Are international happiness rankings reliable?

C P Barrington-Leigh*

5 January 2026

Abstract

Cross-country comparisons of subjective wellbeing underpin a growing body of empirical welfare economics and policy analysis. Yet the empirical foundation for such work — the assumption that life evaluation questions are comparable across countries and survey instruments — is hard to test.

Using the Gallup World Poll, the Global Flourishing Study, and the World Values Survey, this paper evaluates the consistency of two widely used cognitive life evaluation items, the life satisfaction and Cantril ladder questions. The analysis compares their national distributions, cross-country rankings, and estimated associations with life circumstances. The results reveal large and systematic discrepancies that cannot be explained by sampling error, focal-value behaviour, or obvious framing effects.

These findings suggest that current international rankings of wellbeing are not robust to alternative question formulations or survey implementations, even within the same populations. On the other hand, response distributions show stability over time for a given survey and question, and, importantly, estimated within-country determinants remain broadly consistent across questions. The paper therefore suggests some limits to cross-national inference from life evaluation data and outlines priorities for an ongoing research agenda on the measurement of welfare and quality of life in economics.

*This work was supported by SSHRC grant 435-2024-1219. I am grateful for comments from Jan-Emmanuel De Neve, John Helliwell, and others, including audiences at Oxford, McGill, Schiedam, Winnipeg, Johannesburg, London, and Reno.

Please use the latest version of this paper online.

Author contact at `https://wellbeing.research.mcgill.ca/contact`.

Contents

1 Introduction **4**

2 Data **6**

3 Country ranks **7**

4 Univariate response distributions **17**

5 Joint response distributions **19**

6 Model inference **21**

7 Discussion and Conclusion **25**

References **26**

A Supplementary figures and tables **31**

List of Figures

1 Country ranks: life satisfaction (WVS) and Cantril ladder (GWP) . . . 9
2 Country ranks: life satisfaction from the WVS and GWP 10
3 Country ranks: GWP life satisfaction and Cantril ladder 11
4 Country ranks: GFS life satisfaction and Cantril ladder 12
5 Country ranks: life satisfaction from WVS and GFS 13
6 Country ranks: life satisfaction (WVS) and Cantril ladder (GFS) . . . 14
7 Country ranks: Cantril ladder from GFS and GWP 15
8 Consistency of country means, and of survey differences, across time (GWP and GFS) 16
9 Response distributions for a selection of countries, for each survey and question. 18
10 Joint distribution of individuals' responses to life satisfaction and Cantril ladder 20
11 Estimated coefficients using different life evaluation questions in the GFS 23
A1 Response distributions by cultural region, for each survey and question. 32
A2 Response distributions for a selection of countries, by year. 33

List of Tables

1 Summary of correlations for rank comparisons 8
2 Average differences by cultural group for Figure 1 9

3 Average differences by cultural group for Figure 2 10
4 Average differences by cultural group for Figure 3 11
5 Average differences by cultural group for Figure 4 13
6 Average differences by cultural group for Figure 7 14
A1 Average differences by cultural group for Figure 5 31
A2 Average differences by cultural group for Figure 6 31
A3 Variables used in the GFS models of Section 6 34

1 Introduction

The rising international interest in subjective wellbeing can be attributed in part to the existence of large international surveys which ask the same life evaluation question, suitably translated into local languages, across starkly different countries and cultures (Barrington-Leigh, 2022; Helliwell and Wang, 2012). The World Happiness Report has, annually since 2012, reported a ranking across nearly 160 countries in the average answer to a single life evaluation question from one such survey, the Gallup World Poll (e.g., Helliwell et al., 2025). This paper begins by addressing a straightforward question: are those rankings reproducible, for instance by running a similar survey, or by asking a slightly different question?

In principle, the rankings are not an object of primary interest to researchers in the field. Due for instance to the desirable convergence of transition or developing countries catching up to richer ones in various ways, one may expect rankings of some more developed countries to go down even when their average reported numerical life evaluations are going up. Academic researchers are more interested in identifying causal effects of policy environments and other circumstances, changes, and choices on populations' life evaluations. These effects are inferred by estimating coefficients in models that explain differences and changes in life evaluations at the individual level and averaged over groups. Such estimates are thought to be a key ingredient to begin detailed policy design for wellbeing (Frijters et al., 2020; Happiness Research Institute, 2020; Frijters and Krekel, 2021; UK Treasury, 2021; Blodgett et al., 2024; Frayman et al., 2024).

On the other hand, the researcher's ideal can be elusive. Not all salient circumstances change on observable time scales or vary independently of other circumstances. Therefore, population-wide values of a summative, multi-faceted metric like a life evaluation cannot be fully accounted for by specific, disaggregated causes and marginal effects. As a result, international rankings remain important for addressing the broadest policy questions about the success of different political and economic systems, as judged by the experience of their populations. For instance, knowledge of the high life evaluations in Nordic countries has brought attention to their overall policy environment, lending credence to the idea that a high trust, cohesive, supportive, high productivity, individualist, social democratic society is the best model so far for generating a happy populace.

At the same time, some have questioned whether the subjective life evaluation approach could somehow be biased in favor of Western respondents (Lomas et al., 2022). This would contradict the widespread assumption among economists in the field that responses are internationally and interculturally comparable. One problem could be cultural differences in conceptions of wellbeing (Diener, Oishi, Lucas, et al., 2003; Diener and Suh, 2003; Oishi et al., 2013; Oishi, 2010). However, early work on life evaluations from the Gallup World Poll suggested strong comparability around the world, in the sense that similar coefficient estimates for the correlates of life satisfaction were found in regions and countries around the world (Helliwell et al., 2010) and that cultural effects play a limited role in explaining cross-country differences compared with

objective life circumstances (Exton, Smith, and Vandendriessche, 2015). Findings in such studies imply that the life evaluation question taps into something universal about human experience, a premise which underlies the country rankings of the prominent annual World Happiness Reports.

Another possible problem relates to the reporting function rather than problems of translating the meaning of the question. Using surveys with a battery of subjective questions rather than one key one, Likert-scale responses have been explained in part based on individual tendencies towards *moderate responses* (the central option) or *extreme responses* (top and bottom options; see Hamamura, Heine, and Paulhus, 2008; Khorramdel, Davier, and Pokropek, 2019), as well as to norms related to the appropriateness of particular feelings (Diener, Oishi, Lucas, et al., 2003). In such models, psychologists have emphasized the role of *personality* and *culture* in shaping subjective wellbeing reports (e.g., Diener, Oishi, Lucas, et al., 2003; Oishi, 2010; Heine et al., 2002; Hamamura, Heine, and Paulhus, 2008). Although this branch of *item response theory is* not directly applicable to the interpretation of individual questions with largely-numeric response scales as used in many life evaluation questions, a related approach models the *focal value rounding* tendency of individuals answering just one life evaluation question, based on their known personal characteristics, such as education (Barrington-Leigh, 2024). A number of other recent studies have begun to reexamine the possibility of bias due to differences in reporting functions (Bond and Lang, 2019; Kaiser and Vendrik, 2023; Oparina and Srisuma, 2022; Goff, 2025). Such bias, if empirically significant, would be a threat to the possibility of ranking subgroups' mean life evaluations, as well as to make inference about the determinants of quality of life.

Despite these efforts, no obvious problematic pattern in responses has been isolated and found to be large enough to disrupt the global rankings, especially at its top. Moreover, country averages of life satisfaction tend to be relatively well explained by differences in other measured circumstances which are *not* thought to be subject to the same potential problem (Exton, Smith, and Vandendriessche, 2015).

This paper presents evidence of striking gaps in our understanding of international rankings of life evaluations. It begins by revisiting a comparison of two forms of life evaluation questions in the Gallup World Poll. The comparison is then made for a recent survey, the Global Flourishing Study, which poses the same two questions. These four measures are compared with one life evaluation question from a third source, the World Values Survey. Section 2 describes these surveys, the life evaluation questions, and the construction of matched samples. Section 3 presents pairwise comparisons of country ranks and argues that these qualitatively refute seven possible, relatively straightforward hypotheses about why incongruities might arise when comparing ranks from different questions or surveys. Section 4 investigates the shape of response distribution in the 10- or 11- point scales, finding further evidence of nontrivial problems in the reproducibility of distributions of national samples of life evaluations. Section 5 shows examples of the bivariate (joint) response distributions from individuals who answered two life evaluation questions in the same survey, and Section 6 returns to the

approach of modeling responses in order to compare estimated coefficients on predictors across two life evaluation questions. Section 7 concludes.

2 Data

Three survey sources provide the data for this study. Wave 4 (1999–2004), Wave 5 (2005–2009), Wave 6 (2010–2014), and Wave 7 (2017–2022) of the **World Values Survey** (**WVS**, 76 countries; see Inglehart et al., 2020) conducted in one calendar year per wave in each country, are freely available for research. Annual waves of the repeated cross-section **Gallup World Poll** (**GWP**, 157 countries; see Gallup Organization, 2012; Gallup, 2014) from 2006--2022 were purchased from Gallup. The first wave of the **Global Flourishing Study** (**GFS**, 22 countries; see Johnson et al., 2024) in 2023 is freely available for research.

The WVS poses the **life satisfaction question** (**LS**) on a 1–10 scale, while the other two surveys use 0–10 scales for both LS and **Cantril ladder** (**CL**). The GFS poses both questions to all respondents. However, the GWP, which always includes CL, only fielded the LS question between 2007 and 2010, and only in a subset of countries. Indeed, only 114 countries ever received both questions in the GWP, and, except for five (Belgium, Belarus, Denmark, Sri Lanka, and Singapore) which received LS twice,

each received it only once. The number of countries with LS as well as CL in the GWP were 37 in 2007, 67 in 2008, and just 9 in 2009 and 6 in 2010.

Matched pairs

In order to remove time variation as much as possible in what follows, matched pairs are constructed for comparisons of life evaluation samples across countries.

For comparing LS and CL from the GWP, all individuals who have answered both are selected for each country. As mentioned above, these are all in one year per country except for five countries. When WVS (LS) is compared with CL from GWP, only exact year matches are used. When LS is compared between WVS and the GWP, the following approach is used: for each country available in the WVS, the closest WVS year available to the GWP year with LS responses is chosen. When WVS (LS) is compared with either measure from the GFS, the most recently available wave from WVS is used for each country in common. For comparing CL between the GWP and the GFS, the most recent year available from 2021 (5 countries) or 2022 (17 countries) is used for the GWP.

Question order

Because of their superlatively open scope, and in order to minimize framing effects, life evaluation questions are often placed at the very beginning of a survey. In the GFS, the Cantril ladder question opens the survey, phrased as follows:

> Please imagine a ladder with steps numbered from zero at the bottom to ten 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? [0=Worst possible, 10=Best Possible]

Immediately following it is a question on the same scale about life in 5 years, followed by a question about happiness, followed by the life satisfaction question:

> Overall, how satisfied are you with life as a whole these days? [0=Not satisfied with your life at all, 10=Completely satisfied with your life]

However, according to the questionnaire, the interviewer introduces the happiness and life satisfaction questions, along with several subsequent ones, by reiterating the vertical ladder image for use in the response scale: "Now continue to think about the ladder with the top of the ladder at ten being the best possible state or arrangement and the bottom of the ladder at zero being the worst possible state or arrangement" (Johnson et al., 2024). This is an unusual approach.

