Economics Bulletin

Volume 38, Issue 2

Grading happiness: what grading systems tell us about cross-country wellbeing comparisons

Fernanda Marquez-Padilla
CIDE

Jorge Alvarez
IMF

Abstract

Self-reported wellbeing measures have been widely used in cross-country studies. However, there are concerns about the sensitivity of these measures to country-specific factors that affect the interpretation of questions and scales without affecting wellbeing itself. Using a novel database on international grading systems, we find evidence that differences in numerical grading systems affect self-reported wellbeing. In particular, countries with a higher threshold for passing grades tend to report higher levels of life satisfaction. Since grading systems are unlikely to affect wellbeing itself, we conclude that grading systems affect the interpretation of scales--probably by providing reference points that anchor individuals' responses.

The authors thank Janet Currie, Angus Deaton, Marc Fleurbaey, Ilyana Kuziemko, and Tom Vogl, and seminar participants at the Princeton Third-Year Graduate Workshop for their useful comments and helpful discussions. Diana Flores provided helpful research assistance. The views expressed in this study are the sole responsibility of the authors and should not be attributed to the International Monetary Fund, its Executive Board, or its management. All errors are our own.
Citation: Fernanda Marquez-Padilla and Jorge Alvarez, (2018) "Grading happiness: what grading systems tell us about cross-country wellbeing comparisons", *Economics Bulletin*, Volume 38, Issue 2, pages 1138-1155
Contact: Fernanda Marquez-Padilla - fernanda.marquezp@gmail.com, Jorge Alvarez - jalvarez@imf.org.
Submitted: April 11, 2018. **Published:** June 21, 2018.

1 Introduction

It has been argued that measures of life satisfaction may effectively aggregate over different dimensions of wellbeing, and may thus be well suited for comparing wellbeing across countries or over time (Deaton 2008). Moreover, the use of cross-country wellbeing comparisons as a policy-evaluation tool has become increasingly common in the literature (Blanchflower and Oswald 2004; Clark et al. 2008)—even the UN has commissioned an annual World Happiness Report. Veenhoven (2012) alone identifies more than 4,500 survey findings on happiness across nations, which have been used in some 500 scientific publications on happiness and society.

Most studies focusing on subjective wellbeing (SWB) rely on *self-reported* measures of SWB[1] that are sensitive to biases arising from survey design, personal traits, and cultural differences (Deaton and Stone 2016; Ferrer-i- Carbonell and Frijters 2004; Conti and Pudney 2011; Heffetz and Rabin 2013). Additionally, these biases often affect different subgroups differentially. For instance, Deaton and Stone (2016) find that life satisfaction reporting is affected by political questions asked before wellbeing questions, with particular subgroups being more influenced than others.[2] Additionally, the distribution of personality traits—which may vary across countries—appears to be potentially important for cross-country analysis (Ferrer-i- Carbonell and Frijters 2004). Using changes in question design and interview modes in the British Household Panel Survey, Conti and Pudney (2011) find that women tend to rely more on text-labels as anchors and be more sensitive to social-desirability bias[3] in the context of face-to-face interviews. Additionally, Heffetz and Rabin (2013) find that bias from non-response rates may differ by subgroup—they find that hard-to-reach women tend to be different than hard-to-reach men in terms of SWB. Overall, there is evidence of differences between subpopulations that affect self-reported measures.

Survey questions on SWB may be particularly affected by response styles (RS)—the systematic tendency to answer questions in one way or another, regardless of their content—which have been found to vary across nations and cultures (Van Vaerenbergh and Thomas 2013; Bertrand and Mullainathan 2001; Hui and Triandis 1989; Krueger et al. 2009). This type of RS variation can undermine comparisons of cross-cultural surveys (Van Vaerenbergh and Thomas 2013; Tellis and Chandrasekaran 2010; Angelini et al. 2014), including those that relate to SWB. Moreover, understanding such variation could contribute to the literature on discrepancies between cross-country and within-country patterns of SWB such as Easterlin's happiness–income paradox (Easterlin 1974).

Our study is most concerned with differences in RS due to the interpretation of scales. Bertrand and Mullainathan (2001) have argued that scales cause cognitive problems that affect subject responses, ant that these effects can vary across cultural or national groups. For example, Hispanics reduce their tendency to produce extreme responses when 10-point as opposed to 5-point scales are used, while non-Hispanics' RS tend to be neutral to the type

[1]Some attempts have been made to rely on more *objective* measures of life-satisfaction such as suicide rates, frequency of smiling, or neurological measures (Perez-Truglia 2015)

[2]The authors find that African Americans reported life satisfaction increased when first asked about President Obama's performance (Deaton and Stone 2016).

