University of Oxford | Wellbeing Research Centre | March 2024

Assessing data quality in a Big convenience sample of work wellbeing

March 2024

UNIVERSITY OF OXFORD

Wellbeing Research Centre

Nuffield Foundation

Assessing data quality in a Big convenience sample of work wellbeing

William Fleming, George Ward and Jan-Emmanuel De Neve

Wellbeing Research Centre, University of Oxford

Acknowledgments

We thank members of our advisory group for their time and expert advice on the research: Alan Felstead (University of Cardiff), Ingrid Huitfeldt (BI Norwegian Business School) and George MacKerron (University of Sussex). We would also like to thank the Wellbeing Research Centre seminar series attendees for their feedback on the report, which we have included.

We are grateful to the Nuffield Foundation for funding this research (Grant Ref: WEL/FR-000024070). Our particular thanks go to Mark Franks for his support and feedback on the project and report throughout the process.

The Nuffield Foundation is an independent charitable trust with a mission to advance social well-being. It funds research that informs social policy, primarily in Education, Welfare, and Justice. The Nuffield Foundation is the founder and co-funder of the Nuffield Council on Bioethics, the Ada Lovelace Institute and the Nuffield Family Justice Observatory. The Foundation has funded this project, but the views expressed are those of the authors and not necessarily the Foundation.

Nuffield Foundation

www.nuffieldfoundation.org
@NuffieldFound

The data for this report was provided at no cost by *Indeed* and we are very grateful for generous data sharing. The employee wellbeing data used are publicly available at the employer-level on the *Indeed* website or available through licence agreement. De Neve acknowledges working on this report while being under contract with *Indeed*. The authors had full discretion over the content of this report.

This research is supported by the National Institute for Health and Care Research (NIHR) Oxford Health Biomedical Research Centre [NIHR203316]. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Summary

Survey research is facing a multitude of challenges to its validity, especially for the study of labour and organisations. Online surveys with non-probability, convenience samples are simultaneously seen as part of the problem and a promising solution. Methodological literature argues that researchers should not think of data quality of online surveys in terms of ‘good’ and ‘bad’ but in degrees, with a series of recommendations scattered across disciplines for assessing and managing data limitations. We present a case study of a Big, multi-level, online, convenience sample of subjective work wellbeing, the *Indeed* Work Wellbeing Score survey (IWWS). IWWS is an ongoing international survey of subjective work wellbeing, with over 20,000,000 responses and growing. In this study we evaluate the UK subsample collected by October 2023 (N = 1,463,503). While a prima facie valuable source of data, the data generation process raises concerns of selection bias and inattentive responses. We evaluate the extent of bias, variation in bias, response rates, internal consistency and employer cluster-level reliability. We then turn to considering what types of research questions a researcher may want to answer with the data, especially unit comparisons at different survey units and inter-item relationships. Overall, we suggest that at the individual, employee level, the survey suffers from selection and binary bias in responses, but that at the employer-level IWWS offers a valuable resource to supplement existing random probability surveys of work and wellbeing. In our conclusions we offer practical methodological recommendations for others using Big, online convenience samples. Finally, we provide commentary on the strengths and limitations of the IWWS for ongoing and future research, as well as the value for businesses, jobseekers and policy-makers.

Introduction

Crises and directions in survey research

Survey research in social science is in crisis, or at least facing a multitude of challenges to its survival (Krosnick et al., 2015; Meyer et al., 2015; Smith, 2013). The main problem is the increasing difficulty of conducting and maintaining surveys with random probability samples. Response rates and public funding for surveys are declining; whereas the effort and costs of maintaining sampling frames and producing valid, representative samples is increasing. Conversely, the demand for survey data from academic and market researchers is only increasing, with demand being met by a proliferating market of online surveys and relatively low-cost survey platforms with variable quality. Driving these trends are small to drastic changes in communication technologies, primarily the shift from landline to mobile, and social changes, such as a decline of social trust in institutions and rising data privacy concerns (Couper, 2017).

The stakes and challenges are more pronounced for labour and organisational research. Accurate labour data is vital for academic and policy-makers' understanding of national economies and labour market fluctuations; for example it is used in central banking decisions on interest rates. This information is frequently sourced from large scale random probability surveys. In the UK, the Labour Force Survey (LFS) has long been used when estimating employment rates. However, in autumn 2023, the Office for National Statistics (ONS) declared there was 'increased uncertainty around the LFS', and that they were shifting towards administrative tax data having identified significant divergence between data sources for key labour market indicators (ONS, 2023b).

The movement towards using administrative and full population sources is one approach for addressing the current predicament in labour and organisational data. This is of course a promising avenue for many research and policy needs, but brings with it significant limitations for many organisational research questions. Data access, reproducibility and a loss of the role for researchers in the data generation process are all well documented concerns (Connelly et al., 2016; Playford et al., 2016). An especial concern is losing the ability to develop survey items for specific research questions that would attend to gaps in empirical knowledge and theory.

The alternative developing avenue for labour and organisational survey research is the use of crowdsourced and large online surveys. Online convenience samples are very appealing because they save researchers time and money; offer access to larger and international samples; can target traditionally less well-represented groups; offer flexibility, convenience and simpler analysis; and provide a 'live' or dynamic data source (Evans & Mathur, 2018; Gosling & Mason, 2015; Tonidandel et al., 2015; Wenzel & Van Quaquebeke, 2018). While online samples offer many opportunities, they also bring several challenges and limitations for high-quality research because of the absence of a traditional sampling frame. Both the internal response quality and external generalisability are the key concerns (Couper, 2017; Van Quaquebeke et al., 2022). Selection and nonresponse bias therefore remain 'the core challenge facing survey research in the future' (Couper, 2017).

Many commentators are critical of convenience samples, arguing they should not be used at all (e.g. Bethlehem, 2010; Walter et al., 2019). However, samples in organisational research are almost always convenience, with no clear distinction between 'good' and 'bad' data. Instead researchers must adopt a reflexive approach to the various limitations in each source, working in degrees of quality (Landers & Behrend, 2015) and often adopting multi-source and multi-modal approaches (Lehdonvirta et al., 2021). A pragmatic approach such as this, integrating data limitations with appropriate research questions and transparency, avoids 'throwing the baby out with the bath water' (Stedman et al., 2019).

There is a nascent body of literature advising on the integration of large online samples into the repertoire of research methods. However, until now this literature has remained fragmented around specific concerns. Among several important dimensions, examples include missing data (Newman, 2014), internal response quality (M. K. Ward & Meade, 2018), response rate (Fulton, 2018; Holtom et al., 2022), transparency (Aguinis et al., 2018) or representativeness (Felstead, 2021). There has also been much discussion of online panel data or crowdsourced panel data from platforms like Prolific, Mechanical Turk and Facebook (Aguinis et al., 2021; Behrend et al., 2011; Hays et al., 2015; Lovett et al., 2018; Porter et al., 2019; Schneider & Harknett, 2022). There is a scarcity of evidence on the validity of online samples (Landers et al., 2019). Many methods textbooks rarely cover in detail approaches to data quality (e.g. Sue & Ritter, 2012) or they are too broad to grapple with the specific challenges of online modes (e.g. Blasius & Thiessen, 2012). We address this gap by evaluating the data quality of a novel, multilevel survey of employees and employers, *Indeed*'s Work Wellbeing Score survey (IWWS). This study continues a long tradition in survey methods literature for exploring biases in survey modes (e.g. Deming, 1944; Rosenthal, 1965; Simsek & Veiga, 2000; Suchman, 1962; Suchman & McCandless, 1940).

Report outline

We begin by presenting an integrative review of the methodological literature on assessing data quality. Our initial aim was to identify common and appropriate strategies. We group guidance into categories based on the stage of data collection and analysis: before data collection (*a priori*), after data collection (*post hoc*) and when analysing (*mitigation*).

We then fully introduce IWWS as our case study, describing its origin and key variables. We emphasise why this data source is analytically valuable and why this report is necessary for labour and organisational researchers generally and for those specifically interested in work wellbeing data. Following, the main body of analysis evaluates IWWS in a series of ten analytical 'steps' that apply general recommendations for assessing overall data quality. Our analysis is centred on post-hoc evaluation, but we offer suggestions at the a priori and mitigation stages. All steps are technically simple and designed to be easily applicable for researchers with only intermediate quantitative training.

In the discussion and conclusion sections, we summarise results, provide general recommendations and offer future pathways. We find that IWWS suffers sampling bias at the individual-level if the target population for the survey is taken to be the total UK workforce. Yet bias appears consistent across survey-levels and therefore remains especially valuable for employer-level analysis whether comparative or predictive. We also discuss whether setting the target population for IWWS as the total workforce is the correct analytical decision, proposing that instead the IWWS offers a unique insight on job seekers, a subgroup of the workforce especially relevant to science and policy. Our general recommendations encourage as much involvement in the data generation process as possible, flexibility and pragmatism in post-hoc evaluations and research design and caution regarding correction methods. Ultimately, research questions must be commensurate with the limitations of any data source and data quality must be considered in degrees. We finish by reflecting on the value of IWWS for labour and organisation studies as well as for individual workers and organisations concerned for wellbeing.

Literature Review

To identify existing guidance on assessing data quality with special attention to online surveys, we conducted structured and unstructured literature reviews. The structured search strategy was not productive, producing few relevant articles. Instead we found guidance across fields from organisation and management studies, psychology, economics, sociology, epidemiology and statistics, piecing together recommendations from various perspectives. Managing and assessing bias in survey samples can be undertaken at several points in the data generation, collection and analysis stages.

A priori strategies for improving survey data quality

The quest for high quality data in online surveys naturally begins at the earliest stages of study design and data collection. Researchers and survey designers must theorise the data generation process, reflecting on the types of themes that will be covered, the types of respondents that will be accessing the survey and the resultant biases that will be introduced as a result of both. It follows that researchers are advised to be as involved in the data collection processes as possible, using the early vantage points to include survey items that will act as quality checks and response enhancing strategies. This

It is important for researchers to involve relevant stakeholders in the data collection process (Fulton, 2018) and issue pilot or feasibility surveys if appropriate (Thabane et al., 2010). Some researchers also suggest using multiple methods of data collection to provide a point of comparison for the primary online survey (Lehdonvirta et al., 2021). Researchers are encouraged to consider the selection of survey items as well as the visual elements of the survey to ensure short and effective surveys that reduce response burden and fatigue (Tourangeau et al., 2013).

Inattentive or careless responses should be a key concern for online surveys. Consideration should be put into the design of the survey to enhance response quality. This may involve survey instructions that demand attention, such as adding a signature to the response or asking for specifically formatted answers (e.g. dates) (Ward & Meade, 2023; Zickar & Keith, 2023). Ward & Pond (2015) even show that using visual signs for the user, by including an image of a virtual human, reduces inattentive responses. Other attention checks, such as bogus items (e.g. 'Answer 1 to this question') are effective methods (Brühlmann et al., 2020; Kung et al., 2018; Zickar & Keith, 2023). Even simple steps such as using reverse scaling on some survey items can provide tests of response quality or identify inattentive respondents (Kung et al., 2018). Somewhat surprisingly, another effective technique is directly surveying respondents on the reliability of their response. Meade & Craig (2012) suggest 'should we use your data?' or a similar multi-item scale, whereas Dohmen & Jagelka (2023) show that including 'I am someone who is sure that my answers to these questions describe me accurately' and 'Please indicate on the scale below how reliable are your answers to this survey' can act as effective controls for response bias. Measuring individuals risk-taking personality traits is another method researchers have used to assess likelihood of inattentive responses (Peer et al., 2022).

Beyond the response patterns and attitudes of respondents, the choice of substantive theoretical items is naturally a vital decision for ensuring high quality responses. Psychologists recommend that researchers use established and validated multi-item scales where possible, as these offer more avenues for assessing reliability and have been shown to have stronger criterion validity (Schriesheim et al., 1991). However, single item scales minimise response burden, reduce criterion contamination and increase face validity, but require more creative processes of reliability and validity assessment (G. G. Fisher et al., 2016).

Finally, researchers should carefully consider the utility and applicability of incentives for online survey data collection, as they have been shown to enhance response rate and quality (Fulton, 2018; Holtom et al., 2022; Van Quaquebeke et al., 2022). Incentives can be monetary, with many online panel survey platforms such as Mechanical Turk and Prolific operate in this fashion, offering minimal fees for participation in each study or survey. Van Quaquebeke et al. (2022) argue that financial rewards are the most effective strategy for enhancing responses, as opposed to non-financial rewards. However, give-to-get options on employer review sites such as *Indeed* and *Glassdoor*, where respondents only see previous reviews if they provide their own, elicit responses with less bias (Marinescu et al., 2022). However, incentives are not always successful at eliciting higher numbers of responses or higher quality responses, and comparisons should be made between those who have responded voluntarily and those who have responded when offered an incentive, pecuniary or otherwise (Cycyota & Holtom, 2008; Holtom et al., 2022). Much of the guidance on incentives and careless responding have been developed alongside the growth of online panels, where survey respondents are highly adept at identifying response enhancing and attention checking questions while completing the survey in inattentive ways (Porter et al., 2019).

Post hoc evaluations of data quality

In this report we primarily focus on what we categorise as post hoc evaluations of data quality. A priori methods offer preventative solutions, but as much survey research is conducted as

Response rates in organisational research are taken as a key indicator of data quality. However, rates vary greatly with consistent concerns in methodological literature (Baruch & Holtom, 2008; Cycyota & Harrison, 2006; Fulton, 2018; Holtom et al., 2022). Meta-analyses of published organisational research show that studies analysing individual-level data have average response rates of between 30-50%, depending on method (Fulton, 2018; Holtom et al., 2022), whereas organisational-level analysis saw average response rates of 35.7% (Baruch & Holtom, 2008). For Fulton (2018) response rate is a major issue in organisational research because of declining rates. Holtom et al. (2022) are more positive, seeing average reported response rates increasing in recent years, but do argue that response rate itself is not an indicator of the quality of a survey study. Both Fulton (2018) and Holtom et al. (2022) present frameworks and guidance for evaluating response rate. Holtom et al. urge reflection on sampling techniques, researcher-participant relationships, participant qualifications and motivation, survey make-up and research context. Fulton suggests 'nonresponse analysis' by investigating variation in subgroups and waves, re-surveying nonresponders, using response patterns and response time, and assessing stability over time. Celhay et al. (2024) make an important contribution in showing and arguing that a high response rate should not be taken as a prima facie sign of quality, but rather data collection and analysts should strive for survey accuracy.

In contrast to the discussions from the likes of Fulton and Holtom et al., which discuss target populations of organisations and internal workforces, online surveys that are open to all internet users do not define a target population effectively (Bethlehem, 2010). The sample therefore becomes a proportion of the whole population. Bethlehem (2010: 172-174) presents expressions for estimating the absolute maximum bias of sample statistics based on the response rate $|Bmax| = S(Y)\sqrt{\frac{1}{\bar{p}} - 1}$ where *Bmax* is the maximum bias, *S(Y)* is the standard deviation of the population and $\bar{p}$ is the response rate. For standard household random probability surveys, response rate from the sampling frame is typically 70%, producing a maximum bias of 0.65 of the standard deviation. As a standalone value this expression offers little, but when compared with other surveys, a relative bias can be inspected. As an example, Bethlehem shows the maximum bias for an online survey of the Dutch population was 13 times higher.

The process for comparing bias in a sample shows the importance of having alternative survey modes and samples to compare estimates between. Comparisons with random probability samples are one of the important and common steps of assessing the external validity and representativeness of an online sample, but is done too infrequently (Callegaro et al., 2014; Fulton, 2018). However, there are several examples where this has been done in labour and organisational research. Felstead (2021) compares point estimates and demographic distributions for an online job quality quiz with a probability sample of the same survey and with the UK Labour Force Survey. Spencer et al. (Spencer et al., 2022) do not compare with a probability sample but compare estimates for an online sample of platform workers with a telephone sample to assess the extent of bias in the online mode. Winton & Sabol (2022) compare between survey modes, finding that four different kinds of sampling (student, crowdsourced, professional panel and social network snowballing) produced similar estimates. As a result they advocate pooling data sources when this is the case. Elsewhere, Lehdonvirta et al. (2021) compare online sampling techniques with administrative data. While they find the online convenience samples produce biased estimates because of 'topical self-selection', selecting in because of the themes included in the survey, they argue that online convenience samples remain a valuable tool for researchers, especially for exploratory studies with hard-to-reach or emerging social groups. Comparisons with other samples can be conducted at the individual level as these studies do, but organisational researchers can compare with other sources of data at the organisation level as well (Landers et al., 2016).

