey designers.

2 Measuring voter ideology

Voter ideology, political beliefs, or political attitudes represent an inherently ambiguous concept. Exactly what it is, what label to give it, how many dimensions it's composed of, and which of those dimensions we should be interested in are all contested. Even once researchers have agreed on a set of answers for the purpose of a given research project,

it follows that it is not straightforward how to capture a given definition among survey respondents.

One solution to this issue is the use of Likert scales. For the purposes of this paper, a Likert scale is an aggregate scale constructed from several individual survey items. For each survey item, respondents are shown a set of statements and given a range of responses, often five ranging from ‘strongly disagree’ to ‘strongly agree’. Scores from these responses for each statement are then tallied to produce a final measurement of the concept of interest. The basic idea is to show the respondent many items intuitively related to the concept of interest, then to aggregate over the responses to reveal the respondents’ overall preference on the underlying concept of interest.

Likert scales can be balanced, partially unbalanced, or fully unbalanced. Balanced scales are constructed from an equal number of survey items from each end of the dimension of interest (e.g. for a left-right dimension there will be left-wing and right-wing statements). Partially unbalanced items will have more items from one end than the other (but at least one from both), while fully unbalanced items will have items from only one end of the dimension of interest. However, since the individual agree-disagree items will result in acquiescent respondents being more likely to agree than they should be given their position on the underlying dimension, this results in acquiescence bias where scales are unbalanced.

In this section, I express this problem in terms of the common factor model.

2.1 The common factor model

For a given set of indicators of the same concept of interest, we can adopt a generative understanding of the measure. This means the indicator is assumed to be ‘generated’ by the target concept, and variation in the target concept causes variation in the indicator. One method of expressing this is via the common factor model (Brown 2015):

$$x_{ij} = \lambda_{j1}\eta_{i1} + \dots + \lambda_{jm}\eta_{im} + \epsilon_{ij} \tag{1}$$

where x_{ij} is the j th observed indicator for respondent i , η_{im} is the m th latent factor for respondent i , λ_{jm} is the loading on the m th factor for indicator j , and ϵ_{ij} is the unique factor for the j th observed indicator for respondent i .

The latent factors underlie the observed measurements. They can be further split between content factors capturing substantive variation and measurement factors capturing variation due to the measurement method of choice (Kenny and Kashy 1992). The unique factor captures variation in that indicator not found in any other indicators, which will be a mix of random noise and unique substantive variation. For my theoretical purposes in this paper, I adopt this common factor model.

2.2 Acquiescence bias

Substantively, acquiescence bias can be described as a tendency to be more likely to ‘agree’ with survey statements regardless of their content. In terms of the common factor model, we can express acquiescence as a second common factor:

$$x_{ij} = \lambda_{jc}\eta_{ic} + \lambda_{ja}\eta_{ia} + \epsilon_{ij} \tag{2}$$

here the c subscript denotes the concept of interest, while the a subscript denotes acquiescence. When aggregating indicators with this DGP into a Likert scale, we make several

implicit assumptions. First, we assume that λ_{jc} is constant across indicators *except* that for some indicators its sign ‘flips’ depending on the direction of the statement. If for example we have a left-right factor, then we can imagine its sign being negative for left-wing indicators and positive for right-wing indicators. This assumption tends to be incorrect in practice (Billiet and Davidov 2008, 545), but it need only be a reasonably close approximation to be successful.

Second, in the case of a balanced Likert scale we are assuming that λ_{ja} is also constant across indicators, albeit this time with its sign remaining positive regardless of the direction of the statement. Under these assumptions, acquiescence will ‘cancel out’ once the indicators are aggregated. Given this assumption is untrue in practice, it is worth pausing to consider further what a ‘close approximation’ entails. Broadly, within some degree of tolerance, the sum of the acquiescence loadings on the items in one direction (e.g. left) should be roughly equal to the sum of the acquiescence loadings on the items in the other direction. This will typically be more true in balanced Likert scales comprised of a larger number of items, as with enough such items these sums will likely be closer to one another in expectation. Survey designers can however circumvent the need for a large number of items by running trial surveys to discover the values of these loadings and selecting Likert items with desirable properties as described by the factor model.

When the scale is unbalanced, acquiescence bias will shift the scale in the direction in which the scale is unbalanced. This is because where before the equal number of indicators in opposite directions ‘cancelled’ out the acquiescence in one another (see Cloud and Vaughan 1970; Ray 1979; Evans and Heath 1995), in the unbalanced case there is leftover acquiescence. The more unbalanced the scale, the more bias leftover. This carries both descriptive and causal implications. Descriptively, this bias will shift the mean of the resultant scale in the direction of the imbalance. If we have a Likert scale comprised of more left-wing indicators than right-wing, then the resultant mean will be further to the left than it would be on a comparable balanced scale. Causally, on the same scale since acquiescence will point in the left-wing direction, variables that causally contribute to a respondent’s level of acquiescence will appear to contribute to the scale. This can result in spurious causal associations (if the effect on the concept is 0), inflated causal associations (if the effect on the concept and acquiescence are in the same direction), or hidden causal associations (if the effect on the concept and acquiescence are in opposite directions).

2.3 Fully unbalanced likert scales

A case not typically tackled in the literature on capturing acquiescence bias but which often arises in practice (both in political science and the broader social sciences) is that of a fully unbalanced Likert scale. Based on (2) and the subsequent discussion, it should become clear that in this case the acquiescence factor is impossible to *empirically* identify.² This is because at this point it becomes impossible to know which survey respondents are agreeing with the given statements because they sincerely agree with them; and which survey respondents are agreeing with them because they are acquiescent. Empirical identification

² I use this term to distinguish from statistical identification of the model. The model may be statistically identified (i.e. a unique solution exists), but that is no reason to believe we have successfully captured the acquiescence component.

of a model in this format requires *contradiction* in responses, which does not exist in a fully unbalanced scale.

It is arguably the case that this simple fact has led to some researchers mistakenly arguing that acquiescence bias is not a good explanation for the kind of results described above. For instance, Rodebaugh et al. (2007) find that removing reverse-scored (i.e. opposite) items improves the psychometric performance of their model; and argue that in fact these items were introducing an additional factor. I do not dispute their second claim, but instead point to the above: that an absence of contradiction in responses means acquiescence bias will become difficult if not impossible to directly identify. That the psychometric performance of the model is better after removing the contradicting items should be unsurprising: acquiescence is well-known among other things to be associated with inflated reliability coefficients and correlations (Winkler et al. 1982; Evans and Heath 1995)³.