[3]Social desirability bias refers to the tendency of respondents to over-report favorable attitudes and under-report unfavorable ones.

of scale used (Hui and Triandis 1989). Similarly, Krueger et al. (2009) suggest that cultural differences in reporting lead the French to appear less satisfied with their lives than their American counterparts, as they appear to be less prone to use the extreme ends of a scale. Finally, Angelini et al. (2014) find that using vignettes instead of scales affects self-reported life satisfaction and has an effect on cross-country rankings in European countries.[4]

This paper empirically shows that self-reported measures of wellbeing can be biased by country-specific factors that affect the interpretation of scales, as there is evidence that different groups of respondents might use the number scale differently (Diener et al. 2013). In order to test this, we find a novel source of variation that affects the interpretation of scales but does not affect life satisfaction itself: pass/fail grading thresholds (PFT). These are defined as the first numerical value at which a grade is considered a passing grade for higher education. For example, while some countries define the lowest passing grade as scoring 1 out of 4, others define it as scoring 60 over 100. We present evidence that this variation has no direct effect on an individual's *actual* life satisfaction but can influence his *reported* life satisfaction, probably by serving as a numerical anchor that affects the mapping of mental states to a numerical scale.

2 Data

The main LS measure we analyze is the response to the question 'On a scale from 1 to 10: All things considered, how satisfied are you with your life as a whole these days?' from the World Values Survey (1981–2008). We use additional questions asked in the 'scale from 1 to 10' format included in the WVS, and other questions asking how the respondent feels that use non-numerical answers.[5]

We use data from the World Education Services' international grade conversion guide to compute PFT for grading systems across countries. Our sample includes the 59 countries using numerical grades (we exclude countries such as the US and Canada, for which numerical PFT's cannot be computed as *letter* grading systems are used). The PFT is defined as the passing grade for higher education at the country level, standardized to a 10 point-scale.[6] We use country variables from the World Bank's World Development Indicators and regional dummy variables.[7]

We find ample variation in grading systems as measured by PFT's, which we use as the explanatory variable in order to test whether RS vary across countries. The distribution of the standardized PFT for the countries in our sample is shown in Figure B.2 in the

[4]Anchoring vignettes attempt to clean survey answers from reporting differences due to RS in survey questions across countries (King et al. 2004). Our results are consistent with some of the results in Angelini et al. (2014) (such as the fact that the French tend to underreport their wellbeing, also consistent with Kahneman et al. (2004) which discusses how the French are less likely to describe themselves as "very satisfied").

[5]Full description of these questions in Table A.1 in the Appendix.

[6]We observe one grading system per country. Data normalized to correspond to a 10-point scale, according to: $10 \times \frac{PFT-min}{max-min}$. Details in Table A.3.

[7]Latin America (Argentina, Brazil, Chile, Colombia, El Salvador, Guatemala, Mexico, Peru, Puerto Rico, Uruguay, Venezuela); Former USSR (Latvia, Lithuania, Russian Federation, Ukraine); French Colony (Algeria, Burkina Faso, France, Mali) from Price (2003).

Appendix.[8] We use this variation to detect RS differences.

3 Empirical framework

When facing a scale question, survey respondents must transform their experienced sense of wellbeing—or some other emotional state—to a numerical value. This process can be described as:

$$H_{ic} = f_c(H^*_{ic}) + \varepsilon_{ic},$$

where H_{ic} is the *reported* LS of individual i from country c, H^*_{ic} is the *actual* LS, and $f_c()$ is a country specific function that transforms the individual's situation to a numerical response. This function can be interpreted as a country-specific RS.[9]

We propose using variation in PFT to test for differences in $f_c()$ across countries. In particular, we estimate regressions of the form

$$LS_{i,c} = \beta_0 + \beta_1 PFT_c + \gamma X_{i,c} + \varepsilon_{i,c},$$

where $LS_{i,c}$ is the answer to the WVS's LS 10-point scale question for individual i, from country c, and $X_{i,c}$ are individual controls, namely age, sex, highest level of education, (scale) income, GDP per capita, life expectancy, and literacy rates.

Our assumption, which we later support with empirical results, is that while the PFT does not affect happiness directly, it does affect the way in which individuals respond to questions with a *numerical* scale. Under this paradigm, a non-zero β_1 implies that variation must come from the effect of thresholds on $f_c()$.

Our interpretation would be invalid if the choice of grading systems were systematically correlated with current determinants of life satisfaction that we do not control for. Historical accounts on the origins of grading systems would suggest that this is unlikely.[10] Characteristics of grading systems appear to be both antique and, to a certain extent, arbitrary. These attributes increase our confidence in the lack of correlation between PFT and current determinants of life satisfaction.

[8] Finland has the lowest standardized PFT, 2.0, corresponding to a 1 out of 5 in the raw data. Philippines has the highest, 7.5, corresponding to a grade of 75 out of 100. The average standardized PFT is 4.2 and the variance is 1.2 (full data in the Appendix).

[9] Based on Fleurbaey and Blanchet (2013).

[10] For instance, the French philosophy of grading has its roots in the Jesuit tradition as embodied in the *Ratio Studorium*, an education manual from the 16th century. The grading scale then evolved over time until the education ministry established the current 0–20 scale at the national level in 1890 (Bertrand 2007). We have found no evidence that the choice of these numbers is related to significant developments in France. In Germany, the inverted grading system (with the lowest grade being best) had its origins in the Prussian education system that instituted the notion of schools by grades in the 19th century using a 5-point scale. It was eventually expanded to a 6-point grading system by 1938 in order to avoid bunching at the middle grade, three (Kuss 2003). The Prussian system was adopted and inverted by the Ministry of Education of the Russian Empire in 1837, with the highest grade becoming the best outcome. Letter systems such as the ones used in the United States are excluded from our sample, but their history seems to be rooted in early college practices in the 19th century (Smallwood 1935).

