Accepted Manuscript

On the comparison of group inequalities using subjective data

Xavier Fontaine, Luke Haywood

PII: S0165-1765(17)30446-9
DOI: https://doi.org/10.1016/j.econlet.2017.11.001
Reference: ECOLET 7826

To appear in: Economics Letters

Received date: 16 March 2017
Revised date: 26 October 2017
Accepted date: 1 November 2017

Please cite this article as: Fontaine X., Haywood L., On the comparison of group inequalities using subjective data. Economics Letters (2017), https://doi.org/10.1016/j.econlet.2017.11.001

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
We study the validity of comparisons of subjective health across groups.

We use SHARE data on subjective health reports to rank (levels and variances) for individuals with different characteristics.

Using anchoring vignettes we show that differential reporting is much more problematic for comparisons (across groups) of inequality than for comparisons of levels.

This raises questions about recent contributions comparing levels of inequality in subjective well-being across groups.
On the comparison of group inequalities using subjective data

Xavier Fontaine∗& Luke Haywood †

Abstract

Subjective data offer a simple way of comparing outcomes across populations. However, reporting behavior may differ across groups. We investigate the validity of studies comparing inequality based on such data. Comparing group inequalities appears substantially more problematic than comparing averages.

JEL: I14, I32, C83

Keywords: Differences in reporting; inequality; subjective data

∗Paris School of Economics
†Corresponding author. DIW Berlin, Mohrenstrasse 58, D-10117 Berlin, LHaywood@diw.de.

Subject to the usual disclaimer we would like to thank Andrew Clark, François Fontaine, Claudia Senik, Arthur van Soest, Maarten Lindeboom, Costas Meghir and an anonymous referee.
1 Introduction

Subjective data on health, job satisfaction, trust and other values are commonly used in economics, since objective data are difficult to collect in these areas. While comparisons across individuals based on subjective data (e.g. wellbeing and health) are frequent, their validity has also often been called into question. Gross differences in subjective data across groups indeed confound different reporting scales (what “good” health means) and differences in underlying latent values. We here focus on the validity of comparisons of inequality across groups. Such comparisons have become increasingly popular, in line with concerns about high levels of inequality. We use the standard “vignette” method to account for differences in reporting across groups. While we find that comparing mean levels of subjective values across groups is quite informative, comparing levels of inequality seems much less relevant. This raises questions for the recent literature on wellbeing inequality.

Comparisons of levels of inequality across populations based on subjective data is a burgeoning research area. The previous focus of the literature on subjective well-being on comparing means has indeed shifted to considering the whole distribution. Recent contributions that explicitly compare inequality using subjective measures include Fahey and Smyth (2004), Kalmijn and Veenhoven (2005), Ott (2005), Greve (2010), Delhey and Kohler (2011), OECD (2011), Clark et al. (2014), Jordá and Sarabia (2015) and John Helliwell and Wang (2016) \(^1\). In a prominent recent contribution, Goff et al. 2016 consider the influence of wellbeing inequality on life satisfaction, based inter alia on cross-country comparisons. Representative for this literature, they affirm “[i]n common with other SWL papers, we take interpersonal comparability for granted.” The validity of this comparability has however only been tested with respect to comparing group means. This paper is the first to test the validity of comparisons of variances of subjective data.

\(^1\)For policy documents, see New Economics Foundation (2015) and The Happiness Research Institute (2015). Inequality in wellbeing was analyzed as a factor in Britain’s vote to leave the EU (Abdallah (2015)).
across populations.

Several studies show evidence of reporting bias when comparing group means based on subjective data. A first approach proxies latent health with an objective indicator, and focuses on the discrepancies between subjective and objective measures (Murray and Chen 1992, Kerkhofs and Lindeboom 1995, Mathers and Douglas 1998, Lindeboom and van Doorslaer 2004, Etilé and Milcent 2006, Jürges 2007). A second approach uses ‘anchoring vignettes’ to establish differences in scale interpretation across respondents. Respondents are asked to assess the situation of a hypothetical person on the same response scale they use to evaluate their own situation (Salomon et al. 2004, Kapteyn et al. 2007, d’Uva et al. 2008). This method has been applied to reporting poverty (Beegle et al. 2012), job satisfaction (Kristensen and Johansson 2008), life satisfaction (Kapteyn et al. 2010, Angelini et al. 2014) and income satisfaction (Kapteyn et al. 2013). All researchers find substantial reporting heterogeneity. Unless differences in reporting are uncorrelated with differences in latent values, this will bias comparisons across groups. These contributions consider the role of reporting heterogeneity for comparisons in levels. We here consider the specific issues involved in how differences in reporting may influence comparisons of inequality.

