Evaluating the Cross-National Comparability of Survey Measures of Political Interest Using Anchoring Vignettes†

Seonghui Lee ‡ Nick C. N. Lin ‡ Randolph T. Stevenson ‡
seonghui.lee@rice.edu nicklin@rice.edu stevenso@rice.edu

‡ Rice University, 6100 Main St., Houston, TX 77005, USA

Abstract
Making cross-groups comparisons by using survey instruments has raised substantial scholarly concerns due to the potential risk of incomparability resulting from differential item functioning (DIF). However, not every survey item necessarily suffers from DIF. In this paper, we argue that, unlike many other survey items (e.g., political efficacy), the usual question used to measure political interest is likely to be largely immune to DIF. Our theoretical argument centers on the relative specificity of the item and a corresponding cultural homogeneity (at least in advanced democracies) in what it means to be politically interested or not. Utilizing the anchoring vignettes technique (King et al. 2004, King & Wand 2007) in our original surveys in the UK, France, and the Netherlands, we demonstrate the size of DIF is small for the standard political interest question.

Keywords: anchoring vignettes, political interest
1 Introduction

Political interest is one of the most important concepts in the study of political behavior. It predicts core aspects of democratic citizenship such as political knowledge and information acquisition (e.g., Delli Carpini & Keeter 1996, Luskin 1990, Prior 2005) and participation (e.g., Brady, Verba & Schlozman 1995, Delli Carpini & Keeter 1996, Powell 1986). Further, levels of interest in politics appear to vary dramatically across countries, even when comparisons are confined to countries with similar levels of economic and democratic development (e.g., OECD countries). For example, Figure 1 provides the average responses by respondents in 135 surveys (spanning 34 western countries over ten years) to the question: “How interested would you say you are in politics?”\footnote{There are 4 answer categories ranging from “Very” to “Not at all” interested. The surveys are from the European Social Surveys series.} The y-axis indicates the proportion of the respondents who reported that they were “Very” or “Quite” interested in politics and the graphs are ordered from countries with the least interested populace to those with the most. There is a great deal of variation in the average levels of interest across countries, ranging from 21% of Czech’s interested in politics to 68% of the Danes. Even if one limits the cases to the Western European democracies, the difference between the nations with the most (Denmark) and least interested populace (France at 44% interested) is still about 25 percentage points.

Besides the rather large differences in average political interest across countries, the graph also reveals another important regularity: the vast majority of the variance in political interest depicted in the figure occurs across countries, rather than within the same country over time. This strongly suggests that explanations for these differences are to be found in enduring differences across countries rather than in factors that vary from year to year or election to election. Indeed, given this pattern of variation, several scholars have proposed partial explanations for cross-national differences in political interest and knowledge, rooted in corresponding differences in electoral systems (Gordon & Segura 1997, Grönlund & Milner 2006), government policies (Milner 2002), and the availability of common political
heuristics (Lee & Stevenson 2014).

Of course, this pattern of variation (across rather than within countries) is also consistent with another explanation: perhaps individuals in different countries consistently interpret the political interest question in different ways. So, for example, if a typical Spanish respondent thinks a high level of political interest means attending weekly political meetings while a typical German respondent thinks a high level of political interest means reading about politics in the newspaper occasionally, then the large difference in average interest between the countries that is depicted in Figure 1 would not stem from substantive differences in the way these countries’ cultures and institutions shape individuals’ political interest, but instead from the use of a cross-nationally incomparable question to measure them.

Figure 1: Self-Reported Interest in Politics over Time and across Countries

Note: Y-axis: % Quite Interested + Very Interested; Red lines indicate country means over surveys.
This kind of “Differential Item Functioning” (DIF) across countries has been demonstrated for many other important political and social concepts measured in cross-national surveys (e.g., political efficacy, political trust, health assessment, job satisfaction, etc.). To date, however, none of the work exploring or even simply describing contextual differences in political interest has analyzed this possibility.

This paper is an attempt to remedy this problem. Specifically, we use King et al.’s (2004) “anchoring vignettes” methodology along with original surveys that we commissioned in three countries – France, the Netherlands, and the UK – to explore whether the kinds of differences depicted in Figure 1 are simply the result from the differences in the way survey respondents in different countries interpret typical political interest questions.