In our analysis, we include a series of different specifications to our model to strengthen the interpretation of our results. As the anchoring effect provided by the PFT is arguably more salient for grading systems that use 10-, 20-, or 100-point scales (due to ease of conversion to a 10-point scale), we run our baseline models only for countries that use these types of scales. We also run these models excluding countries with *inverted* grading scales (i.e. best grade corresponds to a lower score, as in Germany) as the anchoring effect might be less salient when scales are inverted.

Additionally, we present some robustness checks of our results. First, we use other variables which are not based on scales (i.e. are not *numerical*) but which indicate some dimensions of individual wellbeing as the dependent variable. Namely, we use the answers to questions regarding whether the individual has ever felt depressed or very unhappy, ever felt on top of the world, or feeling of happiness, in addition to the (subjective) state of his health.[11] If the country's PFT only affects self-reported life satisfaction through its effect on response styles, we would expect no correlation with these other variables.

Second, we also report regressions using additional questions in a 10-point scale format included in the WVS as dependent variables. We would expect the PFT to affect the answer to these questions as well (arguably, through the same anchoring channel that affects responses to the life-satisfaction question).[12]

Third, we analyze whether there exist differential effects for immigrants, as these individuals may be differentially affected by the anchoring effects stemming from PFT.[13] We identify immigrants using two variables from the WVS (namely, questions on whether the individual was born in another country or if the individual speaks a different language at home from the one in which the interview was conducted). A weaker effect of the PFT on self-reported life-satisfaction would be consistent with our hypothesis.

Finally, as an illustration of how scale interpretation may bias the cross-country comparisons of life-satisfaction, we use the estimated coefficients from regressing life satisfaction on the PFT to calculate a measure of "imputed happiness". This measures assumes that every country has the same PFT—namely a PFT equal to 5. We then regress this imputed measure of life satisfaction on national income to see how the relationship is affected after "correcting" for the differences in RS induced by the different grading systems.

[11] The specific questions ask: "Have you ever felt depressed or very unhappy", and "Have you ever felt on top of the world", for example. Note that answers are *categorical* rather than *numerical*

[12] Only three additional questions are asked in this fashion (regarding satisfaction with financial situation, job satisfaction, and home life), and the last two are only included in the WVS for 15 countries. The variation in grading systems is therefore limited making it difficult to identify its effect. Some additional questions are asked in a similar way (also using a 10-point scale) regarding how much freedom of choice and control individuals feel they have, where 1 reflects no freedom at all and 10 reflects a great deal of freedom.

[13] For example, Deaton and Stone (2016) find that different subgroups of the population are differentially affected by context effects.

Table I. Relationship between LS and PFT

A. All countries	(1)	(2)	(3)	(4)	(5)
PFT	0.190*	0.232*	0.268**	0.267**	0.337**
	(0.105)	(0.125)	(0.132)	(0.125)	(0.143)
Mean dept. var.	6.424	6.559	6.534	6.534	6.472
Observations	59	59	196,909	196,909	152,382
No. Countries	59	59	59	59	57
R-squared	0.054	0.057	0.014	0.024	0.070
B. 100-, 20-, or 10-pt scale	(1)	(2)	(3)	(4)	(5)
PFT	0.337**	0.462**	0.312**	0.292**	0.340***
	(0.152)	(0.174)	(0.126)	(0.111)	(0.116)
Mean dept. var.	6.381	6.525	6.475	6.475	6.379
Observations	40	40	136,923	136,923	109,745
No. Countries	40	40	40	40	40
R-squared	0.114	0.157	0.012	0.022	0.073
C. 100- or 10-pt scale					
PFT	0.346*	0.440**	0.319**	0.302**	0.360***
	(0.185)	(0.201)	(0.140)	(0.124)	(0.130)
Mean dept. var.	6.400	6.548	6.493	6.493	6.375
Observations	31	31	111,532	111,532	87,865
No. Countries	31	31	31	31	31
R-squared	0.108	0.142	0.013	0.028	0.079
D. No inverted scales					
PFT	0.205*	0.266*	0.286*	0.289*	0.370**
	(0.116)	(0.138)	(0.154)	(0.147)	(0.171)
Mean dept. var.	6.412	6.537	6.534	6.534	6.465
Observations	54	54	183,742	183,742	141,670
No. Countries	54	54	54	54	52
R-squared	0.056	0.067	0.014	0.024	0.069
Clustered Std.Err.			yes	yes	yes
Wave F.E.				yes	yes
Comments		median			indiv. **X**'s
Unit of Obs.	country	country	indiv.	indiv.	indiv.

Notes: LS is the dependent variable. $^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$
Countries included in Panel C are Albania, Argentina, El Salvador, Latvia, Lithuania, Mexico, Netherlands, Romania, Switzerland, Turkey and Viet Nam

4 Results

We find a positive relationship between the normalized PFT at the country level and average LS. We argue that this effect arises from an upward bias in RS from having a higher numerical anchor, and not from an actual effect of the PFT on individual wellbeing.