The examples above provide comparisons of overall distributions between online convenience and probability samples and between different survey modes, but further comparisons can be made for regression analysis. For example, Thompson & Picketty (2020) compare estimates from a household survey and online panel data, and argue that convenience samples can offer insights for the direction of possible effects but that regression analyses with these samples do not underestimating the size of effects. For running experiments, Lutz (2015) found similar underestimation for online panels when compared to laboratory contexts.

Within surveys, convergent validity is an important consideration. Convergent validity is the extent to which measures of the same construct correlate. Carlson & Herdman (2012) argue that too few studies and samples analyse convergent validity. They assess that convergent validity tests of correlation should be concerning to researchers when estimates are less than 0.85.

In psychometric literature there is widespread concern with 'inattentive', 'careless' or 'insufficient effort' responses reducing the internal quality of a survey sample and undermining the legitimacy of scale validation (M. K. Ward & Meade, 2023). In fact, Zickar & Keith (2023: 322) even state that 'when researchers discuss data quality within online samples, their primary concerns revolve around attention, honesty, and, more recently, bots'. While there are steps that can be introduced to prevent and mitigate poor quality responses in the data collection and survey design stages, there are many techniques for evaluating the prevalence of these issues. Ward & Meade (2018, 2023) offer excellent summaries of multiple techniques.

One approach is to identify and exclude 'straightliners', that is respondents who mark the same response for a list of survey items (Belliveau & Yakovenko, 2022). Straightlining is a simple measure of response pattern that is easily detectable, and is primarily a problem because of the inflation of inter-item correlations. Random, invalid response patterns are harder to detect but pose less of a challenge to internal consistency of a survey or scale (DeSimone et al., 2018). However, Reunig & Plutzer (2020) caution against immediate exclusion, because while it is unlikely that these respondents will be providing carefully considered responses (satisficing), it is not impossible they are valid. They suggest researchers can minimise satisficing straightlining by using well-validated and reliable measures.

Beyond straightliners there are multiple other measures of careless responding that can be inferred from response patterns (M. K. Ward & Meade, 2018, 2023; Zickar & Keith, 2023). The length of string (how many values are the same in sequence) can be used, with individual response variability measured by variance or standard deviation statistics (Dunn et al., 2018) and recommendations for exclusion based on 6-14 repeated responses (Zickar & Keith, 2023). A range of 'consistency indices' can also be used (M. K. Ward & Meade, 2023). There are individual correlation coefficient for items, grouped consistency indices, even-odd consistency (split scales) and theoretical grouping of items in scales. Most of these approaches are applicable in large psychometric surveys with well-validated, theory-informed scales. There are also several complex factor models recently developed for detecting inattentive response patterns (Arias et al., 2020; Kam & Cheung, 2023; Steinmann et al., 2022). While promising, these models are again only especially relevant for large surveys with multi-factor scales. These techniques are also only currently available in MPlus software: popular among psychologists but less so in other social science disciplines and not freely available. Despite the many, and increasingly complex, survey design and statistical techniques for improving response quality, Ward & Meade (2018) suggest these techniques often will not fully address the challenges of inattentive response or be able to improve the overall data quality.

A strength of online surveys is the ability to leverage related para and metadata during the data collection process. One measure that is a potentially useful indicator of response quality is the length of time for survey completion (Belliveau & Yakovenko, 2022; Ward & Meade, 2023). Response time may prove a useful measure because it is not obvious to respondents that it is being measured and more difficult to avoid detection (ibid.). Ward & Meade (2018; 2023) recommend capturing completion time by survey page, rather than for the entire survey.

Another strength of online surveys is that the data collection process is often a live and dynamic process, continually updating as new responses are collected. This 'always-on' aspect of some online surveys is an effective way of collecting large quantities of data and in a longitudinal manner. However, this means that responses are often collected at different dates and may affect responses. As an assessment of stability, Fulton (2018) and Gile et al. (2015) suggest plotting cumulative estimates and identifying stable distributions. Eichstaedt & Weidman (2020) undertake a similar procedure but detect stability over time by assessing responses over days during data collection.

Mitigating sampling bias

Within methodological and applied literature there exists a number of established measures for mitigating bias. The first and most obvious is the exclusion of cases which do not meet various levels established during post-hoc evaluations. Exclusion could be based on survey incompletion, failing attention checks, detection in careless response metrics or based on answers to specific questions (Ward & Meade, 2023). The strictness of exclusion criteria will vary depending on the research questions.

In experimental settings, exclusion of cases may be warranted where it directly undermines the inferential validity of a study. However, in survey research, While exclusion is warranted in many cases, unreflective exclusion may restrict analysis by obscuring issues in data collection, data generation or patterns in specific responses. Literature on missing values highlights what are similar issues when handling missing data and the types of decisions researchers make when deciding whether to exclude, impute values through various methods or analyse multiple subsets of data based on case completion (Newman, 2014). Often best practice is decided by disciplinary norms.

The exclusion of cases only functions as a mitigation strategy when poor quality or biased results are detectible or feasibly removable. In situations where researchers have established the bias in a sample and the extent of bias, various statistical methodologies are more relevant.

For a sample with selection bias and/or non response bias, Heckman correction is the most established method (Heckman, 1979; Vella, 1998). Heckman correction is a two-step approach that relies on the researcher being able to model the functional form of selecting into a specific category, meaning it is not always possible. Other common methods involving modelling the selection is through weighting, either with inverse probability weight or propensity score adjustment, but both rely on a comparative sample (Bethlehem, 2010; S. Lee, 2006; Nohr & Liew, 2018; Tourangeau et al., 2013). Bounded estimates are another method, especially where research questions centre on calculating specific point estimates such as official population statistics and treatment effects (Blundell et al., 2007; Dutz et al., 2021; D. S. Lee, 2009; Manski, 2016). Despite the widespread use of these methods, Dutz et al. (2021), comparing a convenience sample survey with census data for Norway, show that none of these options effectively compensate for non response bias. Felstead's (2021) application of weights in the analysis of the online job quality quiz also does not produce a representative sample for all important demographic variables. Hanley (2017) also shows that statistical corrections can produce spurious results. Seperately, Liu et al. (2023) propose Bayesian estimation and inference methods when using non-random samples, but their solutions rely on having sampling frames and administrative data to determine the probability of inclusion and therefore confidence in any conclusions.

We've briefly discussed multiple options for assessing data quality and its specific dimensions, emphasising that the process of ensuring a valid sample for analysis starts before any data is collected, can then be evaluated once it has been stored and that there are various decisions that can be made for mitigating any bias. The exact steps any researcher takes will depend on the mode of collection, the survey items included, the research questions, as well as other important variables in survey and research design.

Materials and Methods

Indeed Work Wellbeing Score survey

Job sites such as *Indeed* and *Glassdoor* are increasingly popular sources of data for assessing job quality and worker wellbeing. In this report we evaluate the representativenes, validity and reliability of the *Indeed* Work Wellbeing Score survey (IWWS) as a case study of a Big, online, convenience sample. IWWS is a short individual-level survey of subjective work wellbeing measures within which respondents identify their current or former employer. The resultant multilevel structure of IWWS offers a valuable resource for researchers of work and wellbeing, with no comparable survey providing this information. IWWS is open to *Indeed*'s active user base of over 250 million unique users. Data is collated by country using website arrival codes, with the full international sample covering over 20 million responses, the majority of which are US-based. Once a single employer has received over 20 responses to the IWWS, an average score for work wellbeing is presented publicly on the employer page on *Indeed*'s website. For the following analysis we used the UK subsample (N = 1,463,503) with multiple further subsets depending on the analytical procedure and relevant exclusion criteria (Table 1).

We accessed the IWWS data through a sharing agreement with data owners *Indeed* for this methodological study and other substantive research studies (De Neve et al., 2023; G. Ward, 2022). Other researchers have held similar agreements with *Indeed* (e.g. Londakova et al., 2021; Sleeman, 2024), while other teams leverage the public nature of job review data on *Indeed* and *Glassdoor* using web scraping techniques in data collection (e.g. Sainju et al., 2021; Suen et al., 2020). Some researchers even licence private third-party data collectors to analyse job postings and job reviews (e.g. Forsythe et al., 2020; Zhang, 2023).

There is much interest among computational social scientists in Big Data sources. IWWS meets some but not all Big Data characteristics, as set out by Salganik (2017). IWWS is big, always-on, incomplete, nonrepresentative and what Salganik (ibid.) terms 'dirty'. It is big in terms of sample size when compared to traditional surveys; always-on because data collection is an ongoing process; incomplete because of the high amount of missingness; and non-representative because it captures users of a job site as opposed to the total workforce population. IWWS is also noisy or 'dirty' because data collection is intended for crowdsourcing employer reviews rather than for answering scientific questions. However, as it is primarily a survey and not the result of users interacting with an automated digital technology, it is not non-reactive, inaccessible, drifting, algorithmically confounded or sensitive (ibid.). IWWS is therefore a very modern form of survey, utilising the opportunities of online data collection for easily accessible, cheap and dynamic information on questions of interest to science, policy and business.

A central argument in the survey methodology literature for assessing data quality is that researchers must consider the data generation process and the biases it introduces. We hypothesised two primary limitations in the data generation for IWWS. The first, the result of the Big nature of the data, is the noisiness and incompleteness based on how users interact with the survey. Online surveys are known to provide poorer quality and more negative responses in general (Ward & Meade, 2023), partly as a result of less social desirability bias (Sue & Ritter, 2012) but also more malign processes such as bot infiltration (Griffin et al., 2022). *Indeed* use various automated filters to screen for bots on their site so this is presumed not to be an issue for IWWS, but overall poor quality responses remained a key concern at the start.

The second primary limitation is, counter to traditional organisational surveys where respondents are more likely to respond if they have higher job satisfaction (Fauth et al., 2013), we hypothesise that IWWS will capture respondents more dissatisfied with their job. Survey respondents are users of the *Indeed* website, an online job market. As a result, the survey likely captures responses primarily from those engaging with the job market or at least considering job transitions. From the outset, IWWS presumably suffers from selection bias. There is no 'economic self-selection' in the survey, where respondents are materially benefiting from survey completion, but there is 'topical self selection', whereby respondents are more likely to be completing the survey motivated by variables that are measured by the survey items (Lehdonvirta et al., 2021; Suchman & McCandless, 1940).

Figure 1 demonstrates the theoretical model of this sample selection in a directed acyclic graph (DAG), as recommended (e.g. Nohr & Liew, 2018). In Figure 1, R is survey response and is predicted by the outcome of interest, subjective work wellbeing (Y), with more dissatisfied workers more likely to respond. R is also theoretically the result of a range of observable variables (X) and unobservable variables (U), which also predict subjective work wellbeing (Y).

IWWS offers a prima facie valuable data source in providing data that matches individuals and employers. However, with clear limitations brought about by its online presence, easily accessible survey and likely user base of job seekers, it requires analysis of to what extent these problems exist.

FIGURE 1: DAG OF SELECTION INTO IWWS

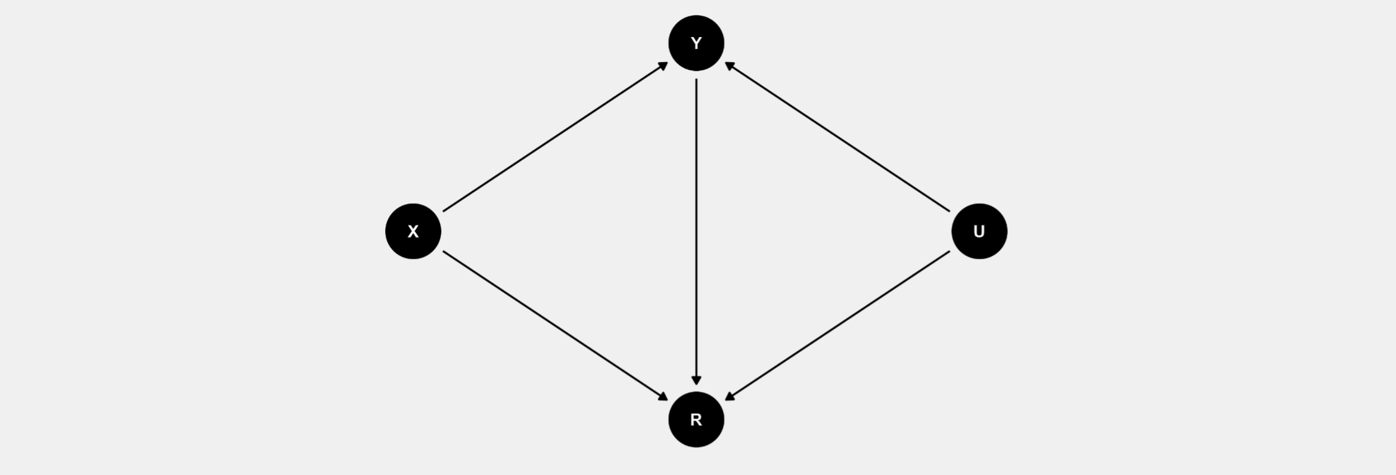


TABLE 1: SAMPLE SUBSETS AND SIZES

Sample	N
Individual-level	
Full UK sample	1,463,503
Replace individual response*	1,152,607
Single item responses excluded	987,240
Single item and straightliners excluded	793,527
Minimum 5 responses per employer	1,141,073
Employer-level	
All employers	241,593
Minimum 5 responses	37,771
Minimum 10 responses	16,798

Note: * Sample if only the most recent response for an individual account ID is included

Variables

IWWS is a short two page survey. The first question requires only the company name of the respondent's current or former employer, which in some cases will already be filled in if the respondent has clicked, for example, on "leave a review" on a company's page. The page continues with 15 items relating to subjective evaluations of the job, with 1-5 ordinal response options (Table 2). In a rudimentary theoretical model (De Neve & Ward, 2023), we frame job satisfaction, work purpose, work happiness and work stress as four 'outcomes', while the following 11 items are 'drivers' of these four outcomes. The survey begins with the four outcomes and asks the drivers afterwards. IWWS does not collect personal demographic data, a limitation for assessing its representativeness.

A strength of online surveys is the related metadata that can be collected during data collection. For the IWWS where there is a scarcity of personal identifiers and demographics at the individual level, paradata takes on greater importance for assessing the quality of the data. Available paradata included date and time surveys were completed, non-identifying variable such as rough location that can be derived from the IP address, and survey arrival route (Table 3). Using automated codes from the *Indeed* website, the data includes the route through which a respondent accessed the survey. Respondents may have accessed the survey unprompted, when uploading their resume or were invited to review through email invitations. One access route that is especially interesting is those who have completed the survey in order to see existing reviews in a give-to-get principle. Give-to-get can reduce bias in online job reviews (Marinescu et al., 2018).

The second page of the survey, if the respondent clicks through to it, collects further information. This includes a written review as well as the respondent's job title and whether they are a current or former employee.

TABLE 2: IWWS ITEMS

Survey item	Short-hand	Response
What is your company name? It can be your current or former employer.	Employer	Open text and suggestions
Overall, I am completely satisfied with my job.	Satisfied	1-5
My work has a clear sense of purpose.	Purpose	1-5
I feel happy at work most of the time.	Happiness	1-5
I feel stress at work most of the time.	Stress	1-5
I am paid fairly for my work.	Paidfair	1-5
There are people at work who give me support and encouragement.	Support	1-5
There are people at work who appreciate me as a person.	Appreciate	1-5
I can trust people in my company.	Trust	1-5
I feel a sense of belonging in my company.	Belonging	1-5
My manager helps me succeed.	Manager	1-5
My work environment feels inclusive and respectful of all people.	Inclusive	1-5
My work has the time and location flexibility I need.	Flex	1-5
In most of my work tasks, I feel energized.	Energized	1-5
I am achieving most of my goals at work.	Achieve	1-5
I often learn something at work.	Learn	1-5

TABLE 3: VARIABLES FROM METADATA AND PARADATA

Variable	Data collection method
Current or former employee	Derived from dates for reviews
Survey completion route	Website paradata
Date of completion	Website paradata
Survey completion time	Website paradata
Standard Occupational Classification	Job title categorisation automated by *Indeed* website engineers
Industry sector	Website metadata
Location	IP address normalisation
Employee number	Linked from employer-level Indeed.com data

Analytical strategy

To assess overall quality of the IWWS data, we present results in a series of analytical steps. The structure of the results is designed to present a logical and, within reason, chronological process of evaluation. These steps are mostly designed to be technically simple and followable by researchers with limited to intermediate methodological training. Table 4 offers a summary of these steps and Table 5 summarises which dimension of data quality, reliability and validity the analysis is relevant for.

Step 1 investigated the overall frequency distributions of key variables in the sample. We present results for both the individual and employer levels of the survey to identify possible sampling errors.

In **Step 2**, we explore the response rate at the employer level as a proportions of employer size. IWWS is linked with an employer-level variable of each organisation's size. Possible responses are a range (2-10, 11-50, 51-200, etc.) so we inspected minimum and maximum responses based on the upper or lower limit of these ranges.