In the GWP, the survey opens with four questions about household geography and size, followed by a dichotomous satisfaction question, and then the Cantril ladder, worded nearly identically to the GFS version. By contrast with the GFS, when the life satisfaction question was fielded, it appeared near the end of the survey, but before questions on household income. It is phrased:

> All things considered, how satisfied are you with your life as a whole these days? Use a 0 to 10 scale, where 0 is dissatisfied and 10 is satisfied.

In the WVS, life satisfaction is asked near the end of the first long module, roughly 20% of the way through the interview, and phrased very similarly to that of the GWP:

> All things considered, how satisfied are you with your life as a whole these days? Please use this card to help with your answer. [1 'Dissatisfied', ... 10 'Satisfied']

Other variables

In Section 6, survey questions are included in predictive models of life evaluations. Each asks about an objective fact concerning respondents' lives. These variables are detailed in Appendix Table A3.

3 Country ranks

There are several sensible pairwise comparisons of country rankings among the five life evaluation sources described above. These comparisons are summarized in Table 1 and detailed below.

Surveys	Metrics	r(ranks)	r(indiv)	Fraction Δ>1 quartile	N
WVS (2006–2022) vs GWP (2006–2022)	LS vs CL	0.66		0.37	76
GWP (2007–2010)	LS vs CL	0.92	0.63	0.06	124
GWP (GFS countries, 2006–2022)	LS vs CL	0.94	0.62	0	22
GFS (2023)	LS vs CL	0.59	0.56	0.27	22
GWP (2021–2022) vs GFS (2023)	CL	0.80		0.14	21
GWP (2007–2010) vs GFS (2023)	LS	0.23		0.59	17
WVS (2001–2022) vs GFS (2023)	LS vs CL	0.75		0.19	21
WVS (2001–2022) vs GFS (2023)	LS	0.56		0.24	21
WVS (2006–2022) vs GWP (2007–2010)	LS	0.69		0.28	74

Table 1. Summary of correlations for rank comparisons. The table lists survey year ranges resulting from each match, the correlation of ranks, the fraction of countries for which ranks differ by more than a quartile, and number of countries in the matching pair, and, for the surveys which posed both questions, the correlation across individuals between LS and CL reported values.

The World Values Survey and Gallup World Poll

To start, Figure 1 compares the rank order of country means from the two most prominent global surveys of life evaluations. This is a comparison of different questions from different surveys: the WVS' life satisfaction (LS) question and the GWP's Cantril ladder (CL). The overall correlation between the ranks is 0.66, yet there are very stark inconsistencies. The group of countries that has been somewhat famously at the top of the World Happiness Report's annual ranking is not at the top according to the life satisfaction question in WVS. Moreover, there appear to be systematic differences across cultural groups between the two data sets. Colors show Inglehart and Baker's (2000) classification of countries into nine cultural groups. Table 2 quantifies these systematic differences across country groups.[1] The difference in means is small for English Speaking countries and Protestant Europe, slightly higher for Catholic Europe, increasingly higher for Orthodox, Confucian, and Islamic, and truly large (≥ 1.7) for Latin America, Africa, and South Asia.

Before investigating what might be leading to some of the extreme discrepancies, by looking at response distributions within a country, let us see how other surveys and questions compare. The disagreements in rank shown in Figure 1 could in principle be due to differences in how the *surveys* are implemented (question order, framing, interview mode, sampling, etc) or in how people interpret and respond to the two different *questions*.

To address the former possibility, Figure 2 compares rankings calculated from re-

[1]The means and confidence intervals for differences in ranks are calculated by Monte Carlo simulation, assuming normally distributed beliefs about the mean of each country.

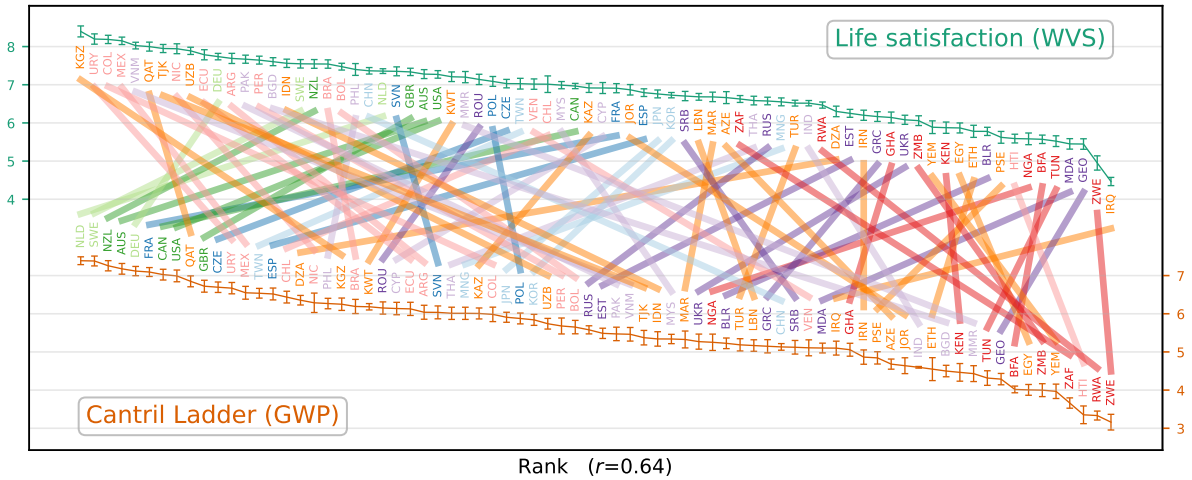


Fig. 1. Comparison of country ranks using mean life satisfaction from the World Values Survey (WVS) and mean Cantril ladder from the Gallup World Poll (GWP). Countries are labelled using their ISO3 codes, and colored according to the cultural groupings shown in Table 2. The range of source years is shown in Table 1.

Life satisfaction (WVS) versus Cantril ladder (GWP)

Country group	ΔRank (%ile)	ΔMean
Africa	$3.3_{\pm 1.0}$	$1.7_{\pm 0.04}$
Catholic Europe	$-21.0_{\pm 1.6}$	$0.6_{\pm 0.04}$
Confucian	$-9.9_{\pm 1.7}$	$1.0_{\pm 0.04}$
English speaking	$-26.1_{\pm 1.1}$	$0.2_{\pm 0.04}$
Islamic	$4.8_{\pm 0.7}$	$1.4_{\pm 0.03}$
Latin America	$15.0_{\pm 1.1}$	$1.7_{\pm 0.03}$
Orthodox	$-14.1_{\pm 1.2}$	$0.9_{\pm 0.03}$
Protestant Europe	$-19.3_{\pm 1.3}$	$0.3_{\pm 0.04}$
South Asia	$18.8_{\pm 1.3}$	$1.9_{\pm 0.03}$

Table 2. Average country rank differences by cultural group for life satisfaction responses from WVS, as compared with Cantril ladder from the GWP. Positive ΔRank means a higher ranking. See Figure 1 for rankings.

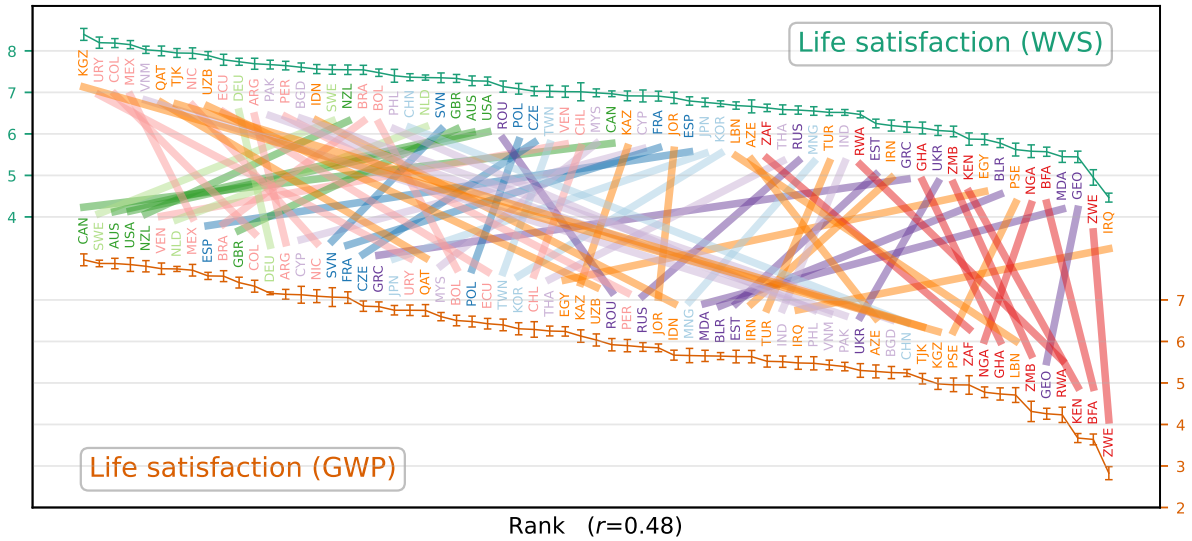


Fig. 2. Comparison of country ranks using mean life satisfaction from the WVS versus the GWP.

Life satisfaction: WVS versus GWP

Country group	ΔRank (%ile)	ΔMean
Africa	$8.5_{\pm 0.6}$	$1.8_{\pm 0.04}$
Catholic Europe	$-22.4_{\pm 1.5}$	$0.0_{\pm 0.04}$
Confucian	$-3.7_{\pm 1.6}$	$0.8_{\pm 0.04}$
English speaking	$-31.3_{\pm 1.2}$	$-0.5_{\pm 0.04}$
Islamic	$13.3_{\pm 0.8}$	$1.2_{\pm 0.03}$
Latin America	$7.0_{\pm 0.9}$	$0.7_{\pm 0.03}$
Orthodox	$-17.8_{\pm 0.9}$	$0.5_{\pm 0.03}$
Protestant Europe	$-14.2_{\pm 1.6}$	$-0.0_{\pm 0.03}$
South Asia	$17.6_{\pm 1.3}$	$1.3_{\pm 0.03}$

Table 3. Average differences by cultural group for Figure 2

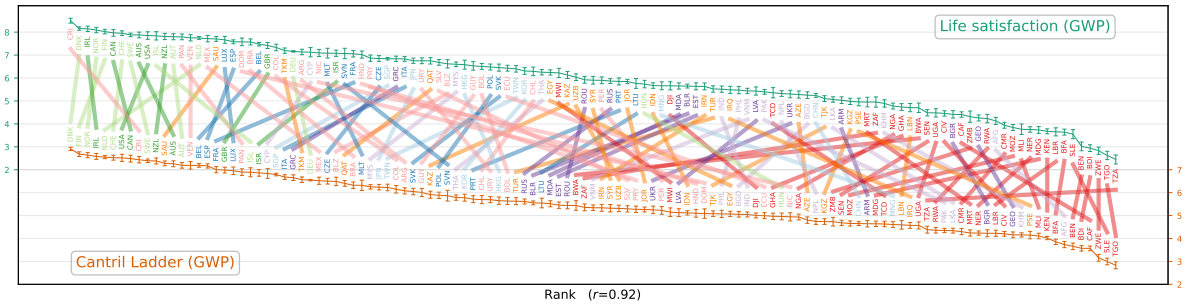


Fig. 3. Comparison of country ranks using mean life satisfaction and mean Cantril ladder, both from the GWP.