Table I shows the main results from an OLS regression of reported LS on the PFT.

Table II. Relationship with LS remains when adding controls

	(1)	(2)	(3)	(4)	(5)	(6)
PFT	0.190*	0.193**	0.182*	0.307**	0.301**	0.138
	(0.105)	(0.092)	(0.100)	(0.123)	(0.117)	(0.088)
Log GDP p.c.		0.347***			0.387**	0.360***
		(0.075)			(0.177)	(0.129)
Life expectancy			0.042***		0.003	-0.011
			(0.014)		(0.026)	(0.019)
Literacy rate				0.007	-0.010	-0.005
				(0.006)	(0.010)	(0.008)
French Col.						-0.142
						(0.444)
Latin Am.						1.185***
						(0.232)
Former USSR						-1.140***
						(0.364)
Mean dept. var.	6.424	6.421	6.421	6.276	6.276	6.276
Observations	59	58	58	46	46	46
R-squared	0.054	0.321	0.193	0.167	0.276	0.663

Notes: Notes: LS is the dependent variable. $^{*}p < 0.1$,** $p < 0.05$,*** $p < 0.01$

When looking at all countries, every specification is statistically significant ($\leq$10% level). The results are larger and stronger when focusing on the subset of countries using 10-, 100- or 20-point scales in grading (Panels B and C) and the subset excluding countries with *inverted* grading scales (i.e. best grade corresponds to a lower score) as the anchoring effect is arguably more salient in these groups of countries. This result adds validity to our hypothesis that the interpretation of scales affects self-reported LS.[14] Additionally, the effect of PFT on LS remains significant in most specifications adding controls, as shown in Table II.

Table III shows that there is *no* relationship between the PFT and other wellbeing measures that are *not* asked in a 1 to 10 (i.e. numerical) format and that there *is* a positive relation between the PFT and other questions that *are* asked in a 10-point scale question besides the one related to LS.

Panel A of Table III shows the lack of correlation between the PFT and SWB-related questions not asked in a 10-point scale format. We interpret that the scale bias induced by PFT is less prominent in this type of questions as they do not rely on a *numerical* scale. Panel B of Table III shows a positive relationship between the PFT and other 10-point scale questions in the WVS. The effect of the PFT on respondents' answers is generally significant (with $\leq$16 countries there is insufficient variation in PFT's to find an effect).

Table IV shows mild evidence of a weaker effect of the PFT for immigrants. Although the interaction term is not statistically different from zero, it does have a negative sign, which provides suggestive evidence of the effect of PFT on RS, as immigrants are likely to be less

[14]Table A.2 in the Appendix categorizes all responses to the LS question as either positive (> 5), neutral ($= 5$), or negative (< 5)—thereby avoiding the issue of whether some people use extreme responding more—as suggested in Diener et al. (2013) and runs equivalent regressions at the individual level. Results are robust to this this correction.

Table III. PFT effect only for questions asked in numerical scale format

Panel A: No correlation with other indicators of wellbeing not asked in numerical scale format

	(1) Life Sat.	(2) Depressed	(3) Top World	(4) Subj. Health
PFT	0.267**	-0.008	0.028	-0.028
	(0.125)	(0.011)	(0.027)	(0.029)
Mean dep. var	6.534	0.243	0.302	2.235
Obs.	196,909	20,473	20,321	192,709
Countries	59	14	14	58
R-sq.	0.024	0.000	0.003	0.011

Panel B: A positive correlation with other indicators asked in numerical scale format

	(1) Life Sat.	(2) Fin. Sit.	(3) Free: Choice	(4) Free: Job	(5) Job Sat.	(6) Home
PFT	0.267**	0.246**	0.191**	0.189	0.082	0.136
	(0.125)	(0.120)	(0.094)	(0.154)	(0.109)	(0.105)
Mean DV	6.534	5.583	6.698	6.657	7.323	7.622
Obs.	196,909	190,244	184,624	15,072	14,141	21,637
Countries	59	58	58	16	15	15
R-sq.	0.024	0.020	0.013	0.003	0.001	0.003

Notes: Description of WVS questions in Table A.1 in the Appendix. Standard errors clustered at the country level and wave fixed effects, for waves: 1989–1993, 1994–1999, 1999–2004, and 2005–2007.
$^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$

exposed to local grading systems.

To assess how PFT affects cross-country comparisons, we create an 'imputed happiness' measure by using the estimated coefficients from regressing LS on the PFT and GDP per capita and imputing a constant PFT of 5 to every country. Figure 1 (Panel A) shows the actual and imputed measures of self-reported LS against the log of GDP. The difference between these two measures ranked from lowest to highest is shown in Panel B. These figures show that the anchoring effect induced by the PFT is large enough to change the ranking of wellbeing across countries. For example, while Finland has a lower self-reported LS than Colombia or Guatemala in the actual data (which may be surprising given Finland's greater economic development), once we correct for the reporting bias introduced by the grading system, Finland actually outperforms both of these countries in terms of LS. France's average reported LS is adjusted upwards, consistent with the evidence suggesting that the French tend to under-report their wellbeing (Angelini et al. 2014; Kahneman et al. 2004).