In **Step 3**, we assessed the representativeness of the frequency distributions presented in Step 1. We present comparisons of these distributions with random sample surveys from the same time period: *Understanding Society* waves 11-13 (University of Essex, Institute for Social and Economic Research, 2023) and the UK sample of the *European Working Conditions Telephone Survey 2021* (Eurofound, 2021). We conducted this for four variables in IWWS (job satisfaction, energy at work, work purpose and manager support) because we found appropriate matches within the other surveys.

Next, in **Step 4** we explored the possible sources and measures of bias that could be measured through IWWS paradata, including response time, survey access route, employment status and response pattern. For response patterns we considered single and partial completion as well as straightline (1, 1, 1 and so on) and zig-zag (1, 2, 3, 4 and so on) response patterns. We also coded a variable from the stress item as a form of attention check. We coded this item to 1 if the answer to the subjective stress question was equal to the questions above (work happiness) and below (fair pay). Stress was reverse ordered compared to the other items, with higher values negatively representing higher stress. We present descriptive counts of all these variables, the sample size if exclusion criteria is met and how distributions change for subsets of the sample.

In **Step 5**, we analysed whether the prevalence of these potential sources of bias are explained at different levels of the survey. To achieve this we estimated a series of 4-level variance components models (VCM) with random intercepts for employer, industry and region to extract variance partition coefficients, i.e. explained variance, at each level. The regression equation for these models is:

$$y_{ijkl} = \gamma + u_{jkl} + v_{kl} + w_{l} + e_{ijkl}$$

Where y_{ijkl} is the outcome for the *i*-th individual in the *j*-th employer, the *k*-th industry and the *l*-th region. γ is the overall intercept, *u* is the employer intercept, *v* is the industry intercept, *w* is the region intercept and *e* is the error term. From estimating this equation, the proportion of variance explained at each level (variance partition coefficients) can be calculated. The aim of this analysis was to investigate whether sources of bias different across multiple units and levels of analysis.

While convenience sampling is primarily an issue for external validity and representativeness, rather than internal quality, a common critique of online surveys is the higher likelihood of careless or inattentive responses. In **Step 6** we established the internal consistency and convergent validity of the survey items by estimating inter-item Pearson correlations. In **Step 7** we evaluate the reliability as stability of responses by employers by using bootstrapped means. Bootstrapping draws random subsets from the sample, operating as a cluster-level test-retest. This analysis also provides a possibly more reliable way to inspect differences between employers, with comparisons of average levels a key possible question researchers, practitioners and job seekers might have of IWWS. For this reason we present bootstrapped means distributions for nine supermarket chains in the UK.

Having assessed the bias in IWWS, in Steps 8 and 9 we moved on to considering possible research questions that researchers and practitioners would be expected to have of a large survey of workers' wellbeing. **Step 8** compares the aggregate values and rankings for six work wellbeing items in UK civil service departments, comparing IWWS to the UK Civil Service People Survey (ONS, 2023a). In **Step 9**, we evaluated whether coefficient estimates for multivariate regression models were comparable to other samples by estimating models for IWWS and the *Skills and Employment Survey 2017* (Felstead et al., 2019).

In the final step, we link to secondary and substantive research we have conducted elsewhere. We propose these analyses as tests of predictive validity and catalytic validity. Predictive validity is a criterion-oriented validity procedure which assess to what extent a test or measure correlates with another outcome it would theoretically relate to (Cronbach & Meehl, 1956). Catayltic validity refers to 'the degree to which the research process re-orients, focusses, and energizes participants' (Lather, 1986: 67).

TABLE 4: ANALYTICAL STEPS AND DATA USED

Step	Aim	Data subset
1	Explore overall distributions at survey levels	Full sample; multiple subsets
2	Identify response rates by employer	Single responses removed; No repeat observations; minimum cluster size
3	Compare distributions with random probability samples	Full sample
4	Identify possible sources of bias	Full sample
5	Convergent validity and internal response quality	No straightliners
6	Assess whether bias is consistent across survey levels	Full sample
7	Employer-level reliability through bootstrapping	6 largest employer samples
8	Comparisons of average levels between employers	UK civil service employees, no single response
9	Comparisons of multivariate regression outputs	Multiple subsets
10	Predictive validity for theoretically consistent relationships	Full sample

TABLE 3: DATA QUALITY ASSESSMENT AND ANALYTICAL STEPS

Data quality assessment	Analytical step
External validity/representativeness	1, 3, 8
Sampling	1, 3, 4, 5
Convergent validity	6
Predictive validity	9, 10
Response rate	2
Reliability	6, 7

Analysis

Step 1 – Overall distributions

As with any quantitative data set, the first stages of any analysis consist of sense checks of key variables, their distributions and sample statistics (Sue & Ritter, 2012). Figure 2 presents the overall frequency distributions of the 15 work wellbeing items in IWWS. Immediately, a 'binary bias' (M. Fisher et al., 2018) is visible in the distributions, with inflated counts at 1 and 5 revealing a bimodal and right skewed distribution. In this case, the distribution appears to have a 'missing middle' to what would theoretically be a normal, censored distribution. The high counts for 1 offer obvious confirmation to the prior concerns of selection and non-response bias.

The high counts for 5 emphasise the binary bias in responses. Binary bias is common in market and consumer research, where respondents appear to make categorical distinctions between 'good' and 'bad', rather than responding on a normal, ordinal distribution. These counts suggest that, when completing the IWWS questionnaire, respondents are completing the survey within a similar mindset to customer and product reviewers. An alternative explanation is that there is social desirability bias in responses, where despite a statement regarding anonymity in survey responses, respondents may be self-consciously completing the survey to give positive reviews of their current or former employers. A final possibility is that respondents are aware of the online publication of the data and actively manipulate results to give certain employers inflated overall scores. On a small scale this seems plausible, but seems unlikely to be consistent across such a large individual and employer-level sample.

The overall distributions for the individual-level data suggest there is considerable bias in response patterns in IWWS. However, the primary strength of IWWS is in offering a matched sample of employees and employers. Figure 3 presents distributions of the mean job satisfaction by employer with different employer-level sample cut offs. When all employers are included, a similar bimodal distribution is clear, with the most common values at 1 and 5. However, as the cut-off for minimum company cluster sizes was increased, the distribution resembles a right-skewed bell-shape, getting closer to a theoretical normal distribution. Average wellbeing scores (an index of job satisfaction, work happiness, purpose and stress) are made public on *Indeed*'s job board for each employer when 20 responses have been collected, and publicly presenting only this sample seems a reasonable strategy. Figure 4 shows the distributions for all the work wellbeing items, showing that this bell-shape is consistent across the 15 key work wellbeing variables once the cut off is specified at a minimum of 10 employees.

While the individual employee-level sample appears biased, the employer-level distributions offer more promise for analysis of work wellbeing at an aggregate employer level, especially considering the current gap in data coverage at this level of analysis.

FIGURE 2: WORK WELLBEING SURVEY ITEM DISTRIBUTIONS

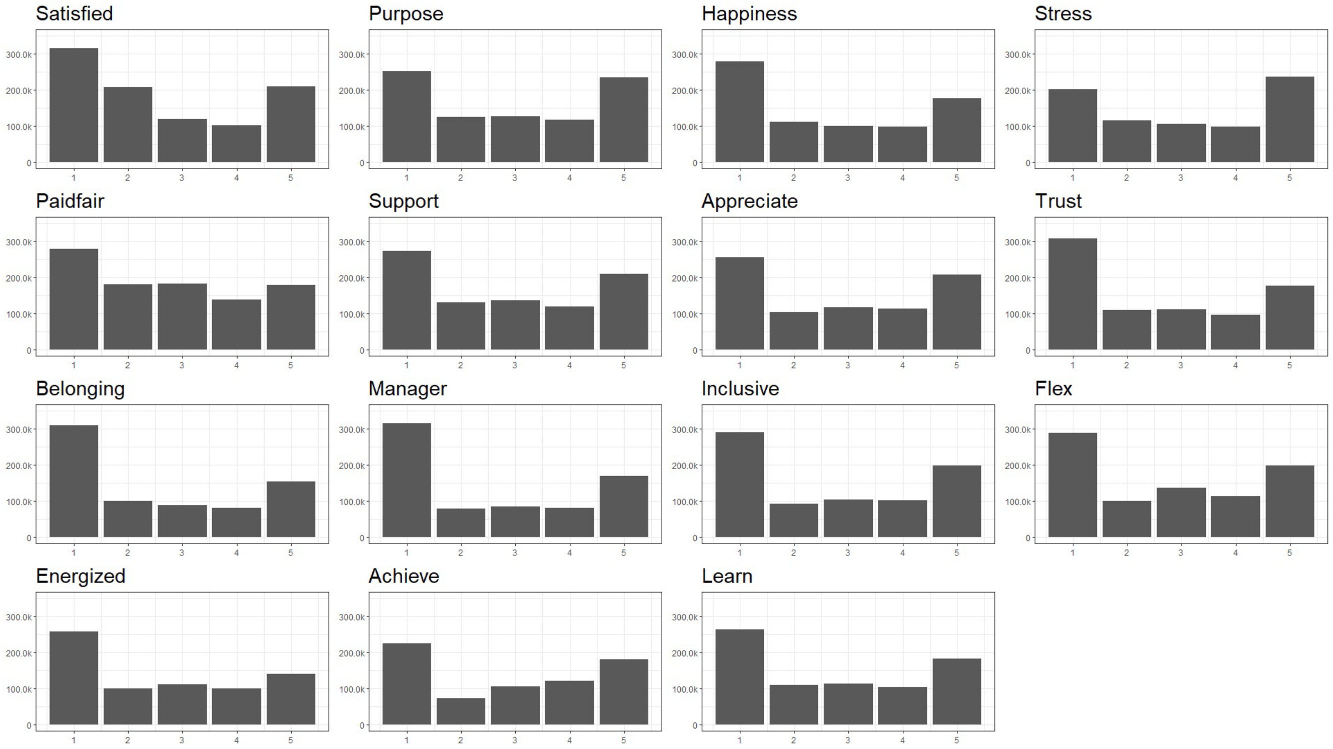


FIGURE 3: EMPLOYER-LEVEL DISTRIBUTIONS OF JOB SATISFACTION WITH MINIMUM CLUSTER SIZE

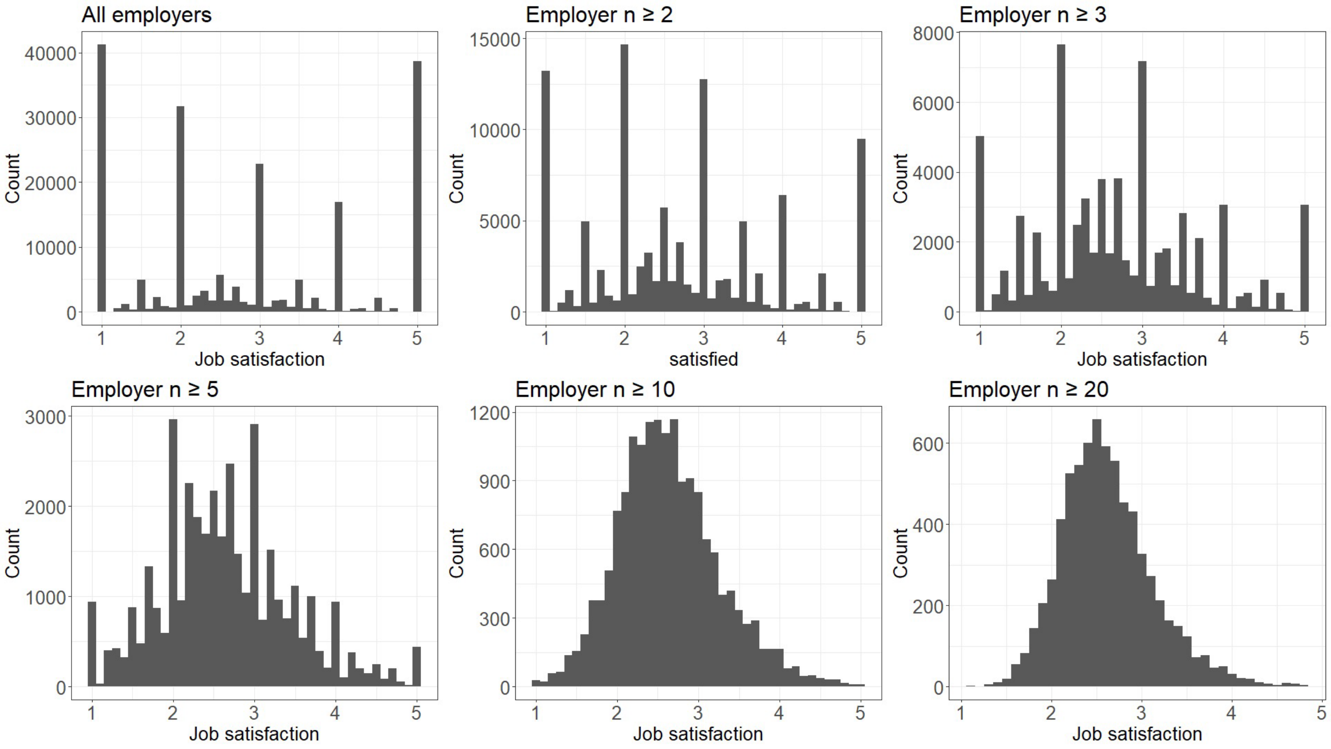
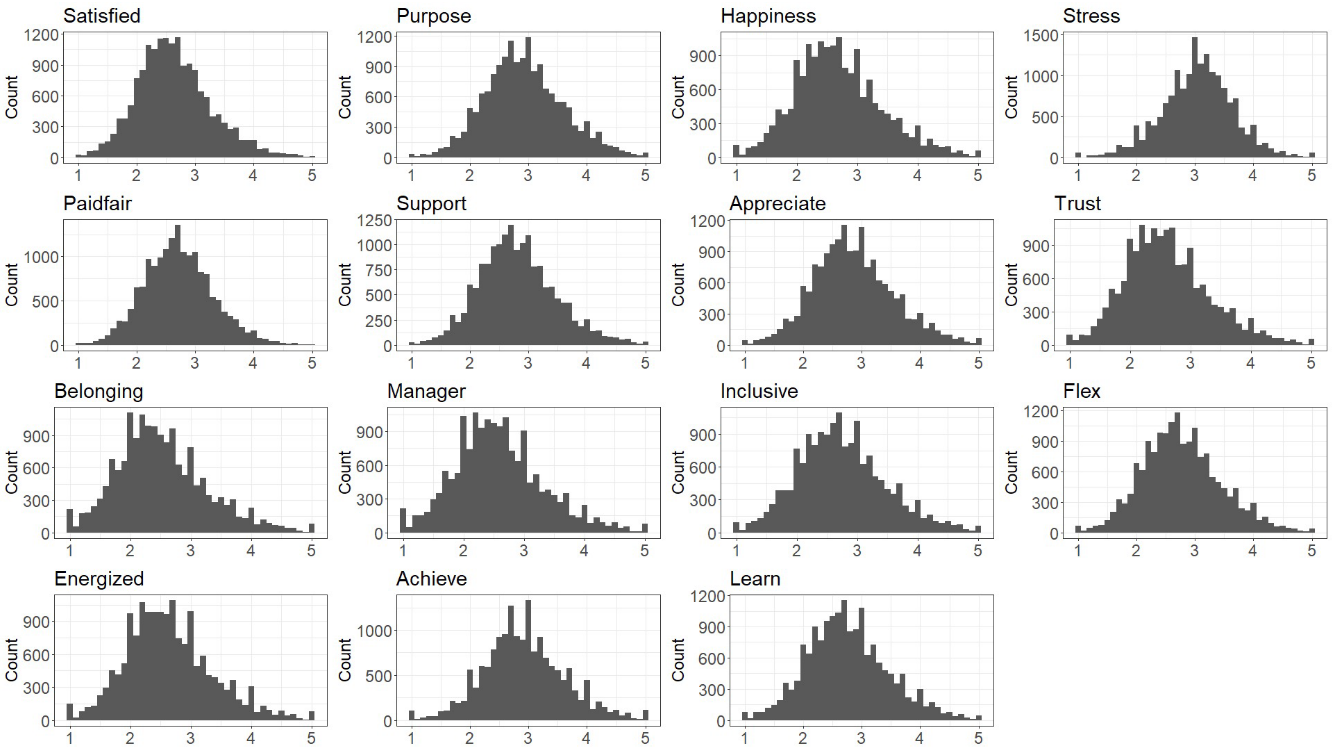