A separate but related line of argument argues that negatively worded items are responsible for the item misresponse, rather than acquiescence bias (Swain et al. 2008). The argument is that negatively worded items introduce additional cognitive complexity (Swain et al. 2008). This point is not straightforwardly wrong: acquiescence bias is not the only potential problem that can occur in the survey design stage. The need for items on both sides of a scale does not mean that a simple negation route should be taken. However, as the demonstration below makes clear - acquiescence bias will still remain if a fully unbalanced scale is utilised. Researchers must therefore be prepared to tackle acquiescence bias within scales they are using.

In this paper, I tackle the specific problem of fully unbalanced Likert scales and offer some solutions for researchers utilising historical data. Given the constraint introduced by the problem of non-contradiction in fully unbalanced scales, these solutions are necessarily data dependent. They do nonetheless offer at minimum a starting point for researchers struggling with the problem of acquiescence in fully unbalanced likert scales.

3 Case selection and datasets

I use public opinion datasets from Great Britain as my case study. This is because since the 1990s, almost all GB datasets contain ‘left-right’ and ‘libertarian-authoritarian’ Likert scales based on the work of Evans and Heath and their coauthors (see Heath et al. 1994; Evans and Heath 1995; Evans et al. 1996). Several operationalisations of the same core concept therefore exist, creating an opportunity to assess how variations in measurement produce variations in research results. I use three datasets - two for demonstration, and one for developing a correction.

For demonstration, I use the British Social Attitudes survey (BSA) (NatCen-Social-Research 2017) and the British Election Study (BES) face-to-face survey (Fieldhouse et al. 2017). Both surveys were collected in almost entirely the same time periods. Both surveys were collected in person (although the BSA also used self-completion questionnaires). Given this and the shared conceptual basis of their Likert scales, it is not unreasonable to expect reasonably similar distributions of attitudes in both surveys. Notably however, in the BSA, all items are left-wing or authoritarian. Insofar as acquiescence bias affects these

³ Indeed, users of measures of psychometric performance should be careful in how they interpret such results. Cronbach’s alpha for instance measures *internal consistency* - but as the common factor model shows consistency can be a function of both measurement variance and content variance!

scales, they should have a left-wing and authoritarian bias. The third dataset used is wave 14 of British Election Study internet panel (BESIP) (Fieldhouse et al. 2020), which is used as a cross-sectional dataset. This wave was collected in May 2018. Similar to the BSA, all items are worded in left-wing and authoritarian directions and it should therefore display a similar bias. The item wordings for all three datasets are as follows:

3.1 BSA likert scales

The statements utilised in the BSA left-right dimension (ranging from Disagree Strongly to Agree Strongly) are as follows:

- Government should redistribute income from the better off to those who are less well off
- Big business benefits owners at the expense of workers
- Ordinary working people do not get their fair share of the nation's wealth
- There is one law for the rich and one for the poor
- Management will always try to get the better of employees if it gets the chance

The statements utilised in the BSA libertarian-authoritarian dimension (ranging from Disagree Strongly to Agree Strongly) are as follows:

- Young people today don't have enough respect for traditional British values
- People who break the law should be given stiffer sentences
- For some crimes, the death penalty is the most appropriate sentence
- Schools should teach children to obey authority
- The law should always be obeyed, even if a particular law is wrong
- Censorship of films and magazines is necessary to uphold moral standards

3.2 BES likert scales

The statements utilised in the BES left-right dimension (ranging from Strongly Disagree to Strongly Agree) are as follows:

- Ordinary working people get their fair share of the nation's wealth (right)
- There is one law for the rich and one for the poor (left)
- There is no need for strong trade unions to protect employees' working conditions and wages (right)
- Private enterprise is the best way to solve Britain's economic problems (right)
- Major public services and industries ought to be in state ownership (left)
- It is the government's responsibility to provide a job for everyone who wants one (left)

The statements utilised in the BES libertarian-authoritarian dimension (ranging from Strongly Disagree to Strongly Agree) are as follows:

- Young people today don't have enough respect for traditional British values (auth)
- Censorship of films and magazines is necessary to uphold moral standards (auth)
- People should be allowed to organise public meetings to protest against the government (lib)

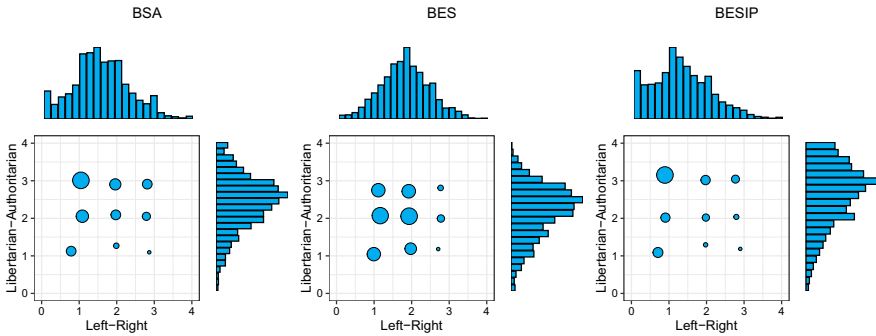


Fig. 1 Joint Distribution of BSA and BES Scales. For each plot, the x-axis ranges from left (0) to right (4). The y-axis ranges from libertarian (0) to authoritarian (4). Respondents were grouped along each scale from left/libertarian (0 to 1.6 inclusive), centrist (1.6 to 2.4 exclusive), and right/authoritarian (2.4 to 4 inclusive). The position of the points correspond to the mean position on both scales of the group, and the size of the point corresponds to the size of the group in the sample. The histograms opposite each axis show the marginal distribution of that particular scale

- People in Britain should be more tolerant of those who lead unconventional lives (lib)
- For some crimes, the death penalty is the most appropriate sentence (auth)
- People who break the law should be given stiffer sentences (auth)

3.3 BESIP likert scales

The BESIP survey items are shown with their labels from the dataset, which are used in some reporting during the appendix for this paper. The statements utilised in the BESP left-right dimension (ranging from Strongly Disagree to Strongly Agree) are as follows:

- **lr1:** Government should redistribute income from the better off to those who are less well off
- **lr2:** Big business takes advantage of ordinary people
- **lr3:** Ordinary working people do not get their fair share of the nation’s wealth
- **lr4:** There is one law for the rich and one for the poor
- **lr5:** Management will always try to get the better of employees if it gets the chance

The statements utilised in the BESP libertarian-authoritarian dimension (ranging from Strongly Disagree to Strongly Agree) are as follows:

- **al1:** Young people today don’t have enough respect for traditional authority
- **al2:** For some crimes, the death penalty is the most appropriate sentence
- **al3:** Schools should teach children to obey authority
- **al4:** Censorship of films and magazines is necessary to uphold moral standards
- **al5:** People who break the law should be given stiffer sentences

Figure 1 presents both the joint and marginal distributions of the Likert scales from all three datasets. The BSA, BES, and BESIP scales are presented from left to right. The scales were constructed to range from 0 to 4. On each x-axis is the left-right scale and on each y-axis is the libertarian-authoritarian scale. The histograms opposite each axis

capture the marginal distributions of these scales. To visualise the joint distribution of the scales, respondents were divided into ‘groups’. Those with scores ranging from 0 to 1.6 were placed in the ‘left’ and ‘libertarian’ groups of the respective dimensions. Correspondingly, those with scores ranging from 2.4 to 4 were placed in the ‘right’ and ‘authoritarian’ groups of the respective dimensions. Finally, those in-between these values were placed in the ‘centre’ group for each dimension. These groupings are of course arbitrary, but were chosen in part to resemble similar groupings utilised in research using these scales (see [SurrIDGE 2018](#)). The mean of each group was plotted, while the size of the group’s dot corresponds to the number of respondents in that group. Survey weights were used for the graphs.

Several notable differences emerge between the scales in [Fig. 1](#). In line with the above predictions BSA and BESIP scales show clear left and authoritarian slants as compared to the BES. Indeed, the similarities between the BSA and BESIP plots are striking, both in terms of the marginal and joint distributions of the scales. This is doubly the case given that the BES and BSA surveys share both a time period and method of collection, while the BESIP survey was collected on-line and in a different time period. By contrast, the BES plot shows both less left-wing and less authoritarian respondents. It retains a left-wing slant, but this is driven by the absence of right-wing respondents - it is still more balanced towards the center of its scale relative to the BSA and BESIP plots. While the BES data therefore would offer firmer grounds for believing that the British electorate in 2017 was left-wing, some caution is still required. First, this may plausibly be a quirk of the sample in question. Second, it may be a function of the statements used to construct the Likert scale. Per [\(1\)](#), smaller loadings in the right-wing statements could produce a skewed result. Nonetheless, it is better evidence than that available in either of the other scales.

4 Demonstration

The impact of acquiescence bias on descriptive inference is straightforwardly demonstrated by the above graphs. However, its impact also extends to explanatory research. It is well-established that acquiescence has a strong negative relationship with education level ([Ware and John 1978](#)); [Winkler et al. 1982](#)), and so I take this as my example. Given the similar collection dates and conceptual overlap of the BSA and BES, I regress the scales contained within on a measure of education level. Since education also has some well-established results showing it has a negative relationship with authoritarianism (see e.g. [Stubager 2008](#); [SurrIDGE 2016](#)) but no well-established association with left-right attitudes, some predictions can be made. First, in the BES scales the results will be as described here. Second, in the BSA scale, a spurious positive association⁴ between education level and left-right attitudes will be observed, while the negative association⁵ between education level and libertarian-authoritarian attitudes will be stronger.

My interest here is not in offering some causal explanation of these scales, but rather to offer a clear example of results changing from unbalanced to balanced scales. I therefore do not include control variables as they do not add anything for the purposes of the

⁴ Since the scale ranges from left (negative) to right (positive) and acquiescence points in the negative direction.

⁵ Since the scale ranges from libertarian (negative) to authoritarian (positive) and acquiescence points in the positive direction.

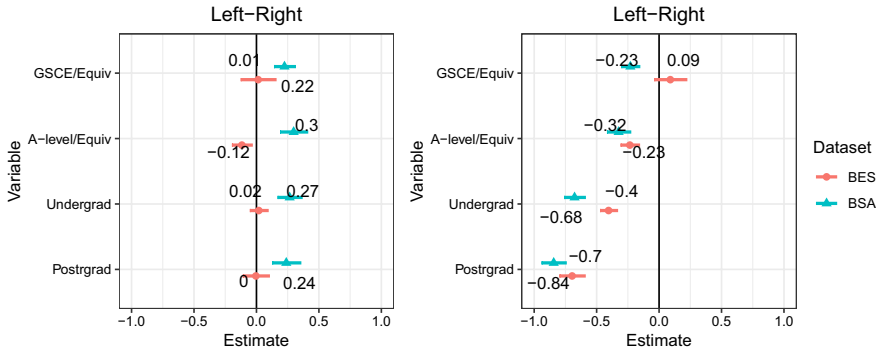


Fig. 2 Coefficient Plot of Demonstration Regressions. Plots showing regression of BSA and BES likert scales on education level. No education is the reference category for all coefficients in the plot. Results are demonstrative only, so no further controls were included in the model

demonstration. The education variables in both surveys were recoded such that the categories would match [full details of the recodes are in Appendix A (available online)]. Most of the recodes will be uncontroversial and thus I do not discuss them further here, but the ‘foreign’ category in the BSA had to be treated as missing as it had no clear placement. Figure 2 shows the results of regressing the scales from the BSA and the BES on the recoded education variable. The left coefficient plot shows the results for the left-right scale, while the right coefficient plot shows the results for the libertarian-authoritarian scale. The reference category is possessing no education. 95% confidence intervals are included for each estimate. A full table of regression results is available in Appendix B (available online).

Figure 2 shows that the pattern of differences between the BSA and BES is as we’d expect given the expectations laid out above. First and least dramatically, the absolute size of the point estimates for the libertarian-authoritarian results are larger in the BSA results. Moreover, two of the confidence intervals for BES coefficients (GSCE/Equiv and Undergrads) have no overlap with those of the BSA, indicating that they are significantly different from one another at the 95% confidence level. By contrast, the decision to use the BSA or BES dataset carries profound consequences for the results a researcher will find. The parameters for the BSA are all significant at the 95% confidence level and positive. By contrast, the parameters for the BES are not significant at the 95% confidence level with the exception of the A-levels parameter, which is negative. Only the parameters for GSCEs have overlapping 95% confidence intervals. The point of these results is not to offer some causal interpretation, but rather to highlight how scale construction can be the primary driver of research results.