Table IV. A weaker effect of local grading scales for immigrants

PFT	0.265*	0.271**	0.350**
	(0.138)	(0.131)	(0.151)
Immigrant	-0.219	0.182	0.620
	(0.698)	(0.680)	(0.774)
PFT × Immigrant	-0.016	-0.091	-0.188
	(0.151)	(0.148)	(0.178)
Mean dept. var.	6.534	6.534	6.472
Observations	196,909	196,909	152,382
No. Countries	59	59	57
R-squared	0.015	0.024	0.071
Clustered Std.Err.	yes	yes	yes
Wave F.E.		yes	yes
Comments			indiv. **X**'s
Unit of Obs.	indiv.	indiv.	indiv.

Notes: Includes wave fixed-effects. Standard errors clustered at country level.
$^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$

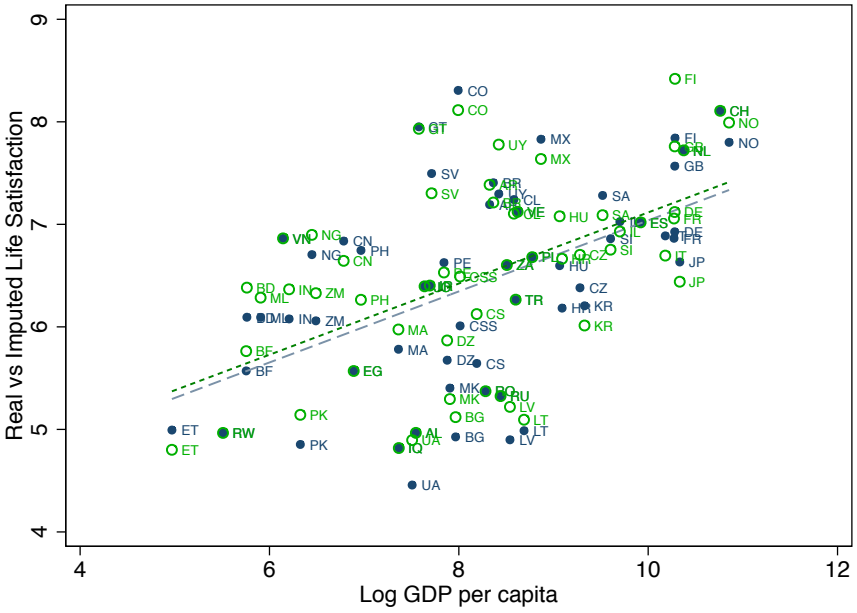
5 Discussion

PFTs shows a significant positive correlation with numerical measures of LS but not with non-numerical measures of wellbeing. We interpret this as evidence of the biasing effects of PFTs on the interpretation of numerical scales. While we are primarily interested in individuals' self-reported wellbeing, the type of bias we study may exist in any self-reported qualitative attribute. Moreover, the grading effect documented here is only one of the potential sources of bias affecting life-satisfaction comparisons across countries.

We find no correlation between PFT's and other indicators of wellbeing, such as feeling depressed or subjective health when these questions are not asked using numerical scales. However, we do find a positive correlation between the level of PFT and answers to other questions from the WVS that use a 10-point scale, suggesting that the channel identified is indeed related to cross-country differences in RS. Furthermore, this effect is stronger for countries with 10-, 20-, and 100-point scales, as the anchor from the PFT is arguably more salient, which also strengthens our interpretation. Finally, there is suggestive evidence that the effect is weaker for immigrants, as they have arguably been less exposed to the grading systems.

It is important to highlight that our research empirically tests the existence of only one of the many potential sources of bias in the measurement of life-satisfaction across countries. Moreover, the effect of grading-systems is likely to be milder than other potential biases arising from more salient linguistic, cultural, or social norm differences. Our results suggest that the interpretation of cross-country comparisons of self-reported LS measures should be interpreted with care, and that in order for these comparisons to serve as useful policy-guiding tools, the biases resulting from differences in RS across countries should be acknowledged and dealt with.

Panel A.



Panel B.