A study like this one is, we think, also particularly valuable because of the largely unappreciated fact that there is a wealth of cross-national survey data on political interest that goes back decades. Further, when one looks closely at the survey questions used in these surveys one finds a remarkable degree of consistency in the wording of the different questions used to tap political interest, as well as in the ordinal answer categories provided to respondents. Thus, we think that there is at least the possibility that these questions can provide a substantial and largely untapped, evidentiary basis for cross-national and over-time studies of political interest (and its closely related concepts like political knowledge and participation) – if it can be shown that survey respondents across contexts actually interpret these questions and their ordinal answer categories in the same way.

In the rest of this paper, we begin with a brief summary of the existing corpus of political interest questions and highlight their apparent similarity in different surveys. Next, we review the “anchoring vignette” methodology that has been proposed as a way both to diagnose and correct for cross-contextual differences in the way individuals answer self-assessment questions across contexts (even when these questions are asked in exactly the same way). We then introduce our original surveys (conducted in the UK, France, and the Netherlands) and the vignettes we constructed. Finally, we use these to apply the anchoring
vignette methodology to access the cross-national comparability of the standard political interest question (which has been used in well over a thousand existing surveys).

2 The Untapped “Treasure Trove” of Survey Data on Political Interest

Table 1 provides a summary of the cross-national surveys that have included questions to tap respondents’ levels of political interest. There are several points to take away from this table. First, there are a lot of surveys: 1051, spanning 125 countries, and 40 years (the first started in 1970). Further, this number is a minimum, since it does not include the hundreds of country specific surveys that have also asked about political interest.

The second thing to notice about these surveys is the remarkable consistency in the way that survey researchers in different regions and across time have asked about the concept. In all cases the question asks the respondent to rate him or herself without reference (or in comparison to) any external referent. Likewise, in none of the cases is there a long preamble or a filtering question. Each question includes one of only three interrogatives: How interested, how much interest, and to what extent interested; and includes one of only three verb phrases: are, would you say, and do you have. Finally, the object of this verb phrase takes only one of three forms: in politics, in politics and government, and in public affairs. We think this is a remarkable degree of consistency across surveys and likely results from both the simplicity of the concept of interest itself as well as the efforts of cross-national survey researchers to craft simple unidimensional questions that can be translated unambiguously in different countries.

This consistency of the questions probing interest is important for our study because it lets us talk sensibly about exploring the measurement properties of “the standard interest question” and makes our analysis relevant to researches contemplating using any of the

---

2 Of course, we realize that in translation, these exact constructions are changed to fit the grammar and usage of a given language. Our point is simply that the basic structure of the question is very similar across surveys.
Table 1: Summary of the Cross-National Surveys that Include Political Interest Questions

<table>
<thead>
<tr>
<th>Survey</th>
<th>Question Wording</th>
<th>Answering Categories</th>
<th>Countries</th>
<th>Time Frame (Num. waves that include the question)</th>
<th>Num. Surveys</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afrobarometer</td>
<td>How interested are you in politics and government? (w1) How interested are you in public affairs? (w2)</td>
<td>Very/ Somewhat/ Not</td>
<td>36 (Africa)</td>
<td>1999-2013 (5+ waves)</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>How interested would you say you are in public affairs? (w3-w5)</td>
<td>Very/ Somewhat/ Not very/ Not at all</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASES (Asia Europe Survey)</td>
<td>How interested would you say you are in politics?</td>
<td>Very/ Fairly/ Little/ Not at all</td>
<td>18 (Asia &amp; Europe)</td>
<td>2001</td>
<td>18</td>
</tr>
<tr>
<td>ABS (Asian Barometer Survey)</td>
<td>How interested would you say you are in politics?</td>
<td>Very/ Somewhat/ Not very/ Not at all</td>
<td>13 (Asia)</td>
<td>2001-2013 (3 waves)</td>
<td>34</td>
</tr>
<tr>
<td>CEEB 1 (Central &amp; Eastern Eurobarometer)</td>
<td>In general, how much interest do you have in politics?</td>
<td>A great deal/ A fair amount/ Not very much/ No interest at all</td>
<td>7 (Central &amp; Eastern Europe)</td>
<td>1990-1991</td>
<td>9</td>
</tr>
<tr>
<td>Eurobarometer</td>
<td>To what extent would you say you are interested in politics?</td>
<td>A great deal/ To some extent/ Not much/ Not at all</td>
<td>16 (Europe)</td>
<td>1970-2002 Trendfile (10 waves)</td>
<td>133</td>
</tr>
<tr>
<td>EES (European Election Survey)</td>
<td>To what extent would you say you are interested in politics?</td>
<td>Very/ Somewhat/ A little/ Not at all</td>
<td>27 (Europe)</td>
<td>1989-2009 (5 waves)</td>
<td>91</td>
</tr>
<tr>
<td>ESS (European Social Survey)</td>
<td>How interested would you say you are in politics?</td>
<td>Very/ Fairly/ Little/ Not at all</td>
<td>36 (Europe)</td>
<td>2002-2012 (6 waves)</td>
<td>161</td>
</tr>
<tr>
<td>ISSP 2004, 2007</td>
<td>How interested would you say you personally are in politics?</td>
<td>Very/ Fairly/ Little/ Not at all</td>
<td>72 (Worldwide)</td>
<td>2004, 2007</td>
<td>72</td>
</tr>
<tr>
<td>Latino-barometro</td>
<td>How interested are you in politics?</td>
<td>Very/ Fairly/ Little/ Not at all</td>
<td>19 (Central &amp; Latin America)</td>
<td>1995-2010 (12 waves)</td>
<td>202</td>
</tr>
<tr>
<td>WVS-EVS</td>
<td>How interested would you say you are in politics?</td>
<td>Very/ Somewhat/ Not very/ Not at all</td>
<td>101 (Worldwide)</td>
<td>1981-2008 (5 waves)</td>
<td>231</td>
</tr>
</tbody>
</table>