All of the results demonstrated in this section have rely on an assumption that there should be no predictable differences from the BES to the BSA other than those caused by acquiescence. Given the importance of this assumption to my analysis above, I have performed two robustness checks to verify that these differences are likely driven by acquiescence bias, rather than any particular quirks of the samples.

First, I merged the five indicators common to the BES and BSA into a single dataset and created a binary variable denoting whether a respondent belonged to the BSA. I then regressed this binary variable on the five common indicators, and ran an OLS, Logit, and Probit model to check against model dependency. In all three cases, two indicators were

significant.⁶ However, their point estimates pointed in opposite directions, strongly suggesting that these were quirks of the samples and not evidence of anything systematic that would explain the observed differences in results.

Next, to verify that there was no temporal instability in results, I regressed the scales from the BES on the survey month of the respondents. The result showed no association between interview month and scale score. I did not do the same for the BSA as the interview date is not included in the publicly available version of the dataset. Taken together, these two checks offer strong evidence that my assumption that systematic differences between these two surveys are primarily driven by acquiescence bias is correct. Regression tables for both of these checks are available in Appendix B (in online).

The results of this demonstration - both descriptive plots and differences in regression coefficients - are a striking example of how acquiescence bias can drive research results. This has frequently occurred in the UK context from which these example scales are drawn, with many researchers drawing on the fully imbalanced scales to argue that there is the UK electorate is on average left-wing and authoritarian in the UK (see e.g. Webb and Bale 2021, p. 158). This argument has made it as far as the public, where researchers have passed on these findings via mainstream newspapers (see e.g. Surridge 2019). There is therefore a clear need for all survey researchers to be careful with the measurements they use, and where possible to correct them for their biases.

5 Methodology

I now turn to the primary task of this paper, which is developing a methodology for the case of fully unbalanced Likert scales. Past research reviewing competing methodologies for modelling acquiescence bias have concluded that one of the most effective is an approach that treats acquiescence as a person-specific intercept across the scale items (Savalei and Falk 2014; Primi et al. 2019; Primi et al. 2019). This model was developed in a Confirmatory Factor Analysis (CFA)/Structural Equation Modelling (SEM) context (Mirowsky and Ross 1991; Billiet and McClendon 2000) but later extended to a unidimensional Item Response Theory (IRT) context (Primi et al. 2019). For the sake of simplicity I focus on the CFA specification in this paper, but the general intuition translates to an IRT context. Here, I only briefly discuss the person intercept CFA model. A more complete background to CFA and its extension to include the person intercept is given in Appendix C (available online).

5.1 Person intercept CFA

First used in Mirowsky and Ross' paper *Eliminating Defense and Agreement Bias from Measures of the Sense of Control: A 2 x 2 Index* (1991), the best exposition of the unit-intercept model is in Maydeu-Olivares and Coffman's paper *Random Intercept Item Factor Analysis* (2006). The model is based on (2) and traditionally is estimated by setting λ_{ja} to 1. This essentially treats acquiescence bias as a form of *differential person functioning* (DPF). Where the more common study of differential item functioning (DIF) focusses on

⁶ For some crimes, the death penalty is the most appropriate sentence; People who break the law should be given stiffer sentences.

how differences in survey items cause variation in responses, differential person functioning instead focusses on how differences in individuals cause variation in responses (Johanson and Alsmadi 2002). Application of methods for detecting differential person functioning have proven productive at detecting acquiescence bias (Johanson and Osborn 2004).

This assumption is therefore kept from the balanced Likert scales, but the assumption that each item equally captures the concept of interest is relaxed. Although the assumption of equal loadings for the acquiescence component is a strong one, simulations do suggest that the model is robust to violations of this assumption (Savalei and Falk 2014). Similarly, since the model is being estimated, the unique variation is also stripped from each item - a further relaxation relative to the balanced Likert scale. To identify the scales, the variance of η_{ic} is constrained to 1 while the variance of η_{ia} is freely estimated, producing the following model (Maydeu-Olivares and Coffman 2006):

$$x_{ij} = \lambda_{jc}\eta_{ic} + 1\eta_{ia} + \epsilon_{ij} \tag{3}$$

For the purposes of this paper I label this version of person intercept CFA as CFA1. An alternative specification can be achieved by constraining the variances of both η_{ic} and the η_{ia} to 1, while freely estimating their loadings. However, a constraint is still placed on λ_{ja} , in that it must be equal across indicators. The linear form of this version of the model can thus be given as:

$$x_{ij} = \lambda_{jc}\eta_{ic} + \lambda_a\eta_{ia} + \epsilon_{ij} \tag{4}$$

For the purposes of this paper I label this version of person intercept CFA as CFA2. The full set of assumptions for both CFA in general and person intercept CFA are given in Appendix C (in online) of this paper. However, one crucial difference in my definition of the model to Maydeu-Olivares and Coffman's is that I drop the language of 'random intercepts'. Here, they are drawing a parallel with hierarchical regression modelling in their description of the person intercept. However, the comparison is not necessary and more importantly undermines the utility of the model. In a random-intercepts regression model, the random intercepts are estimated as an error component. The unit-intercept here is not being estimated as an error term - it is being estimated as another common factor. The orthogonality assumption is thus not required for identification purposes (as other assumptions in the model are), but rather is made for the purpose of this comparison. This unnecessarily confuses things and potentially reduces the desirability of the model. In their review, Savalei and Falk (2014) suggest more work is required to explore potential relaxations of the orthogonality assumption. This assumption however is unnecessary to begin with, and I therefore drop it and utilise the terminology person-intercept instead of random-intercept.

In theory, the main difference between the specifications in (3) and (4) is their interpretability. Since the variances of both factors are the same in (4), the main advantage is that the model allows more direct comparison of the respective loadings - it is immediately clear how acquiescence bias compares to the content factors of interest in its effect on the scales. An advantage of the person intercept approach in general is that it does not require a balanced scale to work. Instead, the person intercept merely acts to capture inconsistency in observed responses and thus *in theory* only requires at least one opposite-worded indicator in order to successfully capture acquiescence bias. How many additional indicators are required *in practice* will need to be discovered in future research. I also consider ordinal versions of CFA1 and CFA2 in this paper, and I label them as OCFA1 and OCFA2 respectively. Their specifications are also detailed in Appendix C (in online).

