WHITEPAPER

Removing Bias from a Hiring Survey for a Diverse Applicant Pool

Michelle Tiu
Bryan Matlen
Elli Suzuki

WestEd StellarEmploy

Introduction

StellarEmploy is on track to become the pre-eminent, highest-quality recruiting tool for employers of the 60M frontline, hourly workers that make up the US workforce. Every month, **thousands of people take our 15-minute survey as a first step in securing a job where they stay, thrive and grow**. The companies that use StellarEmploy as part of their hiring process regularly see a 20% decline in early turnover, which saves large employers millions of dollars in recruiting costs annually.

The StellarEmploy job survey learns about applicants' preferences for a work environment, and maps those preferences onto the unique characteristics of each job to determine which applicants are best fits for which jobs. Since all types of employers use our survey to hire, we prioritize a survey that is user-friendly for all types of job applicants.

The frontline, hourly workforce represents the diversity of this country: they are racially diverse; range in age from teenagers through decades past retirement; speak a variety of languages at home; and can have post-graduate degrees or may not have completed high school (*we recommend Brookings Institute's recent report, Meet the Low-Wage Workforce*). It's important that all these different groups understand the StellarEmploy job survey questions in the same way so that we can draw conclusions about each applicant uniformly.

StellarEmploy worked with educational research firm, WestEd, to review the best ways to design surveys to be reliable across all types of people. In this document, you will find descriptions of the ways that different types of people might change their answers to questions in order to "please" the survey provider; that one group of people might think a given quantity is "a lot" and another might think that same quantity is "a little"; and the recommended design tricks to avoid these -- and other -- pitfalls.

As AI becomes more prevalent in HR, **recruiters are right to be concerned about the risks of bias in poorly-designed recruiting automation**. This document describes one of the many efforts recruiters can make (and StellarEmploy makes) to avoid implementing biased HR technology.

WestEd is a nonpartisan, nonprofit research, development, and service agency that works with education and other communities throughout the United States and abroad to promote excellence, achieve equity, and improve learning for children, youth, and adults. WestEd has more than a dozen offices nationwide, from Massachusetts, Vermont, Georgia, and Washington, DC, to Arizona and California, with headquarters in San Francisco.

Overview of Literature Review Process

From September to October 2019, WestEd staff researched and drafted a literature review for StellarEmploy. WestEd collaborated with StellarEmploy to determine the focus areas for the literature review, as well as to identify any existing company resources or documents that could serve as the basis for the literature review.

The WestEd process entailed first identifying search strings based on the following focus areas:

- Development of equitable items for a Job Preferences Survey

The following search strings were then identified:

- *Universal design AND survey, universal design AND questionnaire, universal design AND survey research, universal design AND questionnaire research, universal design assessment AND survey, universal design assessment AND questionnaire, survey bias, response bias, biases in questionnaires, acquiescence bias, social desirability bias, response bias in survey AND determinants, acquiescence bias AND determinants, social desirability bias AND determinants, culturally sensitive survey, reporting heterogeneity*

The frontline, hourly workforce represents the diversity of this country... It's important that all these different groups understand the StellarEmploy job survey questions in the same way so that we can draw conclusions about each applicant uniformly.

A WestEd researcher spent approximately two days searching these key words in academic databases, such as EBSCO Host (includes APA PsycARTICLES and ERIC), the SAGE Premier Journal Collection, and Google Scholar. When selecting references, the following factors were taken into account:

- Data of publication: Priority was given to references published in the past 10 years.
- Quality of publication: Priority was given to peer-referenced articles.
- Quality of research: Priority was given to the most rigorous study types, such as randomized controlled trials, quasi-experimental designs, correlational designs, descriptive analysis, mixed methods, and literature reviews. Other considerations included the target population and sample, including their relevance to the question, generalizability, and general quality.

All key search terms and references were recorded to draw themes on the focus areas. The literature review on the identified focus areas is presented below.

Development of Equitable Items for a Job Preferences Survey

Surveys that target diverse populations must take into consideration cross-cultural implications. Within the same country and same language, survey respondents could have different answers to the same question, not because respondents had differing responses, but rather due to the way the question was understood (Fowler & Cosenza, 2008). Even if the respondents understood and interpreted the survey questions in the same way, respondents from different sub-groups may still answer the survey questions systematically differently depending on their individual-level factors, such as gender, socioeconomic status, race, age, and education level. This could create response bias, which ultimately affects the reliability and validity of the survey data. For instance:

- Respondents from different sub-groups may have different tendencies around their agreeableness. This tendency to agree irrespective of item content or direction is referred in the literature as “acquiescence bias.” Research shows that acquiescence tends to be more frequent among people with lower level of educational attainment (e.g.,Narayan & Krosnick, 1996,Rammstedt et al., 2010,Rammstedt & Kemper, 2011), older age(e.g.,Meisenberg & Williams-Shillingford, 2008, Weijters et al., 2010), and women (e.g., Weijters et al., 2010).
- Respondents from different sub-groups may also have different tendencies around how

they report their attitudes and behaviors depending on the social expectations and norms. This tendency to respond in a desirable way is referred in the literature as "socially desirable responding." This is often observed in questions that ask about sensitive topics where truthful responding poses an internal or external threat to the respondents. The magnitude of this threat can vary across sub-groups. Research shows that populations on the lower end of the socioeconomic spectrum perceive more topics as being sensitive because they have more to lose (Johnson & van de Vijver, 2003). Studies also show that even after controlling for education and income, African Americans and Mexican Americans reveal higher levels of socially desirable responding than non-Hispanic Whites (Warnecke et al, 1997).

- Respondents of different sub-groups may also systematically differ in their use of the response categories. For instance, one group's standards for what constitutes "a lot" may represent the same standards as what another group may consider "a little." Alternatively, when one group happens to have comparatively higher standards for what constitutes "strongly agree," they may report systematically lower levels of agreement than another group. This phenomenon is referred in the literature as "reporting heterogeneity" (e.g., Bago d'Uva et al., 2011). Evidence of cross-cultural difference in reporting health has been found across gender (e.g., Grol-Prokopczyk, 2014; Grol-Prokopczyk et al., 2011), socioeconomic status (e.g., Grol-Prokopczyk, 2014; Dowd & Zajacova, 2007), race/ethnicity (e.g., Grol-Prokopczyk, 2014, Menec et al., 2007; Shetterly et al. 1996), education level (Grol-Prokopczyk, 2014), and marital status (Grol-Prokopczyk, 2014). Evidence of cross-country difference in reporting job satisfaction has also been found (Kristensen & Johansson, 2008).

Best Practices

A review of the literature uncovered the following best practices that may be followed in order to develop equitable survey items:

- Anchoring vignettes have become an increasingly popular technique among survey researchers to adjust for reporting heterogeneity (e.g., King et al. 2004; King and Wand 2007). Anchoring vignettes are brief texts describing a hypothetical character or situation that exemplifies a certain level of the trait of interest. Respondents are asked to rate the vignette character's level of the trait using the same response categories that they would use to rate themselves. A growing body of literature points to how to optimize vignette wording and implementation. One study empirically demonstrated the importance of avoiding the use of anchoring vignettes that contain highly gendered connotations (Grol-Prokopczyk, 2014). Another study found that switching the question order so that

self-assessments follow the vignettes primes respondents to define the response scale in a common way (Hopkins & King, 2010).

- A scattered body of evidence suggests a few methods for minimizing response bias. Pew Research Center (n.d.) found that, in an experiment conducted in 1999, changing the format of the value- based questions from an agree-disagree format to a forced choice between two alternative statements not only yielded a different overall result, but also changed the pattern of answers among more-educated vs. less-educated respondents.
- There is a growing body of literature around best practices for writing effective survey questions, specifically around sensitive topics (Lensvelt-Mulders, 2008; Pew Research Center, n.d.). Some of the best practices include using simple and conversational tone rather than a formal register; including at least a few response options that indicate unfavorable attitudes or infrequent participation in a favorable activity; writing a question in the form of a short story to better elicit an honest response; and ordering the survey questions like a parabola, starting with simple, unthreatening and easy-to-answer questions, advancing to the more difficult and sensitive questions, and again ending with easy and friendly questions.
- There is a large body of literature around best practices for writing effective survey questions (Dillman et al., 2014; Rea & Parker, 2014; Fowler, 2014; Fowler & Cosenza, 2008; Pew Research Center, n.d.). General recommendations include ensuring clear and concise questions; providing definitions of key terms; avoiding unfamiliar, uncommon, complex, or technical words and phrases; avoiding imbedded assumptions about respondent's situations or their view on certain topics; and limiting each question to one idea. Some of these best practices are also reinforced by a growing body of literature on survey design for diverse and culturally complex populations (Goegan et al., 2018; Mertens, 2020).
- Biases in surveys are well-documented and widely-recognized. Choi & Park (2005) catalogues these biases and offer a checklist that helps survey researchers identify potential problems before pre-testing survey items and conducting cognitive labs. The checklist includes items related to question design such as problems with wording, leading questions, and intrusiveness. It also includes items related to questionnaire design such as formatting problems and flawed questionnaire structure. Finally, it includes items related to the administration of questionnaire such as cultural differences, respondent's subconscious reaction, and respondent's learning.

