

## ARE AMERICANS REALLY LESS HAPPY WITH THEIR INCOMES?

BY ARIE KAPTEYN

*Rand Corporation*

JAMES P. SMITH

*Rand Corporation*

AND

ARTHUR VAN SOEST\*

*Netspar, Tilburg University, and Rand Corporation*

Recent economic research on international comparisons of subjective well-being suffers from several important biases due to the potential incomparability of response scales within and across countries. In this paper we concentrate on self-reported satisfaction with income in two countries: the Netherlands and the U.S. The comparability problem is addressed by using anchoring vignettes. We find that in the raw data, Americans appear decidedly less satisfied with their income than the Dutch. It turns out however that after response scale adjustment based on vignettes, the distribution of satisfaction in the two countries is essentially identical. In addition, we find that the within-country cross-sectional effect of income on satisfaction—a key parameter in the recent debate in the economic literature—is significantly underestimated, especially in the U.S., when differences in response scales are not taken into account.

**JEL Codes:** I30, J30

**Keywords:** happiness, life satisfaction, reporting bias, vignettes

### 1. INTRODUCTION

Economists have become increasingly interested in economic and non-economic determinants of subjective well-being and satisfaction with life and its various domains, like health, daily activities, or income. See Van Praag and Ferrer-i-Carbonell (2008) and Clark *et al.* (2008). Particularly since the publication of the report by Stiglitz *et al.* (2009), policy makers have become more interested in how measures of subjective well-being vary across time, countries, and socio-economic groups.

Research on well-being has touched on important themes such as the Easterlin paradox whereby average happiness remains relatively constant over time in spite of large increases in income per capita (Easterlin, 1974, 1995), while within-country cross-sectional and panel data almost always show that rising incomes

*Note:* This research was funded by NIA. The authors are grateful to the editor, two anonymous referees, and participants at conferences organized by Cemmap and Netspar for helpful comments on earlier versions.

\*Correspondence to: Arthur Van Soest, Tilburg University, PO Box 90153, 5000 LE Tilburg, Netherlands (avas@uvt.nl).

© 2012 The Authors

Review of Income and Wealth © 2012 International Association for Research in Income and Wealth  
Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main St, Malden, MA, 02148, USA.

“buy” additional satisfaction. Resolving this paradox generated many studies attempting to reconcile these findings, for example, adding relative incomes (of others or of oneself in the past) in the utility function (Van de Stadt *et al.*, 1985; Clark *et al.*, 2008) or a process of adaptation to new circumstances (Di Tella *et al.*, 2003). A contrary view is provided by Deaton (2008) who finds that the positive association of income with subjective life satisfaction reappears if a much wider range of countries is considered. Several studies have focused on cross-country differences in subjective well-being, in particular comparing Europe and the U.S., where the U.S. appears to rank lower in satisfaction than many European countries with lower per capita incomes (Alesina *et al.*, 2004; Blanchflower and Oswald, 2004).

The measurement and analysis of the determinants of life satisfaction is a major issue for social policy. For instance, if one’s satisfaction with income is partly relative (that is, my income satisfaction changes if the incomes or consumption of others around me change), then this may have implications for tax policy. If part of consumption is conspicuous (I buy some goods only to impress others), there may be grounds for taxing that part of consumption differently from other goods. For an excellent discussion of the implications of happiness research for policy, see the recent special issue of the *Journal of Institutional Comparisons* (Frey and Stutzer, 2010).

The literature on satisfaction with life emphasizes the role of income (cf., e.g., Clark *et al.*, 2008) but often analyzes life satisfaction directly, without considering satisfaction with income (see Schyns, 2002). A more detailed picture can be obtained by considering satisfaction with several domains of life. Van Praag *et al.* (2003) introduced a two-stage model where satisfaction with life is a function of satisfaction with several domains of life, including satisfaction with income or financial situation (as well as satisfaction with job, housing, health, leisure, and the environment), and where the domain-specific satisfaction variables are determined by socio-economic characteristics including income. They find that satisfaction with the household financial situation is one of the most important determinants of satisfaction with life of the adult German population. Satisfaction with income has also been studied in the context of poverty (Stanovnik, 1992) or household equivalence scales (e.g., Kapteyn and Van Praag, 1976; Van Praag and Van der Sar, 1988; Charlier, 2002; Schwarze, 2003). Bonsang and Van Soest (2012) compare income satisfaction of the 50+ population in ten European countries.

A fundamental problem in most international comparisons of subjective well-being is that one has to assume that response scales are the same across countries. This critical and largely untested assumption becomes even more tenuous if question phrasings change or differ across surveys, as is often the case (see Stevenson and Wolfers, 2008). Here we address these problems head on. In view of the specific interest of economists in the relation between life satisfaction and income, we focus specifically on satisfaction with income.

The population distribution of reported income satisfaction in a country will first of all depend on the distribution of incomes. Second, residents of different countries can differ in the way they translate any given income amount into a true level of satisfaction with that income. Third, they may differ in the thresholds used

in demarcating income satisfaction into discrete categories like “very satisfied” or “not satisfied.” Holland and Wainer (1993) refer to the latter as *differential item functioning* (DIF).

The latter two factors are often confused in the existing literature on income or life satisfaction. To disentangle them, King *et al.* (2004) proposed to use anchoring vignettes—respondents are asked to evaluate not only themselves, but also hypothetical people whose situation is described in a survey question. This additional information helps to identify interpersonal differences in response scales (DIF), even with cross-section data. True differences in satisfaction (and the effect of actual income on true satisfaction) are then identified by comparisons, keeping response scales constant. Anchoring vignettes have been used to analyze cross-country differences in many domains of well-being, such as political efficacy (King *et al.*, 2004), health (Salomon *et al.*, 2004; Bago d’Uva *et al.*, 2008), job satisfaction (Kristensen and Johansson, 2008), work disability (Kapteyn *et al.*, 2007), or satisfaction with life (Kapteyn *et al.*, 2010). These studies typically show that correcting for DIF alters the country ranking or substantially changes differences across socio-economic groups. Bonsang and Van Soest (2012) find that corrections using vignettes bring the cross-country differences in satisfaction with income more in line with objective income differences.

We analyze data from the U.S. and the Netherlands on satisfaction with income using anchoring vignettes. Our analysis indicates that the biases that flow from not taking into account differences in response scales are very large. In the raw data, Americans are decidedly less satisfied with their income than the Dutch. However, after response scale adjustment based on vignettes, the distribution of satisfaction in the two countries is essentially identical. In addition, we find that the within-country cross-sectional effect of income on satisfaction—a key parameter in the recent debate in the economic literature—is significantly underestimated in the U.S. when differences in response scales are not taken into account.

The remainder of this paper has the following structure. Section 2 describes the data. In Section 3, we summarize the vignette methodology and sketch our statistical model that corrects for DIF. Section 4 presents the empirical results and their implications for interpreting observed differences in income satisfaction in the two countries. In Section 5, simulations are used to ascertain what the counterfactual Dutch distribution of income satisfaction would be if the Dutch would use U.S. thresholds. Section 6 focuses on testing one of the identifying assumptions and checks the robustness of the results when this assumption is relaxed. Section 7 concludes.

## 2. DATA SOURCES AND VIGNETTES

Our analysis is based on information obtained from two internet surveys, which we designed and implemented in the Netherlands and the United States. The samples are not restricted to households with internet access. Respondents are initially recruited by telephone; if they agree to participate and do not have internet access, they are provided with internet access (and if necessary, a set-top box). For the Netherlands, we use the CentERpanel, comprising about 2,250 households who get questions every weekend over the internet. It is representative of the adult

Dutch population except the institutionalized. In 2006, we collected vignette evaluations concerning satisfaction with several domains of life including income (described below).