5.2 Fully unbalanced scales

Past simulation studies suggest that unit-intercept models are robust to unbalanced scales where other acquiescence-correction methods require balanced scales (Savalei and Falk 2014). However, the crucial point made above is that the unit-intercept requires *contradiction* in order to empirically identify the acquiescence component, which is lacking in fully unbalanced scales. If for instance we take the BSA left-right scale, it is impossible to try and tell apart those who are agreeing with left-wing statements because they agree with them and those who are agreeing with the same statements because they are acquiescent. *There is no information available to distinguish the two kinds of agreement.*

To solve this problem and empirically identify the acquiescence component, I use Watson's idea of introducing further information in the model (1992). Specifically, if a scale which contains statements for which it would be contradictory to agree to all of them, it can be used to identify the acquiescence component in itself and thus also in the fully unbalanced scale. It is unfortunate that a strategy does not exist based on the fully unbalanced scale alone, but it should be clear that it is not possible to identify acquiescence bias in such a scale *without additional information being introduced in some form*. While the same simulations suggest that the person intercept CFA model is robust to differing levels of acquiescence bias in each indicator (Savalei and Falk 2014), this approach necessarily strengthens the assumption, as it assumes not only the same level of acquiescence for each respondent on one scale, but on all scales in the model.

5.3 Identifying scales in BESIP

The reason I chose the fourteenth wave of BESIP is that it contains two balanced Likert scales which could be used to identify the acquiescence component in the manner described above. This is the May 2018 wave of BESIP and thus some comparability to the other two surveys in this paper is lost. However, as seen in Fig. 1 there is nonetheless enough similarity in the scales in BESIP and the BSA for the dataset to be suitable for my purposes. The two additional scales cover zero-sum approaches to life and second on personal empathy respectively. The individual item wordings are as follows:

5.4 BESIP extra likert scales

The statements on the zero-sum scale are:

- **zero1:** One person's loss is another person's gain (zero-sum)
- **zero4:** There's only so much to go around. Life is about how big a slice of the pie you can get. (zero-sum)
- **zero5:** Life isn't about winners and losers, everyone can do well (everyone can win)
- **zero7:** The only way to make someone better off is to make someone else worse off (zero-sum)
- **zero9:** There are ways to make everyone better off without anyone losing out (everyone can win)
- **zero11:** Everyone can be a winner at the same time (everyone can win)

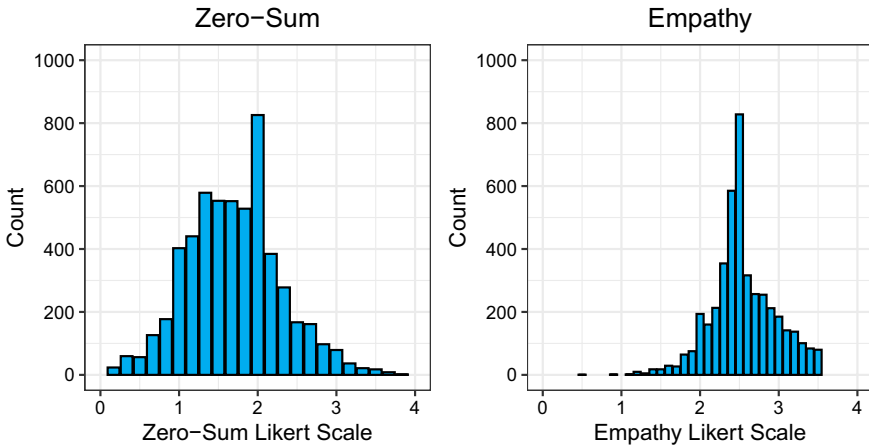


Fig. 3 Bar Plots of the Balanced Scales

The statements from the empathy scale are:

- **empathy1:** I can usually figure out when my friends are scared (empathetic)
- **empathy2:** I can usually realize quickly when a friend is angry (empathetic)
- **empathy3:** I can usually figure out when people are cheerful (empathetic)
- **empathy4:** I am not usually aware of my friends’ feelings (unempathetic)
- **empathy5:** When someone is feeling ‘down’ I can usually understand how they feel (empathetic)
- **empathy6:** After being with a friend who is sad about something, I usually feel sad (empathetic)
- **empathy7:** My friends’ unhappiness doesn’t make me feel anything (unempathetic)
- **empathy8:** Other people’s feelings don’t bother me at all (unempathetic)
- **empathy9:** I don’t become sad when I see other people crying (unempathetic)
- **empathy10:** My friends’ emotions don’t affect me much (unempathetic)

The two scales were asked in two separate subsamples of BESIP wave 14. This creates two separate opportunities to test the model, and so I test the four model types across the two BESIP subsamples. Figure 3 shows bar plots of the two balanced scales. The zero-sum scale ranges from everyone can win (0) to zero-sum (4), while the empathy scale runs from unempathetic (0) to empathetic (4). After filtering for missing data, the zero-sum subsample has 5836 respondents while the empathy subsample has 4478 respondents.

The empathy scale in Fig. 3 is notably less dispersed than the other scales discussed in this paper. This may be a function of the fact that it is comprised of a higher number of indicators than any of the others (10, as opposed to 5 or 6). It may also be however that given individuals are generally predisposed to view themselves as empathetic that there is less noise - and overall acquiescence - in the empathy scale. To test this second point, I ran person intercept CFA models on each of these scales alone, the full results for which are available in Appendix D (in online). The estimated variance for the acquiescence component in the zero-sum was larger than in the empathy model, suggesting that there is less acquiescence in the empathy scale. The extent to which the corrections are successful

likely depends in part on which of these scales is a closer match in terms of acquiescence to the acquiescence in the left-right and libertarian-authoritarian scales. I therefore test all four variations of the person intercept on both subsamples of BESIP wave 14, using the respective additional scales to empirically identify the model.

