USC Dornsife Center for Economic and Social Research

USC Schaeffer Leonard D. Schaeffer Center for Health Policy & Economics

Reversing the gender gap in happiness

Mallory Montgomery

Paper No: 2022-003

CESR-SCHAEFFER
WORKING PAPER SERIES

cesr.usc.edu **healthpolicy.usc.edu**

Journal of Economic Behavior and Organization 196 (2022) 65–78

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Reversing the gender gap in happiness

Mallory Montgomery[1]

Center for Economic and Social Research, University of Southern California, Los Angeles, CA, United States

ARTICLE INFO

Article history:
Received 10 January 2021
Revised 27 December 2021
Accepted 5 January 2022

JEL classification:
I310
J16

Keywords:
Gender gap
Life satisfaction
Happiness

ABSTRACT

Life satisfaction surveys are increasingly being used as a measure of welfare (Stiglitz et al., 2009), and even proposed as a primary measure (Layard, 2005). On average worldwide, surveys consistently find that women report higher life satisfaction than men. Yet, women are worse off in many ways: less education, lower incomes, worse self-reported health, and fewer opportunities. Why do they report higher life satisfaction? Using Gallup World Poll survey data from 102 countries including anchoring vignettes, I show that the gap is consistent with women and men systematically using different response scales, and that once these scales have been normalized, women appear *less* happy than men on average. I find that the effects of other characteristics commonly studied (income, education, marital status, etc.) are at least directionally the same after vignette adjustment, reinforcing previous findings.

1. Introduction

On a worldwide scale, women are granted fewer freedoms, get worse representation, experience more discrimination, and are more frequent victims of violence than men (UN Women and others, 2015). As shown in Table 1, they have, on average, worse self-reported health, lower incomes, and less education – all strong correlates of reduced life satisfaction.[2] Women are more likely to be diagnosed with anxiety or depressive disorders, and to feel fear, anxiety, and sadness (Nolen-Hoeksema and Rusting, 2003).[3]

Yet, women report higher life satisfaction than men; much higher. In the data presented here, all else equal, being female increases life satisfaction reports more than moving a decile higher in the income distribution, equivalent to thousands of dollars annually in the U.S. Although this pattern does not exist everywhere, it is true in much of the world, and of the world on average. Many authors have documented this gap, using data from the Gallup World Poll, the General Social Survey, the DDB Needham Life Style Survey, the World Values Survey, and the European Social Survey (Helliwell et al., 2015; Graham and Chattopadhyay, 2013; Easterlin, 2003; Zweig, 2015; Plagnol and Easterlin, 2008; Herbst, 2011; Arrosa and Gandelman,

E-mail address: m.montgomery@gmail.com

[1] Present: Amazon.com, Seattle, WA.

[2] Though there are some differences, I will follow much of the literature and use “life satisfaction,” “subjective well-being,” and “happiness” interchangeably; in all cases, I am referring to satisfaction with one’s life as a whole. In this study, the life satisfaction measure is Cantril’s ladder, defined in Section 2.

[3] This is also true in the present data: women report all negative day-to-day emotions significantly more than men, including stress, worry, anger, and physical pain. On the other hand, other authors have found that men suffer more often from externalizing psychological disorders, such as antisocial personality disorder and addiction (Nolen-Hoeksema and Rusting, 2003).

https://doi.org/10.1016/j.jebo.2022.01.006

Table 1
Summary statistics.

	Overall	Female	Male	Difference	Std. Err.
Equivalized household income	6,212.04	5,673.21	6,763.58	-1,090.37***	113.47
Primary education or less	42.0%	45.0%	39.0%	6.0%***	0.6%
Secondary education	46.8%	44.7%	49.0%	-4.3%***	0.6%
Tertiary education & up	11.2%	10.3%	12.0%	-1.7%***	0.3%
Single/never married	29.8%	25.8%	33.8%	-8.0%***	0.5%
Married/domestic partner	60.3%	60.3%	60.3%	0.0%	0.6%
Separated/divorced/widowed	9.9%	13.8%	5.9%	8.0%***	0.3%
Health problems	24.1%	26.2%	22.0%	4.2%***	0.5%
Unemployed	7.0%	7.0%	7.1%	-0.1%	0.3%
N	45,332	24,545	20,787	3,758	-
Life satisfaction self-report	2.945	2.957	2.932	0.025**	0.013
Vignette A1 rating	2.525	2.542	2.508	0.034**	0.012
Vignette A2 rating	3.483	3.506	3.459	0.047***	0.013
Vignette A3 rating	2.324	2.345	2.303	0.042***	0.012
Vignette A4 rating	2.927	2.938	2.917	0.020	0.012
Vignette A5 rating	2.271	2.288	2.254	0.034***	0.012
Vignette A6 rating	3.573	3.595	3.551	0.044***	0.014

Table includes final sample only, using survey weights provided by Gallup (except N, which is the true count of the final sample). See Appendix Table A1 for a version including the same countries but before dropping observations. Life satisfaction self-reports and vignette ratings are recoded from 0–10 to a 1–5 scale. Equivalized household income is calculated by taking reported household income and weighting each household member according to OECD weights (1 for the first adult, 0.5 for each additional adult, and 0.3 for each child). See Section 2 for details. Significance: 1% (***), 5% (**), 10% (*).

2016).[4] This gap implies that men are in fact at a disadvantage compared with women – at least in terms of subjective well-being.

Simultaneously, a robust debate is occurring in the literature on the validity of life satisfaction measurements. Much of the established life satisfaction literature treats self-reports cardinally, by converting responses to numerical scales, comparing averages across groups. Recently, several authors have debated the merits of this approach, and discussed potential adjustments, such as using panel data with individual fixed effects to remove time-invariant factors (Bond and Lang, 2019; Chen et al., 2019; Kaiser and Vendrik, 2019). Anchoring vignettes are another way to understand how different individuals use response scales.[5]

Anchoring vignettes allow researchers to consider not only an individual's self-report, but also the scale she uses to report it, by asking her to rate hypothetical vignette subjects on the same scale as herself. By using the same vignettes for many people, we can see how rating scales systematically differ with individual characteristics such as gender, education level, or country. Anchoring vignettes have been used to study differing response scales in many applications, particularly health (Kapteyn et al., 2007; d'Uva et al., 2008; King et al., 2004; Grol-Prokopczyk et al., 2011; Molina, 2016), but also political efficacy (King et al., 2004), job satisfaction (Kristensen and Johansson, 2008), drinking behavior (Van Soest et al., 2011), and corruption (Grzymala-Busse, 2007). They are commonly used when self-reports are subjective, and an objective measure is unavailable or would be difficult to obtain (King et al., 2004). In life satisfaction, several existing studies have explored scale differences among European countries (Angelini et al., 2014), the relationship between income and happiness in different countries (Kapteyn et al., 2010), and how survey response scales vary with age (Angelini et al., 2012). By asking respondents to rate hypothetical vignette characters on the same scale as their self-reports, we "anchor" their scales, making their responses comparable across heterogeneous groups.

Why would men and women use the same response scale differently in their self-reports? There are two primary explanations: response bias and differing standards. Response bias implies that men and women would use different ratings to describe the same latent life satisfaction, with all circumstances identical, including their standards for what makes a good life. This could be due to social desirability bias if one gender feels greater pressure to report high satisfaction (Dalton and Ortegren, 2011), extreme response bias if one gender avoids the top/bottom of the scale (Brulé and Veenhoven, 2017), or gender differences in bias from interviewer characteristics (Davis et al., 2009). Differing standards imply that men and women have different expectations for their lives, so that the same objective circumstances provide more satisfaction to one gender than the other. Differing standards are difficult to measure, although there is evidence of adaptation to changing life circumstances manifested in life satisfaction self-reports, which implies that standards may change (Clark et al., 2008; Stevenson and Wolfers, 2009). Stutzer (2004) shows that, all else equal, higher aspirations reduce utility. Green et al. (2018) find that in the UK, women's job satisfaction *declined* between the early 1990s and the early 2010s until it *converged* with men's,

[4] Although it is a common finding, it is not universal. Batz and Tay (2018) report mixed results on whether there are gender differences.