Life satisfaction versus Cantril ladder, both from GWP

Country group	ΔRank (%ile)	ΔMean
Africa	-3.9$_{\pm 0.4}$	-0.2$_{\pm 0.02}$
Catholic Europe	-4.8$_{\pm 0.7}$	0.4$_{\pm 0.03}$
Confucian	-0.6$_{\pm 1.0}$	0.5$_{\pm 0.04}$
English speaking	-1.6$_{\pm 0.8}$	0.4$_{\pm 0.04}$
Islamic	-0.4$_{\pm 0.7}$	0.5$_{\pm 0.02}$
Latin America	12.5$_{\pm 0.6}$	1.2$_{\pm 0.03}$
Orthodox	-6.2$_{\pm 0.9}$	0.2$_{\pm 0.03}$
Protestant Europe	1.0$_{\pm 0.6}$	0.5$_{\pm 0.03}$
South Asia	0.9$_{\pm 0.7}$	0.5$_{\pm 0.03}$

Table 4. Average differences by cultural group for Figure 3

sponses to the *same* LS question but asked in the two different repeated survey series of Figure 1. The matching of survey years in Figure 2 is not as precise as in Figure 1; see Data. Overall, the correlation of ranks across countries, 0.69, is very similar and moderately strong, yet the disagreements in rank order are once again dramatic and statistically significant and again vary by cultural group. Table 3 shows these systematic differences across groups. In addition to the systematic differences in rank, African countries responded in WVS on average 1.8 higher on the 0–10 scale than when faced with the same question in GWP.

Could both Figure 1 and 2 be explained instead by differences in the administration of the *surveys*? Comparing responses to the two questions answered by the same individuals in the same survey is possible for the GWP using the years 2007–2010, and the resulting country rankings are compared in Figure 3. Here the correlation of country ranks, 0.92, is much higher, indicating relative consistency. Nevertheless, some individual countries — and the majority of the Latin American ones — still exhibit large systematic differences (Table 4).

The Global Flourishing Study

In 2024 data from the first wave of a new global survey, the Global Flourishing Study (GFS), became available. The GFS also fielded both LS and CL life evaluation questions. Figure 4 compares ranks for the 22 countries in the GFS, when average life

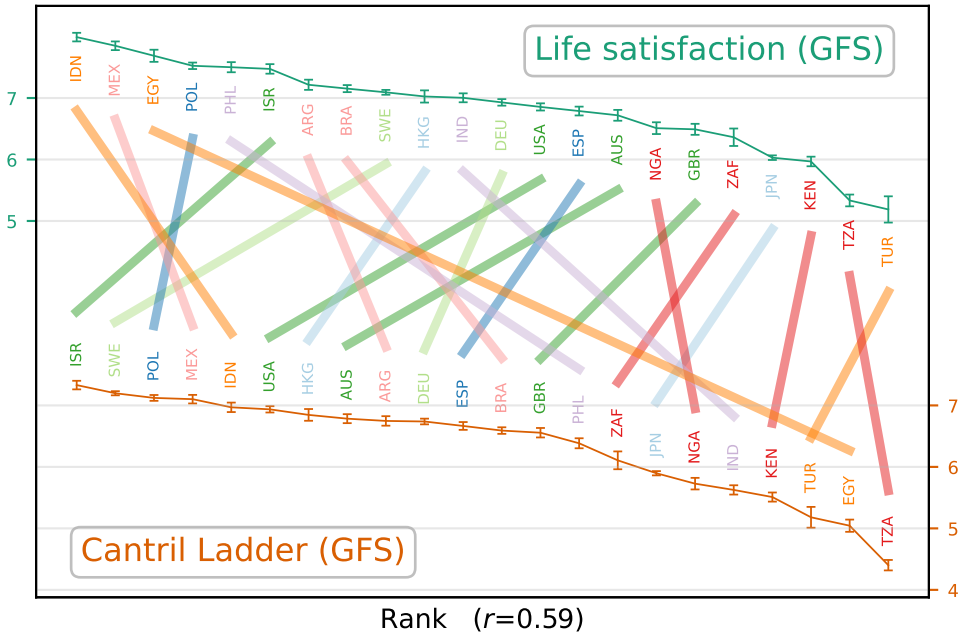


Fig. 4. Life satisfaction versus Cantril ladder, both from GFS.

evaluations are calculated using the two different questions. In this case, the correlation of ranks at the country level is only 0.57. With fewer countries, generalizing across cultural or other groups of countries is harder, but in Table 5 we see that there are nevertheless statistically significant differences by country group. Undoubtedly, there are some stark differences, including that of Egypt, which is near the top of the LS scale but near the bottom of the CL scale.

In the absence of other evidence, Figure 4 might once again lead one to form the hypothesis that differences in the interpretation of the two life evaluation questions are behind the incongruity of ranks. Under this explanation, differences in ranks would reflect substantial differences across countries in certain dimensions of life which also mattered differently for determining answers to the two questions. However, a straightforward interpretation of this sort can again be rejected using GFS data, as it was above for the WVS – GWP pair. Using the most recent wave of the WVS to compare with GFS's 2023 life satisfaction responses (Figure 5), we find a nearly identical, low correlation of ranks and equally stark discrepancies at the individual country level, even though in this case both rankings are based on the life satisfaction question.

Even more remarkably, when life satisfaction from WVS is compared with Cantril ladder (rather than life satisfaction) from GFS, the agreement is better, with a correlation of 0.75 (Figure 6). Egypt, which is one of the extreme cases in Figures 4 and 5, is, in Figure 6, in good agreement across these differing measures on different surveys.

One final comparison is shown in Figure 7. When the Cantril ladder responses are compared from GWP and GFS, the agreement is relatively good, with an overall correlation of 0.80. This is the kind of consistency one might hope for when publishing international rankings of these measures, although even if this were the full picture one

Life satisfaction versus Cantril ladder, both from GFS

Country group	ΔRank (%ile)	ΔMean
Africa	-2.7$_{\pm 0.9}$	0.6$_{\pm 0.04}$
Catholic Europe	-10.1$_{\pm 2.4}$	0.3$_{\pm 0.03}$
Confucian	-14.7$_{\pm 2.6}$	0.2$_{\pm 0.04}$
English speaking	-26.6$_{\pm 1.7}$	-0.0$_{\pm 0.03}$
Islamic	32.3$_{\pm 0.9}$	1.2$_{\pm 0.06}$
Latin America	12.6$_{\pm 1.9}$	0.6$_{\pm 0.03}$
Protestant Europe	-22.8$_{\pm 2.2}$	0.0$_{\pm 0.02}$
South Asia	38.8$_{\pm 2.4}$	1.2$_{\pm 0.04}$

Table 5. Average differences by cultural group for Figure 4

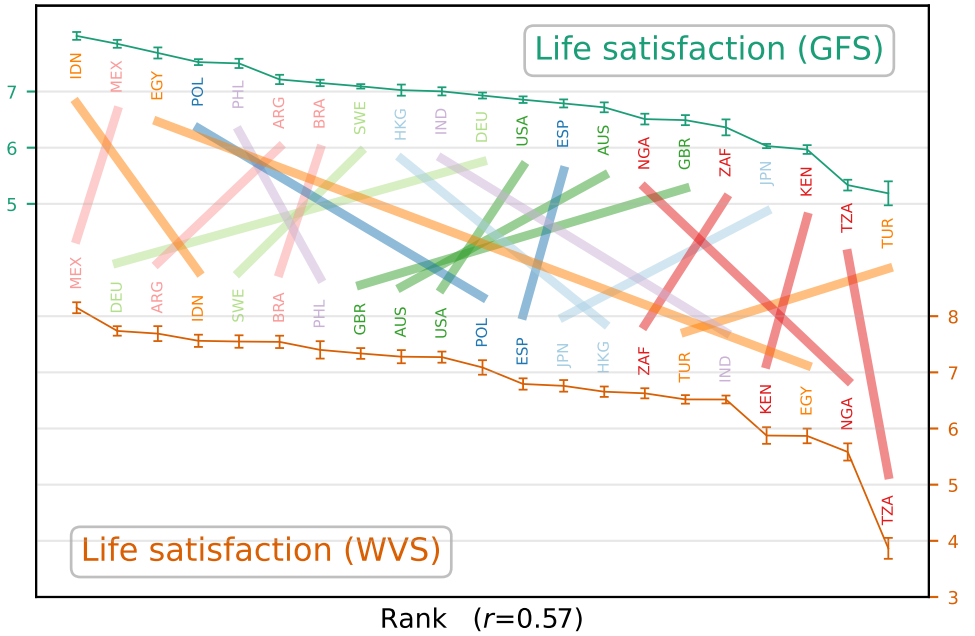


Fig. 5. Rankings calculated using life satisfaction from WVS versus GFS. See Table A1 for average differences by country group.

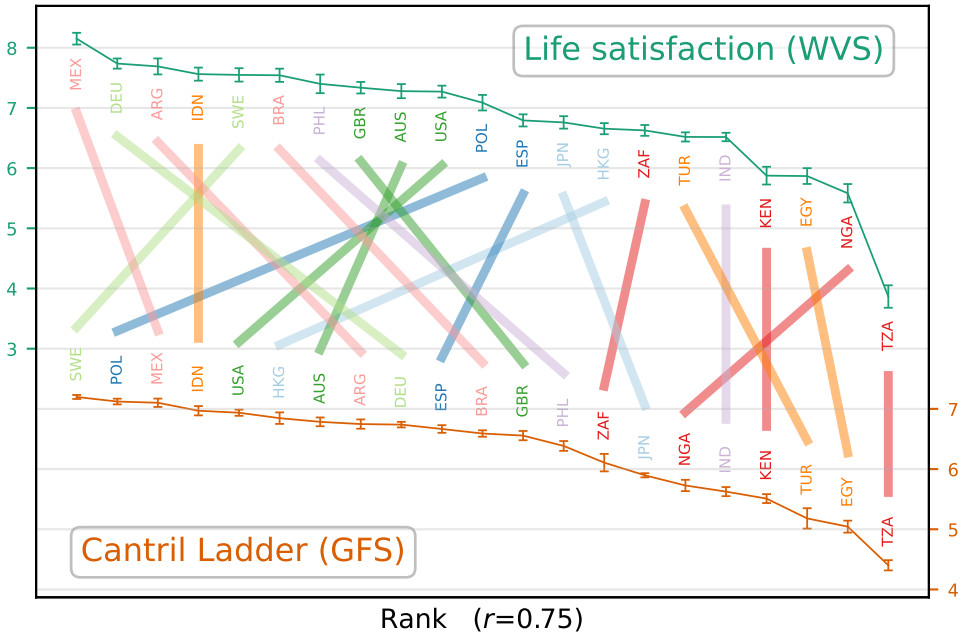


Fig. 6. Ranks calculated from life satisfaction (WVS) versus Cantril ladder (GFS). See Table A2 for average differences by country group.

Cantril ladder (GWP versus GFS)

Country group	ΔRank (%ile)	ΔMean
Africa	-1.2$_{\pm 1.1}$	0.9$_{\pm 0.06}$
Catholic Europe	4.5$_{\pm 3.3}$	0.4$_{\pm 0.05}$
Confucian	10.3$_{\pm 2.6}$	0.6$_{\pm 0.05}$
English speaking	-13.3$_{\pm 1.9}$	-0.1$_{\pm 0.04}$
Islamic	11.0$_{\pm 1.7}$	1.1$_{\pm 0.07}$
Latin America	-0.9$_{\pm 2.7}$	0.3$_{\pm 0.05}$
Protestant Europe	-5.2$_{\pm 2.9}$	-0.1$_{\pm 0.06}$
South Asia	4.3$_{\pm 1.5}$	1.1$_{\pm 0.07}$

Table 6. Average differences by cultural group for Figure 7

would need to investigate the several statistically significant exceptions.