Our internet survey for the United States is the RAND American Life Panel (ALP). This panel was initially recruited from respondents aged 40 plus in the Monthly Survey (MS) of Michigan's Survey Research Center but has been subsequently supplemented with younger respondents.<sup>1</sup> Similar background information was collected for these respondents as was available for Dutch respondents. The U.S. sample that we use for estimation consists of 1,113 respondents interviewed during 2006–07. In both samples, higher socio-economic status respondents are over-represented due to selective unit non-response. We correct for this using sample weights based upon basic socio-economic and demographic characteristics.

In both samples, respondents were given self-assessment questions and vignettes that cover four life domains that have figured prominently in the happiness and life satisfaction literature—income, family relations, work, and health. In each domain, they were asked to rate themselves on the same five-point scale as they rate the vignette person. This scale is the same for all domains: very satisfied, satisfied, not satisfied or dissatisfied, not satisfied, and very dissatisfied. In this paper we only investigate satisfaction with income. Individual respondents were first asked: “How satisfied are you with the total income in your household?” The vignette questions that were asked after that have the form “(*Name*) is married and has two children; the total after tax household income of his/her family is (*Income<sub>i</sub>*). How satisfied do you think (*Name*) is with the total income of (his/her) household?” Once again, the response categories are very satisfied, satisfied, not satisfied or dissatisfied, not satisfied, and very dissatisfied. *Name* can be either a male or a female name assigned randomly across vignettes.<sup>2</sup> *Income<sub>i</sub>*,  $i = 1, \dots, 4$ , can take four different values corresponding to half the median, the median, twice the median, or four times the median income in the country where the respondent is located. These incomes are also assigned randomly across vignettes. Specifically:

- *Income<sub>1</sub>*: €15,000 (\$23,000)
- *Income<sub>2</sub>*: €30,000 (\$46,000)
- *Income<sub>3</sub>*: €60,000 (\$92,000)
- *Income<sub>4</sub>*: €120,000 (\$184,000).

In each country, *Income<sub>2</sub>* is equal to the median after tax household income in that country.

### 3. THEORY OF VIGNETTES AND ECONOMETRIC MODEL

In this section, we first provide an intuitive description of the use of vignettes for identifying response scale differences and then sketch our econometric model. The basic idea behind the use of vignettes is illustrated in Figure 1, which presents

<sup>1</sup>The MS, the leading consumer sentiments survey, produces the widely used Index of Consumer Attitudes. MS respondents are asked if they have internet access and, if yes, if they are willing to participate in internet surveys. Those who agree are added to our household panel to be interviewed regularly over the internet.

<sup>2</sup>Vignettes were presented in random order to eliminate any order effects.

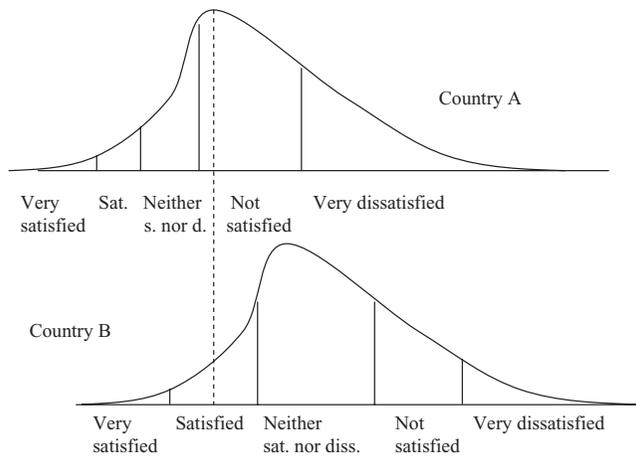


Figure 1. Comparing Self-Reported Income Satisfaction

the distribution of actual living standards or genuine satisfaction (or rather dissatisfaction) with income in two hypothetical countries. The density of the continuous income satisfaction variable in country B is to the right of that in country A, so that on average, people in country B are less satisfied with their income than in country A.

Residents of the two countries also differ in another important sense: they use different response scales if asked to report their satisfaction on the five-point scale that we introduced above. In the example in Figure 1, people in country B use much more positive labels to express their income satisfaction than in country A. Someone in country B with the satisfaction indicated by the dashed line would report to be satisfied, while a person in country A with the same living standard would report “not satisfied.” The frequency distribution of self-reports in the two countries would suggest that people in country B are more satisfied (or less dissatisfied) with their income than those in country A—the opposite of the true income satisfaction distribution. This is DIF. Correcting for DIF is essential to compare the actual income satisfaction distributions in the two countries.

Vignettes can be used to do the correction and to obtain an unbiased estimate of the translation from income to true satisfaction. As explained in Section 2, a vignette question describes the living standard of a hypothetical person and then asks the respondent to evaluate the satisfaction of that person on the same five-point scale that was used for the self-report. Since the vignette descriptions are the same in the two countries, the vignette persons in the two countries have the same genuine income satisfaction. For example, respondents can be asked to evaluate the income satisfaction of a vignette at the dashed line. In country A, this will be evaluated as “not satisfied.” In country B, the evaluation would be “satisfied.” Since the actual level of satisfaction is the same, the difference between the evaluations must be due to DIF.

Vignette evaluations thus help to identify differences between the response scales. Using the scales in one of the two countries as the benchmark, the distribu-

tion of evaluations in the other country can be adjusted by evaluating them on the benchmark scale. The corrected distribution of the evaluations can then be compared to that in the benchmark country—they are now on the same scale. In the example in the figure, this will lead to the correct conclusion that, on average, people in country A are more satisfied with their incomes than people in country B.

The assumptions underlying the vignette corrections are twofold. The first is *response consistency*: a given respondent uses the same scale for self-reports and vignette evaluations. King *et al.* (2004) and Van Soest *et al.* (2011) provide evidence supporting this assumption for vignettes on vision and drinking behavior, respectively, by comparing vignette corrected self-reports with an objective measure. The second assumption is *vignette equivalence*: no systematic differences in the interpretation of a given vignette between the different groups of respondents (so that systematic differences in evaluations are due to DIF only). Since we have given the vignette households an income that relates to the country specific median, this assumption is valid if respondents evaluate on the basis of relative income compared to the country median (in line with relativity of satisfaction; e.g., Van de Stadt *et al.*, 1985). At the time of the survey the median income in the U.S. was, in terms of purchasing power parity, higher than the median income in the Netherlands. As a consequence, it may be the case that U.S. vignettes are interpreted more positively than Dutch vignettes if absolute income level also plays a role. In Section 6, we will discuss how the direction and magnitude of this bias can affect our results.

### 3.1. Econometric Model

We will apply the vignette approach to income satisfaction, using vignettes not only to obtain international comparisons corrected for DIF, but also for comparisons of different groups within a given country. Our model explains respondents' self-reports on satisfaction with their own household incomes as well as their reports on income satisfaction of hypothetical vignette persons. Self-reports are modeled as a function of respondent characteristics  $X_i$  (including household income, a country dummy, and interactions of all characteristics with that dummy) and an error term  $\varepsilon_i$  by the following ordered response equation:

$$(1) \quad Y_i^* = X_i\beta + \varepsilon_i; \quad \varepsilon_i \sim N(0, \sigma^2), \varepsilon_i \text{ independent of } X_i$$

$$(2) \quad Y_i = j \text{ if } \tau_i^{j-1} < Y_i^* \leq \tau_i^j, \quad j = 1, \dots, 5.$$

The thresholds  $\tau_i^j$  between the categories are given by

$$(3) \quad \tau_i^0 = -\infty, \quad \tau_i^5 = \infty, \quad \tau_i^1 = \gamma^1 X_i + u_i, \quad \tau_i^j = \tau_i^{j-1} + \exp(\gamma^j X_i), \quad j = 2, 3, 4$$

$u_i \sim N(0, \sigma_u^2), u_i \text{ independent of } X_i \text{ and the other error terms in the model}$

DIF refers to the fact that different respondents  $i$  use different response scales  $\tau_i^j$ . The term  $u_i$  introduces an unobserved individual effect in the response scale. It implies that evaluations of different vignettes are correlated with each other and

with the self-report (conditional on  $X_i$ ), since some respondents will tend to use high thresholds and others will use low thresholds in all their evaluations.