[5] For an alternative approach to correcting scale use differences by asking about respondents' memories of their past life satisfaction, as well as a discussion of scale use bias, see Kaiser (2020).

and that the decline is better explained by women's changing evaluations of their jobs rather than the qualities of the jobs themselves. The anchoring vignettes in this analysis cannot distinguish between these two explanations; any systematic gender difference in vignette ratings, whether caused by response bias or differing standards, are treated in the same way.

If the cause is response bias, then vignette adjustment appropriately works to even out those differences, making responses comparable across groups. On the other hand, if the difference is due to differing standards, the magnitude of the vignette adjustment is an indicator of the standards difference. Either way, in the case of gender, promoting policies that improve the less fortunate (as measured by subjective life satisfaction self-reports rather than objective measures) means favoring men. To the extent that we dislike such purely utilitarian decisions, we should focus policy on improving objective measures, such as ensuring women's equal treatment and access to opportunity – even if, in the case of differing standards, this reduces women's reported happiness.

This paper proceeds as follows. Section 2 describes my data. Section 3 establishes the gender life satisfaction reporting gap and its magnitude using a baseline ordered probit model. Section 4 describes intuitively and then formally the vignette-adjustment model, while Section 5 estimates the vignette-adjusted model and compares it with the baseline (unadjusted) model. Finally, Section 6 discusses the results and Section 7 concludes.

2. Data[6]

Gallup collected this data as part of its World Poll, in the years 2011-2014. It includes a subset of all World Poll waves conducted during this period, one wave from each of 109 countries. All country samples were taken within a single calendar year: respondents in 18 countries were interviewed in 2011, 39 in 2012, 26 in 2013, and 26 in 2014. Most countries have approximately 1,000 observations, though some large countries have up more (up to 5,000 in India) and Haiti has fewer (500), for a total of about 120,000 observations. As part of a National Institute on Aging supported project,[7] Gallup added a module on the international comparison of well-being to this round of surveys. For additional dataset details, see Appendix A.1 Section, or the Gallup World Poll Codebook (Gallup, 2019).

Most country samples are nationally representative of the population aged 15+, with exceptions mostly in areas where safety of interviewing staff is threatened, scarcely populated islands, and areas interviewers can reach only by foot, animal, or small boat.[8] See the appendix for more details, including countries with non-representative sampling and explanations. Running the analysis without these sampling-compromised countries adds selection bias and reduces power, therefore, I primarily report the results including these countries, and show similar results excluding them in the appendix. All included analyses use survey weights provided by Gallup.

Gallup has created a worldwide data set with standardized income and education data. To make education comparable across countries, education descriptions are recoded into one of three relevant categories: "Elementary": completed elementary education or less (up to eight years of basic education); "Secondary": completed some education beyond elementary education (9–15 years of education); "Tertiary": completed four years of education beyond "high school" and/or received a four-year college degree. To ensure income is comparable across countries, Gallup takes the income provided by the respondent and calculates annual household income in international dollars, using the Individual Consumption Expenditure corrected for the Household PPP ratio from the World Bank. These PPP-corrected values correlate strongly ($r=0.94$) with the World Bank estimate of per-capita GDP (PPP-corrected). The result is a household income measure that is comparable across all respondents, countries, and local and global regions. To make small coefficients on age more readable, I divide age by 10 and age-squared by 100 in regression tables.

I calculate log equivalized household income based on reported income, by taking reported household income and weighting each household member according to OECD weights (1 for the first adult, 0.5 for each additional adult, and 0.3 for each child). In the regressions to follow, this may overstate purchasing power of women, because income is reported at the household level.

My primary interest is in the answers to the following question (the so-called Cantril ladder):

> *"Please imagine a ladder with steps numbered from zero at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?"*

After providing a self-report, subjects were asked to rate the life satisfaction of each person in a set of six vignettes. The interviewer randomly asked about one of two possible vignette sets, set A or set B. Although respondents used a 0–10 scale, because there are relatively few responses at the top and particularly the bottom of the scale, I combine ratings into a five-point scale for both self-reports and vignette ratings: ratings 0–2 are recoded as 1, ratings 3 and 4 are recoded as 2, ratings 5 and 6 are recoded as 3, ratings 7 and 8 are recoded as 4, and ratings 9 and 10 are recoded as 5.

As discussed in Section 4, the B set of vignettes fails tests for one of the crucial assumptions of vignette adjustment models: vignette equivalence. I restrict my analysis to respondents who answered the A set, dropping about half of each

[6] Much of this description is also in another (coauthored) paper of mine using the same dataset, not yet published, "Life Satisfaction Within and Across Countries: Societal Capital and Relative Income."

[7] Awarded to Arie Kapteyn, James P. Smith, and Arthur van Soest.

[8] http://www.gallup.com/services/177797/country-data-set-details.aspx

Table 2
Ordered probit regression.

	(1)	(2)
Female	0.034**	0.085***
	(0.017)	(0.018)
Age/10	-0.134***	-0.141***
	(0.018)	(0.022)
Age^2/100	0.011***	0.018***
	(0.003)	(0.003)
Urban		0.039*
		(0.021)
Married		0.020
		(0.024)
Separated		-0.123***
		(0.047)
Divorced		-0.158***
		(0.039)
Widowed		-0.135***
		(0.040)
Domestic partner		-0.073**
		(0.033)
Employed full time for self		0.024
		(0.023)
Employed part time do not want full time		0.082**
		(0.032)
Unemployed		-0.178***
		(0.038)
Employed part time want full time		-0.036
		(0.036)
Out of workforce		0.025
		(0.029)
Secondary education		0.146***
		(0.021)
Tertiary education		0.279***
		(0.027)
Health problems		-0.256***
		(0.029)
Log equivalized household income		0.237***
		(0.014)
Observations	45,332	45,332

Ordered probit regression of life satisfaction on listed variables as well as country fixed effects (102 countries included). Standard errors are clustered at the country level. Significance: 1% (***), 5% (**), 10% (*).

country's observations. Of those, approximately 200 observations are dropped because respondents did not answer at least two vignette questions.[9] I drop about 15% of the sample (18,956 observations) due to missing household income, the most commonly missing variable. I drop all of Georgia, Singapore, and Ecuador because they are missing employment responses, and Egypt and Lebanon because they are missing health responses.

After dropping observations that are missing an essential variable, and those that did not answer vignette set A, I end up with 45,332 observations from 102 countries. Summary statistics for those respondents are given in Table 1.[10] This is the sample used in all analyses in the main text. For context, the standard deviation of the life satisfaction self-report for this sample is 1.09.

3. Baseline model (ordered probit)

I begin with a standard model, ordered probit. Table 2 shows the results of ordered probit regressions of life satisfaction on personal characteristics: column 1 includes indicators for female, age, and age squared, while column 2 adds indicators for urban, marital status, employment status, education level, and whether they have health problems, as well as a continuous measure, log of equivalized income.[11] The reference categories are single, employed full time for an employer, and low

[9] Results are similar if I only use respondents that answered all six vignette questions, although it reduces the sample by about 1000 observations.

[10] See Appendix Table A1 for a version including the same countries but before dropping observations, including respondents who answered vignette set B.

[11] The health measure is the yes/no answer to the question, "Do you have any health problems that prevent you from doing any of the things people your age normally can do?" Higher order polynomial terms for age were tested and found insignificant.

Table 3
Ordered probit marginal effects.