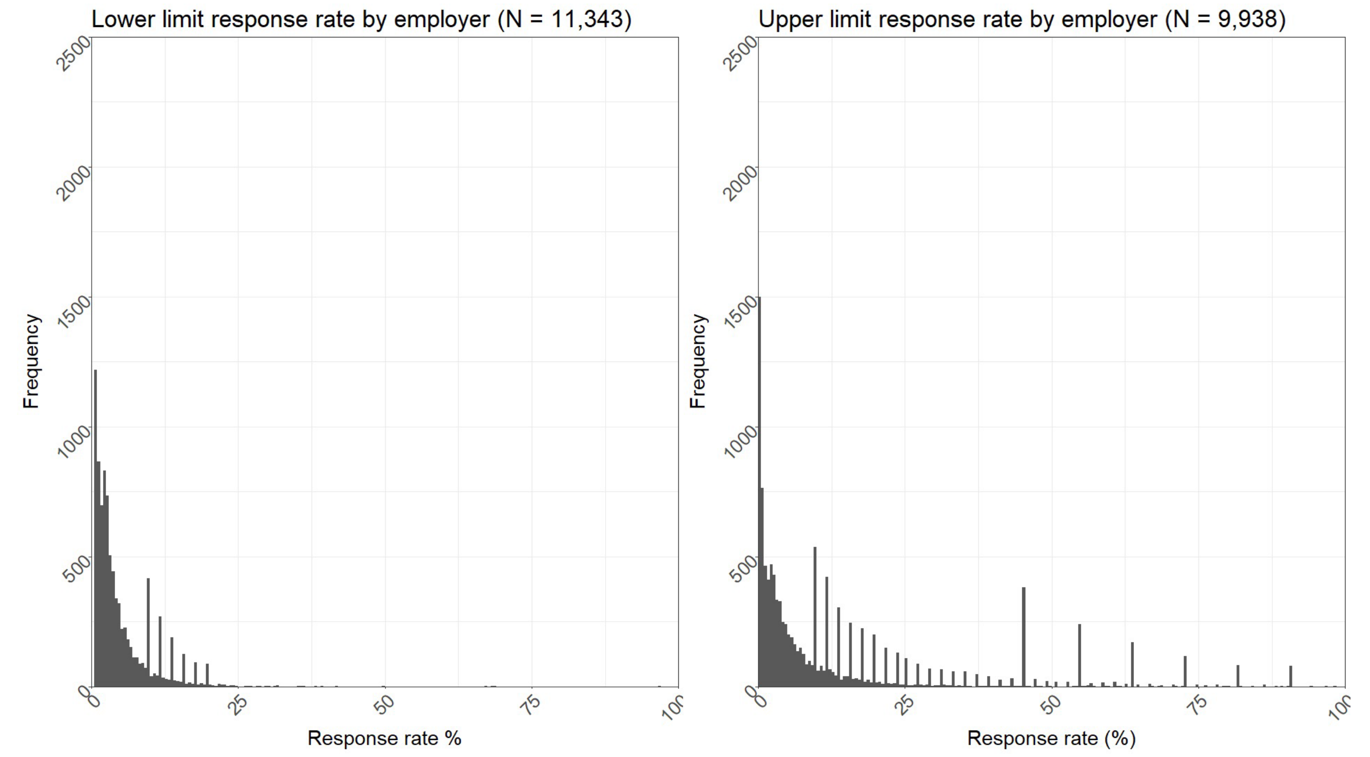
FIGURE 4: IWWS EMPLOYER-LEVEL DISTRIBUTIONS FOR CLUSTER SIZE N ≥ 10



Step 2 – Response rate

Response rate is often considered a key indicator of data quality, but most recommendations argue it is more important to reflect on the data generation process and appropriate research questions than take high response rate as an immediate positive (Holtom et al., 2022). To explore response rate in IWWS, we used linked information on organisation size for each employer. Response rate is calculated for each organisation by taking the cluster N as the numerator and the employer size as the denominator. We present results for upper and lower limits of response rate because *Indeed*'s website uses bracketed size categories (2-10, 11-50, etc.). For example, 11 responses in a range of 11-50 could either indicate a response rate of 100% or 22%. Due to the hypothesised nature of the sample, that it over-represents those who are job seekers, we did not expect response rates to be as high as targeted survey samples. Household surveys collecting data on labour achieve response rates of 50-70%, whereas internal organisation surveys typically achieve response rates of 30-50% (Fulton, 2018; Holtom et al., 2022). Our expectations were confirmed by Figure 5 which shows the response rates for each firm. The distribution of response rates is highly concentrated at less than 1%. The mean response rate ranges from 3.98% to 14.27% for lower and upper limits respectively, and the median ranges from 2.44% to 6.47%.

FIGURE 5: IWWS EMPLOYER-LEVEL DISTRIBUTIONS FOR CLUSTER SIZE N ≥ 10



Step 3 – Comparing with probability samples

Step 1 presents the overall distributions of the variables in IWWS, indicating biased results against a theoretically normal distribution. To better understand the extent of the bias, these distributions can be compared with random sample survey data that is more representative of the total working population. We compared four survey items. That we only able to identify comparative questions for these four survey items across labour surveys of the same time period highlights limitations of existing data sources for exploring work wellbeing from a subjective perspective.

Figure 6a compares job satisfaction responses in IWWS with Understanding Society (UKHLS) for the same multi-year time period. Job satisfaction in the UKHLS is left skewed, and bell-shaped, a completely different shape to IWWS job satisfaction. Presented in Figure 6b similar large discrepancies are found when comparing with the *European Working Conditions Survey* (EWCS) for comparable questions on energy levels at work, purposeful and useful work and manager support. Responses for useful work and manager support in EWCS are left-skewed.

From these basic comparisons it is clear there are limitations for the representativeness of the sample, threatening the external validity at the individual level. We address this in multiple ways in the following steps. First, we identify possible sources of this bias by identifying groups of respondents based on response patterns and survey engagement. We inspect whether the inclusion and exclusion of these affects the overall distributions. We also identify whether this bias extends to the internal validity of the sample.

As, realistically, the magnitude of this sampling bias cannot be corrected for with standard correction methods (e.g. Heckman correction, weighting), we consider to what extent these biases matter for different research questions. We examine whether these biases are randomly spread across the sample or specific to certain organisations, industries or regions.

FIGURE 6A: PROBABILITY SAMPLE ITEM COMPARISONS

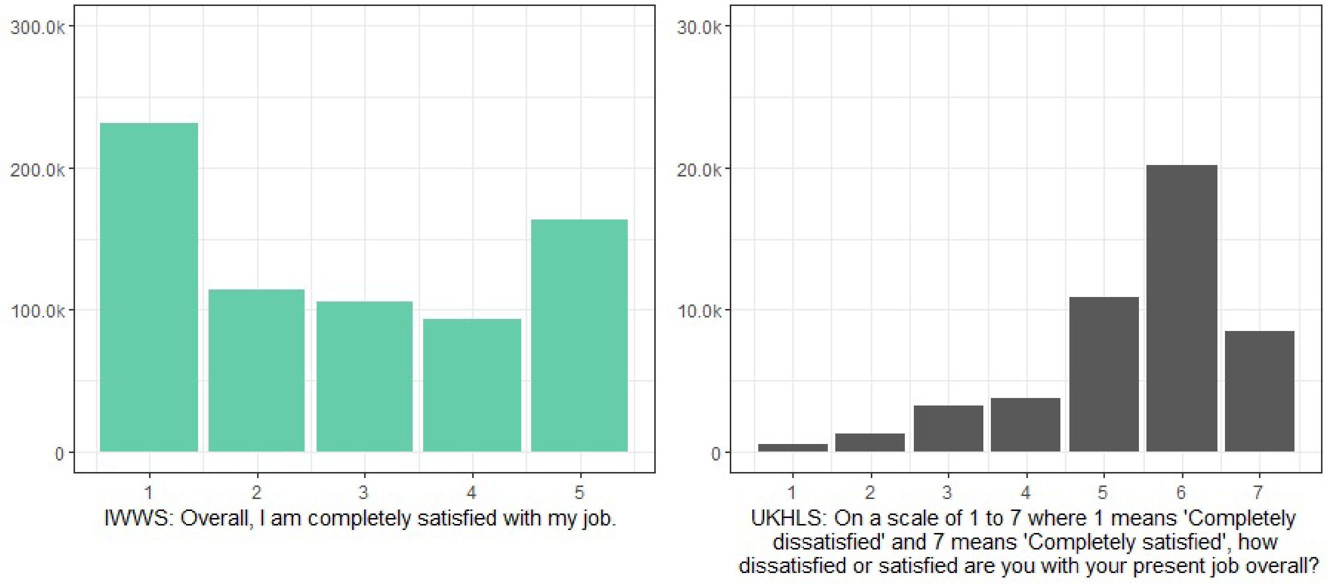
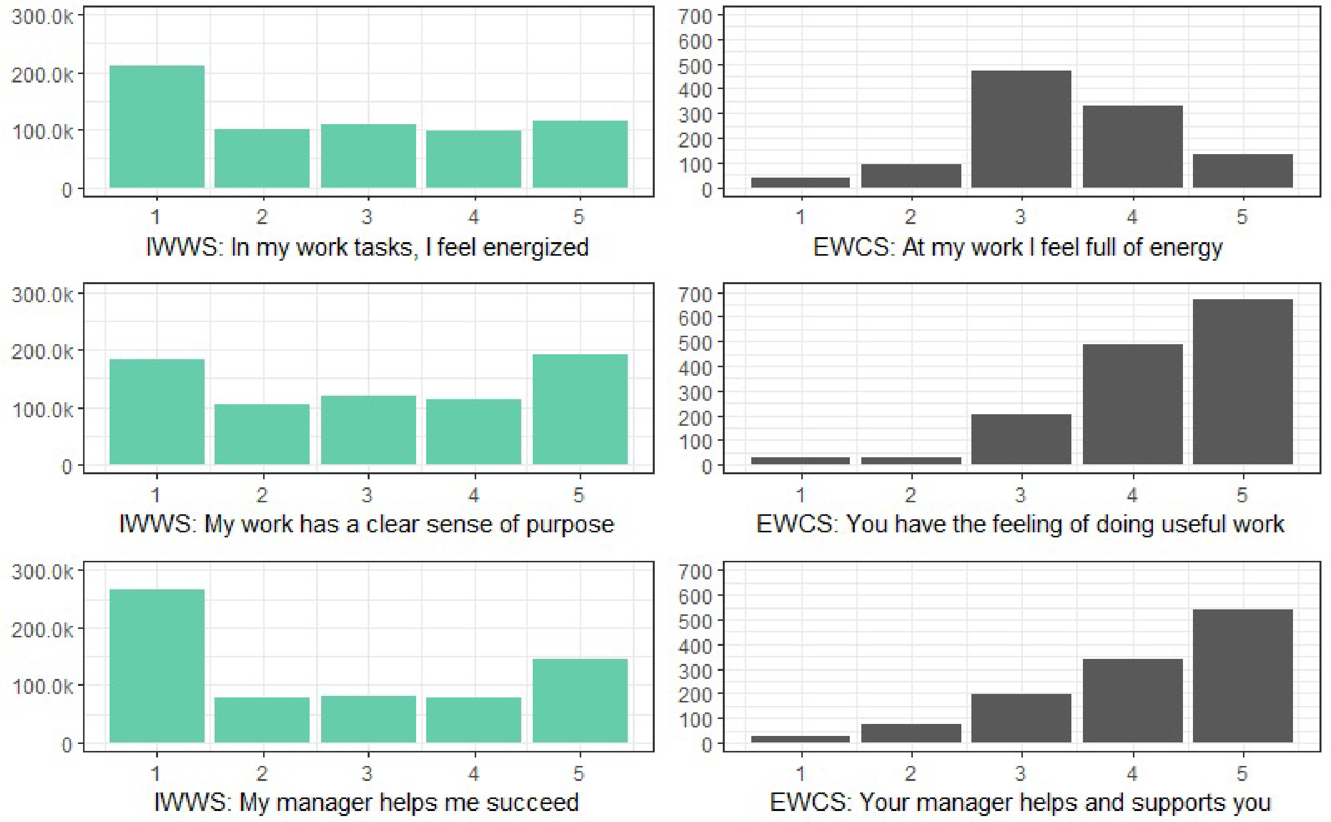


FIGURE 6B: PROBABILITY SAMPLE COMPARISONS



Step 4 – Possible sources of bias

Having established that the distributions are biased at the individual-level when compared with random sample surveys, we moved on to explore possible explanations for this bias that could be empirically accounted for. Using given survey items as well as derived variables from metadata, we identified different categories of respondents. Table 6 shows the counts and average values for each group. Specifically, we looked at employment status, survey arrival route, response time, response patterns and the reverse ordered stress question as a quasi attention check. Despite initial optimism about these measures as items for controlling selection bias, the high number of missing values is concerning for consistent application of these measures. For example, around half of responses do not have data for employment status or survey response route. Similarly, response time has a missing rate of over 90%, meaning it is not a useful item in this case for screening careless responses. This is less problematic for employment status, where the mean job satisfaction is similar for each group, but for the response route variable, there are large differences in mean job satisfaction depending on the access route, meaning missing values are obscuring heterogeneity within this category.

Figure 8 presents the job satisfaction distributions for each observable variable of possible bias. Results are presented as percentages of each category for comparability. Inspecting these distributions does not offer a comprehensive understanding of the source of bias. Surprisingly, the former and current employees distinction does not indicate any meaningfully different underlying distribution. However, some points do reveal notable patterns. Firstly, job satisfaction responses differ depending on access path to IWWS. Those completing surveys as part of a broader review process are more likely to select 1, whereas those completing the survey following an upload of their resume/CV show the reverse with a left skewed distribution.

Of the response patterns, over 65% of those only providing a single item respond with 2, indicating that 2 is the typical answer for the inattentive respondent. The distributions for survey completion rate reveals the same pattern. Whether the stress response is equal to above and below show over 50% reporting 1 for job satisfaction, notably more than for the whole sample. While respondents may be unsatisfied with their job and reporting very low stress, this does suggest the measure can be a potentially effective attention check and for identifying some of the selection bias at 1.

Figure 9 counts job satisfaction responses for subsets of the overall sample to detect whether the possible exclusion criteria suggested in Figure 7 affect these distributions. Overall, the distributions appear very similar, with the selection and binary bias remaining.

TABLE 6: DESCRIPTIVE COUNTS FOR RESPONSE PATTERNS

	Counts		Job satisfaction	
	N	%	Mean	(SD)
Employment status				
Current	395,335	27.01	2.70	1.53
Former	290,393	19.84	2.67	1.58
Missing	777,775	53.14	2.64	1.56
Survey response route				
Alert	427,423	29.42	2.85	1.63
Resume	11,303	0.78	3.77	1.29
Review	10,697	0.74	2.10	1.51
Other	108,119	7.44	2.68	1.73
Missing	429,698	61.62	2.47	1.40
Response time				
Reasonable (30-1200 sec)	109,765	7.50	3.07	1.67
Too short (< 30 sec)	97	<0.01	2.75	1.72
Too long (< 1,200 sec)	11,820	0.81	3.26	1.64
Missing	1,341,821	91.69	2.61	1.53
Response pattern				
Regular	793,119	54.19	2.78	1.57
Incomplete	476,263	32.54	2.18	0.95
Straightliner	193,713	13.23	2.54	1.57
Zig-zag (i.e. 1, 2, 3... or 5, 4, 3...)	408	0.03	1.24	0.94
Stress equal above and below				
Not equal	622,894	42.56	2.81	1.59
Equal	118,411	8.09	2.39	1.69
Missing (incomplete)	722,198	49.35	2.44	1.32