Total Number of Surveys: 1,051
Number of Countries: 125
surveys in Figure 1 (and other similar surveys not found there).

3 The Problem of Differential Item Functioning and Anchoring Vignettes

3.1 DIF in the Interpretation of Ordinal Answer Categories

Self-rating questions ask survey respondents to rate themselves on some underlying concept that is usually conceived of as a continuous trait (e.g., like political interest, general health, or attitudes toward abortion). However, lacking a natural metric for most such ratings, researchers are forced to provide respondents a set of ordinal categories into which the respondent is asked to place herself (i.e., very healthy, healthy, somewhat healthy, somewhat unhealthy, unhealthy, very unhealthy). When asked in different contexts or to different groups of people these items are subject to Differential Item Functioning (DIF). DIF happens when characteristics of the survey item and/or survey situation systematically cause individuals in different groups or contexts to understand the same survey questions differently (Brady 1985) or to have varied interpretations of the end-points and the cut-points of an associated measurement scale (Sadana et al. 2001).

Of course, experienced, careful survey researchers strive to avoid DIF by making questions concrete and specific, asking about only one concept that is at least perceived as unidimensional, and ensuring that the wording of both the question and the answer categories (and any translations of these) are equivalent across contexts and groups. In service to these goals, survey researchers use pilot studies, pre and post cognitive interviews, forward and reverse translations, exemplars, scale descriptions, question preambles, and many other techniques (e.g., Smith 2003, Harkness, Van de Vijver & Johnson 2003).

With respect to the measurement of political interest, it is certainly the case that the survey organizations and researchers responsible for developing the political interest items reviewed in the last section have, over many years, employed all these methods to refine
the political interest question and minimize DIF. The result: simplicity, clarity, and near
uniformity in the form of the “standard” political interest question and answer categories
across hundreds of surveys. Thus, if there were ever a candidate for a question that should
tavel well across contexts, political interest would seem to be it.

Despite this, however, one form of DIF cannot readily be identified and corrected by
simplifying and clarifying the language in a question; and it is this form that King and
his colleagues (2004) and others (Sadana 2000) show can be especially pernicious in cross-
cultural work. Specifically, cross-nationally relevant forms of DIF can happen even for well-
crafted questions when respondents in different countries systematically differ in how they
map the underlying continuous scale of the attitudinal variable to be self-rated to its ordinal
answer categories. In other words, this kind of answer-category DIF occurs when, for exam-
ple, “somewhat healthy” means something different to the typical respondent in Norway
than in Spain.3

If the answer categories for a given question are aligned on a unidimensional scale,
DIF of this kind results entirely from differences across individuals in the unknown cut-
points that individuals apply to determine what levels of the underlying and unobserved
continuous variable corresponds to which answer categories.4 Thus, in the case of political
interest, a potential problem for cross-national comparisons is that despite the rather direct,
simple, and easily translatable question (as well as answer categories that likely align along
a single dimension), the underlying level of political interest that causes someone in one
county to label herself “somewhat interested” in politics may be the same level that causes
a respondent in a different country to label herself “very interested.”