Section (2) extends the traditional anchoring vignette model to include group inequalities. Section (3) describes the subjective health data we use. Section (4) presents results. First, differences in reporting do not invalidate comparisons of group averages. Second, and by contrast, the bias caused by differential reporting makes comparisons of health inequality across populations largely meaningless. Section (5) concludes.

2 Model and Identification

Comparing subjective reports on health or wellbeing across groups raises a number of specific issues. Assume we are interested in comparing underlying latent levels of $H$, but only observe reports $R$. Inference from differences in $R$ to differences in $H$ may be difficult: Individuals’ subjective reports may be relative to group-
level averages. I may rate my health less positively if I am part of a very healthy group. Group-level reference points thus influence the way I translate \( H \) into \( R \). If reporting scales are interpreted differently, reported mean differences in health may not be informative of latent differences across groups. Similarly, health inequality (captured as the variance) may appear larger in a given group only because the health levels associated with the response categories are more distant from each other in one group than another. To reveal how differences in reporting influence comparisons of variance we include heteroskedasticity in the vignette model.

2.1 Group comparisons with reporting heterogeneity

We allow individuals with characteristics \( X_i \) to have both different health outcomes and reporting styles\(^3\). We are interested in the validity of group comparisons. A group is defined in the most flexible possible way: Any set of individuals with the same vector of personal characteristics \( X_i \). Our main focus is the quality of comparisons of mean health, and health inequality, so that our objects of interest are \( \mu_{H_s}(X_i) \equiv E(H_{i,s} | X_i) \) and \( \sigma^2_{H_s}(X_i) \equiv V(H_{i,s} | X_i) \), where the dependence of \( V(.) \) on \( X_i \) indicates heteroskedasticity.

Respondent \( i \) evaluates her health status on a K-point scale. Let \( R_{s,i} \) denote \( i \)'s evaluation of health in domain \( s \) (pain, breathing problems, sleep disorders etc.). This evaluation depends on the genuine and latent health level \( H_{s,i} \). Statement \( k \) indicates that the severity of \( i \)'s condition falls in the interval \( [\tau_{i}^{k-1}; \tau_{i}^{k}] \) that defines \( i \)'s interpretation of this category

\[
R_{s,i} = k \iff \tau_{i}^{k-1} \leq H_{s,i} < \tau_{i}^{k}, \quad \forall k \in \{1, \ldots, K\}.
\]

Our vignette study follows the methodology outlined in King et al. (2004).

\(^2\)Note that other group members may also affect the latent variable. For subjective well-being, the importance of these relative effects is well-documented (Kahneman 2000, see also Clark et al. 2008).

\(^3\)In the empirical section, we consider the following individual characteristics: income, age, number of children, gender, education, job status and country.
Importantly, however, and contrary to previous vignette models, respondents’ characteristics are allowed to affect not only the expectation, but also the variance of health, which leads to the following model:

\[ H_{s,i} = X_i \beta + \sigma_i \varepsilon_{s,i} \]
\[ \sigma_i = \exp(X_i \kappa) \]
\[ \varepsilon_{s,i} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1), \quad \varepsilon_{s,i} \perp X_i. \]

Previous vignette studies implicitly assume \( \kappa = 0 \). In line with previous work we allow the reporting thresholds to also depend on the respondent’s characteristics, using the standard exponential form guaranteeing that the thresholds remain ordered:

\[
\begin{align*}
\tau_0^i &= -\infty, \quad \tau_K^i = +\infty \\
\tau_1^i &= \gamma_0^1 + X_i \gamma^1 \\
\tau_k^i &= \tau_{k-1}^i + \exp(\gamma_0^k + X_i \gamma^k), \quad \forall k \in \{2, \ldots, K - 1\}.
\end{align*}
\]

The condition for differences in reporting scale to bias comparisons across groups is fairly simple: Whenever the reporting scale varies systematically across groups with different latent health outcomes - different means for comparisons of the group means, and in terms of variance for the comparison of inequality levels across groups.