Characteristic	Comparison	LS	Change	x(gender eff.)
Income at 10th percentile	-	2.750	-	-
Income at 20th percentile***	10th percentile	2.832	0.082	1.09
Income at 30th percentile***	20th percentile	2.888	0.056	0.74
Income at 40th percentile***	30th percentile	2.935	0.047	0.63
Income at 50th percentile***	40th percentile	2.979	0.044	0.59
Income at 60th percentile***	50th percentile	3.021	0.042	0.56
Income at 70th percentile***	60th percentile	3.067	0.046	0.61
Income at 80th percentile***	70th percentile	3.120	0.053	0.71
Income at 90th percentile***	80th percentile	3.194	0.074	0.99
Low education	-	2.857	-	-
Medium education***	Low	2.986	0.129	1.72
High education***	Low	3.103	0.246	3.29
No health problems	-	2.999	-	-
Health problems***	No health problems	2.774	-0.225	-3.00
Single	-	2.951	-	-
Married	Single	2.968	0.017	0.23
Separated***	Single	2.843	-0.108	-1.44
Divorced***	Single	2.812	-0.138	-1.85
Widowed***	Single	2.832	-0.118	-1.58
Domestic partner**	Single	2.887	-0.064	-0.85
Employed full time for employer	-	2.941	-	-
Employed full time for self	Employed full time for emp.	2.962	0.021	0.28
Employed part time, don't want full time**	Employed full time for emp.	3.013	0.072	0.96
Unemployed***	Employed full time for emp.	2.785	-0.156	-2.08
Employed part time, want full time	Employed full time for emp.	2.910	-0.031	-0.42
Out of the workforce	Employed full time for emp.	2.963	0.022	0.29
Rural	-	2.932	-	-
Urban*	Rural	2.966	0.034	0.46
Male	-	2.907	-	-
Female***	Male	2.982	0.075	1.00

Full sample (N = 45,332). Marginal effects of changing various characteristics, based on ordered probit regression coefficients in Table 2, column 2; stars indicate significance from that column. Income percentiles are from respondent's own country. *LS* shows estimated life satisfaction with that characteristic, all other characteristics held constant. *x(gender eff.)* compares the magnitude of that marginal effect with the gender effect in the last row.

education (primary or less). Both columns include country fixed effects, and standard errors are clustered at the country level.

The coefficients in both columns fit well with the existing literature. Reported life satisfaction follows a U-shaped function with age, with minimums at 58 and 38 years of age in each specification respectively. Because this is a single cross-section, it is not possible to disentangle age effects from cohort effects: it may be that other factors affecting each generation yield this pattern. Some authors, such as Plagnol and Easterlin (2008), find that the pattern is different for men and women; including interaction terms between gender and age do not affect my findings. Being separated, divorced, or widowed reduces life satisfaction, as does being unemployed, and having health problems. On the other hand, additional education and income increase life satisfaction.

The effect of being female is positive in both models. In the first column, the coefficient on the female indicator is 0.034, showing that conditional only on age, women report higher happiness than men. In the second column, adding the other characteristics increases the coefficient on the female indicator to 0.085.

Table 3 shows the marginal effects of changing individual characteristics. The *LS* column simulates average life satisfaction ratings if the entire population had that characteristic, holding all other individual characteristics constant; stars reflect the significance from Table 3. The *Change* column shows the estimated change in life satisfaction reports versus the listed comparison. The final column, *x(gender eff.)*, shows how large that characteristic's marginal effect is as a factor of the marginal gender effect: the marginal effect of changing only gender, from male to female, appears in the last row. For example, moving from the 30th percentile of one's country's equivalized income distribution to the 40th percentile increases life satisfaction reports by 0.047, which is 0.63 times as large as the gender effect. Note that, because it holds all other variables constant, and women tend to have less desirable individual characteristics, the marginal gender effect is much larger than the raw gap shown in Table 1, 0.075 vs. 0.025.

Moving up a decile in the equivalized income distribution increases life satisfaction reports by about 0.5–1 times as much as the gender effect. For scale, in the U.S., moving up a decile is equivalent to increasing equivalized income by $5000 (10th to 20th percentile) to $14,000 (80th to 90th percentile) in this dataset. Rightly, three momentous possible characteristic changes – moving from low to high education, health problems versus no health problems, and unemployment versus full-time employment for employer – have more than double the impact of the gender effect. Even so, that comparing unemployment to full-time employment shows only twice the impact of the gender effect is a testament to the gender effect's unexpected strength.

4. Anchoring vignette-adjusted model

4.1. Vignettes

After asking respondents on which step of Cantril's ladder they current stand, Gallup also asked them to state on which step of the ladder six hypothetical people stand. The exact vignettes are as follows:

> A1: *Think of a female who is 40 years old and happily married with a good family life. Her monthly family income is about [median income]. She has severe back pain, which keeps her awake at night. On which step of the ladder do you think this person stands?*
> A2: *Think of a male who is 50 years old and divorced. He has a daughter with whom he has a good relationship. He has a secure job that pays about [twice median income] per month. He has no serious health problems. On which step of the ladder do you think this person stands?*
> A3: *Think of a male who is 25 years old and single without many friends. He makes about [half median income] per month. He feels he has little control over his job and worries about losing it. He has no health problems but feels stressed sometimes. On which step of the ladder do you think this person stands?*
> A4: *Think of a female who is 35 years old and married, with no children. Her monthly family income is about [median income]. Her work is a bit dull sometimes, but it is a very secure job. On which step of the ladder do you think this person stands?*
> A5: *Think of a female who is a 70-year-old widow. She receives about [half median income] in income each month. She has many friends. Lately, she suffers from back pain, which makes housework painful. On which step of the ladder do you think this person stands?*
> A6: *Think of a male who is 60 years old. He is single but has many friends his age. He no longer works but is comfortable with his decision to stop working. He receives about [twice median income] in income each month. He is very physically active. On which step of the ladder do you think this person stands?*

The income levels (median, half the median, or twice the median) were filled in with the appropriate value for the respondent's country. Every respondent saw exactly the same six vignettes.[12]

Table 1 shows gender differences in ratings for each vignette. Women rate every single vignette higher, significantly so. This implies that women tend to give higher ratings for the same life circumstances; this is true whether they are rating vignettes of men or women. And, the differences between men's and women's vignette ratings – from 0.035 to 0.047 among significant differences – are similar in magnitude to the difference between women's and men's self-reports, 0.025. The reporting gap was particularly strong for male vignette characters (average: 0.044, compared with 0.029 for female characters).

4.2. Intuition behind the vignette-adjusted model

Consider Fig. 1. Imagine that men actually have higher average life satisfaction than women, as indicated by men's distribution being shifted right of women's, but that they are less generous in their ratings, as indicated by the right-shifted thresholds. Now imagine a hypothetical person with latent life satisfaction at the dashed line. On the men's scale, that's a 2 out of 5. On the women's scale, because the thresholds are shifted left, it's a 4 out of 5. If men and women are using these different response scales in their self-evaluations, it will appear that women are happier, when in fact men are.

To account for this, I use vignette ratings. If respondents agree on the "true" underlying life satisfaction ratings for the vignettes, then if men and women give systematically different ratings, it must be because of differing response scales – what (King et al., 2004) call differential item functioning, or DIF. By understanding how gender moves the thresholds between response categories up and down, it is possible to evaluate the impact of gender on self-reports on a single scale, in this case, based on men's thresholds.

Once men and women are using the same adjusted response scale, we can reevaluate the life satisfaction gender gap by simulating women's life satisfaction ratings when using men's scales. This analysis depends on two key assumptions. First, *response consistency* means that individuals are using the same scale to rate their own life satisfaction as the vignette subjects'. King et al. (2004) and Van Soest et al. (2011) have provided good evidence that response consistency holds in other domains. Unlike some more controversial behavioral survey questions (e.g. regarding alcohol consumption), in which respondents may feel social pressure to change their self-reports but not vignette ratings, there is less incentive to misreport their own life satisfaction. Second, *vignette equivalence* means that men and women interpret the vignettes in the same way – they agree on the "true" underlying life satisfaction ratings. I test for vignette equivalence in two ways in Appendix A.2: first by testing the extent to which respondents adhere to the global ordering based on average ratings, and second by

[12] The B set of vignettes is given in Appendix Section Appendix A.1.