3 And, where this difference is not due to differences in the translation of the actual words – which can be
fixed with better translations – but results because the kinds of people who are considered to be “somewhat
healthy” differ across contexts.
4 We (and King et al. 2004) assume that researchers have endeavored to create a question and answer cate-
gories that correspond to a unidimensional underlying scale (or are perceived by respondents to be unidimen-
sional) and the method requires this to be so. Again, all the various techniques of careful survey construction
can be used to achieve this and we have little doubt that political interest is perceived by most respondents as a
unidimensional scale. That said, by collecting information on carefully constructed vignettes one does get some
evidence that could reveal multidimensionality in the scale – e.g., large proportions of respondents fail to agree
on the relative ordering of the vignettes. We will present some evidence about this below.
3.2 Anchoring Vignettes as a Solution to DIF due to the Interpretation of Ordinal Answer Categories

“Anchoring vignettes” is a technique, developed by Gary King and his colleagues (King et al. 2004, King & Wand 2007), that identifies and ameliorates DIF caused by differing interpretations of the “cut-points” defining answer categories. It does so by utilizing respondent assessments of one or more vignettes. A vignette is simply a description of the kind of person that (in the view of the researcher) exemplifies some level of the underlying concept to be measured. A key goal in developing a set of such vignettes is to have them ordered – that is, when respondents are asked to place them in answer categories, most will do so in the order researchers expect.

For instance, King and his colleagues (2004) used vignettes in a study of political efficacy in Mexico and China. Respondents were asked:

*How much say does ['name'] have in getting the government to address issues that interest [him/her]?

(1) No say at all
(2) Little say
(3) Some say
(4) A lot say
(5) Unlimited say

Three of the five vignettes given to respondents, intended to portray hypothetical individuals from high to low levels of political efficacy, were as follows:

**Alison** lacks clean drinking water. She and her neighbors are supporting an opposition candidate in the forthcoming elections that has promised to address the issue. It appears that so many people in her area feel the same way that the opposition candidate will defeat the incumbent representative.

**Jane** lacks clean drinking water because the government is pursuing an industrial development plan. In the campaign for an upcoming election, an opposition party has promised to address the issue, but she feels it would be futile to vote for the opposition since the government is certain to win.
Moses lacks clean drinking water. He would like to change this, but he can’t vote, and feels that no one in the government cares about this issue. So he suffers in silence, hoping something will be done in the future.

King et al.’s (2004) insight was that vignettes such as these can be used to “nail down” a respondent’s possibly idiosyncratic interpretation of the meaning of a set of ordinal answer categories and to readjust her self-reported category to match a common scale applicable across surveys.\(^5\) The simplest version of their method is illustrated in Figure 2 for the political efficacy question.

In this case, Respondent 2 rates her political efficacy (i.e., Self 2) quite low relative to Respondent 1 (i.e., Self 1), but this is only because she has a different underlying scale – a fact that is revealed by her placements of the hypothetical cases, all of which she rates quite low relative to their placements by Respondent 1. Since King and his colleagues assume that the only thing that differs in the ratings of the vignettes between Respondents 1 and 2 is DIF (while the self-ratings differ by DIF plus any real differences in the underlying levels of political efficacy between the respondents), the vignette scores provide a simple and imme-

\(^5\) This approach is based on two main assumptions: response consistency and vignette equivalence. The former assumes that each individual uses the response scale of a particular survey item in exactly the same way when being asked either about self-evaluation or about assessments on hypothetical vignettes. The latter assumes that all vignettes are perceived by respondents in the same way (King et al. 2004).
diate way to recode the self-ratings to make them comparable across cases (as illustrated on
the right side of Figure 2). Specifically, since the actual level of political efficacy for each vi-
gnette is the same across respondents, one can make the two respondents’ self-assessments
comparable by stretching Respondent 2’s scale so that the vignettes for both match. Stretch-
ing the scale in order to equate vignettes carries Respondent 2’s self-assessment along with
the stretched scale, and Respondent 2’s level of political efficacy ends up higher than Re-
spondent 1’s.