Define a benchmark respondent with characteristics  $X_i = X(B)$ . The DIF adjustment involves comparing  $Y_i^*$  to thresholds  $\tau_B^j$  rather than  $\tau_i^j$ , where  $\tau_B^j$  is obtained in the same way as  $\tau_i^j$  but using  $X(B)$  instead of  $X_i$ . A respondent's counterfactual reported satisfaction level is computed using a benchmark scale instead of a respondent's own scale. This does not give an adjusted score for each individual (since  $Y_i^*$  is not observed) but it can be used to simulate counterfactual *distributions* of  $Y_i$  for the whole population or conditional upon some of the characteristics in  $X_i$ .

Using self-reports on own income satisfaction only, parameters  $\beta$  and  $\gamma^1$  are not separately identified, only the difference between  $\beta$  and  $\gamma^1$ . For example, consider country dummies: people in two different countries can have systematically different income satisfactions, but if the scales on which they report their income satisfaction can also differ across countries, then self-reports are not enough to identify the income satisfaction difference between the countries. The vignettes will be used to identify  $\beta$  and  $\gamma^1$  separately.

The evaluations  $Y_{li}$  of vignettes  $l = 1, \dots, L$  are modeled similarly:

$$(4) \quad Y_{li}^* = \theta_l + \varepsilon_{li}$$

$$(5) \quad Y_{li} = j \text{ if } \tau_i^{j-1} < Y_{li}^* \leq \tau_i^j, \quad j = 1, \dots, 5$$

$$(6) \quad \varepsilon_{li} \sim N(0, \sigma^2), \text{ independent of each other, of } \varepsilon_i \text{ and of } X_i.$$

The systematic part in equation (4) only contains a dummy for each vignette—not the respondent characteristics  $X_i$  (the assumption of *vignette equivalence*). Since the only variation across vignette descriptions is the level of income (see above), one can interpret the dummies as indicating how the income of the vignette person is translated into genuine satisfaction by the respondent. The translation of genuine satisfaction into verbal labels follows the same scheme as for self-reports. The maintained assumption here is that of “*response consistency*,” meaning that the thresholds  $\tau_i^j$  are the same for self-reports and vignettes.

With these assumptions, it is clear how vignette evaluations can separately identify  $\beta$  and  $\gamma (= \gamma^1, \dots, \gamma^4)$ : From the vignette evaluations alone,  $\gamma, \theta, \theta_1, \dots, \theta_4$  can be identified (up to the usual normalization of scale and location). From self-reports,  $\beta$  can then be identified in addition. Thus the vignettes can be used to solve the identification problem due to DIF. In practice, all parameters are estimated simultaneously using maximum likelihood; see, for example, King *et al.* (2004) for details.

#### 4. EMPIRICAL RESULTS

This section highlights our main empirical findings. We first describe what our data imply for satisfaction with own income in the U.S. and the Netherlands and next summarize the distribution of answers given by Dutch and American respon-

TABLE 1  
REPORTED SATISFACTION WITH HOUSEHOLD INCOME

	NL	U.S.
Very satisfied	9.9	6.5
Satisfied	53.6	39.4
Not satisfied or dissatisfied	23.6	21.5
Not satisfied	10.3	27.4
Very dissatisfied	2.7	5.2

dents to the income vignette questions. The third subsection discusses our main parameter estimates determining true satisfaction with income and the thresholds.

#### 4.1. Levels of Self-Reported Income Satisfaction

Table 1 summarizes how respondents rate satisfaction with their own incomes. In spite of the fact that on average incomes are higher in the U.S. than in the Netherlands (compared in terms of purchasing power parity), Americans appear to be much less satisfied with their incomes than the Dutch are: 64 percent of Dutch respondents say that they are either satisfied or very satisfied with their total household income, compared to 46 percent of the Americans. Similarly, a much larger fraction of Americans respond that they are either not satisfied or very dissatisfied—a third of Americans compared to 13 percent amongst the Dutch. This avoidance of the extremes and rush to the middle in comparison to Americans is a common feature of Dutch responses to subjective scale questions and is similar to what we have found in prior work on other outcomes (Kapteyn *et al.*, 2007).

#### 4.2. Responses to Vignette Questions

Table 2 summarizes responses obtained for both countries per vignette.<sup>3</sup> There are four vignettes which we index 1 to 4, with the lowest number representing the lowest income used in the vignette. As with the ratings of their own incomes, Americans assign lower satisfaction to the vignettes than the Dutch at all four income levels. Note that in terms of PPP comparison, the American vignette households have higher incomes than the Dutch, making the fact that the Americans evaluate the vignette incomes as less satisfactory even more striking.

The between-country differences narrow substantially as we increase incomes of the hypothetical vignette person. For example, at both twice and four times the median country incomes, there are relatively small differences between the Americans and the Dutch. Much larger differences appear at lower country incomes. In the model introduced in Section 3, this can be explained by differences across thresholds in the differences between U.S. and Dutch respondents—these differences would, for example, be larger for the threshold distinguishing dissatisfied and very dissatisfied (which is particularly relevant for the low income vignette)

<sup>3</sup>Comparing the rank ordering of vignette evaluations across respondents shows that different respondents tend to order vignettes in the same way—in less than 0.5 percent (13 cases) of all pairs of vignettes evaluated by the same respondent, the evaluation of the higher income vignette is worse than that of the lower income vignette.

TABLE 2  
VIGNETTE EVALUATIONS FOR INCOME SATISFACTION IN UNITED STATES AND NETHERLANDS

Income Satisfaction Vignettes Income in Vignette	1		2		3		4	
	Half Median		Median		Twice Median		Four Times Median	
	NL	U.S.	NL	U.S.	NL	U.S.	NL	U.S.
Very satisfied	0.8	1.1	5.1	2.7	39.6	38.0	74.9	69.5
Satisfied	6.2	0.7	32.2	23.3	50.0	46.0	20.1	22.7
Not satisfied or dissatisfied	15.4	9.1	40.3	28.9	7.7	10.9	3.1	4.9
Not satisfied	50.7	42.4	20.0	37.9	2.2	4.7	1.0	2.2
Very dissatisfied	27.0	46.8	2.4	7.1	0.6	0.4	1.0	0.7

than for the threshold between satisfied and very satisfied (which drives most of the difference in evaluations of the vignettes with the incomes exceeding the median).

An alternative explanation, however, would be that vignette evaluations depend on country-specific institutions that determine the living standard for a given income. This would particularly affect the evaluation of the low income vignette (half the median), since the support for low income households is more extensive in the Netherlands than in the U.S., for example because of housing subsidies and waivers for local taxes. If this is the case, the assumption of vignette equivalence (VE) is not satisfied for the low income vignette. We will investigate whether this is indeed a problem in Section 6. We will find there that although the model imposing VE in Section 3.1 is formally rejected by a statistical test, more general models that relax this assumption give essentially the same conclusions. We therefore present the parameter estimates of the more parsimonious model imposing VE.

#### 4.3. Parameter Estimates

The model presented in Section 3.1 was estimated by maximum likelihood using the self-evaluations and the vignettes for all four income levels. The equations for genuine income satisfaction and for the thresholds include a complete set of interactions with a country dummy for the U.S. We also estimated the simpler model that does not allow for DIF, which for the self-assessments is similar to a standard ordered probit for self-assessed income satisfaction. Table 3 lists estimated parameters and associated standard errors for genuine satisfaction with income (equation (1)) in the models with (DIF) and without (no DIF) the correction for response scale differences. Differences between these two show the impact of incorporating DIF.