Discussion

We might characterize the preceding evidence as addressing the following hypotheses:

1. Major life evaluation metrics are generally consistent with each other at the country level.

2. The two major life evaluation questions tap into different characteristics of life to different degrees, but consistently across surveys.

3. The two major life evaluation questions are interpreted differently across countries, but consistently across surveys.

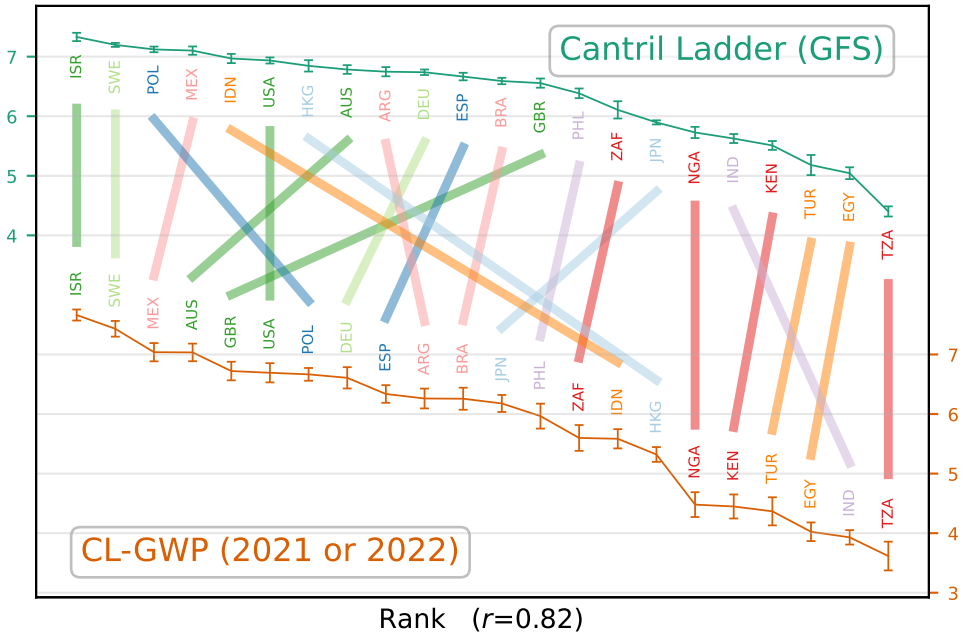


Fig. 7. Ranks calculated from Cantril ladder (GFS versus GWP).

4. The three major surveys have differences in implementation and survey content or framing which induce distinct but consistent impacts on any life evaluation questions.

5. The two major life evaluation questions tap into different characteristics of life to different degrees, but when both life evaluation questions are posed in the same survey, that which is posed first will have a dominant influence on the responses to both.

6. Large differences in life evaluations reported across surveys can usually be explained by the countries in question undergoing rapid change, coupled with subtle differences in survey timing.

The first is rejected by any of Figures 1–7. Hypothesis 2 is rejected by the strong disagreements shown in Figure 2 and Figure 5 and, arguably, the weaker disagreements in Figure 7. Hypothesis 3 is rejected by the same evidence, as cultural or country-specific effects qualify as unmeasured life conditions. If there was a problem translating the questions, or different conceptions of the two questions across countries, then we might still expect to see consistency across surveys asking the same question with the same population sampling frame. On the contrary, we see stark differences across surveys.

Hypothesis 4 predicts that GWP and GFS, which have fielded both questions in the same survey, should find consistent results from the two questions. This is patently false in the GFS (Figure 4) and, to a somewhat lesser degree, the GWP (Figure 3), which exhibits a strong overall correlation despite some coherent cultural differences.

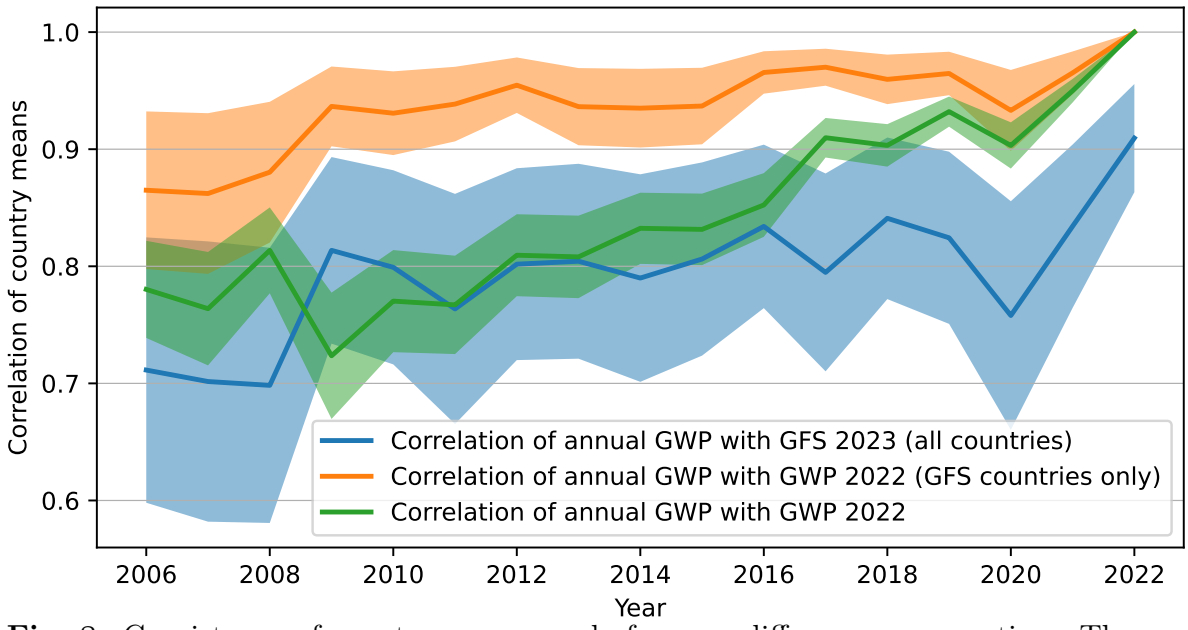


Fig. 8. Consistency of country means, and of survey differences, across time. The green trace shows the correlation of each year's Cantril ladder country means with those from 2022. The orange trace is the same but restricted to the 20 countries surveyed in the GFS. The blue trace is the correlation of those GWP annual country means with the GFS mean responses of 2023. The high values of the orange trace prior to 2022 show that measured means are highly reproducible. Comparing the blue and orange traces shows that GWP values are consistently different from those of GFS.

One way in which those surveys could differ is in the order of and framing around the two questions. However, if one question influences the other, it does not seem to be in any simple way as suggested in Hypothesis 5. In both the GFS and the GWP, CL is asked before LS. The GFS poses the LS question nearly immediately after the CL question, and indeed even refers to using the same scale to answer about LS. In spite of this, the answers in the GFS disagree more strongly than in the GWP, which posed the CL question near the beginning and the LS question near the end of the survey, with only questions about income and some demographics following it. Below I investigate the joint distribution of responses across individuals more explicitly.

Hypothesis 6 represents an objection that even if two surveys are carried out within the same calendar year, outlier results could arise due to rapid changes in circumstances within countries. To better address the temporality of the discrepancies highlighted so far, Figure 8 shows that differences between survey measures are persistent over time. Focusing on the 22 GFS countries, the orange and blue traces show that past GWP country means have been consistently different from the recent GFS values, but consistently similar to the recent GWP values. Thus, the scale of discrepancies across surveys is larger, and broader, than the changes in mean country responses over time.

In light of the rejection of (1) – (6), it appears that any explanation of the non-

reproducibility of country ranks must involve the interpretation or reporting behavior varying not only between the two wordings of questions, but *simultaneously* across survey contexts and *simultaneously* across countries.

4 Univariate response distributions

The straightforward evidence above is based on mean responses at the country level, and appears to reject any simple explanation of the differences between life satisfaction and the Cantril ladder. A next natural step in descriptive evidence is to look at the individual response distribution and how it varies across questions, surveys, and countries.

Consider the first pair of columns in Figure 9, which shows responses from a selection of countries, each notable for at least one large discrepancy in the previous section. The distributions are qualitatively different, with some systematic patterns across countries. In most cases, the distribution for CL in the GWP is shifted to the left as compared with that of LS in the WVS. In addition, in most countries (but not in South Africa, and hardly in the English speaking countries) there is a more pronounced enhancement of the "5" response for CL (GWP) than for LS (WVS). In some cases, this is quite extreme, for instance in the Tajikistan case. These qualitative differences lead to lower means for the CL, but the differences in means (tabulated in Table 2) vary considerably. For instance, it is an enormous +3.1 for Bangladesh, +2.5 for Tajikistan, +2.4 for China, +2.3 for Indonesia, and +2.2 for Colombia, but only 0–0.1 for Canada and the USA.

The tendency to prefer the "5" response seems in most cases to coincide with a relative preference for the bottom (0 or 1) and top (10) values as well. In fact one might hypothesize that the prevalence of 5 in the CL (GWP) could be entirely explained by an overall lower distribution of latent (i.e., true) wellbeing, coupled with the focal value rounding (FVR) response function behavior identified and characterized by Barrington-Leigh (2024). For instance, in Indonesia, both the CL (GWP) and LS (WVS) distributions show enhancements both for 5 and for 10. However, with distributions centered around 5 and 7–8, respectively, there are few 10s for CL and few 5s for LS.

In some cases, however, there appear to be countries without much FVR (South Africa) or where the enhancement for "10" seems less strong in LS (WVS) than might be expected based on that of "5" in CL (GWP) — for instance, in Japan, Bangladesh, Czech Republic, and Poland.

Figure A1 shows the same columns but with all countries now pooled across cultural groups. The patterns in the CL (GWP) and LS (WVS) responses, described above for individual countries, generalize to these pooled distributions.