### 2.2 Identification using Vignettes

Individual characteristics \( X_i \) may influence both actual health outcomes and the reporting scale. The identification challenge consists in separating two channels: underlying differences in latent health across groups, via \( \beta \) and \( \kappa \), from differences across groups in the reporting scale, given by \( \gamma \). We achieve identification through anchoring vignettes. This technique combines subjective self-assessments of health with a subjective report of the health of a hypothetical person (the vignette) described in the questionnaire (King et al. 2004). The vignette model can eliminate response bias under two hypotheses: Response consistency asserts that respondents use the K-point response scale in the same way in their self-reports.
and for vignette evaluation. *Vignette equivalence* states that only reporting, but not perception of the vignette varies across respondents.

As an illustration, assume individuals from groups A and B rate the level of respiratory distress of a hypothetical person. On average, the scores are higher in A than B. Vignette equivalence implies that this difference only stems from reporting heterogeneity. Response consistency implies that the same bias affects self-reports: Members of A should overrate their own health the same way they overrate the vignette.

Since this is to the best of our knowledge the first paper to introduce heteroskedasticity into a vignette model, we focus on the specific assumptions related to this feature. The perception of the v-th vignette is labeled $H_{v,i}$, where as usual $R_{v,i} = k \iff \tau_{k-1}^{i} \leq H_{v,i} < \tau_k^i, \forall k \in \{1, \ldots, K\}$. Respondents share the same understanding of the vignette, up to a random perturbation

$$H_{v,i} = \theta_v + \sigma_v \varepsilon_{v,i}$$

$$\varepsilon_{v,i} \sim N(0, 1), \quad \varepsilon_{v,i} \perp X_i$$

Estimation of the whole model (1)-(4) requires normalizations of location and scale. The location normalization is standard for vignette models, but the scale normalization is related to our introduction of heteroskedasticity. The econometric literature on heteroskedastic ordered response models usually achieves both at the same time by excluding the constant from $X_i$ (Greene and Hensher (2010)). Note that if there is a constant in the equation of the conditional variance, we can write

$$\sigma_i = \exp(\kappa_0 + X_i \kappa) = \exp(\kappa_0) \exp(X_i \kappa),$$

and $\exp(\kappa_0)$ cannot be separately identified from the other coefficients in the reporting model. All the model parameters can then be estimated consistently by maximizing the conditional log-likelihood $l^c_i$ with respect to parameter vectors $\beta$, $\kappa$, $\gamma$, $\theta_v$ and $\sigma_v$. 
\[
I_c^i = \sum_{k=1}^{K} I(R_{i,s} = k) \times \ln \left[ \Phi \left( \frac{1}{\exp(X_i \kappa)} (\tau^k - X_i \beta) \right) - \Phi \left( \frac{1}{\exp(X_i \kappa)} (\tau^{k-1} - X_i \beta) \right) \right] \\
+ \sum_{v=1}^{V} \sum_{k=1}^{K} I(R_{i,v} = k) \times \ln \left[ \Phi \left( \frac{1}{\sigma_v} (\tau^k - \theta_v) \right) - \Phi \left( \frac{1}{\sigma_v} (\tau^{k-1} - \theta_v) \right) \right]
\]

with \(I(.)\) the indicator function.

### 3 Data and Specification

This paper uses the Survey of Health, Ageing and Retirement in Europe (SHARE) which asks a large number of respondents about their health outcomes and their assessment of health vignettes. Our sample consists of individuals born before 1954 and of their spouses. Evaluations of own and hypothetical health were collected during the first two waves of the survey which we pool for our analysis. Respondents rate the severity of their health problems on a scale from 1 (‘none’) to 5 (‘extreme’) for seven health problems (see table (1)). The dataset contains at least three vignettes for each health domain.