FIGURE 7: FLOW CHART OF POSSIBLE EXCLUSION CRITERIA

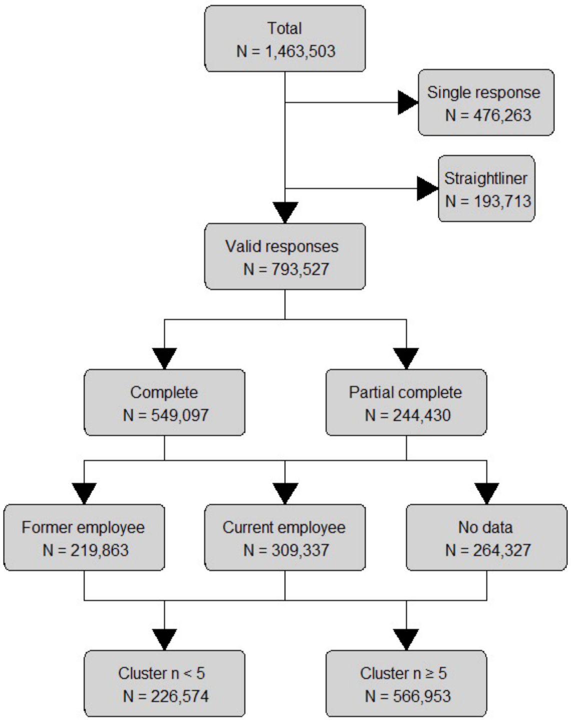


FIGURE 8: JOB SATISFACTION DISTRIBUTIONS AND POSSIBLE SOURCES OF BIAS

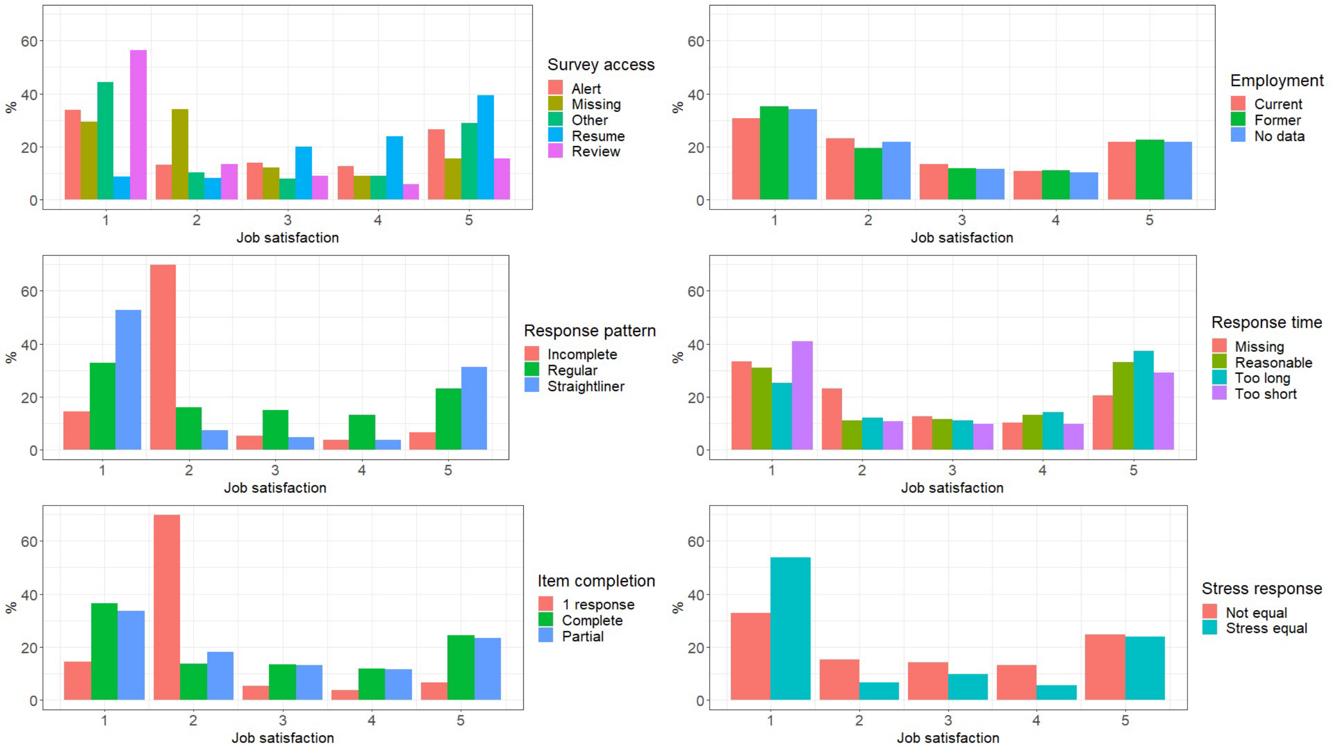
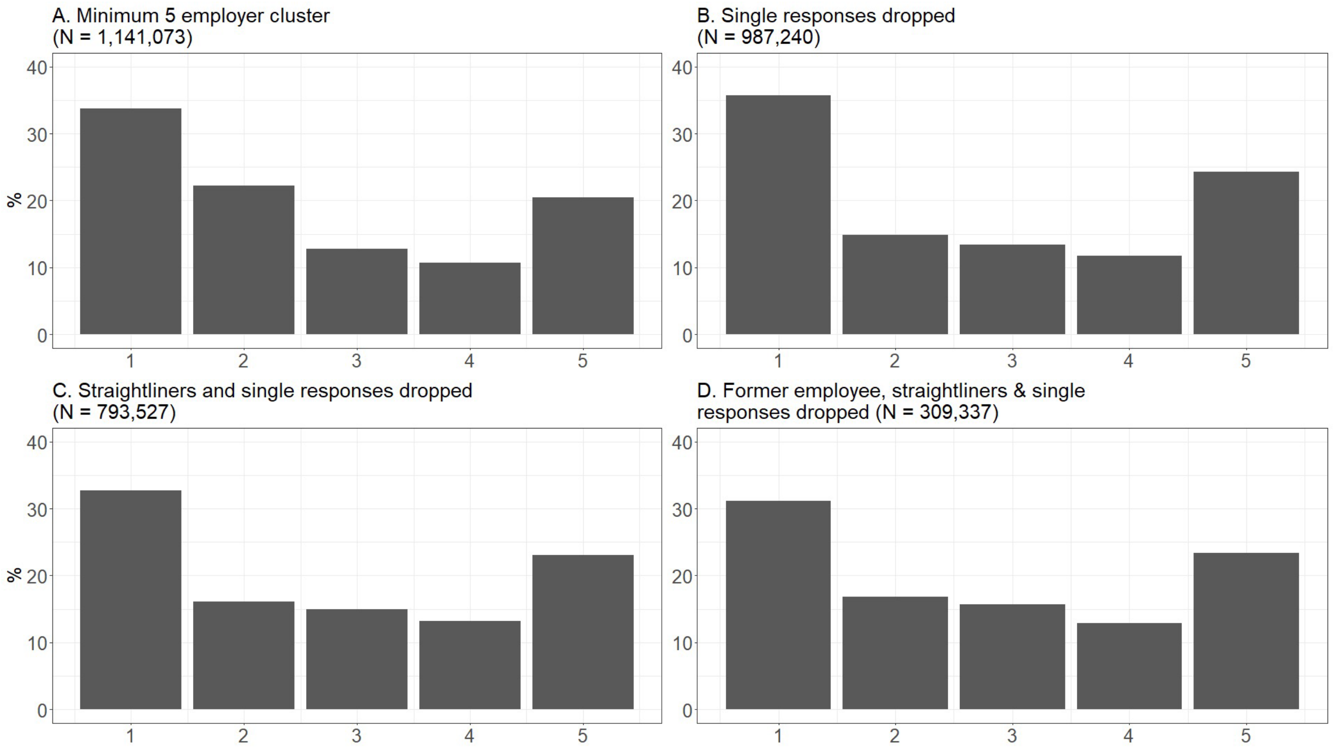


FIGURE 9: JOB SATISFACTION DISTRIBUTIONS FOR SAMPLE SUBSETS



Step 5 – Bias across levels

We considered what types of questions researchers, businesses, workers and policy-makers would be most interested in answering with a data source such as the IWWS. One key aim is the comparison of average levels between employers and across industries. This is primarily how *Indeed* currently makes use of the data, producing aggregate average work wellbeing scores for each employer and including this on public job profiles. As established in the prior steps, there are significant sampling biases in responses at the individual-level. While this undermines the ability to make generalised claims about trends in survey items, it does not automatically discount the ability to compare aggregate scores. If bias in response is consistent across unit levels of analysis, results are still comparable.

To establish whether possible sources of bias varied across levels of interest, we estimated a series of multi-level variance component models. We estimated 4 level models, incorporating random intercepts at the employer (level 2), industry (level 3) and region level (level 4) to identify variance at each. Table 7 shows the results of these models.

We estimated the model for several dependent variables: job satisfaction being equal to 1 and 5; respondent reviewing a former employer or not (current or missing); if response time was too long (greater than 1,200 seconds) or two short (less than 20 seconds); if responses were in a straight line on the survey page; if answers to the item measuring subjective stress was equal to the above and below questions; and whether the survey response was incomplete. We also used two measures of bias as dependent variables. A measure of binary bias for if responses were too long, too short, incomplete, straightlining, or an equal stress response. The measure of additive bias added 1 for each of these variables that were met.

Region explains negligible variance in all of the possible bias outcomes: all level 4 variance partitions were under 0.1%. Industry similarly explained minimal variance when incorporated as a random effect, providing explanation of less than 1% for all outcomes. Random effects at the employer-level offered greater variance explanation, but remained under 5% for all outcomes except the probability of job satisfaction being equal to 5. Careless response items, straightlining and Stress responses, were captured by less than 1% by employer-level variance. For the two most frequent responses for job satisfaction, 1 and 5, variance was explained by 3.7% and 7.5% respectively at the employer-level. However, these estimates likely indicate genuine variance in worker job satisfaction, rather than exclusively bias.

Overall, these results suggest that while the sample as a whole may suffer from bias along these response categories, results are consistently biased across employers, industries and geography. Such a finding gives confidence to comparisons at each level.

TABLE 7: VARIANCE PARTITION ESTIMATES FOR RANDOM INTERCEPT MODELS OF POSSIBLE SOURCES OF BIAS

Outcome	Level 2 (Employer)	Level 3 (Industry)	Level 4 (Region)
Job satisfaction = 1	0.0376	0.0055	0.0003
Job satisfaction = 5	0.0750	0.0072	0.0004
Employment status = former	0.0389	0.0026	0.0006
Response time = too long | too short	0.0469	0.0011	< 0.0001
Straightliner	0.0026	0.0015	< 0.0001
Stress equal to below and above	0.0075	0.0016	< 0.0001
Incomplete response	0.0496	0.0054	0.0003
Binary poor quality response	0.0308	0.0034	0.0004
Additive measure of poor quality	0.0159	0.0025	0.0003

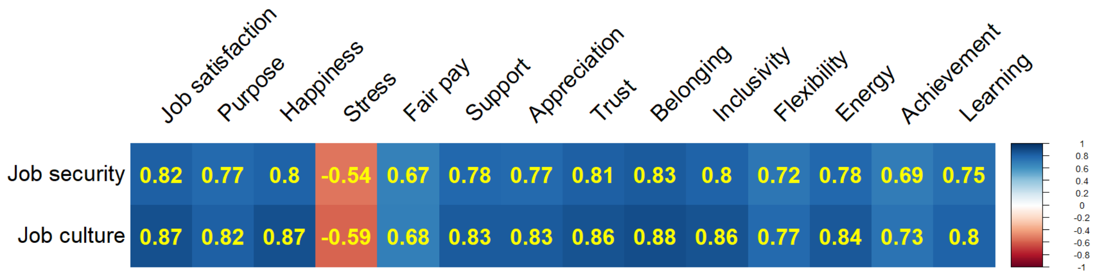
Step 6 – Internal consistency

The analytical steps so far were concerned with the distributions of important items, to what extent these are comparable to random samples and to what extent the sample is then generalisable. Another dimension that must be considered is the internal quality and consistency of the sample. In recent years there has been increased attention to response patterns and inattentive or careless responses, especially in online surveys (e.g. Ward & Meade, 2023). To explore the internal consistency of the survey items measuring subjective work wellbeing, Figure 11 is a correlation plot visualising the inter item Pearson correlations (*r*). Specific correlations also act as tests of convergent validity, the extent to which a measure correlates to another measure of the same construct. Results in this section are estimated from the sample subset with single item responses and straightliners removed.

As would be expected, estimates show consistent high positive correlations across all items, except the reverse-ordered stress item. The highest correlations are estimated for the first three 'positive' questions we have positioned as outcome measures of subjective work wellbeing – job satisfaction, work happiness and purposeful work. The overall consistent correlation supports the theoretical interrelatedness of the items. All estimates for correlations with stress are moderately negative, a result of the reverse scale on the item. That the r estimates for stress are less strong than the other correlation combinations is also theoretically consistent. The relationship between subjective stress and other dimensions of work wellbeing is not unidirectional. A job may be unsatisfying if it is extremely boring or undemanding, or stress may be experienced negatively if someone cannot manage that stress sustainably.

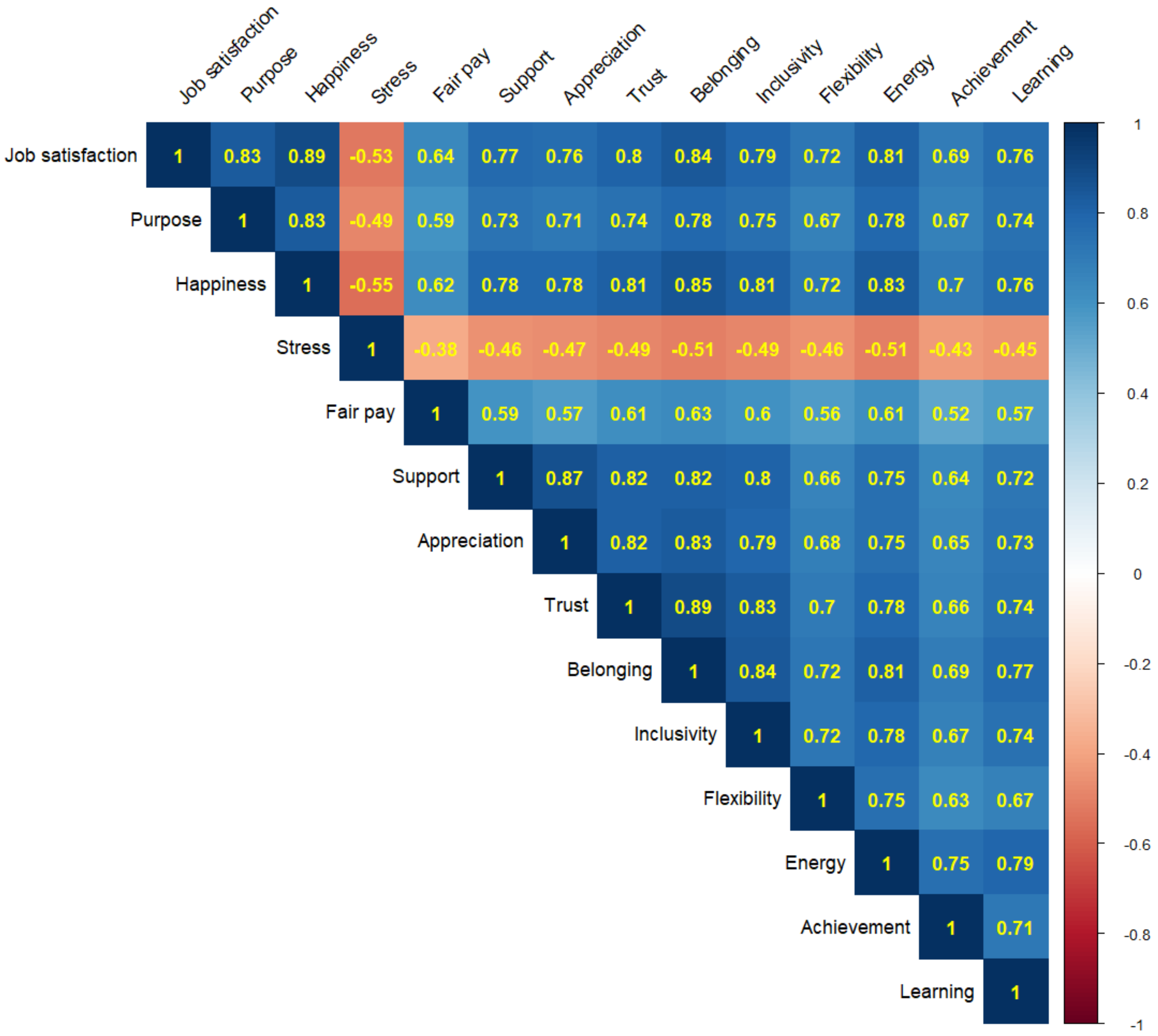
Next, Figure 10 shows correlations for the primary dimensions of subjective work wellbeing and the six employer rating items. This correlation matrix offers further assessment of convergent validity as well, by examining correlations between some of the variables and additional measures of the same constructs, such as comparing fair pay and the compensation item. Estimates are similar across the additional employer rating variables, with high positive correlation and especially for comparisons with the overall rating item.