Returning to Figure 9, the next two columns compare LS and CL from matched years in the GWP. They show that the qualitative differences just described are not intrinsic to the question wording. That is, (as foreshadowed by Figure 3) response patterns are quite similar in the two questions when they are asked together in the

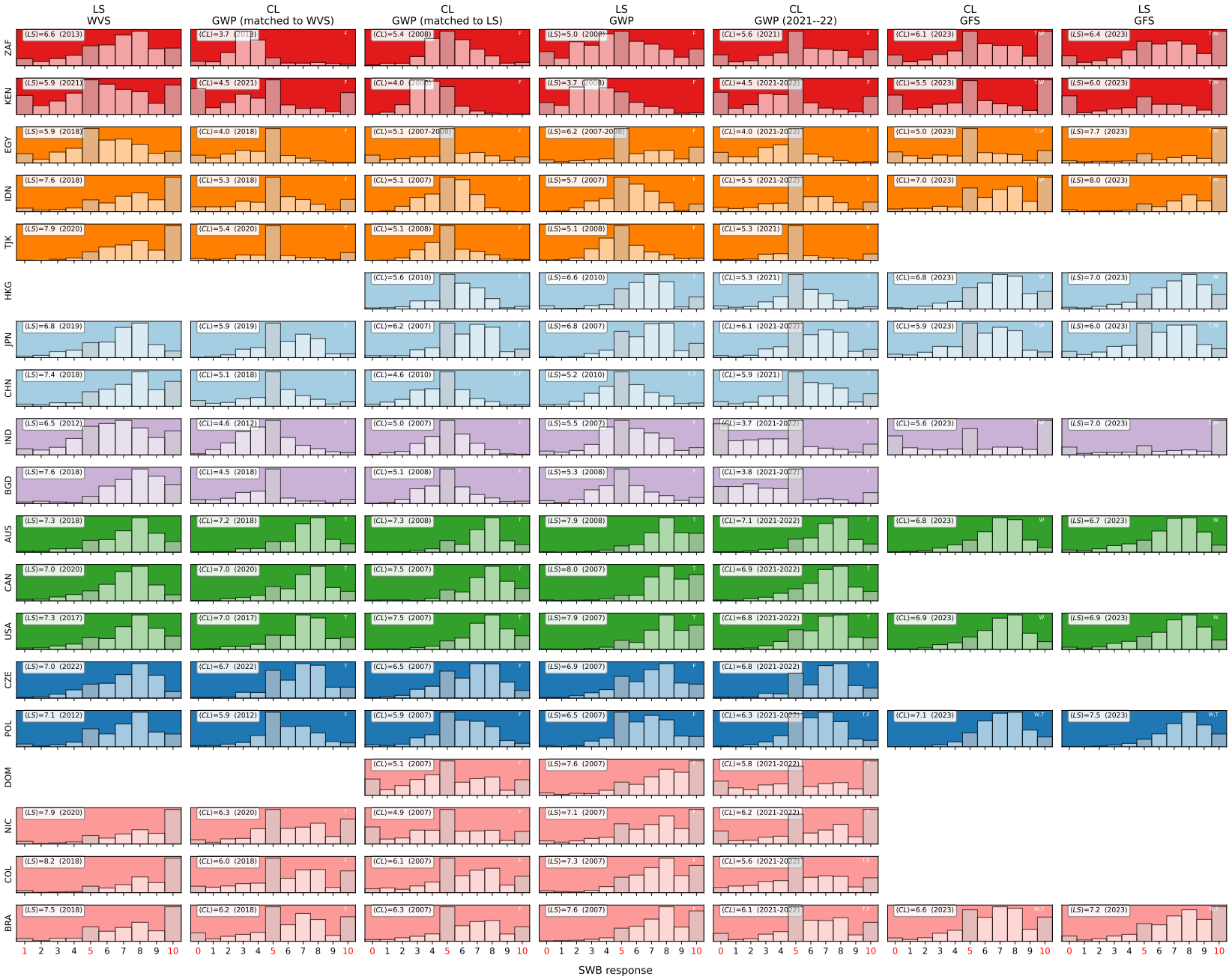


Fig. 9. Response distributions to life satisfaction (LS) and Cantril ladder (CL) from the WVS, GWP, and GFS, for a selection of countries.

same survey. Although FVR varies markedly across countries, it is exhibited similarly across this pair of questions. However, there are exceptions even to this generalization — for instance, the Latin American countries and maybe South Africa. For Latin America, the prevalence of 5s in the CL (GWP) responses and to a lesser degree, 10s in the LS (GWP) responses may account for the larger average difference in means (Table 4 and Figure A1).

Some countries (Kenya, Indonesia, Tajikistan) exhibit qualitative differences between the year matched to WVS (column 2) and that matched to LS (column 3), though these are unusual. Figure A2 addresses the question of reproducibility of qualitative and quantitative patterns across years within a single survey, and shows generally stable or slowly changing distributions for CL in the GWP over time.

Turning next to the analogous comparison for GFS, columns 6 and 7 (the rightmost two columns) in Figure 9 compare responses to LS and CL, once again from the same individuals in the same survey. The relationship between these two columns is different from that of the corresponding pair for the GWP. There is more of a shift to the right for LS in India and Egypt, as compared with the wider distribution for CL. As a result of the very strong FVR, this shift gives Egypt a stark difference (2.7) in means. The strong FVR also leads to exceptionally high mean response for LS in Egypt and Indonesia. It is not clear by inspection whether the degree of FVR can be said to vary across questions, for instance in India and to some extent in South Africa and Brazil. Figure A1 shows that these shifts and strong FVR are features that generalize to the rest of the cultural group for South Asia and Islamic countries.

By examining columns 5 and 6 of Figure 9, a final comparison may be made between CL in the GFS (in 2023) and in the GWP (2021–2022). Some cases, like Japan, look remarkably similar, while others, like Hong Kong and Indonesia, look qualitatively dissimilar. In most countries except for the English Speaking group, however, the use of "10" in CL responses is lower in the GWP than in the GFS. These patterns are evident also in the pooled data of Figure A1.

5 Joint response distributions

The previous section, focusing on individual response distributions, shows that differences across question wording or across surveys, and indeed over time (Figure A2), all vary by country, and that in some cases these differences are consistent within cultural groups of countries. Because respondents in the the GFS and early waves of the GWP answered both forms of the life evaluation question, there is one other kind of descriptive evidence to examine for hints to the mysteries presented so far: the joint response distribution across the two questions. Rather than pursuing these exhaustively, Figure 10 presents a small selection of six countries from the GFS, which serves to illustrate the complexity of response behavior.

In some countries, including the anglophone nations (not shown), Japan, Argentina and, to a lesser extent, South Africa, respondents tend to give the same or similar

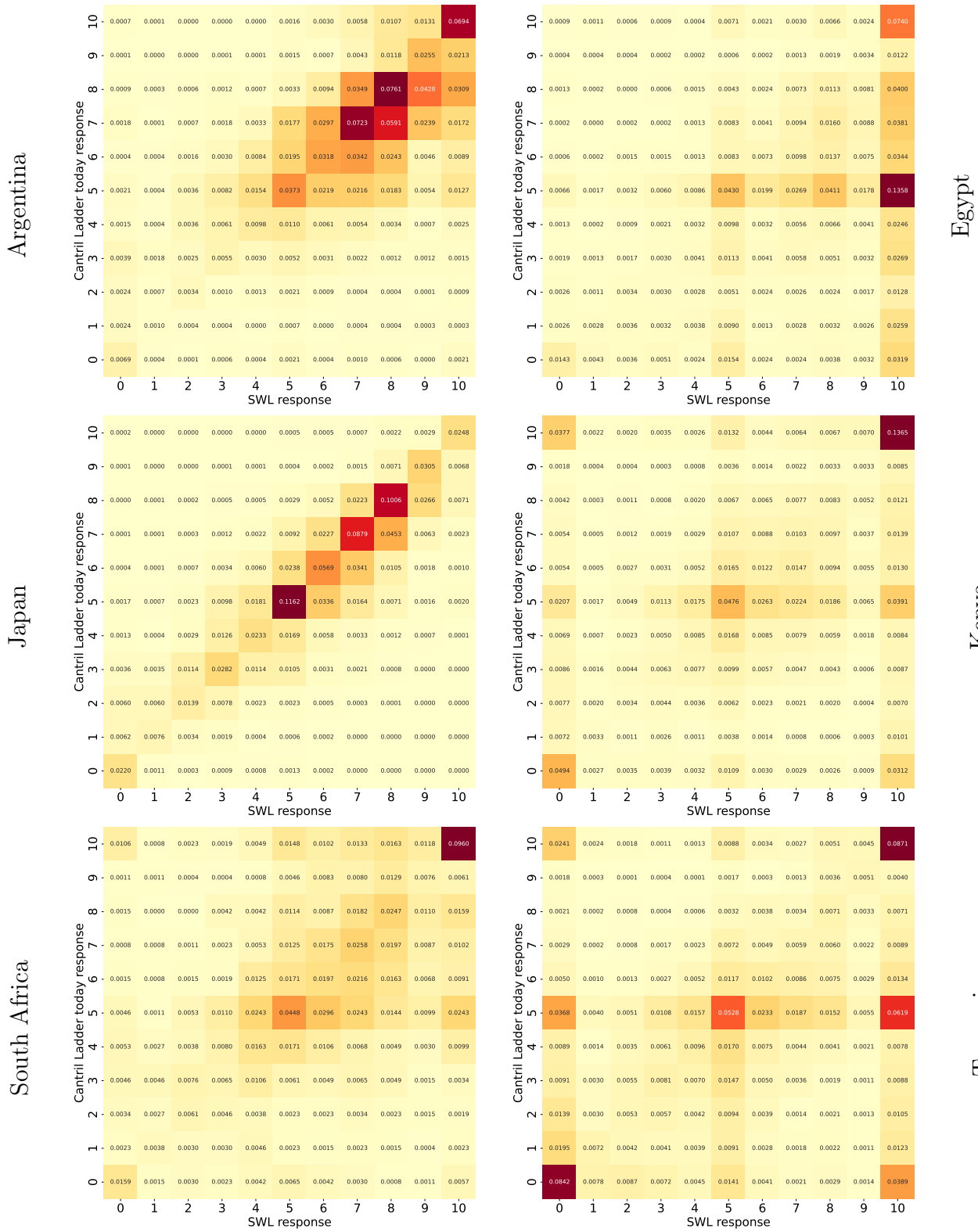


Fig. 10. Cross-tabulations for responses to the life satisfaction question (horizontal axis) and Cantril ladder (vertical axis) for six countries in the GFS (2023), in which respondents answered both questions. The fraction of total responses is noted in each cell. Within each grid (country), darker colors correspond to higher fractions.

responses to the two questions. In Egypt, by contrast, no matter what answer a respondent gave for Cantril ladder, their most likely response for life satisfaction was "10". Conversely, no matter what response they gave for life satisfaction — with the exception of "0" — their most likely response for Cantril ladder was "5". Fourteen percent of respondents accordingly responded with "5" and "10" for the two questions. For those who gave a "10" for life satisfaction, nearly half as many gave a "0" as gave a "10" for Cantril ladder.

While, as mentioned earlier, one might hypothesize a combination of some kind of offset between answers to the questions, combined with strong FVR, to explain the behavior in Egyptian responses, the situation is still stranger in Kenya and Tanazania. There, not only is the preference for 0, 5, and 10 highly dominant, but the answers from an individual are nearly half as likely to be opposite as the same across questions. For example, in Tanzania, among those who responded with "0" for Cantril ladder, the most common response to the life satisfaction question was also "0", yet nearly half as many responded with "10". Among those who responded with "5" for Cantril ladder, similar numbers responded with each of 0, 5, and 10 for life satisfaction.

6 Model inference

Many departures from the ideal in how people respond to subjective wellbeing questions have been identified and found to introduce noise or even bias into responses, yet shown not to be large enough to threaten the usefulness of these data (Krueger and Schkade, 2008; Cheung and Lucas, 2014; Lucas and Lawless, 2013; Lucas, Freedman, and Cornman, 2018; Ferrer-i-Carbonell and Frijters, 2004). As mentioned in the introduction, analysts are most interested in the marginal effects estimated from models accounting for differences or changes in life evaluations across large samples. A significant difference in average responses to LS and CL need not imply a difference in coefficients of interest estimated in an explanatory model. For example, a detectable bias in (influence on) average life evaluation from something like the current weather might introduce noise that is uncorrelated with other causal influences of interest. Even if such differences varied somehow by culture or language, they might not threaten inference on the importance of physical security, positive relationships with friends and family, productive employment, and so on. Thus, the question arises: Are the anomalies identified so far in this paper a threat to the kinds of inference needed to inform policy?