The model is estimated separately on each health domain. We have approximately 12,000 observations for each domain (see table (2) for details). All estimations include sociodemographic control variables: household income per capita (in logs), age, number of children, gender, education, labor market status and country. Due to the large number of estimated parameters, complete estimation results are relegated to an online appendix.

### 4 Results: Reliability of comparing group means & variances

We now report to what extent rankings of group means and variances are affected by reporting heterogeneity. To do this we compare ranking across groups with and without correcting for reporting heterogeneity. Rankings are a natural way
### Table 1: Self-evaluation questions

<table>
<thead>
<tr>
<th>Domain</th>
<th>Self-evaluation question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breathing</td>
<td>In the last 30 days, how much of a problem did you have because of shortness of breath?</td>
</tr>
<tr>
<td>Sleeping</td>
<td>In the last 30 days, how much difficulty did you have with sleeping such as falling asleep, waking up frequently during the night or waking up too early in the morning?</td>
</tr>
<tr>
<td>Pain</td>
<td>Overall in the last 30 days, how much of bodily aches or pains did you have?</td>
</tr>
<tr>
<td>Sadness</td>
<td>Overall in the last 30 days, how much of a problem did you have with feeling sad, low, or depressed?</td>
</tr>
<tr>
<td>Concentration</td>
<td>Overall in the last 30 days how much difficulty did you have with concentrating or remembering things?</td>
</tr>
<tr>
<td>Mobility</td>
<td>Overall in the last 30 days, how much of a problem did you have with moving around?</td>
</tr>
<tr>
<td>Work disability</td>
<td>Do you have any impairment or health problem that limits the kind or amount of work you can do?</td>
</tr>
</tbody>
</table>

### Table 2: Number of individual contributions to partial likelihoods

<table>
<thead>
<tr>
<th></th>
<th>Self-eval</th>
<th>Vignette 1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breathing</td>
<td>11932</td>
<td>11833</td>
<td>4407</td>
<td>4415</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sleeping</td>
<td>11954</td>
<td>4412</td>
<td>4409</td>
<td>11850</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain</td>
<td>11953</td>
<td>11846</td>
<td>4407</td>
<td>4414</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sadness</td>
<td>11927</td>
<td>4417</td>
<td>4404</td>
<td>11822</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concentration</td>
<td>11948</td>
<td>11853</td>
<td>4413</td>
<td>4416</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mobility</td>
<td>11943</td>
<td>4416</td>
<td>4413</td>
<td>11848</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Work disability</td>
<td>8902</td>
<td>4395</td>
<td>8820</td>
<td>4390</td>
<td>4377</td>
<td>4376</td>
<td>8820</td>
<td>8818</td>
<td>4396</td>
<td>4397</td>
</tr>
</tbody>
</table>
of testing the validity of interpersonal comparisons - there is no other obvious
way to compare ordinal data. Intuitively, we are testing how often one group is
actually healthier than another, yet reports lower health levels, viz. actually has
more equal health outcomes but reports a higher level of inequality. Formally, we
compare
\[ \rho_{\text{Spearman}} \left[ \mu_{R_s}(X_i), \mu_{H_s}(X_i) \right] \]
\[ \rho_{\text{Spearman}} \left[ \sigma^2_{R_s}(X_i), \sigma^2_{H_s}(X_i) \right] \]
where \( \rho_{\text{Spearman}}(.) \) denotes the Spearman rank correlation coefficient and \( \mu_{R_s} \) and \( \sigma^2_{R_s} \) are group means and variances of health declarations. Both the corrected and
the uncorrected means and variances are predicted after estimation of the complete
model. The expressions for the predicted latent group mean and variances are
simply:
\[ \hat{\mu}_{H_s}(X_i) = X_i \hat{\beta} \]
\[ \hat{\sigma}^2_{H_s}(X_i) = \exp \left( 2 \left( X_i \hat{\kappa} \right) \right) \]
The uncorrected values are
predicted parametrically as follows\(^4\):