References

- Choi, B. C., & Pak, A. W. (2005). A catalog of biases in questionnaires. Preventing chronic disease, 2(1), A13.
- Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). Internet, phone, mail, and mixed-mode surveys: the tailored design method. Hoboken, NJ: Wiley.
- Dowd, J.B., & Zajacova, A. (2007). Does the predictive power of self-rated health for subsequent mortality risk vary by socioeconomic status in the US? International Journal of Epidemiology, 36(6), 1214– 1221. https://doi.org/10.1093/ije/dym214
- d'Uva, T. B., Lindeboom, M., O'Donnell, O., & van Doorslaer, E. (2011). Slipping Anchor? Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity. The Journal of human resources, 46(4), 875–906.
- Fowler, F. J. (2014). Survey research methods. London: Sage Publication.
- Fowler, F. J., & Cosenza, C. (2009). Writing effective questions. In International Handbook of Survey
- Methodology (pp. 136–160). New York, NY: Psychology Press.
- Goegan, L.D., Radil, A.I., & Daniels, L.M. (2018). Accessibility in Questionnaire Research: Integrating
- Universal Design to Increase the Participation of Individuals with Learning Disabilities. Grol-Prokopczyk H. (2014). Age and Sex Effects in Anchoring Vignette Studies: Methodological and
- Empirical Contributions. Survey research methods, 8(1), 1–17.
- Grol-Prokopczyk, H., Freese, J., & Hauser, R. M. (2011). Using Anchoring Vignettes to Assess Group Differences in General Self-Rated Health. Journal of Health and Social Behavior, 52(2), 246– 261. https://doi.org/10.1177/0022146510396713
- Hopkins, D., & King, G. (2010). Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability. The Public Opinion Quarterly, 74(2), 201-222. https://doi.org/10.1093/poq/nfq011
- Johnson, T. P., & Van de Vijver, F. J. R. (2003). Social desirability in cross-cultural research. In J. A. Harkness, F. J. R. van de Vijver, & P. Mohler (Eds.), Cross-cultural survey methods. Hoboken, NJ: Wiley.
- StellarEmploy: Literature Review
- King, G., Murray, C., Salomon, J., & Tandon, A. (2004). Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research. American Political Science Review, 98(1), 191- 207. doi:10.1017/S000305540400108X
- King, G., & Wand, J. (2007). Comparing Incomparable Survey Responses: Evaluating and Selecting Anchoring Vignettes. Political Analysis, 15(1), 46-66. doi:10.1093/pan/mpl011
- Kristensen, N & Johansson, E. (2008). New Evidence on Cross-Country Differences in Job Satisfaction Using Anchoring Vignettes. Labour Economics. 15. 96-117. 10.1016/j.labeco.2006.11.001.
- Lensvelt-Mulders, G. (2009). Surveying sensitive topics. InInternational Handbook of Survey Methodology (pp. 461–478). New York, NY: Psychology Press.

- Meisenberg, G. & Williams-Shillingford, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education?. Personality and Individual Differences. 44. 1539-1550. 10.1016/j.paid.2008.01.010.
- Menec, V. H., Shooshtari, S., & Lambert, P. (2007). Ethnic Differences in Self-Rated Health Among Older Adults: A Cross-Sectional and Longitudinal Analysis. Journal of Aging and Health, 19(1), 62– 86. https://doi.org/10.1177/0898264306296397
- Mertens, D. M. (2020). Research and evaluation in education and psychology: integrating diversity with quantitative, qualitative, and mixed methods. Thousand Oaks, CA: SAGE.
- Narayan, S. & Krosnick, J. A. (1996) Education Moderates Some Response Effects in Attitude Measurement, Public Opinion Quarterly, 60(1), 58–88. https://doi.org/10.1086/297739
- Pew Research Center (n.d.). Questionnaire design. Retrieved October 12, 2019, from https://www.pewresearch.org/methods/u-s-survey-research/questionnaire-design/.
- Rea, L. M., & Parker, R. A. (2014). Designing and conducting survey research: a comprehensive guide. San Francisco: Jossey-Bass.
- Rammstedt, B., Goldberg, L. R., & Borg, I. (2010). The measurement equivalence of Big Five factor markers for persons with different levels of education. Journal of research in personality, 44(4), 53–61. doi:10.1016/j.jrp.2009.10.005
- Rammstedt, B., & Kemper, C. J. (2011). Measurement equivalence of the Big Five: Shedding further light on potential causes of the educational bias. Journal of Research in Personality, 45(1), 121-125. http://dx.doi.org/10.1016/j.jrp.2010.11.006
- Shetterly, S. M., Baxter, J., Mason, L. D., & Hamman, R. F. (1996). Self-rated health among Hispanic vs non- Hispanic white adults: the San Luis Valley Health and Aging Study. American journal of public health, 86(12), 1798–1801. doi:10.2105/ajph.86.12.1798
- Warnecke, R.B., Johnson, T.P., Chavez, N.A., Sudman, S., O'rourke, D.P., Lacey, L., & Horm, J.W. (1997). Improving question wording in surveys of culturally diverse populations. Annals of epidemiology, 7 5, 334-42 .
- Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. Psychological Methods, 15(1), 96-110. http://dx.doi.org/10.1037/a0018721