Helliwell et al. (2010, see columns 2 and 3 in Table 10.1) already compared coefficients estimated using the same model and set of respondents to explain life satisfaction or Cantril ladder from the GWP. This was a relatively rich model, with both individual and country-level explanatory factors, yet coefficients match quite closely between the two outcomes variables.

In a similar spirit, estimates for the two life evaluation questions in the GFS are compared below. In order to assess whether LS and CL provide the same guidance about marginal effects, a selection of relatively objective life circumstances contained

in the GFS is used as explanatory variables. These are: log of household income; two education indicators for at least secondary and post-secondary; gender; a quadratic of age; two marriage status indicators; indicators for whether the respondent lives in an urban area, is unemployed, has donated to charity in the last month, drinks alcohol, and prays to a deity; a measure of how many days per week the respondent exercises, and the log of the frequency which which they attend religious services (see Table A3 for descriptions).

Individual respondents' life evaluations, either LS or CL, are modeled as a continuous normal variable,

$$y_i \sim \mathcal{N}\big(\alpha_j + \mathbf{X}_i^\top \boldsymbol{\beta}_j,\, \sigma\big) \tag{1}$$

where individual i lives in country j, and $\mathbf{X}_i$ is a vector of K individual characteristics. Intercepts α_j and slopes $\boldsymbol{\beta}_j$ vary by country (sometimes called *random effects*) but are estimated with Bayesian "partial pooling" in order to achieve the most precise and reliable estimates across countries. To accomplish this, the country-level parameters are drawn from global (Normal) distributions:[2]

$$\begin{aligned} \alpha_j &\sim \mathcal{N}(\mu_\alpha,\, \sigma_\alpha), \\ \beta_{jk} &\sim \mathcal{N}(\mu_{\beta_k},\, \sigma_{\beta_k}), \quad k = 1, \ldots, K, \end{aligned}$$

The partial pooling means that, when there is commonality across countries, data from all countries can inform the estimates of the parameters in these distributions. The "hyperparameters" characterizing those distributions are in turn drawn from the following starting distributions:

$$\begin{aligned} \mu_\alpha &\sim \mathcal{N}(0,\, 5), \\ \sigma_\alpha &\sim \text{HalfNormal}(0, 5), \\ \mu_{\beta_k} &\sim \mathcal{N}(0,\, 3), \\ \sigma_{\beta_k} &\sim \text{HalfNormal}(0, 3), \quad k = 1, \ldots, K \end{aligned}$$

where the HalfNormal distribution is like the Normal distribution $\mathcal{N}(0, \cdot)$ but 0 for values less than 0.

Lastly, the scalar variance of individual response noise (i.e., the error term), σ, in Eq. (1) is estimated from another HalfNormal distribution as follows:

$$\sigma \sim \text{HalfNormal}(0, 5)$$

The large widths of the hyperpriors, compared with the eventual size of estimated coefficients, start the model relatively uninformed, relying on the large sample size of the GFS data set to achieve good convergence.

[2]Formally, $\boldsymbol{\beta}_j \equiv (\beta_{j1}, \ldots, \beta_{jK})^\top$.

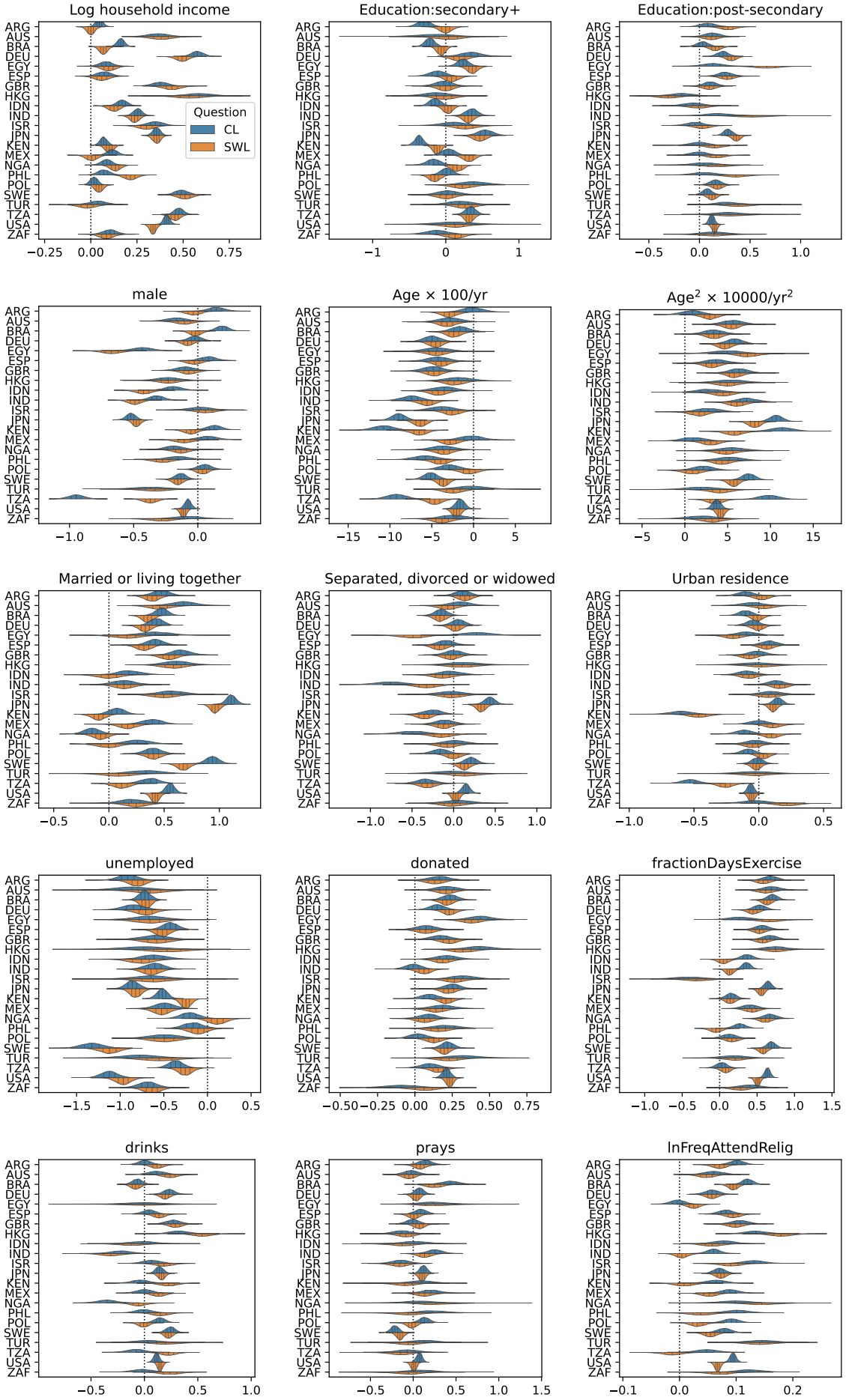


Fig. 11. Estimates from two multivariate models, one for CL and one for LS, both from the GFS.

Figure 11 compares two estimates of the model, one for Cantril ladder and one for life satisfaction. Each subplot compares estimated country-specific coefficients for one predictor across the two models.

The main qualitative finding is that there is a high degree of agreement between the two estimates. Coefficients are in many cases resolved to vary widely across countries, yet they show strong consistency between measures of life evaluation. Even for countries, like Egypt, with highly divergent average responses (see Figures 4 and 9), the structure of determinants (i.e., the coefficients in the model) is relatively consistent. There are, of course, some strong exceptions, such as the coefficient of gender in Tanzania. However, this case of strong statistical exclusion is only one out of 330 pairs (22 countries $\times$ 15 predictors) of estimated coefficients shown.

The particular specification of explanatory variables used for Figure 11 is driven largely by the content of the GFS questionnaire. That is, for the present purposes the interest is more in comparing the two questions rather than evincing policy-relevant marginal effect sizes. Nevertheless, some comment on the magnitude of estimates is appropriate, especially insofar as they reflect the comparability of life evaluation questions across countries. It is important to note that the country-specific marginal effects in Eq. (1) and Figure 11 are estimated allowing for country-specific fixed effects α_j. Therefore, the discrepancies across countries in β_k do not necessarily represent disagreements — for instance, due to cultural differences in the interpretation of life evaluation questions — in the importance of aggregate levels of education, unemployment, and so on at the national level. Rather, the coefficient differences across countries are likely to reflect the environment within each country, which may differ structurally.

Household income is estimated to have a typically-sized effect on wellbeing, though it appears rather muted in some countries. By comparison with the effect of doubling income,[3] the negative impact of being unemployed is enormous — and relatively consistent across countries — as are the positive impacts of marriage and, in most countries, regular exercise. Social activity associated with religion also looms large, as compared with the more individual religious activity of prayer. Education shows a typically small or even negative impact on wellbeing after controlling for other factors, though post-secondary education is significantly beneficial in more developed countries. Overall, age shows a typically U-shaped relationship to wellbeing, through a negative coefficient on age and a positive one on the square of age.

While consistency across life evaluation questions is the dominant pattern, there may be some subtle differences in structure evident using the present model. The estimates suggest in some places a smaller benefit of religious attendance on SWL as compared with CL, as well as a smaller or more negative benefit of secondary education on CL as compared with SWL.

[3] A coefficient of 0.5 implies a life evaluation boost from doubling income of ~ 0.34 on the eleven-point scale.

7 Discussion and Conclusion

Several interconnected questions have been addressed in this study. In more general terms than the hypotheses posed in Section 3, they encompass the following three dimensions: Are the two cognitive life evaluation questions equivalent? Is each question interpreted the same way when asked in different surveys? Is each question interpreted the same way in different countries? The strict answer to each is No, but both the reasons for and the importance of the discrepancies remain unresolved. Clearly, then, this paper does not solve all the puzzles it highlights. By documenting both the robustness and the fragility of global life-evaluation measures, this study presents a fundamental challenge for the comparative study of human well-being and provides a foundation for future methodological advances.

In principle, the factors affecting the interpretation of life evaluation questions are unavoidably complex. They include the conceptual content of a given question wording, the translation of the question into a local language, the framing or context given by the rest of the questionnaire, cultural values around the concept of a good life, cultural norms about self-expression, and other non-cultural influences on the ability to report on the ten or eleven point scale. These issues resonate with a recent resurgence of interest in psychometric and econometric investigations concerning subjective wellbeing (Bond and Lang, 2019; Kaiser and Vendrik, 2023; Oparina and Srisuma, 2022; Goff, 2025).

Generally these efforts hypothesize some underlying functional dependence of well-being on experienced circumstances, along with a reporting function of some kind. However, the joint distributions presented in Section 5 are difficult to reconcile with a single underlying distribution of well-being, even accounting for focal value rounding. Qualitative debriefing of respondents, especially those with unusual response patterns, could shed light on how to interpret and model them.

Amid the considerable degree of reproducibility across questions within surveys and across surveys, a subset of countries appear as outliers in the ranking comparisons. The commonalities within cultural groups suggest a need for more nuanced interpretation, though they do not yet point to a clear prescription for changing how such data are presented. Comparisons within more homogeneous groups would appear to be safer, yet there are cases of neighboring or culturally related countries showing qualitatively distinct response patterns and differences across questions and surveys.