For interpreting the parameter estimates it should be kept in mind that the scale is from good to bad (1: very satisfied, . . . , 5: very dissatisfied) so that a minus sign implies a higher satisfaction level. Demographic variables include dummies for whether the respondent is female, married, and for age categories 40–50, 51–64, 65+ (the omitted group is those under 40 years old). Education is separated into three groups—low, medium, or high, with the low-education group as the omitted

TABLE 3  
SATISFACTION WITH OWN HOUSEHOLD INCOME

	Model without DIF		Model with DIF	
	$\beta$	S.E.	$\beta$	S.E.
Constant	5.555*	0.277	5.252*	0.405
Female	-0.037	0.050	-0.038	0.062
Married	-0.516*	0.086	-0.772*	0.105
Ln(family size)	0.046*	0.075	0.243*	0.092
Age 40–50	-0.085	0.065	-0.085	0.084
Age 51–64	0.007	0.069	0.018	0.085
Age 65+	-0.112	0.108	-0.059	0.131
Ed med	-0.023	0.060	-0.074	0.074
Ed high	-0.280*	0.063	-0.422*	0.078
Unemployed/disabled	0.690	0.098	0.759*	0.116
Retired	-0.053	0.104	0.009	0.124
Other non-working	-0.016	0.067	-0.031	0.085
Ln(income/famsize)	-0.282*	0.025	-0.350*	0.033
<i>Interactions with dummy U.S.</i>				
Constant	5.457*	0.524	7.982*	0.705
Female	0.061	0.081	-0.113	0.106
Married	0.274*	0.118	0.492*	0.150
Ln(family size)	-0.358*	0.112	-0.785*	0.146
Age 40–50	0.140	0.113	0.295*	0.142
Age 51–64	-0.026	0.119	0.104	0.151
Age 65+	-0.137	0.181	-0.027	0.224
Ed med	0.064	0.113	-0.100	0.146
Ed high	0.083	0.111	-0.207	0.148
Unemployed/disabled	-0.601*	0.151	-0.759*	0.185
Retired	-0.374*	0.154	-0.356†	0.190
Other non-working	-0.200†	0.121	-0.137	0.160
Ln(income/famsize)	-0.486*	0.050	-0.743*	0.067

\*Significant at the 5% level; †significant at the 10% level.

category.<sup>4</sup> Income is measured as the logarithm of equivalized family income, using the OECD equivalence scale. Dutch incomes are transformed to US\$ incomes using the same transformation as in the income vignettes—based upon equalizing median incomes in the two countries.<sup>5</sup> Log family size is a separate regressor to account for increasing returns to scale in household consumption. Finally, dummy variables are included, indicating whether the respondent is unemployed or disabled, retired, or not participating for a different reason (e.g., homemaker); working is the omitted category. These dummies serve as proxies for household income sources; the data do not contain information on whether the respondent is the only income provider, or on personal income of the respondent or other household members.

<sup>4</sup>In the U.S., High School Graduate or less is coded as low education, Post College degrees as high education, and the medium group includes all others between these two. In the Netherlands, the medium group has intermediate vocational or general training, and the high education group has higher vocational training or any university degree. The low education category has everyone with primary school only or lower vocational training.

<sup>5</sup>Since incomes are entered in log-form this only affects intercepts in estimated equations.

The model with DIF (adjusting for threshold differences) clearly outperforms the model without DIF according to a likelihood ratio test.<sup>6</sup> It is presented in the last two columns. The first panel shows that in the Dutch sample there are no significant differences in income satisfaction by gender or age. Keeping income and other characteristics constant, respondents drawing unemployment or disability benefits are less satisfied with their income than workers, retired respondents, or other non-participants. Perhaps this is because these respondents were used to receiving a higher income in the past or because they have an income that is lower than what they expected. As expected, higher equivalized family income makes the Dutch significantly more satisfied with their income. Conditional on income and other characteristics, higher education also makes the Dutch more satisfied with their income. One interpretation of this is that those with higher education enjoy higher permanent income. Alternatively, our self-reported household income measure may be imperfect, and education proxies the deviation between this measure and actual family income. Conditional on other factors, partnered Dutch respondents are more satisfied with their income than respondents without a partner. One interpretation is that marriage is an insurance against negative income shocks.<sup>7</sup> Keeping equivalized income constant, larger families in the Netherlands are less satisfied with their income, suggesting that economies of scale are smaller than as implied by the OECD equivalence scales.

The bottom panel presents the estimates of differences in parameters between the U.S. and the Netherlands. There are no statistically significant gender differences in income satisfaction among Americans, similar to what was found for the Dutch, and the estimated education pattern in income satisfaction is not all that different either. The age group 40–50 in the U.S. is found to be significantly less satisfied with their income (keeping income and other factors constant). In contrast to the Dutch, unemployed or disabled Americans are equally satisfied with their income as working respondents with the same income and other characteristics. There are substantially more economies of living in larger families in the U.S. than in the Netherlands, as indicated by the sharply negative estimate for log family size among the Americans. The difference between single people and married couples without children is much smaller than in the Netherlands.

The most important variable for comparing the two countries is income. The impact of log equivalized income on income satisfaction is more than twice as large in the U.S. than in the Netherlands. Conditional on income, higher education makes Americans even more satisfied than it does the Dutch, but the differences in the education effects are not significant.

A central question is how important the corrections for threshold differences within and across countries are in our interpretation of these relationships with income satisfaction. They turn out to be quite important. Comparing the parameter estimates in the model without DIF to the model with DIF shows that the DIF correction changes our conclusion with respect to levels of income satisfaction in both countries. The Dutch main effects mostly retain their sign, but the magni-

<sup>6</sup>The value of the LR test statistic is 1167.5; the p-value is 0.0000.

<sup>7</sup>We added an interaction of the married/living together dummy with gender but this was neither significant in the Netherlands nor in the U.S.

tudes are sometimes quite different. For example, for the Dutch the estimate of high education is  $-0.28$  without DIF and  $-0.42$  with DIF. Similarly, for the Dutch the estimated impact of marital status increases by 50 percent when taking DIF into account. For the Dutch the effect of income changes from  $-0.28$  in the model without DIF to  $-0.35$  in the model with DIF. In the U.S., it changes from  $-0.77$  to  $-1.10$ . Thus particularly in the U.S., the effect of income on income satisfaction is much steeper than conventional models would indicate.

The (interaction of the) U.S. dummy (with the constant term) is difficult to interpret, since other regressors do not have mean zero. Instead, it is better to look at predicted systematic parts for “average” Dutch and U.S. respondents. According to the results allowing for DIF, these predictions are very similar using Dutch and U.S. parameters (1.20 and 1.21 for the Dutch average respondent characteristics and 1.40 and 1.38 for the U.S. average). This indicates that, once a correction for DIF is made, the own income evaluations of the average respondent are very similar in the two countries.

Not correcting for DIF, however, would lead to a different conclusion. The difference between the average Dutch respondent’s predicted systematic part with the Dutch and U.S. parameters would be 0.385 and for the average U.S. respondent it would be 0.392, in contrast to the differences of 0.01 and 0.02 in the DIF case noted above. This would suggest that in the U.S., much higher incomes are needed to achieve a given level of income satisfaction. In fact, this is the interpretation in the existing literature. The result allowing for DIF suggests that this is completely due to the fact that the average U.S. respondent uses a less positive response scale than the average Dutch respondent.<sup>8</sup>

The impact of threshold differences on reported satisfaction with income becomes clear from the threshold parameter estimates in Table 4. In these models a negative coefficient means that a respondent sets a tougher standard on income satisfaction—that is, it takes a higher income for a respondent reporting to be satisfied with that income. Because of this, if an estimated coefficient is positive, more people with this trait will report to be satisfied with their income.