(b)
\[ \hat{\mu}_{R_s}(X_i) = K - \sum_{k=1}^{K-1} \Phi \left( \frac{\hat{\tau}^k_i - X_i \hat{\beta}}{\exp(X_i \hat{\kappa})} \right) \]
\[ \hat{\sigma}^2_{R_s}(X_i) = \sum_{k=1}^{K} k^2 \times \ln \left[ \Phi \left( \frac{1}{\exp(X_i \hat{\kappa})} \left( \hat{\tau}^k_i - X_i \hat{\beta} \right) \right) - \Phi \left( \frac{1}{\exp(X_i \hat{\kappa})} \left( \hat{\tau}^{k-1}_i - X_i \hat{\beta} \right) \right) \right] \]
\[ - \left[ \sum_{k=1}^{K} k \times \ln \left[ \Phi \left( \frac{1}{\exp(X_i \hat{\kappa})} \left( \hat{\tau}^k_i - X_i \hat{\beta} \right) \right) - \Phi \left( \frac{1}{\exp(X_i \hat{\kappa})} \left( \hat{\tau}^{k-1}_i - X_i \hat{\beta} \right) \right) \right] \right]^2 \]

We find that comparing means of reported health is informative. Comparing
levels of inequality across populations, by contrast, is problematic. The left col-
umn of table (3) shows correlations between rankings of group averages based on
reported health values compared to rankings of the latent values (i.e. after cor-
rection via the vignettes). These correlations appear quite high - with an average

\(^4\)For the uncorrected analogues, we may be tempted to directly use reported values. However,
first, since \( X \) contains continuous variables, we would have very few observations in every group,
and estimating variances would be infeasible. Second, since we must predict the corrected
measures, a comparison between reported (uncorrected) and predicted (corrected) values would
conflate prediction errors with the effect of differential reporting we are interested in.
Table 3: Correlations between the conditional expectations / variances of latent and reported health

<table>
<thead>
<tr>
<th>Correlations of corrected &amp; uncorrected means</th>
<th>Correlations of corrected &amp; uncorrected variances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breathing</td>
<td>0.674</td>
</tr>
<tr>
<td>Sleeping</td>
<td>0.898</td>
</tr>
<tr>
<td>Pain</td>
<td>0.901</td>
</tr>
<tr>
<td>Sadness</td>
<td>0.903</td>
</tr>
<tr>
<td>Concentration</td>
<td>0.745</td>
</tr>
<tr>
<td>Mobility</td>
<td>0.963</td>
</tr>
<tr>
<td>Work disability</td>
<td>0.946</td>
</tr>
</tbody>
</table>

\( \rho_{\text{Spearman}} \) of 0.861, close to 0.95 in three cases (sleep, mobility, work disability), and below 0.8 only for breathing (0.674) and concentration (0.754). Rankings based on group averages of reported health are mostly a good proxy of the true rankings and can be considered as a valuable tool for research.

By including heteroskedasticity in the vignette model, we can now also consider the impact of reporting heterogeneity on comparisons of variances. The second column correlates within-group variances in reported health with variances in latent health (i.e. after the vignette correction). The correlations are much lower - on average the correlation coefficient is 0.328, and close to zero for mobility (concentration is the only domain where the correlation is above 0.5). Variances in reported health are a weak and unreliable proxy for variances in true health.\(^5\) Comparing variances of self-assessed health or well-being appears very challenging in light of these results.

\(^5\)What is the cause of this lack of reliability? Additional analysis – not reported here in the interest of brevity – reveals significant correlations between the variance of latent health and the spaces between reporting cut-offs, giving rise to the current pattern.
5 Conclusion

Reporting heterogeneity is a pervasive issue in subjective data. Previous methodological work has focused on the validity of comparisons of means. This article tests the effect of including differences in reporting of variances, in line with a recent literature comparing inequality across groups. To do this, we add heteroskedasticity to a standard anchoring vignette approach: Different individuals’ assessments of the health of common vignettes (health scenarios) are used to identify two types of reporting bias of subjective reports, covering the mean and the variance of the scale used to translate latent to reported health. The analysis is performed using answers to questions about seven health domains. Our results show that whilst reporting heterogeneity biases all comparisons across groups, the effect is small for comparisons of means but substantial for comparisons of health inequalities. This raises questions about the validity of comparisons of inequality based on subjective data which do not correct for individual differences in reporting.
References