Worldwide, the vast majority of statistical agency surveys asking for life evaluations use the life satisfaction question, in accordance with recommendations from the OECD and U.S. National Academies (OECD, 2013; Stone, Mackie, et al., 2014). Specific criticisms of the Cantril ladder around its propensity to evoke a focus on status and wealth rather than other life dimensions, and to elicit higher reports than life satisfaction (Nilsson et al., 2024), do not fully account for the cultural variation observed here.

Notably, this investigation does not contradict the use of life satisfaction regressions to infer marginal effects of policy-relevant circumstances. Indeed, much of the literature is based on panel studies, in which changes over time to individuals' lives are the object of study. Because panelists' life evaluations are compared with their own earlier

reports, cultural variation in the interpretation of questions is less likely to introduce measurement problems. There is also no reason, based on evidence presented here, to prefer multivariate indices over direct life evaluations as a basis for ranking countries. Life evaluation data can be used to suggest weights denoting the importance of indicators in a dashboard of statistics, but choosing a weighted multivariate index as a primary indicator risks multiplying the uncertainties investigated above, in addition to the accountability problem inherent to non-empirical weighting schemes.

One role that international rankings continue to play in the World Happiness Report is to bring the world's attention to the deeper literature on what contributes to a satisfying life, and to the prospect of improving lives through more human-oriented policy objectives and correspondingly better policies. Thus, the significant challenges for cross-national rankings highlighted in this paper represent a call for new ideas to understand these data more deeply, but do not warrant abandonment of current reporting practices without that understanding. As the OECD revises its international guidelines for measurement of subjective wellbeing (Mahoney, 2023), there is a clear need for renewed innovation, drawing on both qualitative and quantitative approaches, to strengthen confidence in measuring what has been called the ultimate objective of the social sciences.

References

[1] C.P. Barrington-Leigh. "Trends in conceptions of progress and wellbeing". In: *World Happiness Report 2022*. Ed. by John Helliwell et al. Sustainable Development Solutions Network, 2022. Chap. 3, pp. 53–74. URL: https://worldhappiness.report/ed/2022/trends-in-conceptions-of-progress-and-well-being/.

[2] C.P. Barrington-Leigh. "The econometrics of happiness: Are we underestimating the returns to education and income?" In: *Journal of Public Economics* 230 (Feb. 2024). ISSN: 0047-2727. DOI: https://doi.org/10.1016/j.jpubeco.2023.105052.

[3] Jo Blodgett et al. "What works to improve life satisfaction in intervention and observational research: a technical report of two systematic rapid reviews". What Works Centre for Wellbeing. Apr. 2024. URL: https://whatworkswellbeing.org/wp-content/uploads/2024/04/Life-Satisfaction_what-works_Full-technical-report_Final-april-2024.pdf.

[4] Timothy N. Bond and Kevin Lang. "The Sad Truth about Happiness Scales". In: *Journal of Political Economy* 127.4 (2019), pp. 1629–1640. DOI: 10.1086/701679. eprint: https://doi.org/10.1086/701679. URL: https://doi.org/10.1086/701679.

[5] Felix Cheung and Richard E. Lucas. "Assessing the validity of single-item life satisfaction measures: results from three large samples". In: *Quality of Life Research* 23.10 (Dec. 2014), pp. 2809–2818. ISSN: 1573-2649. DOI: 10.1007/s11136-014-0726-4. URL: https://doi.org/10.1007/s11136-014-0726-4.

[6] E. Diener, S. Oishi, R.E. Lucas, et al. "Personality, culture, and subjective well-being: Emotional and Cognitive Evaluations of Life". In: *Annual Review of Psychology* 54.1 (2003), pp. 403–425.

[7] Edward Diener and Eunkook M Suh. *Culture and subjective well-being*. MIT press, 2003.

[8] C. Exton, C. Smith, and D. Vandendriessche. "Comparing Happiness across the World: Does Culture Matter?" In: *OECD Statistics Working Papers* (Apr. 2015). DOI: 10.1787/5jrqppzd9bs2-en.

[9] A. Ferrer-i-Carbonell and P. Frijters. "How Important is Methodology for the estimates of the determinants of Happiness?" In: *The Economic Journal* 114.497 (2004), pp. 641–659. DOI: 10.1111/j.1468-0297.2004.00235.x.

[10] David Frayman et al. *Value for money: How to improve wellbeing and reduce misery*. CEP Reports 44. Centre for Economic Performance, LSE, Sept. 2024. DOI: None. URL: https://ideas.repec.org/p/cep/cepsps/44.html.

[11] Paul Frijters and Christian Krekel. *A Handbook for Wellbeing Policy-Making: History, Theory, Measurement, Implementation, and Examples*. Oxford University Press, May 2021. ISBN: 9780192896803. DOI: 10.1093/oso/9780192896803.001.0001. URL: https://doi.org/10.1093/oso/9780192896803.001.0001.

[12] Paul Frijters et al. "A happy choice: wellbeing as the goal of government". In: *Behavioural Public Policy* 4.2 (2020), pp. 126–165. DOI: 10.1017/bpp.2019.39.

[13] Gallup. *Understanding How Gallup Uses the Cantril Scale*. 2014. URL: http://www.gallup.com/poll/122453/understanding-gallup-uses-cantril-scale.aspx (visited on 06/05/2014).

[14] The Gallup Organization. "World Poll Methodology". (available on request). Apr. 2012.

[15] Leonard Goff. *Identifying causal effects with subjective ordinal outcomes*. Tech. rep. 2025. arXiv: 2212.14622 [econ.EM]. URL: https://arxiv.org/abs/2212.14622.

[16] Takeshi Hamamura, Steven J Heine, and Delroy L Paulhus. "Cultural differences in response styles: The role of dialectical thinking". In: *Personality and Individual differences* 44.4 (2008), pp. 932–942. DOI: 10.1016/j.paid.2007.10.034.

[17] Happiness Research Institute. *Wellbeing Adjusted Life Years: A universal metric to quantify the happiness return on investment*. Berlin, 2020. URL: https://www.happinessresearchinstitute.com/waly-report.

[18] Steven J Heine et al. "What's wrong with cross-cultural comparisons of subjective Likert scales?: The reference-group effect." In: *Journal of personality and social psychology* 82.6 (2002), p. 903.

[19] J.F. Helliwell et al. "International evidence on the social context of well-being". In: *International Differences in Well-Being*. Ed. by E. Diener, J.F. Helliwell, and D. Kahneman. Oxford University Press, 2010, pp. 213–229. URL: https://books.google.ca/books?id=m99aqwLFrGoC&pg=PA291.

[20] John Helliwell and Shun Wang. "The State of World Happiness". In: *World Happiness Report 2013*. Ed. by John Helliwell, Richard Layard, and Jeffrey Sachs. Sustainable Development Solutions Network, 2012. Chap. 2, pp. 10–57. URL: http://worldhappiness.report/ed/2012/.

[21] John Helliwell et al. "Caring and sharing: Global analysis of happiness and kindness". In: *World Happiness Report 2025*. Ed. by John Helliwell et al. Sustainable Development Solutions Network, 2025. Chap. 2, pp. 11–52. URL: http://worldhappiness.report/ed/2025/.

[22] R. Inglehart et al., eds. *World Values Survey: All Rounds – Country-Pooled Datafile*. Madrid, Spain & Vienna, Austria, 2020.

[23] Ronald Inglehart and Wayne E. Baker. "Modernization, Cultural Change, and the Persistence of Traditional Values". In: *American Sociological Review* 65.1 (2000), pp. 19–51. URL: http://search.proquest.com/docview/60084910?accountid=12339.

[24] B. R. Johnson et al. *The Global Flourishing Study*. Feb. 2024. DOI: 10.17605/OSF.IO/3JTZ8. URL: https://doi.org/10.17605/OSF.IO/3JTZ8.

[25] Caspar Kaiser and Maarten Vendrik. "How much can we learn from happiness data?" In: *doi* 10 (2023), p. 31235.

[26] Lale Khorramdel, Matthias von Davier, and Artur Pokropek. "Combining mixture distribution and multidimensional IRTree models for the measurement of extreme response styles". In: *British Journal of Mathematical and Statistical Psychology* 72.3 (2019), pp. 538–559. DOI: 10.1111/bmsp.12179.

[27] Alan B. Krueger and David A. Schkade. "The reliability of subjective well-being measures". In: *Journal of Public Economics* 92.8-9 (Aug. 2008), pp. 1833–1845.

[28] Tim Lomas et al. "Insights from the First Global Survey of Balance and Harmony". In: *World Happiness Report 2022*. Ed. by John F Helliwell et al. New York: Sustainable Development Solutions Network, 2022. Chap. 6. URL: https://worldhappiness.report/ed/2022/.

[29] Richard E Lucas and Nicole M Lawless. "Does life seem better on a sunny day? Examining the association between daily weather conditions and life satisfaction judgments." In: *Journal of personality and social psychology* 104.5 (2013), p. 872.

[30] Richard E. Lucas, Vicki A. Freedman, and Jennifer C. Cornman. "The short-term stability of life satisfaction judgments". In: *Emotion* 18.7 (2018), pp. 1024–1031. ISSN: 1528-3542. DOI: 10.1037/emo0000357.

[31] Jessica Mahoney. *Subjective Well-being Measurement: Current Practice and New Frontiers*. Working Paper 17. JEL Classification: I31, I38, D91. OECD, 2023. DOI: 10.1787/4ca48f7c-en. URL: https://doi.org/10.1787/4ca48f7c-en.

[32] August Håkan Nilsson et al. "The Cantril Ladder elicits thoughts about power and wealth". In: *Scientific Reports* 14.1 (2024), p. 2642.

[33] OECD. *OECD Guidelines on Measuring Subjective Well-being*. OECD Publishing, Mar. 2013. DOI: 10.1787/9789264191655-en. URL: http://dx.doi.org/10.1787/9789264191655-en.

[34] Shigehiro Oishi. "Culture and well-being: Conceptual and methodological issues". In: *International differences in well-being*. Vol. 34. 2010, p. 69.

[35] Shigehiro Oishi et al. "Concepts of Happiness Across Time and Cultures". In: *Personality and Social Psychology Bulletin* 39.5 (2013). PMID: 23599280, pp. 559–577. DOI: 10.1177/0146167213480042. eprint: https://doi.org/10.1177/0146167213480042. URL: https://doi.org/10.1177/0146167213480042.

[36] Ekaterina Oparina and Sorawoot Srisuma. "Analyzing Subjective Well-Being Data with Misclassification". In: *Journal of Business & Economic Statistics* 40.2 (2022), pp. 730–743. DOI: 10.1080/07350015.2020.1865169. eprint: https://doi.org/10.1080/07350015.2020.1865169. URL: https://doi.org/10.1080/07350015.2020.1865169.

[37] Arthur A Stone, Christopher Mackie, et al. *Subjective well-being: Measuring happiness, suffering, and other dimensions of experience*. National Academies Press, 2014.

[38] UK Treasury. *Wellbeing Guidance for Appraisal: Supplementary Green Book Guidance*. Tech. rep. UK Government, July 2021. URL: https://www.gov.uk/government/publications/green-book-supplementary-guidance-wellbeing.