FIGURE 10: INTER-ITEM CORRELATIONS IWWS RESPONSES AND EMPLOYER STAR RATINGS



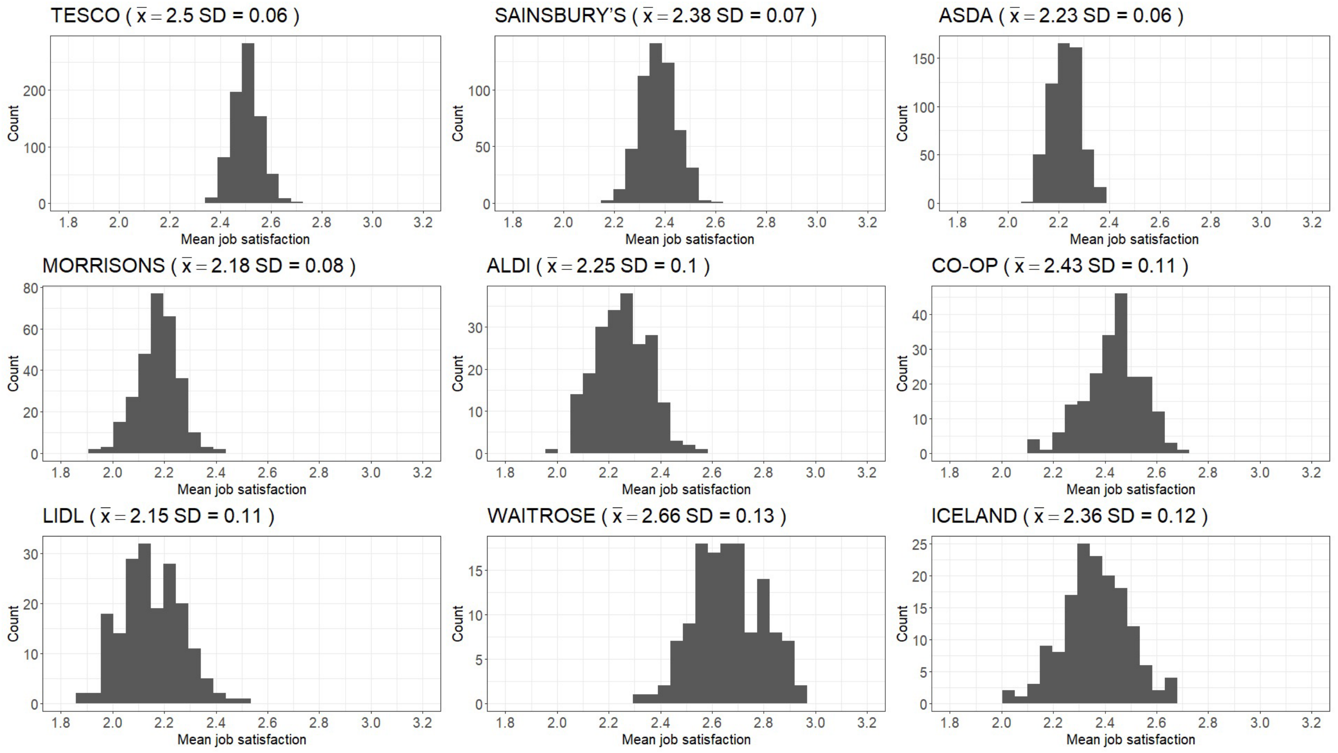
	Job satisfaction	Purpose	Happiness	Stress	Fair pay	Support	Appreciation	Trust	Belonging	Inclusivity	Flexibility	Energy	Achievement	Learning
Job security	0.82	0.77	0.8	-0.54	0.67	0.78	0.77	0.81	0.83	0.8	0.72	0.78	0.69	0.75
Job culture	0.87	0.82	0.87	-0.59	0.68	0.83	0.83	0.86	0.88	0.86	0.77	0.84	0.73	0.8

FIGURE 11: INTER-ITEM CORRELATIONS



	Job satisfaction	Purpose	Happiness	Stress	Fair pay	Support	Appreciation	Trust	Belonging	Inclusivity	Flexibility	Energy	Achievement	Learning
Job satisfaction	1	0.83	0.89	-0.53	0.64	0.77	0.76	0.8	0.84	0.79	0.72	0.81	0.69	0.76
Purpose		1	0.83	-0.49	0.59	0.73	0.71	0.74	0.78	0.75	0.67	0.78	0.67	0.74
Happiness			1	-0.55	0.62	0.78	0.78	0.81	0.85	0.81	0.72	0.83	0.7	0.76
Stress				1	-0.38	-0.46	-0.47	-0.49	-0.51	-0.49	-0.46	-0.51	-0.43	-0.45
Fair pay					1	0.59	0.57	0.61	0.63	0.6	0.56	0.61	0.52	0.57
Support						1	0.87	0.82	0.82	0.8	0.66	0.75	0.64	0.72
Appreciation							1	0.82	0.83	0.79	0.68	0.75	0.65	0.73
Trust								1	0.89	0.83	0.7	0.78	0.66	0.74
Belonging									1	0.84	0.72	0.81	0.69	0.77
Inclusivity										1	0.72	0.78	0.67	0.74
Flexibility											1	0.75	0.63	0.67
Energy												1	0.75	0.79
Achievement													1	0.71
Learning														1

FIGURE 12: BOOTSTRAPPED MEANS DISTRIBUTION FOR LARGEST SUPERMARKETS



TESCO ($\bar{x} = 2.5$ SD = 0.06)

SAINSBURY'S ($\bar{x} = 2.38$ SD = 0.07)

ASDA ($\bar{x} = 2.23$ SD = 0.06)

MORRISONS ($\bar{x} = 2.18$ SD = 0.08)

ALDI ($\bar{x} = 2.25$ SD = 0.1)

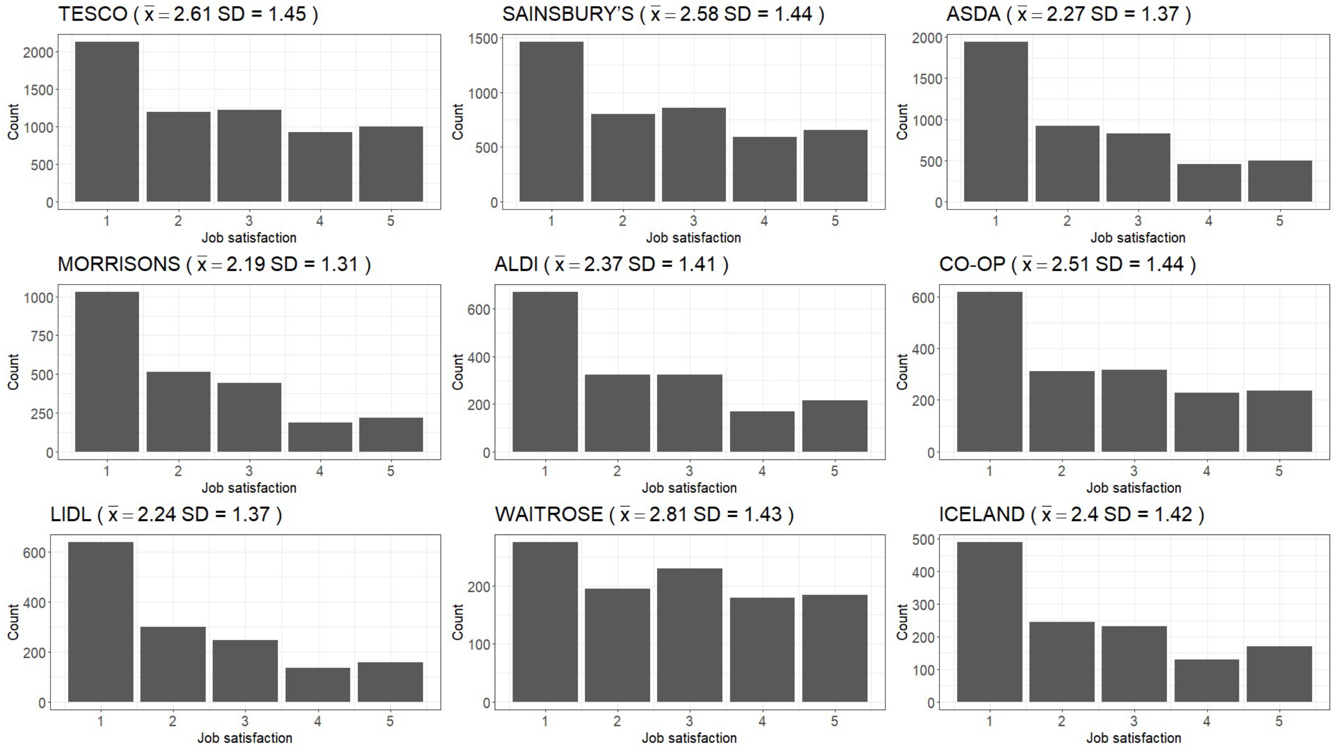
CO-OP ($\bar{x} = 2.43$ SD = 0.11)

LIDL ($\bar{x} = 2.15$ SD = 0.11)

WAITROSE ($\bar{x} = 2.66$ SD = 0.13)

ICELAND ($\bar{x} = 2.36$ SD = 0.12)

FIGURE 13: MOST FREQUENT EMPLOYER JOB SATISFACTION DISTRIBUTIONS



TESCO ($\bar{x} = 2.61$ SD = 1.45)

SAINSBURY'S ($\bar{x} = 2.58$ SD = 1.44)

ASDA ($\bar{x} = 2.27$ SD = 1.37)

MORRISONS ($\bar{x} = 2.19$ SD = 1.31)

ALDI ($\bar{x} = 2.37$ SD = 1.41)

CO-OP ($\bar{x} = 2.51$ SD = 1.44)

LIDL ($\bar{x} = 2.24$ SD = 1.37)

WAITROSE ($\bar{x} = 2.81$ SD = 1.43)

ICELAND ($\bar{x} = 2.4$ SD = 1.42)

Step 7 – Employer-level stability

Data collection for IWWS is 'always-on', in that the survey is always open with responses continually collected since 2020. One concern for data that is collected over time or in clusters is that measures are reliable in continuing to measure constructs. Reliability as stability is therefore the consistency of a measure over time (Drost, 2020: 106-108)

In this step, we used bootstrapped means as a novel method for assessing stability. We resampled with replacement subsets of 10 responses, plotting each subset mean to assess the stability of responses for individual employers. In Figure 12, we present this analysis for nine supermarkets. We chose supermarkets because three of the top five most frequent employers in the whole UK sample were supermarkets, and the whole figure can offer within industry comparisons of job satisfaction. The average wellbeing scores for each of these nine supermarkets can be publicly viewed on *Indeed*, and so we do not anonymise the names of each. For a note of caution, y-axis scales vary for each supermarket because of the different sizes of employer clusters. We also present Figure 13 which is the raw responses to job satisfaction for each supermarket, as a point of comparison to demonstrate the remaining underlying selection bias.

All nine supermarkets demonstrate reliability as stability through the consistency of bootstrapped estimates, with low variance (standard deviation < 0.15) and bell shaped distributions. The larger supermarkets (*Tesco*, *Sainsbury's* and *ASDA*) show greater stability, with lower standard deviation (0.6, 0.7 and 0.6 respectively) and the smaller supermarkets (*Lidl*, *Waitrose* and *Iceland*) demonstrate that as the sample decreases the distribution of the bootstrapped means is less consistent. Comparing the supermarket bootstrapped mean distributions can also provide comparison of the experiences of working at each employer. As these are well-known British employers, job seekers and consumers will have expectations and prejudices against specific supermarkets. For example, *Waitrose* and *Co-Op* operate employee models of partnerships and co-operatives, and as a result, average job satisfaction is higher than corporate and multinational supermarkets.

Reliability by stability acts as a cluster-level test-retest, administering the same test to the same cluster over time. Test-restest is redundant at the individual level. For IWWS, variance at the individual-level is natural and expected, and therefore any repeated measures would likely represent real changes in, for instance, job satisfaction. Additionally, we have already established that at the individual level the response distributions are biased. Testing for reliability through test-retest is redundant in this case at the individual-level. However, distributions at the employer level offer more promise for the value of IWWS.

Step 8 – Between employer comparisons

As demonstrated in the previous analysis step and the more satisfactory distributions of employer-level average values, one possible use of IWWS is in comparing the levels of survey items across employers. There is very little publicly available survey data with employer identifiers, a unique strength of IWWS. One exception is the annual UK Civil Service People Survey (CSPS). CSPS is an internal survey of the UK government workforce covering a range of measures of employee wellbeing. CSPS has a mean response rate of 77% across all departments. Aggregate values for the percentage that respond 'strongly agree' or 'agree' for each government department are publicly available. We compared these percentages with the same percentages responding 4 and 5 in IWWS, calculating Spearman rank and Pearson correlations for government departments based on six subjective work wellbeing items similar across the two surveys. This is a limited analysis due to the lack of coverage in IWWS of many civil service organisations as well as the lack of similarity between many of the items. Nevertheless, the analysis offers some external comparability for employer-level comparisons.

Table 8 reports these rank correlation estimates. These results compare the full IWWS sample (2020, 2021, 2022, 2023) with the CSPS 2022 wave. Using a minimum cluster size of 10 in IWWS, estimates compare ranks for 22 civil service departments. IWWS cluster n ranges from HMRC (n = 448) to Department of Education, OFGEM and The Insolvency Service (all n = 10).

Overall correlations are low to moderately positive. It is likely that the correlations are stronger for the all years IWWS correlations because of the larger cluster sizes offering more reliable comparisons. The pay (0.355 / 0.287) and flexibility (0.325; 0.356) items have the strongest positive correlation. The flexibility estimates are not a confident comparison because the CSPS item more closely resembles a measure of autonomy than one of flexibility. For all years of IWWS, manager (0.194), learning (0.209), trust (0.274) and inclusive (0.162) also showed low but positive rank correlations. In just the 2022 IWWS sample, the relationship for manager did not remain (-0.021). While the overall comparisons are not a perfect match in questions and IWWS offers a limited sample, the moderate positive associations in these ranks offers positive signs for comparisons within an industry even with small sample size. Considering the difference in sample coverage and the specific wording of questions, that these correlation estimates are even low to moderately positive is encouraging for the employer-level comparability of IWWS.

TABLE 8: RANK CORRELATIONS BETWEEN IWWS AND CIVIL SERVICE PEOPLE SURVEY 2022

IWWS item	CSPS item	Spearman rank	Pearson correlation
Paidfair	'I feel that my pay adequately reflects my performance'	0.355	0.287
Manager	'My manager motivates me to be more effective in my job'	0.198	0.225
Learning	'There are opportunities for me to develop my career in [my organisation]'	0.141	0.327
Trust	'I am trusted to carry out my job effectively'	-0.005	0.005
Flex	'I have a choice in deciding how I do my work'	0.325	0.356
Inclusive	'My organisation is committed to creating a diverse and inclusive workplace'	0.281	0.278

Step 9 – Comparing regression outputs

Research questions focussing on comparisons of average levels for companies, occupations or regions are one possible use of IWWS. Another type of research questions we may be interested in would be relationships between key variables in the survey. Interpreting what dimensions of working life are most important to wellbeing is a common type of analysis conducted by social scientists researching labour (e.g. Krekel et al., 2019; Warr, 2007). We estimated a series of OLS regression models for job satisfaction using IWWS to investigate multivariate associations across multiple dimensions of work wellbeing. We then compared the IWWS to a similar model using the *Skills and Employment Survey* (SES) 2017 (Felstead et al., 2019) after min-max normalising all variables.

Results offer some confidence in the predictive validity of IWWS, with coefficients similar across the models from the two surveys. Evaluations of pay have the same coefficient estimates (+0.145 in SES; +0.143 to +0.146 in IWWS), whereas evaluations of management, training and learning opportunities and colleague friendliness and support are similar ($0.103 < B < 0.17$). The differences between the abilities and achievement items and the values and belonging items is larger, but the fact these findings are derived from less similar survey questions and therefore perhaps not valid comparisons.

The IWWS models also have notably higher R^2 estimates (0.77 to 0.8 IWWS; 0.5 SES), indicating less variation in response and higher multicollinearity. Overall, results suggest that IWWS can offer comparable findings for these types of analysis to explore relationships, despite failing to meet the normality assumption for OLS regression. Considering the bias in IWWS, these comparisons may be better summarised as a comparison of the covariates of the workforce as a whole versus a sample of job-seekers.

These results are from a rudimentary analysis and are perhaps not surprising considering the normalised scales of the survey items. Nevertheless, overall they suggest that the structure of the job satisfaction equation is similar in IWWS to the SES.

TABLE 9: OLS REGRESSION RESULTS ON JOB SATISFACTION

SES		IWWS				
	Model 1		Full sample	No Straightliner	Current employee	2-level model
Values	0.093 (0.015)	Belonging	0.382 (0.001)	0.384 (0.001)	0.391 (0.002)	0.377 (0.003)
Pay	0.145 (0.012)	Paidfair	0.143 (0.001)	0.143 (0.001)	0.146 (0.001)	0.146 (<0.001)
Manager relations	0.170 (0.015)	Manager	0.132 (0.001)	0.134 (0.001)	0.127 (0.002)	0.128 (0.001)
Abilities	0.294 (0.017)	Achieve	0.106 (0.001)	0.105 (0.001)	0.107 (0.001)	0.105 (0.001)
Training	0.103 (0.015)	Learn	0.124 (0.001)	0.124 (0.001)	0.124 (0.002)	0.125 (0.001)
Friendliness	0.103 (0.019)	Support	0.123 (0.001)	0.121 (0.001)	0.112 (0.002)	0.125 (0.001)
N	2,767		636,132	569,879	264,290	484,038
R^2	0.50		0.80	0.77	0.77	-

Note: Standard errors in brackets. All coefficients significant at p < 0.001 ***. 2-level model uses random intercepts. Values = 'I find that my values and the organisation's values are very similar'. Pay = 'How satisfied or dissatisfied you are with that particular aspect of you present job: Your pay'. Manager relations = 'How satisfied or dissatisfied you are with that particular aspect of you present job: Relations with your supervisor or manager'. Abilities = 'Opportunities to use your abilities'. Training = 'How satisfied or dissatisfied you are with that particular aspect of you present job: The training provided'. Friendliness = 'How satisfied or dissatisfied you are with that particular aspect of you present job: The friendliness of the people you work with'.