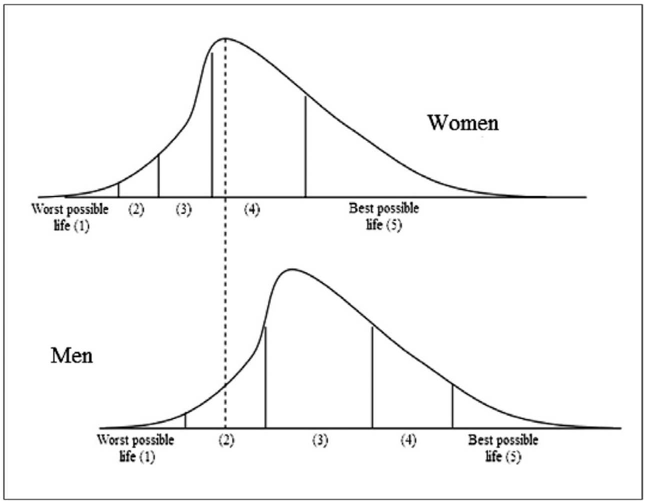

Fig. 1. Illustration of differential item functioning (DIF). Adapted from a similar figure in (Kapteyn, Smith, van Soest, 2007).

testing whether gender has a direct impact on vignette ratings. As noted in the Data section, the B set of vignettes failed both tests. Subjects' rank ordering of the vignettes varied substantially more among the B vignettes, while an analysis based on the HOPIT model found that gender influenced interpretation of those vignettes (and not the A set).

4.3. Model

To carry out the vignette adjustment formally, I use a HOPIT (hierarchical ordered probit) model. HOPIT operates similarly to ordered probit, except that in addition to affecting the underlying latent variable for life satisfaction, individual characteristics can *also* affect the thresholds between response levels. Each respondent answers the life satisfaction question for themselves, as well as for 2–6 of the 6 hypothetical individuals whom they were asked to evaluate.

Y_{ri} is a respondent i's self-report, and it is a function of their characteristics as well as an error term ϵ_{ri}, which is normally distributed with zero mean, and is independent of individual characteristics X_i. As in an ordered probit model, Y^*_{ri} is a latent variable such that $Y_{ri} = j$ if Y^*_{ri} is above the threshold τ_i^{j-1} and below the threshold τ_i^j:

$$\begin{aligned} &Y^*_{ri} = X_i'\beta + \epsilon_{ri}, \\ &\epsilon_{ri} \sim N(0, \sigma_r^2) \end{aligned} \tag{1}$$

$$Y_{ri} = j \text{ if } \tau_i^{j-1} < Y^*_{ri} \leq \tau_i^j \tag{2}$$

This model differs from an ordered probit model in how the thresholds are determined. They are also functions of the subject's characteristics, as well as an idiosyncratic error u_i, which is independent of ϵ_{ri} and X_i:

$$\begin{aligned} &\tau_i^0 = -\infty,\ \tau_i^5 = \infty,\ \tau_i^1 = \gamma^{1\prime} X_i + u_i \\ &\tau_i^j = \tau_i^{j-1} + \exp{(\gamma^{j\prime} X_i)},\ j = 2, 3, 4 \end{aligned} \tag{3}$$

The individual thresholds τ_i^j represent differing response scale usage, or DIF.

To make self-evaluations comparable, take one respondent's scale as the benchmark scale. That respondent has characteristics $X_i = X(B)$, and thus has thresholds τ_B^j. I can now compare all other individuals using thresholds τ_B^j. Because the latent variable Y^*_{ri} is not affected by the thresholds, this does not imply a new level of the latent variable. But it does imply a new rating, Y_{ri}, for which it is possible to simulate the adjusted distribution.

Not all parameters are identified; namely, only the difference between β and γ^1 can be determined using self-reports alone. To identify them separately, I use vignette ratings. Y_{li} is the rating given by i to vignette l:

$$Y^*_{li} = \theta_l + \epsilon_{li} \tag{4}$$

$$\begin{aligned} Y_{li} = j \text{ if } \tau_i^{j-1} < Y^*_{li} \leq \tau_i^j,\ j = 1, \ldots, 5, \\ \epsilon_{li} \sim N(0, \sigma^2),\ \text{independent of } \epsilon_{li}, \epsilon_{ri},\ \text{and } X_i. \end{aligned} \tag{5}$$

where θ_l is an indicator for vignette l. Notice that the equation for Y^*_{li} does not include any personal characteristics X_i, in line with the assumption of vignette equivalence. The assumption of response consistency means that the τ_i^j here are the same as those used with Y_{ri}. With this, I can identify β, $\gamma_1, \ldots, \gamma_5$, $\theta_1, \ldots, \theta_6$ up to a normalization of scale and location.

Table 4
HOPIT regression.

	Life satisfaction		τ^1		$ln(\tau^2 - \tau^1)$		$ln(\tau^3 - \tau^2)$		$ln(\tau^4 - \tau^3)$	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Female	-0.011	0.043**	-0.024	-0.032*	-0.015*	-0.011	-0.004	-0.001	-0.032***	-0.024**
	(0.017)	(0.018)	(0.016)	(0.018)	(0.009)	(0.010)	(0.008)	(0.007)	(0.010)	(0.011)
Age/10	-0.081***	-0.110***	0.078***	0.053***	-0.014	-0.013	-0.001	-0.001	-0.020*	-0.011
	(0.017)	(0.022)	(0.013)	(0.015)	(0.009)	(0.011)	(0.009)	(0.010)	(0.011)	(0.012)
Age^2/100	0.004	0.014***	-0.010***	-0.009***	0.001	0.002	0.001	0.001	-0.001	-0.002
	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)	(0.002)
Urban		0.050**		0.013		-0.009		0.004		0.015
		(0.024)		(0.022)		(0.017)		(0.013)		(0.013)
Married		0.075***		0.041*		0.009		0.017		-0.006
		(0.027)		(0.022)		(0.014)		(0.013)		(0.015)
Separated		-0.108**		0.070		-0.041		-0.002		-0.042
		(0.049)		(0.061)		(0.038)		(0.030)		(0.042)
Divorced		-0.130***		0.052		0.005		-0.017		-0.055*
		(0.045)		(0.044)		(0.029)		(0.024)		(0.030)
Widowed		-0.093**		0.078**		-0.042		0.008		-0.017
		(0.043)		(0.034)		(0.028)		(0.022)		(0.027)
Domestic partner		-0.040		0.059		-0.005		-0.005		-0.040
		(0.041)		(0.041)		(0.023)		(0.024)		(0.032)
Employed FT for self		0.016		0.010		0.002		-0.027*		-0.004
		(0.026)		(0.025)		(0.016)		(0.015)		(0.019)
Employed PT do not want FT		0.073**		-0.016		0.021		-0.026		0.018
		(0.032)		(0.034)		(0.021)		(0.018)		(0.020)
Unemployed		-0.218***		0.013		-0.025		-0.016		-0.022
		(0.043)		(0.026)		(0.021)		(0.015)		(0.026)
Employed PT want FT		-0.082***		-0.026		0.009		-0.019		-0.021
		(0.031)		(0.034)		(0.018)		(0.017)		(0.024)
Out of workforce		0.012		-0.020		0.014		-0.005		0.023
		(0.029)		(0.020)		(0.016)		(0.010)		(0.016)
Secondary education		0.164***		-0.039*		0.017		0.046***		0.067***
		(0.024)		(0.021)		(0.012)		(0.012)		(0.018)
Tertiary education		0.357***		-0.091***		0.073***		0.058***		0.161***
		(0.032)		(0.027)		(0.017)		(0.015)		(0.025)
Health problems		-0.245***		0.084***		-0.053***		-0.025**		-0.036***
		(0.027)		(0.020)		(0.013)		(0.012)		(0.012)
Log equivalized household income		0.287***		0.017		0.009		0.005		0.051***
		(0.017)		(0.012)		(0.008)		(0.007)		(0.007)
Observations	45,332	45,332	45,332	45,332	45,332	45,332	45,332	45,332	45,332	45,332

Regressions include country fixed effects, standard errors are clustered at the country level. τ^i is the threshold between the i^{th} and $(i+1)^{th}$ response levels. Significance: 1% (***), 5% (**), 10% (*).