This approach can be defined more generally as the equation below shows. $y_i$ is the
categorical self-assessment for respondent $i$ ($i = 1, \ldots, n$) and $z_{ij}$ represents the response for
respondent $i$ on vignette $j$ ($j = 1, \ldots, J$). Vignettes are numbered in the order in which they
are expected to fall on the scale (as determined by the researcher). For those respondents
who have ranked all vignettes in their expected order ($z_{i, j-1} < z_{ij}$, for all $i, j$), the DIF-
corrected rating for individual $i$, which we call $C_i$, is:

$$C_i = \begin{cases} 
1 & \text{if } y_i < z_{i1} \\
2 & \text{if } y_i = z_{i1} \\
3 & \text{if } z_{i1} < y_i < z_{i1} \\
& \vdots \\
2J + 1 & \text{if } y_i < z_{ij} 
\end{cases} \quad (1)$$

Elaborations of the approach deal with the issue of respondents who give different
vignettes the same score or rate vignettes out of the expected order. Further cognate meth-
ods have been developed to evaluate the quality and usefulness of different sets of vignettes
(King et al. 2004, King & Wand 2007). We will discuss (and use) these methods below; but
these elaborations do not change the essential intuition captured in Figure 2 that persists
through all its elaborations.

Most importantly, what this example should illustrate are the implications of DIF for
cross-national use of self-ratings on ordinal scales: if Respondent 1 in the example is the
typical respondent from one country and Respondent 2 is the typical respondent from another, the raw self-placements will result in a mistaken inference about the relative level of political efficacy in the two cases, while the corrected self-ratings reverse this mistake. In the sections below, we will develop a set of vignettes that we think is appropriate for exploring DIF in political interest measures across countries. We will use the method illustrated above (and elaborations of it) to determine if there is cross-nationally significant answer-category DIF in the typical political interest question (and so we might account for the cross national difference depicted in Figure 3).6

3.3 Constructing Vignettes

The method described above does not assume that all respondents give the same unique ranks to vignettes; however, it does assume that respondents place all the vignettes on the same, unidimensional scale that they use for their own self placement. How can we know if this assumption holds and how can we construct vignettes to make it more likely to hold?

Of course, it is not reasonable to think that one can write vignettes about complex, inherently multidimensional concepts like political efficacy, general health, or political interest that are unidimensional in the sense that they are not thought of in terms of various different aspects or sub-dimensions. Indeed, vignettes almost always illustrate a target concept (like happiness) by invoking a selection of these sub-dimensions. For example, one might illustrate an unhappy person by saying that he has a loveless marriage and a failing career, thus priming these two sub-dimensions of happiness (satisfying interpersonal relationships and professional success) in favor of other aspects that are not mentioned (e.g.,

---

6 This approach has been applied in a variety of fields, such as in health studies (e.g., Rice, Robone & Smith 2008, Angelini et al. 2013), economics (e.g., Kristensen & Johansson 2008, Beegle, Himlelin & Ravallion 2009), and political science (e.g., King et al. 2004, Ward et al. 2006, Bakker et al. 2012). In political science, the method has been used to study measures of political efficacy, inter-ethnic trust (Ward et. al. 2006), and perceived public sector performance (Rice et al. 2008). In each of these cases, vignettes revealed significant DIF that suggested unadjusted measures are suspect. In contrast, for several other variables of widespread interest in cross-national research, scholars have used the anchoring vignette methodology to show that the measures do not suffer from consequential DIF. For example, Bakker et al. (2012) show that DIF does not bias expert placements of parties on left-right scale in a consequential way. The only exception in their study is that experts in Scandinavian countries perceive left-right scale slightly different from experts in other countries.
spiritual fulfillment). What is required for respondents to perceive the scale to be unidimensional, however, is only that the values of the invoked sub-dimensions move together (or monotonically) and in the same direction, over the vignettes. In this way the various sub-concepts primed in the vignettes aggregate naturally into an overall assessment of the target concept that follows the same pattern across vignettes as its sub-dimensions. Practically, this means not defining intermediate vignettes as mixes of high and low values on different sub-dimensions (e.g., defining an intermediate happiness case as a person with a good marriage and a failing career) but rather as intermediate values on each sub-dimension (e.g., a person with an “ok” marriage who is doing alright, but not great, at work). The latter case promotes the perceptions of unidimensionality (the marriage and career dimensions move together over the examples), while the former provokes the respondent to think about the sub-dimensions as moving separately and perhaps in opposing ways (thus promoting perceptions of multidimensionality).