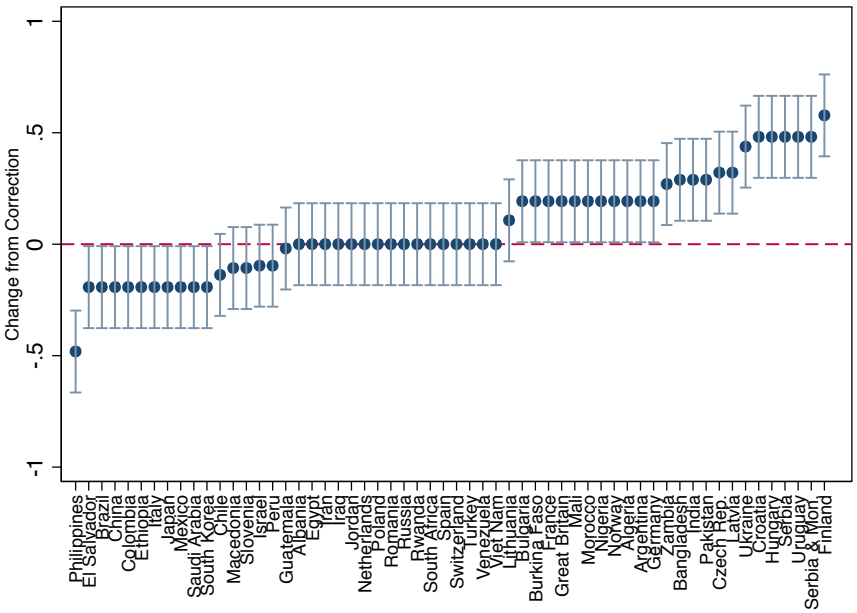


Figure 1. Relationship between income and Imputed LS

Notes: **(A)** Reported (blue) and imputed (green) LS. **(B)** Difference between imputed and reported LS. Country codes correspond to Albania (AL); Algeria (DZ); Argentina (AR); Bangladesh (BD); Brazil (BR); Bulgaria (BG); Burkina Faso (BF); Chile (CL); China (CN); Colombia (CO); Croatia (HR); Czech Republic (CZ); Egypt (EG); El Salvador (SV); Ethiopia (ET); Finland (FI); France (FR); Germany (DE); Great Britain (GB); Guatemala (GT); Hungary (HU); India (IN); Iran (IR); Iraq (IQ); Israel (IL); Italy (IT); Japan (JP); Jordan (JO); Latvia (LV); Lithuania (LT); Macedonia (MK); Mali (ML); Mexico (MX); Morocco (MA); Netherlands (NL); Nigeria (NG); Norway (NO); Pakistan (PK); Peru (PE); Philippines (PH); Poland (PL); Romania (RO); Russian Federation (RU); Rwanda (RW); Saudi Arabia (SA); Serbia and Montenegro (CS); Slovenia (SI); South Africa (ZA); South Korea (KR); Spain (ES); Switzerland (CH); Taiwan (TW); Turkey (TR); Ukraine (UA); Uruguay (UY); Venezuela (VE); Viet Nam (VN); Zambia (ZM)

References

Angelini, V., D. Cavapozzi, L. Corazzini, and O. Paccagnella (2014). "Do Danes and Italians rate life satisfaction in the same way? Using vignettes to correct for individual-specific scale biases." *Oxford bulletin of Economics and Statistics 76*(5), 643–666.

Bertrand, G. (2007). *Les notes à l'école ou le rapport à la notation des enseignants de l'école élémentaire*. Paris: L'Harmattan.

Bertrand, M. and S. Mullainathan (2001). "Do people mean what they say? Implications for subjective survey data." *The American Economic Review 91*, 67–72.

Blanchflower, D. G. and A. J. Oswald (2004). "Well-being over time in Britain and the USA." *Journal of public economics 88*(7), 1359–1386.

Clark, A. E., P. Frijters, and M. A. Shields (2008, March). "Relative income, happiness, and utility: An explanation for the Easterlin Paradox and other puzzles." *Journal of Economic Literature 46*(1), 95–144.

Conti, G. and S. Pudney (2011). "Survey design and the analysis of satisfaction." *Review of Economics and Statistics 93*(3), 1087–1093.

Deaton, A. (2008, June). "Income, health, and well-being around the world: Evidence from the Gallup World Poll." *Journal of Economic Perspectives 22*(2), 53–72.

Deaton, A. and A. A. Stone (2016). "Understanding context effects for a measure of life evaluation: how responses matter." *Oxford Economic Papers 68*(4), 861–870.

Diener, E., R. Inglehart, and L. Tay (2013). "Theory and validity of life satisfaction scales." *Social Indicators Research 112*(3), 497–527.

Easterlin, R. A. (1974). "Does economic growth improve the human lot? Some empirical evidence." *Nations and households in economic growth 89*, 89–125.

Ferrer-i-Carbonell, A. and P. Frijters (2004). "How important is methodology for the estimates of the determinants of happiness?" *The Economic Journal 114*(497), 641–659.

Fleurbaey, M. and D. Blanchet (2013). *Beyond GDP Measuring Welfare and Assessing Sustainability*. Oxford University Press.

Heffetz, O. and M. Rabin (2013). "Conclusions regarding cross-group differences in happiness depend on difficulty of reaching respondents." *The American economic review 103*(7), 3001–3021.

Hui, C. H. and H. C. Triandis (1989). "Effects of culture and response format on extreme response style." *Journal of cross-cultural psychology 20*(3), 296–309.

Kahneman, D., A. B. Krueger, D. Schkade, N. Schwarz, and A. Stone (2004). "Toward national well-being accounts." *The American Economic Review 94*(2), 429–434.

King, G., C. J. Murray, J. A. Salomon, and A. Tandon (2004). "Enhancing the validity and cross-cultural comparability of measurement in survey research." *American political science review 98*(1), 191–207.

Krueger, Alan B.and Kahneman, D., D. Schkade, N. Schwarz, and A. Stone (2009). *Measuring the Subjective Well-Being of Nations: National Accounts of Time Use and Well-Being*. University of Chicago Press.

Kuss, S. (2003). In diesen tagen gibt es zeugnisse: Zur geschichte der noten. *Frankfurter Allgemeine Zeitung*.