Supplementary Material

A Supplementary figures and tables

Life satisfaction (WVS versus GFS)

Country group	ΔRank (%ile)	ΔMean
Africa	$3.5_{\pm 1.4}$	$0.6_{\pm 0.05}$
Catholic Europe	$14.9_{\pm 2.2}$	$0.2_{\pm 0.05}$
Confucian	$-0.8_{\pm 2.8}$	$-0.2_{\pm 0.04}$
English speaking	$-25.0_{\pm 1.8}$	$-0.6_{\pm 0.04}$
Islamic	$24.8_{\pm 2.2}$	$0.3_{\pm 0.05}$
Latin America	$-10.0_{\pm 2.0}$	$-0.4_{\pm 0.04}$
Protestant Europe	$-29.2_{\pm 3.1}$	$-0.6_{\pm 0.04}$
South Asia	$23.4_{\pm 3.5}$	$0.3_{\pm 0.05}$

Table A1. Country group rank differences for life satisfaction responses from WVS, as compared with GFS. Positive ΔRank means a higher ranking. See Figure 5 for the ranks.

Life satisfaction (WVS) versus Cantril ladder (GFS)

Country group	ΔRank (%ile)	ΔMean
Africa	$-6.4_{\pm 1.1}$	$0.1_{\pm 0.05}$
Catholic Europe	$-27.7_{\pm 2.2}$	$0.0_{\pm 0.05}$
Confucian	$-14.7_{\pm 2.6}$	$0.3_{\pm 0.04}$
English speaking	$-4.6_{\pm 2.2}$	$0.5_{\pm 0.04}$
Islamic	$5.8_{\pm 2.2}$	$0.9_{\pm 0.05}$
Latin America	$21.6_{\pm 2.7}$	$1.0_{\pm 0.04}$
Protestant Europe	$5.3_{\pm 3.3}$	$0.7_{\pm 0.04}$
South Asia	$15.6_{\pm 3.1}$	$1.0_{\pm 0.05}$

Table A2. Country group rank differences for life satisfaction (WVS) versus Cantril ladder (GFS). Positive ΔRank means a higher ranking. See Figure 6 for the ranks.

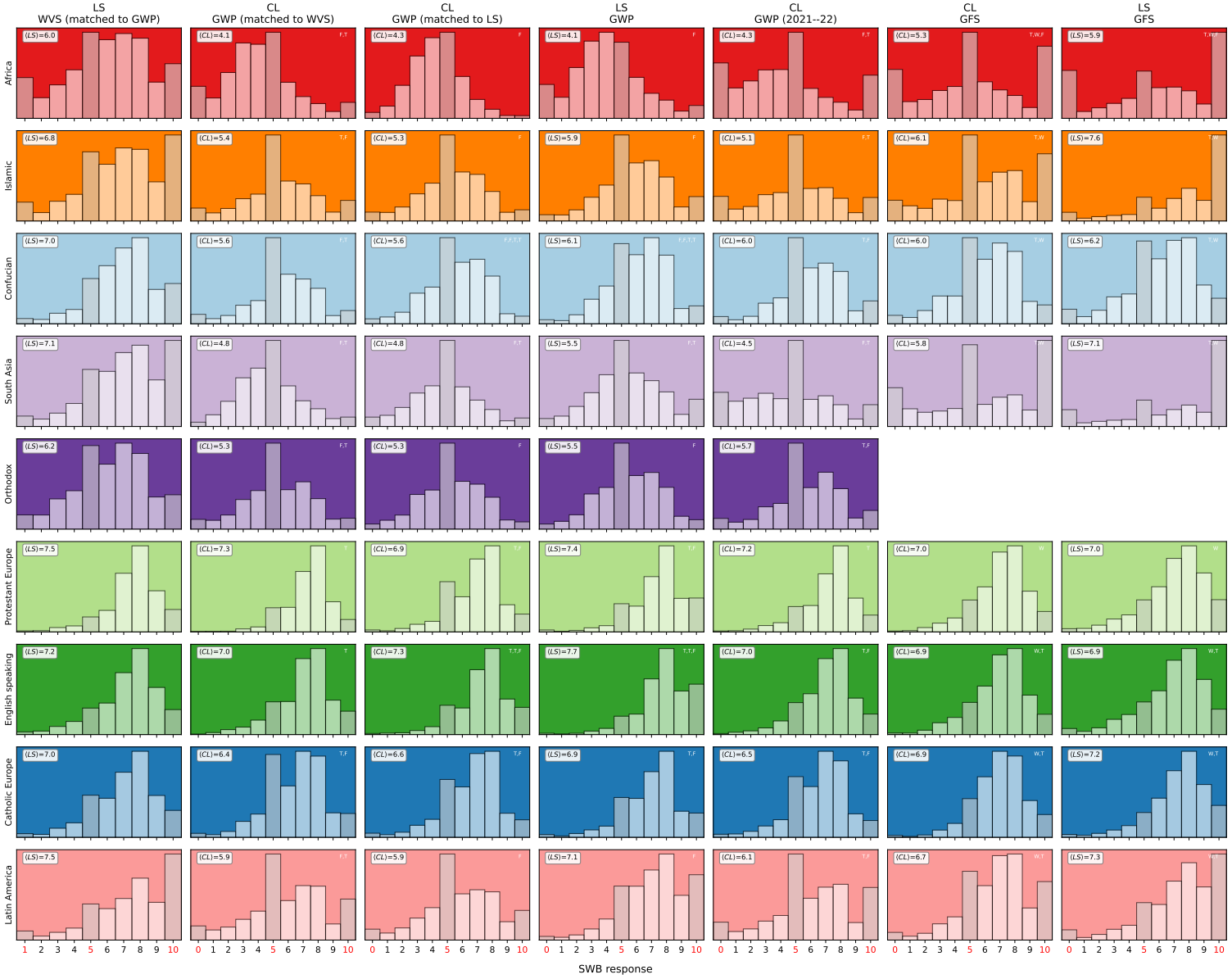


Fig. A1. Response distributions by cultural region, for each survey and question.

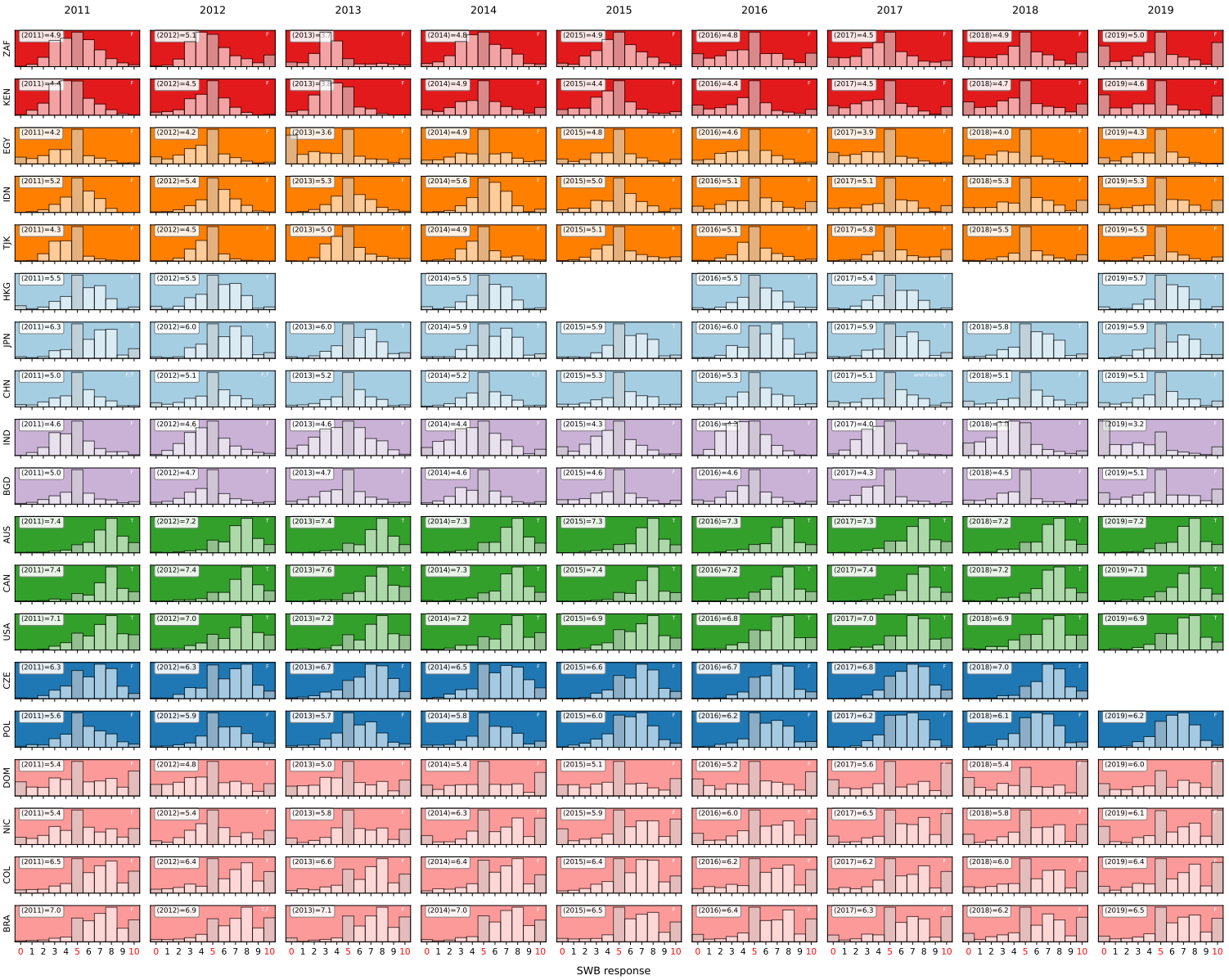


Fig. A2. Response distributions for a selection of countries, by year. Generally, patterns in the distribution are consistent across time.

Variable	count	mean	std	min	25%	50%	75%	max
LS	202199	6.87	2.57	0	5	7	9	10
CL	202412	6.50	2.45	0	5	7	8	10
lnlocalHHincome	202898	1.99	1.26	0	1.24	2.01	2.72	5.30
	Local currencies, so magnitudes not meaningful across countries							
Education:secondary+	202711	0.84	0.37	0	1	1	1	1
Education:post-secondary	202711	0.26	0.44	0	0	0	1	1
male	202649	0.47	0.50	0	0	0	1	1
age	202878	45.83	17.67	18	31	44	60	99
Married / cohabiting	201048	0.62	0.49	0	0	1	1	1
Separated / divorced / widowed	201048	0.13	0.34	0	0	0	0	1
urban	201645	0.46	0.50	0	0	0	1	1
prays	202165	0.74	0.44	0	0	1	1	1
	How often do you pray or meditate? [More than once a day, about once a day, sometimes, never]							
donated	202306	0.39	0.49	0	0	0	1	1
	In the past month, have you donated money to a charity? [Y / N]							
drinks	199925	0.40	0.49	0	0	0	1	1
	Approximately how many full drinks of any kind of alcoholic beverage did you drink in the past seven days, if any? Please enter the number below. A full drink is a glass of wine, a can or bottle of beer, or a shot of hard liquor. [Open-ended response]							
lnFreqAttendRelig	202181	1.07	2.82	-2.30	-2.30	1.10	3.95	4.64
	Wording: How often do you attend religious services? [More than once a week, once a week, one to three times a month, a few times a year, never]							
fractionDaysExercise	199723	0.36	0.35	0	0	0.29	0.57	1
	On how many days did you exercise or engage in vigorous physical activities for 30 minutes or more in the past week? [0=0 days, 7=7 days]							
unemployed	202076	0.08	0.28	0	0	0	0	1

Table A3. Variables used in the GFS models of Section 6