5.5 Estimation

To estimate the CFA1 and CFA2 models, I use robust maximum likelihood (MLR) estimation. MLR returns the same point estimates as ML estimation but adjusts standard errors and test statistics for violations of the normality assumption. A rough rule of thumb suggests it's a reasonable approximation once at least 5 response categories exist. To estimate the OCFA1 and OCFA2 models, I use unweighted least squares estimation (ULS). In a comparison between MLR and diagonally weighted least squares (DWLS) estimation found in favour in DWLS for ordinal data (Li 2016). However, simulations comparing DWLS to ULS have in turn found in favour of ULS, with the caveat that DWLS may converge in situations where ULS does not (Forero et al. 2009), I therefore utilise ULS estimation. All CFA models in this paper were estimated using R version 4.3.1 and the R package lavaan version 0.6-16 (Rosseel 2012) using code adapted from the Appendix of Savalei et al. (2014). Since lavaan does not currently support survey weights for ordinal CFA models I have not used them in the CFA models themselves, but they were used in producing distributions from the predicted factor scores of the models. Since the associations between variables should be reasonably robust to weighting, this is likely unproblematic.

6 Results

In this section I present a series of results demonstrating the comparative performance of the methods. Since my emphasis as throughout the paper is on obtaining corrected measurements, the plots presented here pertain to the predicted factor scores. Tables containing results for the CFA models can be found in Appendix D (available online). To verify that the scales were broadly capturing the same content, their correlation matrices were checked. These tables are also available in Appendix D (available online). Correlation plots showing the correlations between the recovered measures, the likert scales, and the acquiescence factors are also available in Appendix D (available online).

6.1 Distributions

Since my demonstration is preceded by distributional differences, I also begin my presentation of correction results with the distributions of the predicted factor scores. Figure 4 shows two-dimensional binplots of the resultant measurements from the four variations and the two subsets of the BESIP dataset. The extracted measures were rescaled to range from 0 to 4 to facilitate comparability with one another.⁷ The 'left' factor extracted was flipped

⁷ An identifying constraint on the scales is that they are mean 0. However, this won't necessarily be a meaningful midpoint, especially if the distributions are skewed. Rescaling in this way establishes the mid-range point as the central point of each scale, which is no less arbitrary in theory but in practice may be a better approximation to a 'true' midpoint.

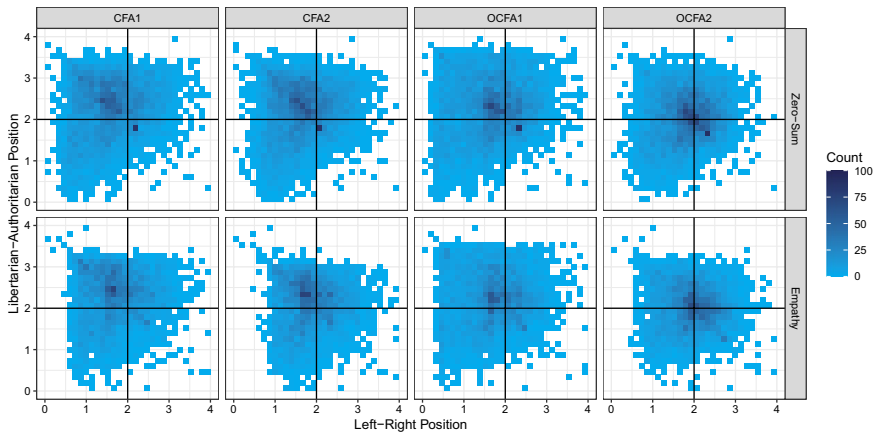


Fig. 4 Two-Dimensional Bin Plot of Voter Beliefs. Each plot uses a heatmap to show the joint distributions of the extracted factor scores. The left-right factor scores are on the x-axis, and the libertarian-authoritarian factor scores are on the y axis. The scores were rescaled to range from 0 to 4

to range from left to right, rather than right to left. Libertarian-authoritarian factors are on the y-axis, and left-right factors are on the x-axis. The colour of the bins change from light blue to dark blue as the count of respondents in that bin increases. The plots are organised in columns for correction method, and by row for the two BESIP subsets. Plots of the marginal distributions of the predicted factor scores can be viewed in Appendix D (available online).

Alongside the marginal distributions available in Appendix D (available online), Fig. 4 shows that for all correction methods, relative to Fig. 1 there is a shift towards a more normal distribution. The scale is broadly more evenly distributed (especially in the case of OCFA2). There is a starker effect for the left-right scale, which in some cases appears to remain somewhat left-leaning. These differences are carried into the association between the two scales. The extent to which the joint distribution is even across the four quadrants varies from correction method to correction method, but in all cases similarly appears more evenly distributed than in Fig. 1. These results would therefore indicate that once acquiescence bias is accounted for, the distribution of voter ideology on both scales is closer to a normal distribution. Caution is required in interpreting the midpoint of these scales, but nonetheless the extract factor scores do appear more evenly distributed.

6.2 Acquiescence factor

The question that remains after the initial positive assessment of Fig. 4 is whether the acquiescence factor has been properly captured by the person intercept. It may be the case that other substantive sources of variation are being captured. To verify the match between the acquiescence factor and the observed indicators, I additively aggregated all indicators, assigning a number for the level of agreement with each item (0 for ‘strongly disagree’ up to 4 for ‘strongly agree’). I then assessed the correlations between this additive count and the acquiescence factor extracted from each model (Table 1).

The correlations offer a mixed pattern. Broadly, those for the zero-sum subset of BESIP show reasonably sized correlations ranging from 0.63 to 0.75. By contrast, the

Table 1 Correlations of Acquiescence Factor with Additive Agreement

Subset	Model	Correlation
Zero-Sum	CFA1	0.71
Zero-Sum	CFA2	0.63
Zero-Sum	OCFA1	0.75
Zero-Sum	OCFA2	0.65
Empathy	CFA1	0.50
Empathy	CFA2	0.31
Empathy	OCFA1	0.69
Empathy	OCFA2	0.49

correlations for the empathy subset sit in a smaller range of 0.39 to 0.5 except for the OCFA1 model, which achieved a correlation of 0.69. Overall, these medium-sized correlations suggest that the model does indeed capture acquiescence in the model. In line with the difference in acquiescence in the zero-sum and empathy scales, it would therefore appear that the Zero-Sum scale was a more effective identifying scale. It would not be a good thing if these correlations were near 1: there is also left-right and libertarian-authoritarian variation in the additive index, alongside noise generated by the zero-sum and empathy indicators (since aggregation was for agreement, these dimensions of variation will cancel each other in the index instead). We can therefore be increasingly confident regarding the efficacy of the model in properly capturing acquiescence.