5. Results: the gender gap with vignette adjustment (HOPIT)

To see how much of the gender gap is due to differences in scale usage, I compare results from the baseline ordered probit regressions to results from HOPIT regressions with vignette adjustment. In both cases, Shannon (1948) entropy[13] is higher in the ordered probit model, indicating more uncertainty than in the HOPIT model.

Table 4 shows coefficients in HOPIT regressions of the same specifications as Table 2: model (1) regresses life satisfaction on a female indicator, age, and age squared; model (2) adds additional individual characteristics. Both columns include country fixed effects, and standard errors are clustered at the country level. Each specification in Table 4 has five sets of coefficients: the first includes coefficients affecting the life satisfaction rating, and the following are coefficients affecting the thresholds between response categories.

A positive coefficient in the life satisfaction equation means increasing that variable increases life satisfaction. A positive coefficient on τ^1 means that increasing that variable is moving the threshold toward higher values of Y_i^*, i.e., making a respondent with Y_i^* just above the threshold change their response to the lowest response category. Similarly, a positive coefficient on $ln(\tau^{i+1} - \tau^i)$ indicates an increased distance between the ith and $(i+1)$th thresholds, showing that conditional on the ith threshold's location, increasing that covariate moves the $(i+1)$th threshold to a higher position, leading a respondent with Y_i^* just above that threshold to indicate the lower rating.

Comparing (1) between Tables 2 and4, using a vignette-adjusted model moves the coefficent on *female* from a significantly positive 0.34 to about zero. Two-sided Z-tests show that the coefficients on female, age, and age-squared are significantly different between the ordered probit and HOPIT models, at the 10%, 5%, and 10% levels respectively. The *female*

[13] Shannon entropy is calculated $\sum_i \sum_j \widehat{P_{ij}} * ln(\widehat{P_{ij}})$, where $\widehat{P_{ij}}$ is the estimated probability that respondent i rated their life satisfaction as response level j.

Table 5
HOPIT vs. ordered probit marginal effects.

Characteristic	Comparison	HOPIT (vignette-adjusted)			Ordered probit (unadjusted)		
		LS	Change	x(gen. eff.)	LS	Change	x(gen. eff.)
Income at 10th percentile	-	2.714	-	-	2.750	-	-
Income at 20th percentile***	10th pctile	2.810	0.095	2.60	2.832	0.082	1.09
Income at 30th percentile***	20th pctile	2.875	0.065	1.78	2.888	0.056	0.74
Income at 40th percentile***	30th pctile	2.930	0.055	1.51	2.935	0.047	0.63
Income at 50th percentile***	40th pctile	2.982	0.052	1.42	2.979	0.044	0.59
Income at 60th percentile***	50th pctile	3.031	0.049	1.34	3.021	0.042	0.56
Income at 70th percentile***	60th pctile	3.086	0.054	1.47	3.067	0.046	0.61
Income at 80th percentile***	70th pctile	3.149	0.063	1.72	3.120	0.053	0.71
Income at 90th percentile***	80th pctile	3.237	0.088	2.41	3.194	0.074	0.99
Low education	-	2.841	-	-	2.857	-	-
Medium education***	Low	2.981	0.140	3.82	2.986	0.129	1.72
High education***	Low	3.146	0.305	8.32	3.103	0.246	3.29
No health problems	-	2.990	-	-	2.999	-	-
Health problems***	No health probs	2.782	-0.208	-5.68	2.774	-0.225	-3.00
Single	-	2.915	-	-	2.951	-	-
Married***	Single	2.979	0.063	1.73	2.968	0.017	0.23
Separated**	Single	2.824	-0.092	-2.50	2.843	-0.108	-1.44
Divorced***	Single	2.806	-0.110	-3.00	2.812	-0.138	-1.85
Widowed**	Single	2.837	-0.079	-2.15	2.832	-0.118	-1.58
Domestic partner	Single	2.882	-0.034	-0.91	2.887	-0.064	-0.85
Employed FT for employer	-	2.947	-	-	2.941	-	-
Employed FT for self	Employed FT for emp.	2.961	0.014	0.38	2.962	0.021	0.28
Employed PT, don't want FT**	Employed FT for emp.	3.009	0.062	1.70	3.013	0.072	0.96
Unemployed***	Employed FT for emp.	2.763	-0.184	-5.02	2.785	-0.156	-2.08
Employed PT, want FT***	Employed FT for emp.	2.877	-0.070	-1.90	2.910	-0.031	-0.42
Out of the workforce	Employed FT for emp.	2.957	0.010	0.28	2.963	0.022	0.29
Rural	-	2.923	-	-	2.932	-	-
Urban**	Rural	2.965	0.043	1.17	2.966	0.034	0.46
Male	-	2.920	-	-	2.907	-	-
Female**	Male	2.957	0.037	1.00	2.982	0.075	1.00

Full sample (N = 45,332). Marginal effects of changing various characteristics, based on HOPIT regression coefficients in Table 4 column 2 and on ordered probit coefficients in Table 2 column 2; stars indicate significance from that column. Income percentiles are from respondent's own country. "LS" shows simulated life satisfaction if all respondents had that characteristic, all other characteristics held constant. "x(gen. eff.)" compares the magnitude of that marginal effect with the gender effect in the last row.

coefficients in the τ equations explain the difference: women's τ thresholds are lower, i.e., they give higher ratings for the same latent Y_i^* near the thresholds. In specification (2), with individual characteristics added, the coefficient on *female* is reduced from 0.085 in Table 2 to 0.043 in Table 4, and the coefficients on *female* in the τ equations are also significantly negative.

Like that on *female*, the coefficients in the HOPIT life satisfaction equation match the signs of their counterparts in the ordered probit regressions in specification (2) of Table 2. Only the coefficients on tertiary education and income are significantly different, at the 10% and 5% levels respectively. The signs and significance levels of these variables are consistent with the literature. These effects are not due simply to DIF, e.g., being widowed does not simply change the way people use response scales, it significantly decreases life satisfaction, as the literature has long reported. In the τ equations of model (2), being more educated reduces the first threshold and increases the gaps between the higher thresholds more than enough to offset the change in τ_1, ultimately pushing respondents toward the middle values. Health problems have the opposite effect, increasing the first threshold and reducing higher τ_js, pushing respondents toward the extreme values. The vignette characters' lives are described with reference to crucial qualities of a person's life, and where possible, I include similar information about the respondents in the model specification (e.g. income, education, health). Although some vignettes also include information on social relationships (e.g. A2: *He has a daughter with whom he has a good relationship*; A5: *She has many friends*), my dataset lacked relevant measures of social relationships with sufficient coverage of the sample.

Table 5 shows the marginal effects of changing individual characteristics in the HOPIT equation for life satisfaction, alongside the ordered probit marginal effects from Table 3. As in Table 3, each row simulates average life satisfaction if every respondent had that characteristic, all other characteristics held constant, in the *LS* column, and stars reflect significance from Table 4. Using the comparison listed, the *Change* column finds the marginal impact of changing only that characteristic, and *x(gen. eff.)* shows how large that characteristic's marginal effect is as a factor of the marginal gender effect in the last row. As expected from the regression results, the marginal gender effect is much smaller in the HOPIT model, less than half as large as that without vignette adjustment. A more realistic pattern emerges, in which most of the included characteristic changes have substantially larger impacts on average life satisfaction than the gender effect. This is partially driven by the reduction in the marginal gender effect, and partially driven by some characteristics' increased marginal effect, notably in-

Table 6
Simulated life satisfaction with own scales and men's scales, global sample.