Let us first examine the parameters that apply to the first threshold—separating the “very satisfied” from the “satisfied.” Judging by the estimated main (Dutch) coefficients, there is not a great deal of heterogeneity amongst the Dutch in how they set this threshold. In particular, Dutch income levels do not seem to affect this threshold very much. There are several significant American interactions on attributes, implying that Americans are much more heterogeneous in this. In particular, the negative coefficient on the income interaction with the U.S. dummy indicates that higher income makes Americans more demanding on this particular threshold. Higher income Americans are less likely to say they are very satisfied with the same level of high income than are low income Americans. Two other attributes that appear to matter are family size and education: an increase in family size or in education (controlling for equalized income) makes an American respondent more demanding.

<sup>8</sup>An equivalent way of showing this would be to define the regressors in deviations from their (overall, Dutch, or U.S.) means. The U.S. dummy would then become insignificant and slightly negative in the model with DIF, and positive and significant according to the model without DIF.

TABLE 4  
THRESHOLD PARAMETERS  $\gamma^1, \dots, \gamma^4$

	Threshold 1		ln (Threshold 2 – Threshold 1)		ln (Threshold 3 – Threshold 2)		ln (Threshold 4 – Threshold 3)	
	$\gamma^1$	S.E.	$\gamma^2$	S.E.	$\gamma^3$	S.E.	$\gamma^4$	S.E.
Constant	0	–	0.466*	0.232	0.732*	0.336	–0.994*	0.434
Female	0.044	0.059	–0.033	0.036	–0.011	0.053	0.065	0.052
Married	–0.146	0.094	–0.031	0.061	0.016	0.089	0.068	0.087
ln family size	0.118	0.082	0.050	0.053	0.000	0.079	0.094	0.079
Age 40–50	0.092	0.077	–0.077†	0.046	0.110	0.068	–0.007	0.067
Age 51–64	0.079	0.078	–0.106*	0.048	0.249*	0.071	–0.043	0.069
Age 65+	0.157	0.122	–0.138†	0.077	0.319*	0.109	0.004	0.125
Ed med	0.051	0.067	–0.078†	0.041	–0.032	0.059	0.161*	0.065
Ed high	–0.076	0.073	0.003	0.043	–0.089	0.064	0.204*	0.068
Unemployed/disabled	–0.037	0.119	0.004	0.076	–0.034	0.098	0.069	0.099
Retired	0.026	0.117	0.059	0.072	–0.029	0.099	0.058	0.119
Other non-working	0.002	0.076	–0.003	0.046	–0.015	0.071	0.021	0.076
ln eq income	–0.034	0.035	0.020	0.022	–0.083*	0.033	0.102*	0.042
<i>Interactions with dummy U.S.</i>								
Dummy	2.590*	0.642	–0.909†	0.549	0.860	0.656	2.149*	0.646
Female	–0.173†	0.096	0.062	0.063	–0.291*	0.090	–0.033	0.077
Married	0.172	0.135	–0.034	0.093	0.108	0.130	–0.114	0.111
ln family size	–0.313*	0.130	–0.020	0.089	–0.050	0.124	–0.130	0.107
Age 40–50	0.166	0.130	0.004	0.087	–0.089	0.124	–0.087	0.105
Age 51–64	0.077	0.134	0.096	0.089	–0.273*	0.130	0.016	0.105
Age 65+	0.054	0.207	0.103	0.137	–0.248	0.205	0.019	0.180
Ed med	–0.419*	0.129	0.167†	0.093	0.150	0.121	–0.170†	0.103
Ed high	–0.447*	0.129	0.047	0.093	0.214†	0.123	–0.232*	0.106
Unemployed/disabled	–0.167	0.193	0.026	0.133	0.056	0.166	–0.130	0.138
Retired	0.153	0.175	–0.100	0.112	–0.049	0.169	–0.171	0.156
Other non-working	0.069	0.172	–0.015	0.104	–0.012	0.156	–0.088	0.127
ln eq income	–0.238*	0.061	0.065	0.053	–0.098	0.065	–0.159*	0.063

\*Significant at the 5% level; †significant at the 10% level.

The other parameters concern differences between thresholds (see equation (3)). For example, the estimates of  $\gamma^4$  in the last column concern the distance between the third and fourth threshold, driving the tendency to answer “dissatisfied.” The significantly positive estimate of the income coefficient in the Netherlands (0.102) shows that the Dutch tendency to answer dissatisfied increases with income. This is not the case in the U.S., where the income effect is negative (0.102–0.159 = –0.057) and not significant.

Since the coefficients on the U.S. dummy are hard to interpret due to the interactions, we also present the predicted thresholds for the average Dutch and U.S. respondents with the Dutch as well as the U.S. parameters; see Table 5. This confirms what we already saw in the data: the thresholds using the U.S. parameters are always significantly lower (and very similar using the average U.S. or the average Dutch sample characteristics), indicating that response scales of U.S. respondents tend to produce less satisfied responses than response scales of Dutch respondents.

The estimates of the additional model parameters are presented in Table 6. The dummies for the vignettes fall significantly with the vignette income, as

TABLE 5  
PREDICTED THRESHOLDS FOR AVERAGE RESPONDENTS (MODEL WITH DIF)

	Threshold 1	Threshold 2	Threshold 3	Threshold 4
Dutch sample mean, Dutch parameters	-0.254	1.525	2.543	3.926
U.S. sample mean, Dutch parameters	-0.250	1.578	2.576	3.917
Dutch sample mean, U.S. parameters	-0.444	1.064	1.840	3.426
U.S. sample mean, U.S. parameters	-0.437	1.067	1.844	3.462

TABLE 6  
ESTIMATES OF OTHER PARAMETERS (MODEL WITH DIF)

	Estimate	S.E.
Vignette dummies (equation 4)		
$\theta_1$	3.379	0.375
$\theta_2$	1.812	0.371
$\theta_3$	0.039	0.370
$\theta_4$	-1.034	0.368
Standard deviations		
$\sigma_i$ (thresholds; equation 3)	0.609	0.021
$\sigma$ (vignette evaluations; equation 4)	1.016	0.023

expected. The variance of the error term in the vignette equation is almost identical to (and not significantly different from) the variance of the error term in the self-assessments (normalized to 1). We find clear evidence of unobserved heterogeneity in the thresholds, with an estimated standard deviation of 0.609, explaining about 25 percent of the total unsystematic variation in self-reports and vignette evaluations.

## 5. MODEL SIMULATIONS

To understand the implications of our approach, we simulated the distribution of satisfaction with income in the two countries for different parameter values. Essentially we first simulate the Dutch distribution of self-reported income satisfaction and then replace various sets of parameters by the corresponding U.S. values. Table 7 presents the results of these simulations by age group—less than 40 years, 40–50 years, 50–64 years, and 65 years or over—and for all ages. The first row for each age group summarizes the distribution of satisfaction with income for the Dutch using their own parameters, which roughly reproduces the Dutch data. The second row simulates the Dutch distribution if we replace the parameters in the Dutch satisfaction equation (i.e., Table 3 with DIF) by the American parameters. The third row replaces Dutch thresholds by American thresholds (cf. Table 4). The fourth row replaces all Dutch parameters by American parameters. The fifth row simulates distributions for the American sample using American parameters (roughly reproducing the U.S. data).