From this example it is clear that vignettes should also be written to prime the most important and relevant sub-dimensions individuals are likely to think about when rating themselves. This helps to get the respondent thinking about the concept in terms of the sub-dimensions in the examples and minimizes the chances that a person applies a different set of sub-dimensions in evaluating themselves (thus violating the assumption that they rate themselves on the same scale as the vignettes).

In addition, Hopkins & King (2010) have shown that vignettes should be written in a way that encourages the respondent to visualize the person as similar to themselves – for example, it is helpful to manipulate gender cues to match that of the respondent. This again promotes the respondent to rate themselves on the same scale (comprised of the same sub-concepts) as they rated the individuals in the vignettes. Finally, King & Wand (2007) have emphasized that the overall set of vignettes will be most helpful when they discriminate

---

7 If this is not possible because the researcher wants respondents to think about a concept in a way that is, while helpful for the research question, not natural to them (e.g., vignettes about happiness that only talk about whether one enjoys one’s food), then the target concept should be redefined (e.g., to “enjoyment of food”).
individuals along the whole range of the scale – that is, they are not all portrayals of someone with very high or low values on the underlying scale.

Usefully, the data on respondents’ rating of vignettes can help us understand whether respondents really do perceive the scale as unidimensional and whether the set of vignettes discriminates well. Below, we introduce the vignettes we wrote for political interest and explain how we used monotonic movement in the most important sub-dimensions of political interest to define the individuals in our vignettes.

4 Application of Vignette Approach to Political Interest

4.1 Country Selection and Sampling

In order to use anchoring vignettes to examine the extent of DIF in the typical political interest question, we commissioned surveys in the UK, France, and the Netherlands. Each sample was a random sample of the internet using population of each country, with appropriate weights for weighting to census values for standard demographics. We chose these cases because they span the range of levels of political interest that we saw for the western European democracies in Figure 1. We can see this clearly in Figure 3 below, which gives the average political interest scores across time in each western European country (derived from the data in Figure 1). This Figure show clearly that, over the last decade, levels of political interest have been highest in the Netherlands (65%), followed by the UK (53%), and

---

8 The Dutch Survey was conducted by Survey Sampling International (SSI) using their internet panel. The French and British Surveys were conducted by the Center for Experimental Social Sciences, Nuffield College, Oxford University.

9 There appeared to be about 10% of the French respondents who did not take the survey as a whole seriously and simple rushed through the answers to all questions. This is apparent in their answers to the vignette questions, in which they gave the same response for all four vignettes. Likewise, on many other questions, these respondents provided meaningless answers and (when possible) don’t know or non-response options. Of the 1073 respondents in the French survey, we determined that for 181 there was substantial evidence that the respondent did not read most of the questions in the survey and with respect to the vignette questions, in which we did not allow the respondent to skip the question or say don’t know, simply chose the same answer for each vignette. We eliminate these individuals from the sample in the analyses below (thus leaving us 892 French responses). While there were far fewer such cases in the UK and Dutch surveys, in order to treat the different surveys equivalently, we also eliminated such respondents from these other surveys as well (28 in the Netherlands and 48 in the UK).
Finally France (44%). Further, these countries cover virtually the full range of variation in the western European countries. We are confident, then, that an analysis of DIF in these three countries will be probative for, at least, the larger set of Western European countries.

Given the relative rankings of average levels of interest in our sample countries that have been found in previous survey work (reviewed in Figure 3), we expected a typical respondent in our new surveys of France, the UK, and the Netherlands, to follow the same pattern. Figure 4 below shows that indeed they did, with mean interest scores (on a 10 point scale) of 5.9 for France, 6.2 for the UK, and 7.3 for the Netherlands. More generally, the figure gives the whole distribution of the respondents’ interest scores. In what follows, we will be especially interested in whether accounting for potential DIF in our political interest question will substantially change these distributions (and their means).

---

10 See Figure 1 for data sources and variable definitions.
4.2 Constructing Vignettes for Political Interest

Following Hopkins & King (2010), we ask respondents to evaluate the political interest of four hypothetical individuals before asking them to assess their own interest. For each vignette, respondents are asked to assess these hypothetical individuals’ levels of political interest on a 10 point scale, with 1 indicating a person that is not at all interested in politics and 10 representing a person who is extremely interested in politics.