Perez-Truglia, R. (2015). "A Samuelsonian validation test for happiness data." *Journal of Economic Psychology 49*, 74–83.

Price, G. N. (2003). "Economic growth in a cross-section of nonindustrial countries: Does colonial heritage matter for Africa?" *Review of Development Economics 7*(3), 478–495.

Smallwood, M. (1935). *Examinations and Grading Systems in Early American Universities*. Harvard University Press.

Tellis, G. J. and D. Chandrasekaran (2010). "Does culture matter? Assessing response biases in cross-national survey research." *International Journal of Research in Marketing, Forthcoming*.

Van Vaerenbergh, Y. and T. D. Thomas (2013). "Response styles in survey research: A literature review of antecedents, consequences, and remedies." *International Journal of Public Opinion Research 25*(2), 195–217.

Veenhoven, R. (2012). "Cross-national differences in happiness: Cultural measurement bias or effect of culture?" *International Journal of Wellbeing 2*(4), 333–353.

A Appendix tables

Table A.1. WVS questionnaire: selected questions

Variable	WVS questionnaire	Values
Life Satisfaction	All thing considered, how satisfied are you with your life as a whole these day?	1–10 scale
Depressed	During the past few weeks, did you ever feel depressed or very unhappy	0=no 1=yes
Top World	During the past few weeks, did you ever feel on top ot the world/feeling that life is wonderful	0=no 1=yes
Subj. Health	All in all, how would you describe your state of health these days?	1=very good 2=good 3=fair 4=poor 5=very poor
Financial Sit.	How satisfied are you with the financial situation of your household?	1–10 scale
Job Satisfaction	Overall, how satisfied or dissatisfied are you with your job?	1–10 scale
Home	Overall, how satisfied or dissatisfied are you with your home life?	1–10 scale
Free: Choice	How much freedom of choice and control do you feel you have over the way your life turns out?	1–10 scale
Free: Job	How free are you to make decisions in your job?	1–10 scale

Source: WVS 1981–2008

Table A.2. Relationship between LS and PFT, recoded LS

	(1)	(2)	(3)
PFT	0.073*	0.073**	0.089**
	(0.037)	(0.034)	(0.038)
Mean dept. var.	0.461	0.461	0.437
Observations	196,909	196,909	152,382
No. Countries	59	59	57
R-squared	0.010	0.022	0.069
Clustered S.E.	yes	yes	yes
Wave F.E.	no	yes	yes
Controls	no	no	yes

Notes: Recoded LS is the dependent variable (equal to 1 if $LS > 5$, to zero if $LS = 5$, and to -1 if $LS < 5$). $^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$

Table A.3. Grading scales by country

	Pass/Fail Threshold	Lowest	Highest	Raw Threshold
Albania	5.0	0	10	5
Algeria	4.0	0	20	8
Argentina	4.0	0	10	4
Bangladesh	3.5	0	100	35
Brazil	6.0	0	100	60
Bulgaria	4.0	1	6	3
Burkina Faso	4.0	0	20	8
Chile	5.7	0	7	4
China	6.0	0	100	60
Colombia	6.0	0	5	3
Croatia	2.5	1	5	2
Czech Rep.	3.3	4	1	3
Egypt	5.0	0	100	50
El Salvador	6.0	0	10	6
Ethiopia	6.0	0	100	60
Finland	2.0	0	5	1
France	4.0	0	20	8
Germany	4.0	6	1	4
Great Britain	4.0	0	100	40
Guatemala	5.1	0	100	51
Hungary	2.5	1	5	2
India	3.5	0	100	35
Iran	5.0	0	20	10
Iraq	5.0	0	100	50
Israel	5.5	0	100	55
Italy	6.0	0	30	18
Japan	6.0	0	100	60
Jordan	5.0	0	100	50
Latvia	3.3	1	10	4
Lithuania	4.4	1	10	5
Macedonia	5.6	1	10	6
Mali	4.0	0	20	8
Mexico	6.0	0	10	6
Morocco	4.0	0	20	8
Netherlands	5.0	0	10	5
Nigeria	4.0	0	20	8
Norway	4.0	6	1	4
Pakistan	3.5	0	100	35
Peru	5.5	0	20	11

Continued on next page

Table A.3 – *Continued from previous page*

	Pass/Fail Threshold	Lowest	Highest	Raw Threshold
Philippines	7.5	0	100	75
Poland	5.0	1	5	3
Romania	5.0	0	10	5
Russia	5.0	1	5	3
Rwanda	5.0	0	100	50
Saudi Arabia	6.0	0	100	60
Serbia	2.5	5	1	4
Serbia and Mont.	2.5	5	1	4
Slovenia	5.6	1	10	6
South Africa	5.0	0	100	50
South Korea	6.0	0	100	60
Spain	5.0	0	10	5
Switzerland	5.0	0	10	5
Taiwan	6.0	0	100	60
Turkey	5.0	0	10	5
Ukraine	2.7	1	12	4
Uruguay	2.5	0	12	3
Venezuela	5.0	0	20	10
Viet Nam	5.0	0	10	5
Zambia	3.6	0	100	36

Notes: Data from the World Education Services' Grade Conversion Guide.