Comparing the first and fifth row in the age group less than 40 years old, for example, shows that 57.5 percent of the Dutch report that they are satisfied or very satisfied with their incomes, compared to 38.1 percent of Americans. The second

TABLE 7  
SIMULATIONS BASED UPON MODEL WITH DIF (USING ALL VIGNETTES): DISTRIBUTION OF INCOME SATISFACTION (%)

	Very Satisfied	Satisfied	Not Satisfied/ Dissatisfied	Dissatisfied	Very Dissatisfied
<i>Age group younger than 40</i>					
Dutch using own thresholds	9.5	49.5	23.9	14.9	2.2
Dutch using U.S. self-assess equation	12.3	49.0	20.5	12.7	5.5
Dutch using U.S. thresholds	5.2	32.4	26.0	32.3	4.2
Dutch using all U.S. parameters	6.7	34.3	23.4	29.9	5.7
U.S. using U.S. thresholds	6.0	30.7	22.2	33.1	8.0
<i>Age group 40–50</i>					
Dutch using own thresholds	12.4	49.1	24.9	12.0	1.6
Dutch using U.S. self-assess equation	11.6	45.4	24.8	13.7	4.5
Dutch using U.S. thresholds	9.2	36.3	26.1	24.9	3.5
Dutch using all U.S. parameters	7.8	33.4	25.1	27.9	5.8
U.S. using U.S. thresholds	8.9	34.6	23.2	27.5	5.8
<i>Age group 50–64</i>					
Dutch using own thresholds	11.3	45.6	28.5	12.4	2.1
Dutch using U.S. self-assess equation	13.1	44.6	26.5	11.7	4.1
Dutch using U.S. thresholds	9.1	37.2	23.4	26.4	3.8
Dutch using all U.S. parameters	9.8	37.6	22.7	26.0	4.5
U.S. using U.S. thresholds	10.3	36.6	20.1	27.0	6.0
<i>Age group 65 and older</i>					
Dutch using own thresholds	15.9	50.8	25.2	7.4	0.7
Dutch using U.S. self-assess equation	23.2	48.6	20.0	6.1	2.1
Dutch using U.S. thresholds	15.5	42.9	22.1	18.1	1.4
Dutch using U.S. thresholds	21.8	43.2	18.2	15.0	1.7
U.S. using U.S. thresholds	16.8	38.1	21.6	21.0	2.5
<i>All age groups</i>					
Dutch using own thresholds	11.9	48.7	25.6	12.0	1.7
Dutch using U.S. self-assess equation	14.6	46.9	23.0	11.3	4.2
Dutch using U.S. thresholds	9.2	36.7	24.6	26.1	3.4
Dutch using all U.S. parameters	10.8	36.8	22.4	25.4	4.6
U.S. using U.S. thresholds	9.2	34.0	21.9	28.7	6.2

row shows that replacing Dutch parameters in the own income satisfaction equation by American parameters does not substantially change this difference. In fact, the distributions in the second row of each panel are similar to those in the first row. Apparently, the source of the difference between the Dutch and the Americans does not fundamentally lie in differences in their respective income satisfaction equations.

However, if the Dutch had American thresholds instead of their own (the third row of each panel), the situation would be quite different: the Dutch distribution of income satisfaction in all age groups then looks almost identical to the American distribution. That conclusion does not change appreciably if we also assign the American satisfaction parameters to the Dutch, as one would expect after comparing the first and second row. Thus, the results strongly suggest that most of the observed differences in the raw data between the Dutch and the Americans lie in the scales they use (the thresholds separating the various verbal labels).

We will illustrate why this happens by first considering the highest “very satisfied” threshold. Our estimates indicate both that own income increases overall

income satisfaction (Table 3) and that high income Americans have more demanding standards than the Dutch on what income is necessary to be very satisfied with income (Table 4).

Since income satisfaction is increasing in income, attributes of respondents around the threshold between very satisfied and satisfied are those of higher income respondents. Thus, we should be using the comparative thresholds of higher income Americans and higher income Dutch in making the Dutch adopt the American thresholds. Our estimates show that higher income Americans are more demanding than higher income Dutch, so having the Dutch look like Americans at the very satisfied threshold basically makes the Dutch set a higher standard (higher income) for claiming to be very satisfied with their incomes. Consequently, fewer Dutch will claim that they are very satisfied with their incomes. Moreover this effect is strong enough to make the hypothetical Dutch distribution of very satisfied almost identical to the American one.

Next examine the other end of the distribution—the threshold between dissatisfied and very dissatisfied. The positive association of own income with income satisfaction now implies that on average attributes of respondents around this threshold are those of lower income Dutch and American respondents. The estimated steeper effect of income on this threshold now implies that Americans would be less demanding than the Dutch. That is, they would be less likely to translate a satisfaction level into the verbal category “very dissatisfied.” On the other hand, however, the coefficient on income in the satisfaction equation is larger for Americans, which implies a lower level of satisfaction with income at low levels. Indeed we see (comparing the first two rows in each panel) that giving the Dutch the U.S. satisfaction parameters leads to an increase in the number of Dutch who are classified as very dissatisfied for most age groups.

## 6. VIGNETTE EQUIVALENCE AND ROBUSTNESS CHECKS

As discussed in Section 4.2, it might be the case that (part of) the cross-country difference in the low income vignette evaluations is not due to response scale differences but to institutional differences, implying that a given income corresponds to a higher living standard in the Netherlands than in the U.S. In that case the assumption of vignette equivalence would not hold and our corrections for response scale differences would be biased. To investigate whether this is a problem, we followed Bago d’Uva *et al.* (2011)<sup>9</sup> who consider an extension of the model of Section 3 that does not impose vignette equivalence.<sup>10</sup>

The extension replaces the part of the model for the vignette evaluations  $Y_{li}$ ,  $l = 1, 2, 3, 4$  (equations (4)–(6) in Section 3.1) by:

<sup>9</sup>Bago d’Uva *et al.* (2011) also discuss a test for response consistency. We do not consider this test since we do not have enough objective information on income satisfaction.

<sup>10</sup>Since it is particularly the first (low income) vignette that might cause the problem, we also simply estimated the standard model not using the low income vignette evaluations. This gave estimation and simulation results similar to those in Sections 4 and 5. The evaluations of the three remaining vignettes should be much less affected by the institutional differences, which mainly relate to support of poor households. This gave estimation and simulation results similar to those in Sections 4 and 5.

$$(4') \quad Y_{li}^* = \theta_l + X_i^- \lambda_j + \varepsilon_{li}$$

$$(5') \quad Y_{li} = j \text{ if } \tau_i^{j-1} < Y_{li}^* \leq \tau_i^j, \quad j = 1, \dots, 5$$

$$(6') \quad \varepsilon_{li} \sim N(0, \sigma^2), \text{ independent of each other, of } \varepsilon_i \text{ and of } X_i$$

Here  $X_i^-$  is  $X_i$  without the constant term (following the notation of Bago d'Uva *et al.*). The standard model in Section 3.1 is a special case of this, imposing  $\lambda_j = 0$ ,  $j = 1, 2, 3, 4$ . In the extended model, one of the  $\lambda_j$  has to be set to zero for identification. Setting  $\lambda_4 = 0$  (that is, using the high income vignette as the benchmark), a test for vignette equivalence proposed by Bago d'Uva *et al.* is a test for the null hypothesis:

$$(VE) \quad \lambda_1 = 0; \lambda_2 = 0; \lambda_3 = 0.$$

This hypothesis can be tested using a likelihood ratio test (estimating the standard model and the extended model with maximum likelihood). If another benchmark vignette is used, the extended model is parameterized differently, but the likelihood ratio test remains the same.

The result of this test is that VE is rejected: the test statistic is 266.3, exceeding all common critical values of a chi squared distribution with 75 degrees of freedom (96.2 for a 5% level; 106.4 for a 1% level). The estimates of  $\lambda_1, \lambda_2, \lambda_3$  (not presented to save space) show that indeed, several covariates are significant, particularly in  $\lambda_1$ . This is to be expected since  $\lambda_1$  refers to the differences between the two most extreme vignettes.