	(1)	(2)
Own response thresholds		
Female average	2.952	2.950
Male average	2.927	2.927
Female - male gap	0.024***	0.023***
Female Pr(Life sat. = 1)	9.7%	9.7%
Female Pr(Life sat. = 2)	24.9%	24.6%
Female Pr(Life sat. = 3)	34.4%	34.6%
Female Pr(Life sat. = 4)	22.5%	23.0%
Female Pr(Life sat. = 5)	8.4%	8.1%
Male Pr(Life sat. = 1)	9.9%	9.9%
Male Pr(Life sat. = 2)	25.5%	25.2%
Male Pr(Life sat. = 3)	34.4%	34.5%
Male Pr(Life sat. = 4)	22.5%	23.0%
Male Pr(Life sat. = 5)	7.8%	7.4%
Men's response thresholds		
Female average	2.912	2.911
Male average	2.927	2.927
Female - male gap	-0.015***	-0.017***
Female Pr(Life sat. = 1)	10.0%	10.2%
Female Pr(Life sat. = 2)	25.8%	25.5%
Female Pr(Life sat. = 3)	34.5%	34.6%
Female Pr(Life sat. = 4)	22.2%	22.6%
Female Pr(Life sat. = 5)	7.4%	7.2%
% of gap explained	**162%**	**172%**

Simulated based on coefficients in a global HOPIT model. Specifications match the columns in Table 4. Significance stars indicate the results of t-tests of average life satisfaction by gender in each panel: 1% (***), 5% (**), 10% (*).

come and education. These results imply that the standard (non-vignette-adjusted) approach is underestimating the impacts of these key characteristics even as it overestimates the impact of gender.

We return now to our motivating puzzle: why do women report higher life satisfaction on average, despite being worse off in important ways? Our HOPIT results show that, because of DIF, women's response thresholds are lower than men's, leading them to choose a higher rating given the same latent life satisfaction. To compare the gap before and after accounting for DIF, Table 6 shows simulations of average life satisfaction ratings for each specification shown in Table 4, based on the estimated probability and cardinal value of each response level (i.e., $\sum_j \widehat{P_{ij}} * j$). In the top panel, I simulate men's and women's responses using their own gender's scale. In the bottom panel, I simulate men's and women's responses both using men's scales, i.e., by setting the coefficient on *female* to zero in the τ equations.[14] The last line shows the percentage of the top panel gap that is eliminated in the bottom panel.

Using men's thresholds, women are less likely to put themselves in the highest response category and more likely to put themselves in the lowest response category. Now that DIF is removed, the large negative effects on life satisfaction from women's disadvantages (e.g. less education, less income, worse self-reported health) outweigh the small positive effect of being female. As a result, in both models, evaluating women's and men's self-reports using men's thresholds completely reverses the gap, by over 150%; if women used the same response scale as men, they would report *lower* life satisfaction than men. The gap would be about half as large as it is now, but in the opposite direction.

6. Discussion

Life satisfaction surveys measure something important, that standard macroeconomic indicators alone, such as GDP, cannot (Stiglitz et al., 2009). Beyond financial matters, people cite family, health, work, and other personal concerns as the building blocks of a good life (Cantril, 1965).

The gender gap in life satisfaction self-reports is mysterious because women are objectively worse off in important ways. Some studies have even found evidence that life satisfaction declines among women following improvements in gender equality (Graham and Chattopadhyay, 2013; Stevenson and Wolfers, 2009). Stevenson and Wolfers (2009) propose that socioeconomic changes, changes in what the measures capture due to large social shifts, or changes in reference group, could explain the decline. In a similar vein, other authors have proposed that life satisfaction self-reports capture aspirations and optimism in addition to life satisfaction (Zweig, 2015; Plagnol and Easterlin, 2008; Arrosa and Gandelman, 2016). Future studies should examine modifiers of the relationships identified here, including GDP and gender rights, and particularly how each gender's average life satisfaction, and gender differences in life satisfaction, evolve following changes in those

[14] This method does not adjust for the fact that women also have other characteristics that are different from men's (e.g. men's higher education levels should shift their τ^1 threshold more left, as shown in Table 4).

modifiers. With a panel of surveys including vignettes tracking these changes, we may find that following improvements in gender equality, women's estimated underlying life satisfaction is constant (or increasing) while their self-reports decline, whether due to changes in response bias or changes to expectations.

Gender gaps due to response bias or differing standards have different implications when using life satisfaction to evaluate policies, as proposed by Layard (2005). As Deaton (2018) points out, based on life satisfaction self-reports alone, "Contrary to what is often recommended, transfers to men will do more to improve social well-being than transfers to women, at least when social well-being is taken to be total utility." On the other hand, if the gender gap is mostly driven by differences in expectations among men and women, we may improve men's happiness through policies designed to lower their standards. In either case, we should take great care in choosing policies based on reported life satisfaction.

If a woman says that she is an 8 on a 0–10 scale, and a man says that he is a 6, should we not take them at their word? In cases where there are objective standards, such as in political efficacy (King et al., 2004) or alcohol consumption (Van Soest et al., 2011), vignettes have generally done a good job in explaining differences in responses across countries, social groups, or individuals. Response bias (her 8 is his 6) and differing standards (the same situation is an 8 to her and a 6 to him) are impossible to disentangle, particularly when no objective measure exists (as in the case of life satisfaction). One interpretation of my findings is that women use different scales; the other is that women have different standards.

Beyond the difficulty of interpretation, the vignette-adjusted model includes assumptions of its own worth considering: vignette equivalence and response consistency. No statistical test can say for sure, based on the information collected, whether all subjects were interpreting the vignettes in the same way. They are only a few lines each, and women may fill in the gaps more optimistically, imagining the character's life as more full and fulfilling. Similarly, the assumption of response consistency may be wrong. As Deaton (2011) noted regarding the use of anchoring vignettes on disability, respondents may have systematic differences in their ability to empathize with vignette characters. Men may be "tougher" in rating the lives of others than they are in rating themselves; women may be more able to imagine themselves in the vignette characters' positions. Research has indicated that women are more empathetic than men (e.g. Macaskill et al., 2002), although it is not clear that more empathy should lead to strictly higher vignette ratings. Women also report feeling more negative feelings, including stress, worry, and physical pain than men; if they empathize especially with these negative aspects of vignette descriptions, they may be harsher in their ratings.

Comparing the gender gaps in vignette ratings to the vignette descriptions, we see that the largest gaps are for male characters (A2, A3, A6) and are not especially small where there are no explicit mentions of bad health or emotions (A2, A4, A6). If women's lower standards for women are solely driving these results, we should expect the gap to be largest among female vignette characters, when in fact we see the opposite. This may reflect that men similarly have low standards for women, or may be due to other differences among the vignettes. Future research could ask participants about their standards and expectations for vignette characters, and randomize gender, health, income, and other key characteristics in vignettes.

7. Conclusion

Taking a simple average of men's and women's life satisfaction ratings implies that women are happier than men, despite their measurable disadvantages. Using an undadjusted ordered probit model, it appears that simply being female increases happiness as much as a significant income bump. This is because a standard ordered probit model assumes that men's and women's responses are on the same scale. Using a HOPIT model with vignette adjustment, I find that the coefficient on *female* is still positive, but much less so. Because women are worse off in their other characteristics (income, health, education, etc.), despite this positive coefficient on *female*, women's life satisfaction is on average lower than men's after vignette adjustment.

Although I focus here on the impact of gender, it is reassuring that the impact of other key determinants of life satisfaction from the literature remain strong determinants even with vignette adjustments.

This study fits into a growing literature assessing the validity of life satisfaction self-reports. Adding vignette ratings to surveys places significant additional burden on respondents, and may be unnecessary when investigating some hypotheses. At least while considering the impact of gender on happiness self-reports, anchoring vignettes provide a useful signal of scale use differences so great that the gap is reversed.

Author agreement

I declare that this manuscript, "Revering the Gender Gap in Happiness," is original, has not been published before, and is not currently beging considered for publicatin elsewhere.

As the sole author, I have read and approved this manuscript, and confirm that there are no other persons who satisfied the criteria for authorship but are not listed.

I understand that I, as the corresponding author, am the sole contact for the editorial process.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

This research was supported by grant 1 R01 AG036784-01 from the National Institute on Aging.

Acknowledgments

I thank Arie Kapteyn, James P. Smith, Richard A. Easterlin, John Strauss, Daniel J. Benjamin, and Kelsey J. O'Connor for their invaluable comments. This work was supported by the National Institute on Aging grant 1 R01 AG036784-01.