Table A.4. Grading system uncorrelated to common determinants of LS

	(1)	(2)	(3)	(4)	(5)
Log GDP p.c.	-0.003			-0.026	-0.019
	(0.109)			(0.233)	(0.234)
Life expectancy		0.004		0.016	0.004
		(0.018)		(0.034)	(0.034)
Literacy rate			0.007	0.004	0.007
			(0.007)	(0.013)	(0.015)
Latin Am.					0.307
					(0.419)
French Col.					-0.490
					(0.804)
Former USSR					-1.067
					(0.639)
Mean dept. var.	4.601	4.601	4.751	4.751	4.751
Observations	58	58	46	46	46
R-squared	0.000	0.001	0.022	0.028	0.133

Notes: The dependent variable is the PFT refers to the first number that is considered a passing grade for higher education at the country level, using data on grades from the World Education Services' international grade conversion guide for higher education. It is normalized to correspond to a 10-point scale. Aggregate country variables from the World Bank's World Development Indicators.
$^{*}p < 0.1,^{**} p < 0.05,^{***} p < 0.01$

B Appendix figures

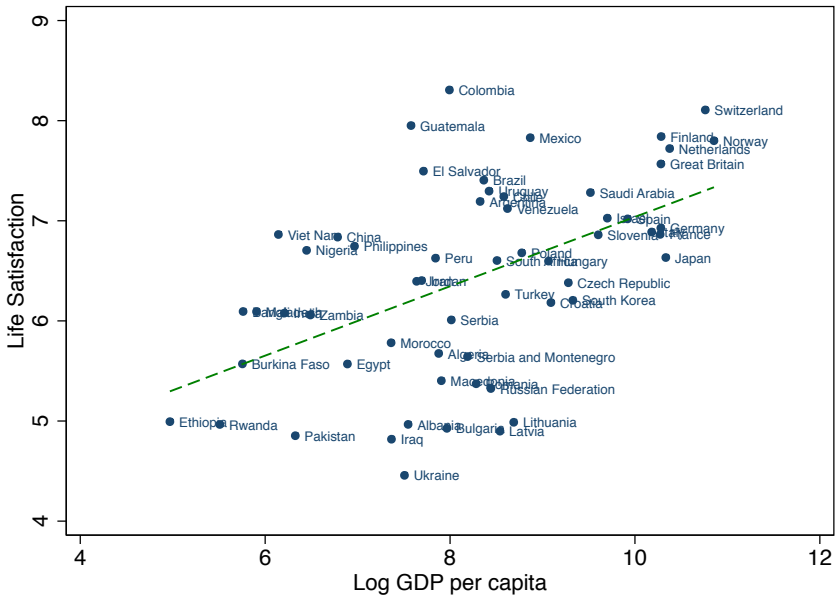


Figure B.1. LS is positively correlated with income

Notes: Measure of LS is the (unweighted) country average from individual answers to the question 'On a scale from 1 to 10: All things considered, how satisfied are you with your life as a whole these days?' from micro data from the World Values Survey 1981-2008 (average is over waves for countries represented in multiple years). GDP from the World Bank's World Development Indicators. Line represents a fitted OLS regression.

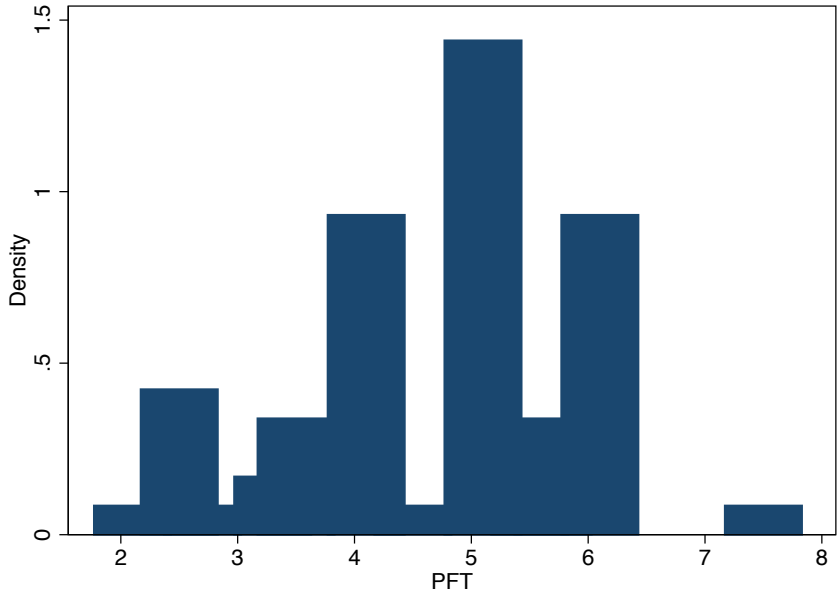


Figure B.2. Histogram: PFT

Notes: The PFT refers to the first number that is considered a passing grade for higher education at the country level, using data on grades from the World Education Services' international grade conversion guide for higher education. It is normalized to correspond to a 10-point scale.
Source: Authors' calculations with data from World Education Services