VE was also rejected for vignettes on mobility and cognition by Bago d'Uva *et al.* (2011). This result raises the question whether the violation of VE also leads to a substantial bias in the main result of interest, the finding that response scale differences explain most of the differences in the distributions of reported income satisfaction in the U.S. and the Netherlands. This finding from Section 5 is reproduced in the first row of Table 8 (from the panel "all age groups" in Table 7). In particular, it shows that the percentages in the column "Dutch sample, Dutch self-assessments, U.S. thresholds" are similar to those in the column "U.S. sample, U.S. self-assessments, U.S. thresholds," while they are quite different from the percentages in the column "Dutch sample, Dutch self-assessments, Dutch thresholds."

The extended model imposes, somewhat arbitrarily, that vignette equivalence holds for the first vignette but not for other vignettes. We can repeat the same counterfactual simulations on the basis of the extended model, essentially using only the fourth vignette to identify response scale differences between countries and socio-economic groups. (The other vignettes identify  $\lambda_1, \lambda_2, \lambda_3$ .) This gives the results in the second row of Table 8. The numbers are different and the corrections for response scale differences are somewhat smaller than in the first row, but the qualitative conclusion remains the same as the one based upon the standard model: correcting for DIF alone brings the Dutch income satisfaction reports much closer to the American reports, since Americans less easily say they are satisfied or very satisfied with their household income.

TABLE 8  
SIMULATIONS BASED UPON VARIOUS MODELS: DISTRIBUTION OF INCOME SATISFACTION (%); ALL AGE GROUPS

Model	Very Satisfied				Satisfied				Not Satisfied/Dissatisfied				Dissatisfied				Very Dissatisfied			
	Dutch	Dutch	U.S.	U.S.	Dutch	Dutch	U.S.	U.S.	Dutch	Dutch	U.S.	U.S.	Dutch	Dutch	U.S.	U.S.	Dutch	Dutch	U.S.	U.S.
Model 1	11.9	9.2	9.2	9.2	48.7	36.7	34.0	34.0	25.6	24.6	21.9	12.0	12.0	26.1	28.7	28.7	1.7	4.6	4.6	6.2
Model 2	12.6	9.4	8.9	8.9	48.2	38.8	34.8	34.8	25.0	24.3	22.0	12.1	12.1	24.5	27.4	27.4	2.2	2.9	2.9	6.9
Model 3	12.6	8.9	8.9	8.9	48.2	33.5	34.8	34.8	25.0	23.9	22.0	12.1	12.1	29.0	27.4	27.4	2.2	4.6	4.6	6.9
Model 4	12.2	9.6	9.4	9.4	48.6	36.6	34.5	34.5	25.1	24.3	21.6	11.9	11.9	25.9	27.2	27.2	2.1	3.6	3.6	7.3
Model 5	13.5	7.9	11.3	11.3	48.1	34.4	38.9	38.9	24.2	22.7	21.8	12.6	12.6	29.4	24.3	24.3	1.6	5.7	5.7	3.7
Model 6	12.7	6.7	9.2	9.2	48.7	34.6	34.6	34.6	24.8	24.8	21.8	12.2	12.2	29.7	27.6	27.6	1.6	4.2	4.2	6.7

Model 1: standard model with DIF presented in Section 4 and used for simulations in Section 5.  
 Model 2: extended model presented in Section 6 using vignette 4 (high income) as the benchmark vignette.  
 Model 3: extended model presented in Section 6 using vignette 1 (low income) as the benchmark vignette.  
 Model 4: extended model of Section 6 with vignette 4 (low income) as benchmark vignette without interactions with U.S. dummy in  $\lambda_1, \lambda_2, \lambda_3$ .  
 Model 5: standard model without regressors except country dummy.  
 Model 6: standard model without regressors except country dummy, In equivalized income, and their interactions.

We repeat the same exercise with a different benchmark vignette. In row 3 we present the counterfactual simulations using the first vignette as the benchmark, implying that corrections for DIF are based upon the low income vignettes only. Again, the numbers are different, but the qualitative conclusion remains the same. The corrections for response scale differences are larger than in the second row, and bring the Dutch even closer to the Americans. Given the institutional differences favoring the poor in the Netherlands, these corrections may be too strong, and we should probably attach more value to the results in row 2 assuming vignette equivalence for the high income vignettes. But the main message is the robustness of the correction.

Why is it that, in spite of the fact that different vignettes lead to significantly different corrections for DIF, the counterfactual distributions comparing the Netherlands and the U.S. are so robust for the choice of vignettes? Part of the explanation seems to be that the coefficients on the interactions with the U.S. dummy in  $\lambda_1$ ,  $\lambda_2$ ,  $\lambda_3$  are much smaller and less often significant than the coefficients on the non-interacted socio-demographics. This suggests that vignettes are often not interpreted in the same way by different socio-economic groups in a given country, but the differences are much smaller between groups with the same socio-economic characteristics in the two countries. Since the counterfactuals compare countries and not socio-economic groups, they are not affected by the former type of violation of vignette equivalence.

This can be explored further by estimating the extended model without the interactions in  $\lambda_1$ ,  $\lambda_2$ ,  $\lambda_3$ . This specification is rejected at the 5% but not at the 1% level against the more general extension with all interactions in  $\lambda_1$ ,  $\lambda_2$ ,  $\lambda_3$  (LR test statistic 60.52; p-value 0.0152). And indeed, the counterfactual simulations using this model lead to results similar to the standard model imposing vignette equivalence, as shown in the fourth row of Table 8.

The final part of Table 8 shows the results of additional robustness checks of our main finding. Our dataset does not contain, for example, information on personal income or the number of income providers in the household, which might drive income satisfaction. This raises the question whether the simulation results are sensitive to the regressors included in the equations. The last two rows show what happens in the extreme case where the set of regressors is minimized to either just the U.S. dummy (row 5) or the U.S. dummy, the log of equivalized family income, and their interaction (row 6). These two models lead to slightly bigger changes in the predicted distribution of income satisfaction in each country, but do not change the main conclusion that most of the reporting difference can be explained by a cross-country difference in reporting scales. Experiments with other subsets of the regressors led to the same conclusion.

Finally, we also performed several tests for VE as discussed in Rice *et al.* (2011). First, they consider systematic reversals of the order of vignette evaluations. Since our vignettes are ordered in the natural way (that is, a higher income vignette is evaluated as more satisfied with income than a lower income vignette), this does not lead to any doubts on VE. Second, they consider the differences in evaluations by male and female respondents of vignettes describing male and female persons. Here we find small but statistically significant differences, implying that VE is formally rejected. Again, however, accounting for this

in the model does not affect the qualitative conclusions (details available upon request).<sup>11</sup>

## 7. CONCLUSIONS

In this paper, we have used vignettes to disentangle determinants of income satisfaction within and across countries from the verbal scales people use to express this satisfaction. We find that the verbal scales are substantially different in the Netherlands and the U.S. and also across respondents within a country, particularly in the U.S. Correcting for the differences leads to very similar distributions of income satisfaction in the two countries, in sharp contrast to the large differences in the raw data. In this case, not adjusting for response scale differences between countries can lead to misleading conclusions about cross-country differences in income satisfaction. At the same time, applying the DIF correction has an appreciable impact on some of the parameters in the income satisfaction equation. For instance, the effect of own income on income satisfaction of Americans increases substantially.

Anchoring vignettes provide additional information, but making use of this requires additional assumptions, as discussed in Section 3. One assumption is VE, which implies that a hypothetical family with U.S. median income is evaluated in the U.S. in the same way as a family with the Dutch median income in the Netherlands, keeping response scales constant. A deviation from VE in the sense that absolute income matters would reinforce our findings, since the PPP adjusted median income in the U.S. is higher than in the Netherlands. Not using the low income vignette—which might be interpreted in different ways because of institutional differences in giving support to poor households—does not change our qualitative conclusions at all. We can therefore be confident that our conclusions are not biased because of violations of VE. It also raises, however, the methodological question of how essential VE equivalence is for the vignette correction approach. Our analysis cannot provide a conclusive answer to this question, since VE for one vignette is still needed as an identifying assumption. More work on testing VE seems fruitful but also requires richer data.