Step 10 – Predictive validity

Researchers should weigh up the strengths and limitations of any data source and consider research questions that are appropriate, i.e. not undermined by any limitations. We've shown that comparisons of average levels can be made when employer sizes are large (e.g. supermarkets in Step 7), and that it is still possible but less effective when samples are small (civil service departments in Step 8). We've also shown that IWWS produces similar estimates for multivariate analysis to random probability samples, offering some confidence in the predictive validity of the data.

A further way to demonstrate predictive validity is to link to external or alternative data sources. Published elsewhere, we provide two further examples to support the predictive validity of IWWS. The first, De Neve et al. (2023), uses IWWS items averaged to the employer level, matching and comparing this aggregate data with available financial data for publicly traded firms in the US. This analysis focussed on whether employee wellbeing measures were predictive of organisational performance over time, tying to a longstanding literature on wellbeing and productivity. Overall the results were positive, showing average worker wellbeing correlated well with financial performance.

The second piece of work demonstrating predictive validity investigates whether job satisfaction responses in IWWS predict job applications on the *Indeed* job boards. Using data from IWWS, Ward (2024, forthcoming) shows a link between job seeker self-reported wellbeing and their subsequent search behaviour on the platform. For example, while those answering 1 to the satisfaction question apply, on average, to around 2.6 jobs in the following 24 hours, those answering 5 instead make, on average, around 1.5 applications during the same time window. Regression analyses that account for date, state, occupation, and industry fixed effects suggest that wellbeing is strongly predictive of job search intensity across all four wellbeing measures including satisfaction, purpose, stress, and satisfaction. Moreover, including company fixed effects in the analysis, he shows that the self-reported wellbeing of workers, even within the same company, predicts how many jobs they subsequently apply to. This provides a behavioural test of the data and is, in some sense, a revealed preference confirmation of the data, showing that unhappy workers are most keen to leave their existing jobs.

In another previous substantive study, we were able to establish the catalytic validity of IWWS. Ward (2023) shows that online users of *Indeed*'s job board use the public presentation of average work wellbeing scores for each employer to make labour market decisions. Ward shows that on average job applicants are willing to take 10.5% lower pay to work for an organisation with an 'above average' work wellbeing score as advertised on *Indeed*. By offering a sense of transparency regarding the experiences of working for specific employers, IWWS provides information for job seekers, a catalyst for affecting job market behaviour.

Seeking to answer the question of validity with substantive research questions is a risky strategy on its own. For example, can positive or negative findings be taken as genuine results and indication of the predictive validity of data; or are the limitations in the data producing spurious results? In this case we would only advocate validity-through-application if approached cautiously, and after multiple prior checks of data quality had been undertaken.

Discussion

Assessing *Indeed*'s Work Wellbeing Score

In this report we have worked through various stages to assess the data quality of IWWS. IWWS appears to be a highly valuable data source for the study of labour and organisations because it collects information on how individuals feel about their job and links that response with a named employer. This multilevel nature of the data is unique in its scale, and provides fruitful opportunities for analysing between employer, industry and geographic differences as well as linking with external organisation data (e.g. De Neve et al., 2023; G. Ward, 2024). However, as it is collected online we expected there to be higher numbers of poor quality or inattentive responses, and as it collected through a job market platform we expected widespread selection bias in the sample as the survey picked up job seekers. Our analytical steps followed general guidance in the methodological literature where possible, but primarily sought to think through the specific limitations of IWWS and the types of research questions we or other researchers may want to address.

Our expectation of sampling bias was confirmed by simple comparisons with random probability samples of the UK workforce. Rather than revealing a left-skewed bell-shape distribution of variables like job satisfaction, IWWS suffers from 'binary bias': respondents are highly likely to report the lowest response option, 1, followed by the highest, 5. The extent of sampling bias entails that we did not consider it an appropriate step to use correction or weighting techniques. These statistical tools have limited utility even in cases where sampling bias is not as pronounced as IWWS (e.g. Bethlehem, 2010; Dutz et al., 2021), so in a case where effectiveness cannot be assured this would be misguided.

We explored whether there were any observable predictors of the sampling biases, but in general distributions appeared similar for different subsets of data. Distributions and point averages were generally similar across possible sources of bias, such as employment status, response time and response patterns. The exception is for specific survey access routes, with higher quality responses from those who had responded following 'give-to-get' practices or when uploading their resume. This finding echoes analysis of *Glassdoor* employer review data (Marinescu et al., 2018). Unfortunately survey access cannot be used consistently as a control for IWWS because of the high prevalence of missing data in the metadata. It does, however, suggest that this mechanism should be leveraged as much as possible during data collection. Adopting more targeted approaches for data collection may be beneficial overall. Response enhancing strategies for IWWS may be best served by focussing on quality over quantity, especially considering the overall low response rate on an employer-by-employer basis.

A more positive result for the validity of IWWS is that in multiple possible sources of bias there was minimal variance across survey levels of employer, industry and region. As a consequence, we argue that while IWWS may suffer sampling bias, it is consistent. This means comparisons between groups can still be made. Such a position supports our next claim: that IWWS offers most promise at the employer level. Distributions of employer average values show close to normal distributions once a minimum employer cluster size of 10 is set. A proposed cluster size cut-off of 10 is reinforced by the inspection of bootstrapped means for supermarket competitors which showed stability in average values. Employer-level comparisons are also supported by the moderate positive correlations between the IWWS and CSPS rankings for civil service department.

If the aim is to provide a representative sample of the total global or national workforce, there is significant sampling bias in IWWS. However, as the sample is collected from users of a job site, evaluating IWWS as if it is a sample of the total workforce may have been a misintended endeavour to begin with. Instead, it may be better to understand IWWS as a sample of job seekers. Attempting to weight or mitigate for sampling bias would likely obscure interpretation of valid survey responses. We prefer to accept that IWWS has sampling bias and captures a non-random sample of job seekers. To reframe IWWS as a sample of job seekers, the prefix to all empirical claims alters: 'among job seekers, belonging has the largest regression estimate when predicting job satisfaction', or 'job seekers who have worked at supermarket A have higher average job satisfaction when compared to supermarket B'. Such a reframing would also flip a perceived limitation of IWWS into a potential strength. IWWS could be viewed to have oversampled the workers who would be of most theoretical interest to research on work and wellbeing. It has comparatively large numbers of the happiest and least happy workers. The happy workers offer insight for what makes for a good working life, and the unhappy workers can identify what makes work most unbearable.

While a decision to consider IWWS a sample of job seekers can reorientate assessments of quality, it would also curtail generalisable empirical claims about the workforce population. Instead, users of IWWS data must consider appropriate research questions. From our analytical steps, these questions may include exploring multivariate relationships through regression, as our findings in Step 9 indicate estimates were very similar to a probability sample survey of the total workforce. We would not recommend solely using IWWS for this question, but it does offer promise as a supplementary source of data for understanding the most important drivers of workers' wellbeing. Crucially, IWWS can incorporate multilevel variance into regression estimates, analysis that is impossible in other survey samples.

Overall, IWWS offers opportunities for studying job seekers and for novel types of analysis not possible in existing labour and organisation surveys. However, researchers must include the underlying data quality in the development of empirical research questions, consider the extent to which empirical and theoretical claims rely on the generalisability of wider working population.

General methodological recommendations

We began by reviewing much of the methodological guidance on assessing data quality across disciplines. Having applied this guidance where possible and adopted our own techniques to the IWWS data, we offer further recommendations.

Ultimately, data quality is not dichotomous in terms of good and bad, but is in degrees. Of course, an online convenience sample will not be as representative of a total workforce as a household panel survey collecting data face-to-face with a reliable sampling frame. However, considering the increasing social, logistical and financial constraints, it is not helpful to the ongoing development of survey research in the social sciences to discard efficient data collection methods and existing large data sources. However, we recommend that research use multiple data sources whenever possible. Every data source has its limitations, and conducting analysis with as many sources as possible will broaden the scope of any estimates produced, enhancing confidence by offering checks for reliability, validity and contextual and modular differences. The important decision for researchers to make is to ensure research questions and resulting knowledge claims are commensurate with the limitations of any data source.

Researchers can ensure higher quality responses by having as much control over the data generation process as possible, a strength over other emerging data sources such as administrative data. Doing so can allow for the inclusion of attention check questions and other survey design techniques for enhancing response quality, as outlined by Ward & Meade (2023) and Zickar & Keith (2023). This can also ensure that survey items are included that are already established, psychometrically validated and which have comparators. In IWWS we were not able to find matches for many questions, and even when matches to other surveys were available, responses had to be normalised for comparability. Adopting a multimodal approach from the beginning of the data collection phase will also allow for a constant point of comparison for survey items. These additional surveys may only act as pilot surveys to ensure the reliability of measures or be an ongoing operation.

Comparing online survey responses to other modes of data collection is just one necessary point of comparison. Survey design should include items that are measured in other surveys where the quality is long-established so that comparisons of frequency distributions can be made. Ideally this will include important demographic variables, which IWWS lacked, as this will provide knowledge on the types of survey respondents and which groups are over or underrepresented. Comparisons with administrative data such as from censuses will provide certainty in this regard.

Beyond comparisons with random probability samples, there are not specific analytical steps that must be implemented after data collection. Researchers must theorise, measure and identify the main sources of bias that can be observed. In the case presented here the major concern was selection into the survey by those more unsatisfied with their job and item response quality meaning this was our primary concern in analysis.

A marked strength of online surveys is the ability to collect metadata for both quality assessments and substantive analysis. Broad (i.e. non-identifying) location can be derived from IP address, for example, and in our case we were able to observe the respondent's region in the United Kingdom. However, survey response times were not captured for all respondents, with large quantities of missing values. As an important lesson, metadata can be effective controls but should not be solely relied upon due to the potential for missingness.

Finally, researchers must reflect on whether mitigation strategies are the correct approach. Increasingly scepticism is growing with regard to correction and weighting techniques when applied to online convenience surveys. In the case we present, correction methods appeared useless in face of the scale of the selection bias, and so we instead assessed whether this bias was consistent across observable variables.

Conclusion

Contributions

This report makes a number of contributions to the survey research methodology literature and to the field of labour and organisational research. First, we reviewed existing guidance for researchers when working with online convenience samples, highlighting that steps can be taken to assure, improve and mitigate for data quality before, during and after data collection. This review should be helpful for a wide social science audience. Second, we turned to applying as much of the guidance as possible to a prominent Big, online convenience sample, IWWS. Data from websites such as *Indeed* is gaining in popularity, but until now the underlying validity and quality of that data was uncertain.

Based on the assessments that we conducted for this report, we show that IWWS suffers from selection bias in individual-level responses. Such a finding is not surprising as users of *Indeed*'s website will more likely be dissatisfied job seekers. This limitation means that IWWS should not be used for generalisable claims about, for instance, the average job satisfaction of the British workforce. At the individual-level, any research would have to proceed with caution, reflecting on the underlying selection bias and the extent to which research questions and empirical claims rely on assured data quality and representativeness. However, IWWS is valuable in providing data clustered by employer. Potential bias is negligibly explained at multiple cluster levels, employer, sector and geography. This leads us to claim that IWWS suffers obvious sampling bias, but consistently so, and as a result can still be used for analysis at the employer level and for comparisons. Therefore it is a valuable relative measure of workforce wellbeing for organisations. Employer comparisons can ensure that IWWS is a valuable resource for both job seekers and employers. We also note the value of IWWS when the target population for research questions is job seekers.

In conclusion, we concur with many commentators in the survey methods literature that view multi-modal approaches to survey research as the future (e.g. Couper, 2017; Lehdonvirta et al., 2021). Administrative data offers promise for delivering full population coverage, and household surveys will continue to offer reliable data on a wide range of social variables. Online convenience samples will continue to offer a valuable tool for survey researchers by providing information not logistically or ethically easy to collect through these other methods. Online convenience samples will also continue to provide affordable and easy data collection with a range of interesting quality measures collected through the digital nature of the survey. As a final call, we do not foresee the growth of large online samples slowing. Instead, these data sources will become more popular and more complex, and we argue researchers must learn how to validate the quality of data sources and be transparent about the limitations of any data when making substantive claims.

Public value beyond research

We highlight several opportunities and limitations for academic research using IWWS, but our findings also illuminate several avenues for public use of work wellbeing data away from traditional research. When we evaluate the quality of data like IWWS, the question we as researchers ask is whether the data can be used for scientific exploration and whether resulting empirical claims will be valid. Yet social science is just one domain for data usage. In the introduction we cited examples of using labour force survey data in labour and banking policymaking. However, many of the investigations we detail are concerned with general empirical questions that would be pertinent for, especially, job seekers and organisations. Such uses of workforce wellbeing, while must still be based on reliable sources, do not require quite the same level of data quality as when making scientific claims. As a result, demands on the quality of data are different depending on needs and contexts.

We have discussed job seekers in Step 9, and the evidence for people making decisions regarding job applications between organisations in the same industry, they can, and do, use public IWWS averages to determine where to apply (Ward, 2022). We argue this shows the catalytic validity of IWWS, providing valuable labour market information for job seekers. IWWS and this assessment of its validity can assist job seekers as they navigate the job market and aim to support their own wellbeing in a changing economy. The public presentation of wellbeing data gives power to individuals.

There are many benefits the IWWS data can bring for employers committed to improving the wellbeing of their workforce. For these enlightened employers, IWWS and its public display on *Indeed* facilitates easy benchmarking with competitors or similar organisations. That we were able to establish consistent bias means that these types of comparisons for employers remain valid, as we demonstrate with UK supermarket chains in Step 8. Over time, organisations should be able to keep track of trends in work wellbeing among jobseekers within their own organisation and competitors. These public trends are drawn from jobseekers, but IWWS questionnaire items remain useful metrics to replicate internally in organisations for tracking trends and evaluating wellbeing related workplace interventions. IWWS and its ecosystem can act as a catalyst for improving wellbeing across the workforce.

References

Aguinis, H., Ramani, R. S., & Alabduljader, N. (2018). What You See Is What You Get? Enhancing Methodological Transparency in Management Research. *Academy of Management Annals, 12*(1), 83–110. https://doi.org/10.5465/annals.2016.0011

Aguinis, H., Villamor, I., & Ramani, R. S. (2021). MTurk Research: Review and Recommendations. *Journal of Management, 47*(4), 823–837. https://doi.org/10.1177/0149206320969787

Arias, V. B., Garrido, L. E., Jenaro, C., Martínez-Molina, A., & Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods, 52*(6), 2489–2505. https://doi.org/10.3758/s13428-020-01401-8

Baruch, Y., & Holtom, B. C. (2008). Survey response rate levels and trends in organizational research. *Human Relations, 61*(8), 1139–1160. https://doi.org/10.1177/0018726708094863

Behrend, T. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods, 43*, 800–813.

Belliveau, J., & Yakovenko, I. (2022). Evaluating and Improving the Quality of Survey Data From Panel and Crowd-Sourced Samples: A Practical Guide for Psychological Research. *Experimental and Clinical Psychopharmacology, 30*(4), 400–408.

Bethlehem, J. (2010). Selection Bias in Web Surveys. *International Statistical Review, 78*(2), 161–188.

Blasius, J., & Thiessen, V. (2012). *Assessing the Quality of Survey Data*. SAGE Publications. https://doi.org/10.4135/9781446251874

Blundell, R., Gosling, A., Ichimura, H., & Meghir, C. (2007). Changes in the Distribution of Male and Female Wages Accounting for Employment Composition Using Bounds. *Econometrica, 75*(2), 323–363.

Callegaro, M., Baker, R., Bethlehem, J., Göritz, A. S., Krosnick, J. A., & Lavrakas, P. J. (Eds.). (2014). *Online Panel Research: A Data Quality Perspective*. Wiley.

Carlson, K. D., & Herdman, A. O. (2012). Understanding the Impact of Convergent Validity on Research Results. *Organizational Research Methods, 15*(1), 17–32.