6.3 Regression results

With the distributional results of the correction methods established, I now turn to examining how explanatory research results are changed by using the predicted factor scores. I regressed the raw Likert scales and the predicted factor scores in each subsample on the education level variable available in the BESIP dataset. Once again, the reference category for the education level variable is 'no qualifications'. 95% confidence intervals are included in the plot. Figure 5 shows coefficient plots for each of these regression results. Tables for each regression are available in Appendix D (available online). For the results displayed in the main body of the paper, I have avoided recoding the education level in the same manner as the demonstration above (i.e. to have the same levels) as this recoding makes some results appear to be somewhat better than they are. I have however included regression results with the recoded education level variable in Appendix D (available online).

In the case of the results for the zero-sum left-right scales, the confidence intervals for each model fully overlap in both datasets. However, the point estimates shift towards 0 once correction methods are used; and more importantly inferential differences emerge. If a correction method is used, it becomes the case that a researcher using null hypothesis significance testing will reach the same conclusions using the BESIP data as they would using the balanced scales in the BES. The correction is sharper in the zero-sum subsample, as predicted by the differences in acquiescence between the zero-sum and empathy scales. However, the OCFA2 model sufficiently shifts the point estimates in both subsamples that the same inferences will be produced in BESIP regardless of the subsample of choice. In terms of the libertarian-authoritarian scales, the gap in point estimates are considerably larger. In several cases, the 95% confidence intervals do not overlap at all. However, in line with the differences between the BES

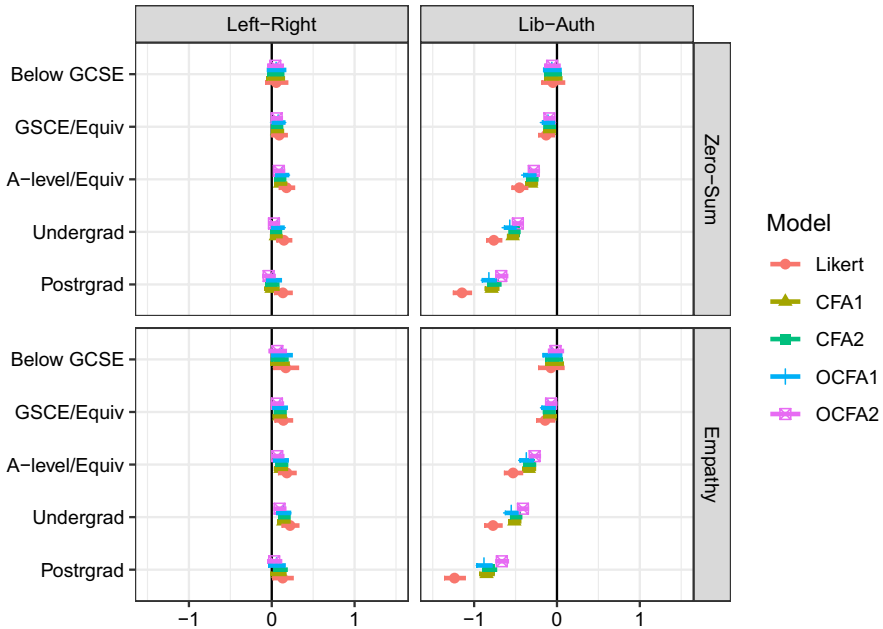


Fig. 5 Coefficient Plots of Scales Regressed on Education. Plots showing the results of a regression of extracted factor scores on education level in BESIP. For all four plots, the reference category is no education. Plots are organised in columns by scale, and by row for subset of BESIP. Results for likert scale included for comparison. Coefficient values available in Appendix

and BSA, a researcher using null hypothesis significance testing will reach the same inferential results - albeit with smaller point estimates for the corrected scales. The correction methods therefore broadly produce the same inferences as the balanced scales in the BES, while utilising biased data as in the BSA.

6.4 Robustness to number of categories

I have developed the correction method outlined in this paper with 5-response category survey items in mind. While the technique has worked for this case, Appendix D (available online) contains the results of a robustness check where all of the items are collapsed to three categories ('strongly agree' and 'agree' into 'agree'; 'strongly disagree' and 'disagree' into 'disagree'). Broadly, the method failed to recover fully normal distributions for the left-right and libertarian-authoritarian scales, but nonetheless succeeded in recover the same inferences as the balanced scales in the BES. Further work should therefore explore improving the performance of the technique with more limited response sets (perhaps switching to using Item Response Theory models in lieu of CFA models), but the correction technique should nonetheless aid researchers seeking to avoid bias in regression models.

7 Conclusion

In this paper, I have set out the specific problem of fully unbalanced Likert scales in the context of wider work on acquiescence bias. Fully unbalanced Likert scales carry the particular problem of rendering the acquiescence within impossible to empirically identify without the introduction of additional information. In this paper, I have further clarified the different versions of person intercept CFA relative to Likert scales, relaxed the unnecessary orthogonality assumption, and developed a strategy for identifying a person intercept CFA model in the case of fully unbalanced Likert scales. The OCFA2 approach appears to work best for fully unbalanced scales, but it is not immediately clear why this should be the case. Researchers using these approaches should run all four and compare the results until further research can be conducted on the relative performance of the four methods.

A clear limitation of the correction methods used in this paper is the data requirements they impose on the user in the case of fully unbalanced scales. However, these limits are necessary: acquiescence bias is not empirically identifiable without some degree of contradiction. Where researchers cannot utilise the corrections, they should at minimum be conscious of the role that acquiescence bias is likely to be playing in their results. Even where Likert scales are fully balanced, their use entails strong assumptions about the data generating process that can be relaxed by the use of person intercept CFA. There is no clear case where the use of these models if possible is not preferable to a raw Likert scale. For survey designers, two main points should be taken from this paper. First, as far as possible they should seek to design Likert scales that are fully balanced. Where this is not entirely possible, whether due to difficulties in designing reverse-keyed items or the need for backwards comparability, they should instead try to include other, substantively unrelated scales for the purposes of identifying person intercept CFA models.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11135-024-01891-0>.

Acknowledgments This paper has benefitted immensely from the advice and support of others. I thank my PhD supervisor Rob Johns, who provided substantial guidance on the framing, structuring, and methods of this paper, and pointed me in the direction of factor analysis as a solution and without whom this paper would not be what it is now. I am grateful to the advisory board members who provided invaluable feedback and suggestions on this paper: Dominik Duell, Paul Whiteley, Roi Zur, and Ryan Bakker. I thank the members of my PhD cohort who proof-read this article and provided comments and suggestions: Samira Diebire, Anam Kuraishi, Sarah Wagner, and Lorenzo Crippa. I am thankful to the anonymous reviewers whose comments have improved this paper. All remaining errors are my own.