Appendix

A.1. Additional Gallup dataset details

With some exceptions, all samples are probability-based and nationally representative of the resident population aged 15 and older. The coverage area is the entire country including rural areas, and the sampling frame represents the entire civilian, non-institutionalized, aged 15 and older population of each country. Exceptions include areas where safety of interviewing staff is threatened, scarcely populated islands, and areas interviewers can reach only by foot, animal, or small boat. Specifically, sampling in the Central African Republic, Democratic Republic of the Congo, Lebanon, Pakistan, India, Syria, Azerbaijan, Georgia, Morocco, Myanmar (Burma), Chad, Madagascar, Moldova, and Sudan was affected by security; some of these as well as Canada, China, Laos, and small parts of Japan had non-representative sampling of some geographic regions. In Arab countries (Bahrain, Kuwait, Saudi Arabia), sampling was of citizens (including Arab expatriates) and those who could complete the survey in Arabic or English; in the United Arab Emirates, all non-Arabs were excluded, i.e. more than half of the population. In the Philippines, urban areas were over-sampled. Israel excludes East Jerusalem (Gallup reports Palestinian Territories separately).

Telephone surveys are used in countries where telephone coverage represents at least 80% of the population or is the customary survey methodology. In Central and Eastern Europe and most of the developing world, an area frame design is used for face-to-face interviewing. In some countries, over-samples are collected in major cities or areas of special interest. In some large countries, such as China and Russia, samples of at least 2,000 are collected. For more details, refer to the Gallup World Poll Codebook (Gallup, 2019).

The excluded B set of vignettes are as follows:

> B1: *Think of a male who is 40 years old and happily married with a good family life. His monthly family income is about [twice median income]. He likes to work but suffers from serious back pain, which keeps him awake at night. On which step of the ladder do you think this person stands?*
>
> B2: *Think of a male who is a 65-year-old widow. She misses her husband a lot but has good relationships with her children and grandchildren. She receives about [half median income] in income each month. She has heart problems, which caused her to stop working. On which step of the ladder do you think this person stands?*
>
> B3: *Think of a male who is 35 years old and married, with no children. His monthly family income is about [half median income]. His work is a bit dull sometimes, but it is a very secure job. On which step of the ladder do you think this person stands?*
>
> B4: *Think of a female who is 50 years old and divorced. She has children from her marriage but has little contact with them. She has an interesting job. Her monthly income is about [twice median income]. She often has trouble sleeping. On which step of the ladder do you think this person stands?*
>
> B5: *Think of a female who is a 70 years old and married. She and her husband lead their own lives and don't do many things together. They have two children but rarely see them. Her monthly family income is about [median income]. She is overweight and gets tired when walking for more than a few minutes. On which step of the ladder do you think this person stands?*
>
> B6: *Think of a male who is 50 years old. He does not exercise and is obese. He has pain in his knees almost all the time. He is very secure in his job. He has been married for a long time, but he and his wife spend very little time together. His monthly family income is about [median income]. On which step of the ladder do you think this person stands?*

A.2. Two tests of vignette equivalence

A.2.1. Vignette ordering

As discussed in Section 4, *vignette equivalence* means that men and women interpret the vignettes in the same way, and is a key assumption of the HOPIT model. Here I discuss two tests of this assumption.

First, I quantify to what extent respondents deviated from the global vignette ordering in their individual orderings. If all respondents interpret the vignettes similarly, then their rank ordering of the vignettes should also be similar. The "global" vignette ordering is defined as the ordering when all respondents' ratings are averaged. (The global ordering is the same if

I define it by the mode or median.) Following (Murray et al., 2003), I call the benefit of the doubt (Spearman) rank order correlation coefficient (BDROCC) the Spearman correlation with ties resolved to favor the overall (global) ordering. Each respondent rated the vignettes on the same 0–10 scale, and thus, they could give the same rating to multiple vignettes. In finding which orderings are consistent with the global ordering, I resolve ties as if they matched the global order. In the extreme case, if an individual rated every vignette the same way, it would be consistent with the global ordering. (In practice, well under 1% of individuals did this in the A set, and about 1% did this in the B set.) A high BDROCC (near 1) means it is close to the global ordering.

Table A2 summarizes the BDROCC for both sets of vignettes, and Table A3 shows every country separately. Four individual countries were especially problematic: Chad (median BDROCC = -0.337 for A set), Palestinian Territories (median BDROCC = -0.143 for B set), Japan (median BDROCC = 0.086 for B set), and United Arab Emirates (median BDROCC = 0.029 for B set). Chad is removed from all A-set analysis that follows.

The median BDROCC for the B set is just 0.543, compared with 0.829 with the A set. While the majority of respondents (59%) were very close to the global ordering in the A set (perfect match, one single-rank inversion, two single-rank inversions, or one double-rank inversion),[15] only about a third were very close in the B set. This could reflect any number of factors, including how the vignettes were written (they may not be different enough), how the individuals interpreted the questions (respondents may "fill in the gaps" differently), or how the survey was administered (the surveyors may have made mistakes in asking the questions, Gallup may have made a mistake in the materials they used, or the data may have been recorded incorrectly). For whatever reason, the people rating the B-set were not as consistent, and thus vignette equivalence seems not to hold for them. Though there is no theoretical cut-off for how high the BDROCC can be while still assuming vignette equivalence, it is clear that the B set is much less consistent than the A set.

A.2.2. The effect of gender on vignette ratings

Next I test whether respondents' gender significantly determine vignette evaluations. If the vignettes are interpreted the same way to all, then the latent vignette rating Y^*_{li} should be determined entirely by the vignette fixed effect, and any difference in the observed Y_{li} should come from differing thresholds τ_i^j, which vary by individual characteristics (e.g. female vs. male). Following (d'Uva et al., 2011), I consider a slightly modified version of the model, replacing Eq. (4) with the following:

$$\begin{aligned} Y^*_{l1} &= \theta_1 + \epsilon_{l1} \\ Y^*_{li} &= \theta_l + \lambda_l' X_i + \epsilon_{li}, l \neq 1 \end{aligned} \tag{6}$$

where l counts $1, \ldots, 12$ when both sets of vignettes are included and $1, \ldots, 6$ when only using one set. For identification, I must omit $\lambda_l' X_i$ from one vignette equation. If vignettes are equivalent, then coefficients λ_l should be all equal to zero.

Table A4 shows coefficients on *female* in (6). Using the pooled sample (columns 1 and 2), the A vignettes do not indicate a vignette equivalence violation, but the B vignettes do: the significant coefficients show that a constant DIF alone does not explain the gender differences in ratings, implying that, at least for B3 and B5, men and women interpreted these vignettes differently even beyond our measured DIF. In the separate samples, for both the full model and the *female* and *age* only model, the pattern holds: *female* does seem to influence interpretation of some vignettes in the B set, but not the A set.

The B set of vignettes underperforms relative to the A set in the first test, and fails the second test, indicating that vignette equivalence is violated. Thus, respondents that received the B set of vignettes, roughly half, are removed from my sample, and I focus my analysis on the sample that received the A vignettes.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jebo.2022.01.006.

References

Angelini, V., Cavapozzi, D., Corazzini, L., Paccagnella, O., 2012. Age, health and life satisfaction among older Europeans. Soc. Indic. Res. 105 (2), 293–308.

Angelini, V., Cavapozzi, D., Corazzini, L., Paccagnella, O., 2014. Do danes and italians rate life satisfaction in the same way? Using vignettes to correct for individual-specific scale biases. Oxf. Bull. Econ. Stat. 76 (5), 643–666.

Arrosa, M.L., Gandelman, N., 2016. Happiness decomposition: female optimism. J. Happiness Stud. 17 (2), 731–756. doi:10.1007/s10902-015-9618-8.

Batz, C., Tay, L., 2018. Gender differences in subjective well-being. Handbook of Well-Being. UT: DEF Publishers. Salt Lake City

Bond, T.N., Lang, K., 2019. The sad truth about happiness scales. J. Polit. Econ. 127 (4), 1629–1640. doi:10.1086/701679.