Celhay, P., Meyer, B. D., & Mittag, N. (2024). What leads to measurement errors? Evidence from reports of program participation in three surveys. *Journal of Econometrics, 238*(2), 105581. https://doi.org/10.1016/j.jeconom.2023.105581

Connelly, R., Playford, C. J., Gayle, V., & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research, 59*, 1–12. https://doi.org/10.1016/j.ssresearch.2016.04.015

Couper, M. P. (2017). New Developments in Survey Data Collection. *Annual Review of Sociology, 43*, 121–145.

Cronbach, L. J., & Meehl, P. E. (1956). Construct Validity in Psychological Tests. *Psychological Bulletin, 52*(4), 281–302.

Cycyota, C. S., & Harrison, D. A. (2006). What (Not) to Expect When Surveying Executives: A Meta-Analysis of Top Manager Response Rates and Techniques Over Time. *Organizational Research Methods, 9*(2), 133–160. https://doi.org/10.1177/1094428105280770

De Neve, J.-E., Kaats, M., & Ward, G. (2023). *Workplace Wellbeing and Firm Performance*.

De Neve, J.-E., & Ward, G. (2023). *Measuring Workplace Wellbeing* (2303). University of Oxford Wellbeing Research Centre Working Paper.

Deming, W. E. (1944). On Errors in Surveys. *American Sociological Review, 9*(4), 359–369.

DeSimone, J. A., DeSimone, A. J., Harms, P. D., & Wood, D. (2018). The Differential Impacts of Two Forms of Insufficient Effort Responding. *Applied Psychology, 67*(2), 309–338. https://doi.org/10.1111/apps.12117

Drost, E. A. (2020). Validity and reliability in social science research. *Education Research and Perspectives, 38*(1), 105–123. https://doi.org/10.3316/informit.491551710186460

Dunn, A. M., Heggestad, E. D., Shanock, L. R., & Theilgard, N. (2018). Intra-individual Response Variability as an Indicator of Insufficient Effort Responding: Comparison to Other Indicators and Relationships with Individual Differences. *Journal of Business and Psychology, 33*(1), 105–121. https://doi.org/10.1007/s10869-016-9479-0

Dutz, D., Huitfeldt, I., Lacouture, S., Mogstad, M., Torgovitsky, A., & van Dijk, W. (2021). *Selection in Surveys: Using Randomized Incentives to Detect and Account for Nonresponse bias*. NBER Working Papers.

Eichstaedt, J. C., & Weidman, A. C. (2020). Tracking Fluctuations in Psychological States Using Social Media Language: A Case Study of Weekly Emotion. *European Journal of Personality, 34*(5), 845–858. https://doi.org/10.1002/per.2261

Eurofound. (2021). *European Working Conditions Telephone Survey, 2021* (3rd Edition) [dataset]. UK Data Service. https://doi.org/10.5255/UKDA-SN-9026-3

Evans, J. R., & Mathur, A. (2018). The value of online surveys: A look back and a look ahead. *Internet Research, 28*(4), 854–887. https://doi.org/10.1108/IntR-03-2018-0089

Fauth, T., Hattrup, K., Mueller, K., & Roberts, B. (2013). Nonresponse in Employee Attitude Surveys: A Group-Level Analysis. *Journal of Business and Psychology, 28*(1), 1–16. https://doi.org/10.1007/s10869-012-9260-y

Felstead, A. (2021). Are online job quality quizzes of any value? Selecting questions, maximising quiz completions and estimating biases. *Employee Relations, 43*(3), 724–741.

Felstead, A., Gallie, D., Green, F., & Henseke, G. (2019). *Skills and Employment Survey, 2017* [dataset]. UK Data Service. https://doi.org/10.5255/UKDA-SN-8581-1

Fisher, G. G., Matthews, R. A., & Gibbons, A. M. (2016). Developing and investigating the use of single-item measures in organizational research. *Journal of Occupational Health Psychology, 21*(1), 3–23. https://doi.org/10.1037/a0039139

Fisher, M., Newman, G. E., & Dhar, R. (2018). Seeing Stars: How the Binary Bias Distorts the Interpretation of Customer Ratings. *Journal of Consumer Research, 45*(3), 471–489. https://doi.org/10.1093/jcr/ucy017

Forsythe, E., Kahn, L. B., Lange, F., & Wiczer, D. (2020). Labor demand in the time of COVID-19: Evidence from vacancy postings and UI claims. *Journal of Public Economics, 189*, 104238. https://doi.org/10.1016/j.jpubeco.2020.104238

Fulton, B. R. (2018). Organizations and Survey Research: Implementing Response Enhancing Strategies and Conducting Nonresponse Analyses. *Sociological Methods & Research, 47*(2), 240–276. https://doi.org/10.1177/0049124115626169

Gile, K. J., Johnston, L. G., & Salganik, M. J. (2015). Diagnostics for Respondent-driven Sampling. *Journal of the Royal Statistical Society. Series A, (Statistics in Society), 178*(1), 241–269. https://doi.org/10.1111/rssa.12059

Gosling, S. D., & Mason, W. (2015). Internet Research in Psychology. *Annual Review of Psychology, 66*, 877–902.

Griffin, M., Martino, R. J., LoSchiavo, C., Comer-Carruthers, C., Krause, K. D., Stults, C. B., & Halkitis, P. N. (2022). Ensuring survey research data integrity in the era of internet bots. *Quality & Quantity, 56*(4), 2841–2852. https://doi.org/10.1007/s11135-021-01252-1

Hanley, J. A. (2017). Correction of Selection Bias in Survey Data: Is the Statistical Cure Worse Than the Bias. *American Journal of Epidemiology, 185*(6), 409–411.

Hays, R. D., Liu, H., & Kapteyn, A. (2015). Use of Internet panels to conduct surveys. *Behavior Research Methods, 47*(3), 685–690. https://doi.org/10.3758/s13428-015-0617-9

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica, 47*(1), 153–161. https://doi.org/10.2307/1912352

Holtom, B., Baruch, Y., Aguinis, H., & Ballinger, G. A. (2022). Survey response rates: Trends and a validity assessment framework. *Human Relations, 75*(8), 1560–1584.

Kam, C. C. S., & Cheung, S. F. (2023). A Constrained Factor Mixture Model for Detecting Careless Responses that is Simple to Implement. *Organizational Research Methods*, 10944281231195298. https://doi.org/10.1177/10944281231195298

Krosnick, J. A., Presser, S., & Fealing, K. H. (2015). *The Future of Survey Research: Challenges and Opportunities*. National Science Foundation.

Landers, R. N., & Behrend, T. S. (2015). An Inconvenient Truth: Arbitrary Distinctions Between Organizational, Mechanical Turk, and Other Convenience Samples. *Industrial and Organizational Psychology, 8*(2), 142–164. https://doi.org/10.1017/iop.2015.13

Landers, R. N., Brusso, R. C., & Auer, E. M. (2019). Crowdsourcing Jo Satisfaction Data: Examining the Construct Validity of Glassdoor.com Ratings. *Personnel Assessment and Decisions, 5*(3), Article 6.

Landers, R. N., Brusso, R. C., Cavanaugh, K. J., & Collmus, A. B. (2016). A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. *Psychological Methods, 21*(4), 475–492. https://doi.org/10.1037/met0000081

Lee, D. S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *The Review of Economic Studies, 76*(3), 1071–1102. https://doi.org/10.1111/j.1467-937X.2009.00536.x

Lee, S. (2006). Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys. *Journal of Official Statistics, 22*(2), 329–349.

Lehdonvirta, V., Oksanen, A., Räsänen, P., & Blank, G. (2021). Social Media, Web, and Panel Surveys: Using Non-Probability Samples in Social and Policy Research. *Policy & Internet, 13*(1), 134–155. https://doi.org/10.1002/poi3.238

Liu, Y., Gelman, A., & Chen, Q. (2023). Inference from Nonrandom Samples Using Bayesian Machine Learning. *Journal of Survey Statistics and Methodology, 11*(2), 433–455. https://doi.org/10.1093/jssam/smab049

Londakova, K., Roy-Chowdury, V., Gesiarz, F., Burd, H., Hacohen, R., Mottershaw, A., Ter Meer, J., & Likki, T. (2021). *Encouraging employers to advertise jobs as flexible*. Behavioural Insights Team & Government Equalities Office.

Lovett, M., Bajaba, S., Lovett, M., & Simmering, M. J. (2018). Data Quality from Crowdsourced Surveys: A Mixed Method Inquiry into Perceptions of Amazon's Mechanical Turk Masters. *Applied Psychology, 67*(2), 339–366.

Lutz, J. (2015). The Validity of Crowdsourcing Data in Studying Anger and Aggressive Behavior: A Comparison of Online and Laboratory Data. *Social Psychology, 47*(1), 38–51.

Manski, C. F. (2016). Credible interval estimates for official statistics with survey nonresponse. *Journal of Econometrics, 191*(2), 293–301. https://doi.org/10.1016/j.jeconom.2015.12.002

Marinescu, I., Klein, N., Chamberlain, A., & Smart, M. (2018). *Incentives Can Reduce Bias in Online Reviews* (Working Paper 24372). National Bureau of Economic Research. https://doi.org/10.3386/w24372

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

Meyer, B. D., Mok, W. K. C., & Sullivan, J. X. (2015). Household Surveys in Crisis. *Journal of Economic Perspectives, 29*(4), 199–226. https://doi.org/10.1257/jep.29.4.199

Newman, D. A. (2014). Missing Data: Five Practical Guidelines. *Organizational Research Methods, 17*(4), 372–411. https://doi.org/10.1177/1094428114548590

Nohr, E. A., & Liew, Z. (2018). How to investigate and adjust for selection bias in cohort studies. *Acta Obstetricia et Gynecologica Scandinavica, 97*(4), 407–416.

ONS. (2023a). *Civil Service People Survey 2022: Technical Guide*. GOV.UK. https://www.gov.uk/government/publications/civil-service-people-survey-2022-results/civil-service-people-survey-2022-technical-guide

ONS. (2023b). *Labour market overview, UK: October 2023*. https://archive.ph/TEQng

Playford, C. J., Gayle, V., Connelly, R., & Gray, A. J. (2016). Administrative social science data: The challenge of reproducible research. *Big Data & Society, 3*(2), 2053951716684143. https://doi.org/10.1177/2053951716684143

Porter, C. O. L. H., Outlaw, R., Gale, J. P., & Cho, T. S. (2019). The Use of Online Panel Data in Management Research: A Review and Recommendations. *Journal of Management, 45*(1), 319–344.

Reuning, K., & Plutzer, E. (2020). Valid vs. Invalid Straightlining: The Complex Relationship Between Straightlining and Data Quality. *Survey Research Methods, 14*(5), Article 5. https://doi.org/10.18148/srm/2020.v14i5.7641

Rosenthal, R. (1965). The Volunteer Subject. *Human Relations, 18*(4), 389–406. https://doi.org/10.1177/001872676501800407

Sainju, B., Hartwell, C., & Edwards, J. (2021). Job satisfaction and employee turnover determinants in Fortune 50 companies: Insights from employee reviews from Indeed.com. *Decision Support Systems, 148*(113582).

Salganik, M. J. (2017). *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.

Schneider, D., & Harknett, K. (2022). What's to Like? Facebook as a Tool for Survey Data Collection. *Sociological Methods & Research, 51*(1), 108–140. https://doi.org/10.1177/0049124119882477

Schriesheim, C. A., Hinkin, T. R., & Podsakoff, P. M. (1991). Can ipsative and single-item measures produce erroneous results in field studies of French and Raven's (1959) five bases of power? An empirical investigation. *Journal of Applied Psychology, 76*(1), 106–114. https://doi.org/10.1037/0021-9010.76.1.106

Simsek, Z., & Veiga, J. F. (2000). The Electronic Survey Technique: An Integration and Assessment. *Organizational Research Methods, 3*(1), 93–117.

Sleeman, C. (2024). *Extracting Dimensions of Job Quality from Online Employee Reviews* (2024–01). ESCoE Discussion Paper.

Smith, T. W. (2013). Survey-Research Paradigms Old and New. *International Journal of Public Opinion Research, 25*(2), 218–229. https://doi.org/10.1093/ijpor/eds040

Spencer, N. H., Syrdal, D. S., Coates, M., & Huws, U. (2022). Assessing bias in online surveys using alternative survey modes. *Work Organisation, Labour & Globalisation, 16*(1), 34–51.

Stedman, R. C., Connelly, N. A., Heberlein, T. A., Decker, D. J., & Allred, S. B. (2019). The End of the (Research) World As We Know It? Understanding and Coping With Declining Response Rates to Mail Surveys. *Society & Natural Resources, 32*(10), 1139–1154. https://doi.org/10.1080/08941920.2019.1587127

Steinmann, I., Strietholt, R., & Braeken, J. (2022). A constrained factor mixture analysis model for consistent and inconsistent respondents to mixed-worded scales. *Psychological Methods, 27*(4), 667–702. https://doi.org/10.1037/met0000392

Suchman, E. A. (1962). An Analysis of 'Bias' in Survey Research. *The Public Opinion Quarterly, 26*(1), 102–111.

Suchman, E. A., & McCandless, B. (1940). Who answers questionnaires? *Journal of Applied Psychology, 24*(6), 758–769.

Sue, V. M., & Ritter, L. A. (2012). *Conducting Online Surveys*. SAGE Publications.

Suen, H.-Y., Hung, K.-E., & Tseng, F.-H. (2020). Employer Ratings through Crowdsourcing on Social Media: An Examination of U.S. Fortune 500 Companies. *Sustainability, 12*(16), 6308.

Thabane, L., Ma, J., Chu, R., Cheng, J., Ismaila, A., Rios, L. P., Robson, R., Thabane, M., Giangregorio, L., & Goldsmith, C. H. (2010). A tutorial on pilot studies: The what, why and how. *BMC Medical Research Methodology, 10*(1), 1. https://doi.org/10.1186/1471-2288-10-1

Tonidandel, S., King, E. B., & Cortina, J. M. (Eds.). (2015). *Big Data at Work: The Data Science Revolution and Organizational Psychology* (1st ed.). Routledge.

Tourangeau, R., Conrad, F. G., & Couper, M. P. (2013). *The Science of Web Surveys*. Oxford University Press.

University of Essex, Institute for Social and Economic Research. (2023). *Understanding Society* (11th Release) [dataset]. UK Data Service. https://doi.org/10.5255/UKDA-Series-2000053

Van Quaquebeke, N., Salem, M., van Dijke, M., & Wenzel, R. (2022). Conducting organizational survey and experimental research online: From convenient to ambitious in study designs, recruiting, and data quality. *Organizational Psychology Review, 12*(3), 268–305. https://doi.org/10.1177/20413866221097571

Vella, F. (1998). Estimating Models with Sample Selection Bias: A Survey. *The Journal of Human Resources, 33*(1), 127–169.

Walter, S. L., Seibert, S. E., Goering, D., & O'Boyle, E. H. (2019). A Tale of Two Sample Sources: Do Results from Online Panel Data and Conventional Data Converge? *Journal of Business and Psychology, 34*(4), 425–452. https://doi.org/10.1007/s10869-018-9552-y

Ward, G. (2022). *Workplace Happiness and Job Search Behavior: Evidence From a Field Experiment*.

Ward, G. (2024). *The Wellbeing of Workplaces: Organization-Level Differences Between and Within Industries*.

Ward, M. K., & Meade, A. W. (2018). Applying Social Psychology to Prevent Careless Responding during Online Surveys. *Applied Psychology, 67*(2), 231–263.

Ward, M. K., & Meade, A. W. (2023). Dealing with Careless Responding in Survey Data: Prevention, Identification, and Recommended Best Practices. *Annual Review of Psychology, 74*, 577–596.

Wenzel, R., & Van Quaquebeke, N. (2018). The Double-Edged Sword of Big Data in Organizational and Management Research: A Review of Opportunities and Risks. *Organizational Research Methods, 21*(3), 548–591. https://doi.org/10.1177/1094428117718627

Winton, B. G., & Sabol, M. A. (2022). A multi-group analysis of convenience samples: Free, cheap, friendly, and fancy sources. *International Journal of Social Research Methodology, 25*(6), 861–876. https://doi.org/10.1080/13645579.2021.1961187

Zhang, L. (2023). The Changing Role of Managers. *American Journal of Sociology, 129*(2), 439–484. https://doi.org/10.1086/727145

Zickar, M. J., & Keith, M. G. (2023). Innovations in Sampling: Improving the Appropriateness and Quality of Samples in Organizational Research. *Annual Review of Organizational Psychology and Organizational Behavior, 10*(1), 315–337. https://doi.org/10.1146/annurev-orgpsych-120920-052946

Nuffield Foundation

This project has been funded by the Nuffield Foundation, but the views expressed are those of the authors and not necessarily the Foundation.

www.nuffieldfoundation.org