Funding This work was supported by the University of Essex Social Sciences Doctoral Scholarship.

Data availability Replication materials are available online at <https://github.com/philswatton/Correcting-Acquiescence-Unbalanced-Replication>

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not

permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Billiet, J.B., Davidov, E.: Testing the stability of an acquiescence style factor behind two interrelated substantive variables in a panel design. *Sociol. Methods Res.* **36**(4), 542–562 (2008)
- Billiet, J.B., McClelland, M.K.J.: Modeling acquiescence in measurement models for two balanced sets of items. *Struct. Equ. Model.* **7**(4), 608–628 (2000)
- Brown, T.A.: *Confirmatory factor analysis for applied research*. Guilford publications, New York city (2015)
- Cloud, J., Vaughan G. M.: “Using balanced scales to control acquiescence.” *Sociometry* pp. 193–202 (1970)
- Evans, G.A., Heath, A.F.: The measurement of left-right and libertarian-authoritarian values: A comparison of balanced and unbalanced scales. *Qual. Quant.* **29**(2), 191–206 (1995)
- Evans, G., Anthony H. and Mansur L. Measuring left-right and libertarian-authoritarian values in the British electorate. *Br J Sociol* pp. 93–112 (1996)
- Fieldhouse, E., Green, J., Evans, G., Schmitt, H., van der Eijk, C., Mellon, J., Prosser, C., (2017). British Election Study 2017 Face-to-Face Post-Election Survey. <https://www.britishelectionstudy.com/data-objects/cross-sectional-data/>
- Fieldhouse, E., Green, J., Evans, G., Mellon, J., Prosser, C.: British Election Study Internet Panel. (2020). <https://www.britishelectionstudy.com/data-objects/panel-study-data/>
- Forero, C.G., Maydeu-Olivares, A., Gallardo-Pujol, D.: Factor analysis with ordinal indicators: a monte carlo study comparing DWLS and ULS estimation. *Struct. Equ. Model.* **16**(4), 625–641 (2009)
- Heath, A., Evans, G., Martin, J.: The measurement of core beliefs and values: the development of balanced socialist/laissez faire and libertarian/authoritarian scales. *Br. J. Polit. Sci.* **24**(1), 115–132 (1994)
- Johanson, G.A., Osborn, C.J.: Acquiescence as differential person functioning. *Assess. Evaluat. Higher Educat.* **29**(5), 535–548 (2004)
- Johanson, G., Alsmadi, A.: Differential person functioning. *Educ. Psychol. Measur.* **62**(3), 435–443 (2002)
- Kenny, D.A., Kashy, D.A.: Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychol. Bull.* **112**(1), 165 (1992)
- Li, C.-H.: Confirmatory Factor Analysis with Ordinal Data: Comparing Robust Maximum Likelihood and Diagonally Weighted Least Squares. *Behav. Res. Methods* **48**(3), 936–949 (2016)
- Maydeu-Olivares, A., Coffman, D.L.: Random intercept item factor analysis. *Psychol. Methods* **11**(4), 344 (2006)
- Mirowsky, J., Ross, C.E., Eliminating defense and agreement bias from measures of the sense of control: a 2 x 2 index.” *Social Psychology Quarterly* pp. 127–145 (1991)
- NatCen-Social-Research. British Social Attitudes Survey, (2017). <https://www.ukdataservice.ac.uk/>
- Primi, R., Santos, D., De Fruyt, F., John, O.P.: Comparison of classical and modern methods for measuring and correcting for acquiescence. *Br. J. Math. Stat. Psychol.* **72**(3), 447–465 (2019)
- Primi, R., Nelson H.F., Felipe V., Daniel S., Carl F.F., Controlling acquiescence bias with multidimensional IRT modeling. In *The Annual Meeting of the Psychometric Society*. Springer pp. 39–52 (2019).
- Ray, J.J.: Is the acquiescent response style problem not so mythical after all? Some results from a successful balanced F scale. *J. Pers. Assess.* **43**(6), 638–643 (1979)
- Rodebaugh, T.L., Woods, C.M., Heimberg, R.G.: The reverse of social anxiety is not always the opposite: the reverse-scored items of the social interaction anxiety scale do not belong. *Behav. Ther.* **38**(2), 192–206 (2007)
- Rosseel, Y.: Lavaan: an R package for structural equation modeling and more. version 0.5-12 (BETA). *J. Stat. Softw.* **48**(2), 1–36 (2012)
- Savalei, V., Falk, C.F.: Recovering substantive factor loadings in the presence of acquiescence bias: a comparison of three approaches. *Multivar. Behav. Res.* **49**(5), 407–424 (2014)
- Stubager, R.: Education effects on authoritarian-libertarian values: a question of socialization1. *Br. J. Sociol.* **59**(2), 327–350 (2008)
- Surridge, P.: Education and liberalism: pursuing the link. *Oxf. Rev. Educ.* **42**(2), 146–164 (2016)
- Surridge, P.: The fragmentation of the electoral left since 2010. *Renewal: J. Labour Polit* **26**(4), 69–78 (2018)
- Surridge, P. (2019). The difficult truth for liberals: Labour must win back social conservatives. <https://www.theguardian.com/commentisfree/2019/dec/06/difficult-truth-labour-social-conservatives>

- Swain, S.D., Weathers, D., Niedrich, R.W.: Assessing Three Sources of Misresponse to Reversed Likert Items. *J. Mark. Res.* **45**(1), 116–131 (2008)
- Ware Jr, John E. (1978). Effects of acquiescent response set on patient satisfaction ratings. *Medical care* pp. 327–336
- Watson, D.: Correcting for acquiescent response bias in the absence of a balanced scale: an application to class consciousness. *Sociol. Methods Res.* **21**(1), 52–88 (1992)
- Webb, P., Bale, T.: *The modern British party system*. Oxford University Press, Oxford (2021)
- Winkler, J.D., Kanouse, D.E., Ware, J.E.: Controlling for acquiescence response set in scale development. *J. Appl. Psychol.* **67**(5), 555–561 (1982)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.