Brulé, G., Veenhoven, R., 2017. The '10 excess' phenomenon in responses to survey questions on happiness. Soc. Indic. Res. 131, 853–870. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5357485/

Cantril, H., 1965. The Pattern of Human Concerns. Rutgers University Press, New Brunswick.

Chen, L., Oparina, E., Powdthavee, N., Srisuma, S., 2019. Have Econometric Analyses of Happiness Data Been Futile? A Simple Truth About Happiness Scales. IZA Discussion Paper 12152. https://ssrn.com/abstract=3349935

Clark, A.E., Diener, E., Georgellis, Y., Lucas, R.E., 2008. Lags and leads in life satisfaction: a test of the baseline hypothesis. Econ. J. 118 (529), F222–F243. doi:10.1111/j.1468-0297.2008.02150.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0297.2008.02150.x

[15] A single-rank inversion means reversing the rankings of two adjacent vignettes, e.g. ranking the global top vignette (A6) as the second-most satisfied with his life, and the global second-place vignette (A2) as the most satisfied. A double-rank inversion is the same, except it compares vignettes that are two ranks away in the global ordering, i.e. it reverses the positions of the first-place and third-place vignettes.

Dalton, D., Ortegren, M., 2011. Gender differences in ethics research: the importance of controlling for the social desirability response bias. J. Bus. Ethics 103, 73–93. doi:10.1007/s10551-011-0843-8.

Davis, R.E., Couper, M.P., Janz, N.K., Caldwell, C.H., Resnicow, K., 2009. Interviewer effects in public health surveys. Health Educ. Res. 25 (1), 14–26. doi:10.1093/her/cyp046. https://academic.oup.com/her/article-pdf/25/1/14/1465507/cyp046.pdf

Deaton, A., 2011. Comment on "Work Disability, Work, and Justification Bias in Europe and the U.S.". University of Chicago Press, Chicago, pp. 312–314. http://www.nber.org/chapters/c11947

Deaton, A., 2018. What do self-reports of wellbeing say about life-cycle theory and policy? J. Public Econ. 162, 18–25. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6135248/

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. J. Hum. Resour. 46 (4), 875–906.

d'Uva, T.B., Van Doorslaer, E., Lindeboom, M., O'Donnell, O., 2008. Does reporting heterogeneity bias the measurement of health disparities? Health Econ. 17 (3), 351–375. doi:10.1002/hec.1269. https://onlinelibrary.wiley.com/doi/pdf/10.1002/hec.1269

Easterlin, R.A., 2003. Happiness of women and men in later life: nature, determinants, and prospects. In: Sirgy, M.J., Rahtz, D., Samli, A.C. (Eds.), Advances in Quality-of-Life Theory and Research. Springer Netherlands, Dordrecht, pp. 13–25. doi:10.1007/978-94-017-0387-1_2.

Gallup, 2019. Worldwide Research Methodology and Codebook. Reference manual. https://news.gallup.com/poll/165404/world-poll-methodology.aspx

Graham, C., Chattopadhyay, S., 2013. Gender and well-being around the world. Int. J. Happiness Dev. 1 (2), 212–232. doi:10.1504/IJHD.2013.055648. PMID: 55648

Green, C.P., Heywood, J.S., Kler, P., Leeves, G., 2018. Paradox lost: the disappearing female job satisfaction premium. Br. J. Ind. Relat. 56 (3), 484–502. doi:10.1111/bjir.12291. https://onlinelibrary.wiley.com/doi/abs/10.1111/bjir.12291

Grol-Prokopczyk, H., Freese, J., Hauser, R.M., 2011. Using anchoring vignettes to assess group differences in general self-rated health. J. Health Soc. Behav. 52 (2), 246–261.

Grzymala-Busse, A., 2007. Rebuilding leviathan: party competition and state exploitation in post-communist democracies. Cambridge Studies in Comparative Politics. Cambridge University Press doi:10.1017/CBO9780511618819.

Helliwell, J.F., Layard, R., Sachs, J., 2015. World Happiness Report 2015. Sustainable Development Solutions Network, New York. https://worldhappiness.report/ed/2015/

Herbst, C.M., 2011. 'Paradoxical' decline? Another look at the relative reduction in female happiness. J. Econ. Psychol. 32 (5), 773–788. doi:10.1016/j.joep.2011.07.001. http://www.sciencedirect.com/science/article/pii/S0167487011000985

Kaiser, C., 2020. Using Memories to Assess the Intrapersonal Comparability of Wellbeing Reports. SocArXiv. https://osf.io/preprints/socarxiv/xpghn/

Kaiser, C., Vendrik, M.C.M., 2019. How Threatening are Transformations of Reported Happiness to Subjective Wellbeing Research?. SocArXiv. https://osf.io/preprints/socarxiv/gzt7a/

Kapteyn, A., Smith, J.P., van Soest, A., 2007. Vignettes and self-reports of work disability in the United States and the Netherlands. Am. Econ. Rev. 97 (1), 461–473. doi:10.1257/aer.97.1.461. https://www.aeaweb.org/articles?id=10.1257/aer.97.1.461

Kapteyn, A., Smith, J.P., Van Soest, A., 2010. Life satisfaction. In: International Differences in Well-Being. Oxford University Press, New York, pp. 70–104.

King, G., Murray, C.J.L., Salomon, J.A., Tandon, A., 2004. Enhancing the validity and cross-cultural comparability of measurement in survey research. Am. Polit. Sci. Rev. 98 (1), 191–207. doi:10.1017/S000305540400108X.

Kristensen, N., Johansson, E., 2008. New evidence on cross-country differences in job satisfaction using anchoring vignettes. Labour Econ. 15 (1), 96–117. doi:10.1016/j.labeco.2006.11.001. http://www.sciencedirect.com/science/article/pii/S092753710600087X

Layard, R., 2005. Rethinking public economics: the implications of rivalry and habit. In: Economics and Happiness: Framing the Analysis. Oxford University Press, pp. 147–169.

Macaskill, A., Maltby, J., Day, L., 2002. Forgiveness of self and others and emotional empathy. J. Soc. Psychol. 142 (5), 663–665. doi:10.1080/00224540209603925.

Molina, T., 2016. Reporting heterogeneity and health disparities across gender and education levels: evidence from four countries. Demography 53 (2), 295–323. doi:10.1007/s13524-016-0456-z.

Murray, C.J., Ozaltin, E., Tandon, A., Salomon, J., Sadana, R., Chatterji, S., et al., 2003. Empirical evaluation of the anchoring vignette approach in health surveys. In: Health Systems Performance Assessment: Debates, Methods and Empiricism, 369. World Health Organization, Geneva, p. 399.

Nolen-Hoeksema, S., Rusting, C.L., 2003. Gender differences in well-being. In: Well-being: Foundations of Hedonic Psychology. Russell Sage Foundation, pp. 330–350. chapter 17

Plagnol, A.C., Easterlin, R.A., 2008. Aspirations, attainments, and satisfaction: life cycle differences between american women and men. J. Happiness Stud. 9 (4), 601–619.

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27 (3), 379–423.

Stevenson, B., Wolfers, J., 2009. The paradox of declining female happiness. Am. Econ. J. Econ. Policy 1 (2), 190–225.

Stiglitz, J. E., Sen, A., Fitoussi, J.-P., et al., 2009. Report by the commission on the measurement of economic performance and social progress.

Stutzer, A., 2004. The role of income aspirations in individual happiness. J. Econ. Behav. Organ. 54 (1), 89–109. https://www.sciencedirect.com/science/article/abs/pii/S0167268103002038

UN Women and others, 2015. Progress of the World's Women 2015–2016: Transforming Economies, Realizing Rights. Technical Report. New York

Van Soest, A., Delaney, L., Harmon, C., Kapteyn, A., Smith, J.P., 2011. Validating the use of anchoring vignettes for the correction of response scale differences in subjective questions. J. R. Stat. Soc. Ser. A (Stat. Soc.) 174 (3), 575–595.

Zweig, J.S., 2015. Are women happier than men? Evidence from the gallup world poll. J. Happiness Stud. 16 (2), 515–541.