In the debate about the cross-national relation between income and income satisfaction, the incomparability of response scales has been long recognized. Vignettes are an obvious instrument to get at the incomparability issue. Our results suggest that after the vignette corrections the distributions of income satisfaction are not all that different across the two countries, although we do find a stronger relation between satisfaction and income in the U.S. than in the Netherlands, which may reflect institutional differences. It will be of interest to expand the analysis to more countries to see what light vignette corrections shed on the Easterlin paradox. Other extensions requiring richer data could consider the role of the income distribution and differences in preferences for income inequality. Such differences might explain why reporting thresholds differ across countries or socio-economic groups within one country.

<sup>11</sup>A third type of tests of vignette equivalence in Rice *et al.* (2011) uses country groupings. We cannot perform these tests since we only have two countries.

Increasingly policy makers appear to be interested in measuring subjective well-being of citizens (cf. Stiglitz *et al.*, 2009), and income is one of the factors driving subjective well-being that is amenable to policy. As with traditional measures like GDP per capita, one expects policy makers to have an interest in calibrating the performance in a given country to other countries. In order to be able to do that, comparability of the measures used is an obvious necessary condition. This paper illustrates that correcting income satisfaction reports for differences in response scales has a substantial effect on the conclusions one might draw from such comparisons.

## REFERENCES

- Alesina, Alberto, Rafael Di Tella, and Robert J. McCulloch, "Inequality and Happiness: Are Europeans and Americans Different?" *Journal of Public Economics*, 88(9–10), 2009–42, 2004.
- Bago d'Uva, Teresa, Eddy Van Doorslaer, Maarten Lindeboom, and Owen O'Donnell, "Does Reporting Heterogeneity Bias the Measurement of Health Disparities?" *Health Economics*, 17(3), 351–75, 2008.
- Bago d'Uva, Teresa, Maarten Lindeboom, Owen O'Donnell, and Eddy Van Doorslaer, "Slipping Anchor? Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity," *Journal of Human Resources*, 46(4), 875–906, 2011.
- Blanchflower, David G. and Andrew J. Oswald, "Well-Being over Time in Britain and the USA," *Journal of Public Economics*, 88(7–8), 1359–86, 2004.
- Bonsang, Eric and Arthur Van Soest, "Satisfaction with Job and Income among Older Individuals across European Countries," *Social Indicators Research*, 105(2), 227–54, 2012.
- Charlier, Erwin, "Equivalence Scales for the Former West-Germany," *Review of Income and Wealth*, 48(1), 99–126, 2002.
- Clark, Andrew E., Paul Frijters, and Michael Shields, "Relative Income, Happiness, and Utility: An Explanation for the Easterlin Paradox and Other Puzzles," *Journal of Economic Literature*, 46(1), 95–144, 2008.
- Deaton, Angus, "Income, Aging, Health and Well-Being around the World: Evidence from the Gallup World Poll," *Journal of Economic Perspectives*, 22(1), 53–72, 2008.
- Di Tella, Rafael, Robert J. McCulloch, and David G. Blanchflower, "The Macroeconomics of Happiness," *Review of Economics and Statistics*, 85(4), 809–27, 2003.
- Easterlin, Richard A., "Does Economic Growth Improve the Human Lot? Some Empirical Evidence," in Paul A. David and Melvin W. Reder (eds), *Nations and Households in Economic Growth: Essays in honor of Moses Abramowitz*, Academic Press, New York, pp. 89–125, 1974.
- , "Will Raising the Incomes of All Increase the Happiness of All?" *Journal of Economic Behavior and Organization*, 27(1), 35–48, 1995.
- Frey, Bruno S. and Alois Stutzer, "Happiness: A New Approach in Economics," *Journal for Institutional Comparisons*, 8(4), 3–7, 2010.
- Holland, Paul W. and Howard Wainer, *Differential Item Functioning*, Lawrence Erlbaum, Hillsdale, NJ, 1993.
- Kapteyn, Arie and Bernard M. S. Van Praag, "A New Approach to the Construction of Family Equivalence Scales," *European Economic Review*, 7(4), 313–35, 1976.
- Kapteyn, Arie, James P. Smith, and Arthur Van Soest, "Vignettes and Self-Reports of Work Disability in the U.S. and the Netherlands," *American Economic Review*, 97(1), 461–73, 2007.
- , "Life Satisfaction," in Ed Diener, John F. Helliwell, and Daniel Kahneman (eds), *International Differences in Well-Being*, Oxford University Press, Oxford, pp. 70–104, 2010.
- King, Gary, Christopher J. L. Murray, Joshua A. Salomon, and Ajay Tandon, "Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research," *American Political Science Review*, 98(1), 567–83, 2004.
- Kristensen, Nicolai and Edvard Johansson, "New Evidence on Cross-Country Differences in Job Satisfaction Using Anchoring Vignettes," *Labour Economics*, 15(1), 96–117, 2008.
- Rice, Nigel, Silvana Robone, and Peter Smith, "Analysis of the Validity of the Vignette Approach to Correct for Heterogeneity in Reporting Health System Responsiveness," *European Journal of Health Economics*, 12(2), 141–62, 2011.
- Salomon, Joshua A., Ajay Tandon, and Christopher J. L. Murray, "Comparability of Self Rated Health: Cross Sectional Multi-Country Survey Using Anchoring Vignettes," *British Medical Journal*, 328, 258–60, 2004.

- Schwarze, Johannes, "Using Panel Data on Income Satisfaction to Estimate Equivalence Scale Elasticity," *Review of Income and Wealth*, 49(3), 359–71, 2003.
- Schyns, Peggy, "Wealth of Nations, Individual Income and Life Satisfaction in 42 Countries: A Multilevel Approach," *Social Indicators Research*, 60(1–3), 5–40, 2002.
- Stanovnik, Tine, "Perception of Poverty and Income Satisfaction: An Empirical Analysis of Slovene Households," *Journal of Economic Psychology*, 13(1), 57–69, 1992.
- Stevenson, Betsey and Justin Wolfers, "Economic Growth and Subjective Well-Being: Reassessing the Easterlin Paradox," *Brookings Papers on Economic Activity*, 39(1), 1–87, 2008.
- Stiglitz, Joseph E., Amartya Sen, and Jean-Paul Fitoussi, *Report of the Commission on the Measurement of Economic Performance and Social Progress*, Commission on the Measurement of Economic Performance and Social Progress, Paris, 2009.
- Van de Stadt, Huib, Arie Kapteyn, and Sara Van de Geer, "The Relativity of Utility: Evidence from Panel Data," *Review of Economics and Statistics*, 67(2), 179–87, 1985.
- Van Praag, Bernard M. S. and Ada Ferrer-i-Carbonell, *Happiness Quantified—A Satisfaction Calculus Approach*, Oxford University Press, Oxford, 2008.
- Van Praag, Bernard M. S. and Nico L. Van der Sar, "Household Cost Functions and Equivalence Scales," *Journal of Human Resources*, 23(2), 193–210, 1988.
- Van Praag, Bernard M. S., Paul Frijters, and Ada Ferrer-i-Carbonell, "The Anatomy of Subjective Well-Being," *Journal of Economic Behavior and Organization*, 51(1), 29–49, 2003.
- Van Soest, Arthur, Liam Delaney, Colm Harmon, Arie Kapteyn, and James P. Smith, "Validating the Use of Vignettes for Subjective Threshold Scales," *Journal of the Royal Statistical Society Series A*, 174(3), 575–95, 2011.