As discussed above, for an inherently multi-dimensional concept like political interest, we need to invoke the same, important sub-dimensions of interest in each vignette and make sure the levels of each of these aspects move monotonically and in the same direction from the low interest vignette to the high interest vignette. In addition, in order to make sure the set of vignettes discriminates individuals along the whole range of the scale, we need to try to describe individuals with various plausible levels of these variables. We attempted to achieve these goals with the following vignettes:11

11 We confine our efforts at exploring DIF in political interest to the modern democracies and so these vignettes reflect that choice. Specifically, they refer to such things as the nightly news on television that might not be a
How interested in politics would you say each of the people described below is? When you finish answering for each of these people, please place yourself on the same scale.

[Individual A] reads about the government and politics in the newspaper every morning on [his/her] way to work. When at home, [he/she] often reads political blogs on the internet and watches television programs about politics. [He/She] likes to talk to [his/her] friends and coworkers about politics and is eager to share [his/her] views on politics with other people.

[Individual B] watches the nightly news on television and usually pays attention when the news is about the government and politics. From time to time [he/she] will read a political blog and occasionally talks about politics with [his/her] friends at work, especially during election campaigns or when the government is dealing with an important issue.

[Individual C] rarely pays attention to politics and government on [his/her] own, but will read an article about politics or visit a political blog if a close friend recommends it to [him/her]. Likewise, [he/she] will participate in conversations about politics when [his/her] friends initiate them, but almost never starts such conversations [himself/herself].

[Individual D] is turned off by politics. [He/She] only pays attention to news updates about the government and politics when [he/she] is waiting for [his/her] favorite program to come on and [he/she] avoids discussing politics with [his/her] friends and family.

In our view, these anchoring vignettes can be arranged on an ordered scale, ranging from the most interested in politics (Individual A), to some interest (Individual B), then to little interest (Individual C), and finally to no interest at all (Individual D). Of course, we will be able to test whether our respondents ranked them as we expected and if these rankings differed across countries.

Notice that each taps several different dimensions of interest (e.g., reading newspapers or political blogs on the Internet or watching TV news, and talking about politics with friends) but varies only the levels of these activities over vignettes and does not mix and relevant experience in some less developed countries. A broader exploration of DIF in questions of political interest would need to construct vignettes more appropriate for these kinds of contexts or, ideally, ones that applied across very different developmental and cultural contexts.
match them, so that intermediate categories do not have, for example, high levels of reading about politics but talking about politics with friends. This kind of monotonicity in the construction of the vignettes is essential if we want to encourage the respondent to think of the scale as unidimensional. Also notice that we repeat the sub-concepts in each vignette so they encourage the respondent to think about the same aspects of interest in each vignette and so perceive the scale as the same for each vignette. Also, by asking the self-placement last, we have primed reading about news and discussing with friends as the relevant sub-criteria for the respondent to think about when they ask themselves whether they are politically interested. This all promotes the perception of political interest as a unidimensional scale and one that is the same for each vignette and the self-rating.

4.3 Do the Vignettes Fall on a Unidimensional Scale?

If respondents are placing the vignettes on the same scale, which they also perceive to be unidimensional scale, and if we have written the vignettes to clearly capture relative positions on that scale, then the primary evidence of that should be found in the percentage of respondents who actually rate the vignettes in the order we expect.

Table 2 gives the top 10 orderings of the vignettes by all the respondents in all three surveys. Almost 50% of respondents placed the four vignettes in exactly the order we expected and 68% percent of the respondents had no order violations – they place all four vignettes in the order we expect but had some ties – and 85% had only one order violation. In addition, there was no particular pair of vignettes that respondents seemed to have a lot more trouble ordering than any other, so mis-orderings seem fairly randomly distributed across the possibilities. Table 3 gives the proportion of cases in which a given vignette (row) is ranked less than another (column) with the vignettes listed in our expected order. Adjacent pairs were ordered correctly between 73 and 80% of the time and non-adjacent pairs (as we would expect) much higher than that.

\[12\] If we use three vignettes the total percentage of responses that do not violate order rises to 74%.
Table 2: Frequency of Vignette Ordering, Top 10

<table>
<thead>
<tr>
<th>Ordering</th>
<th>Frequency</th>
<th>Proportion</th>
<th>Cumulative Proportion</th>
<th>Order Violation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A &gt; B &gt; C &gt; D</td>
<td>1385</td>
<td>.49</td>
<td>.49</td>
<td>0</td>
</tr>
<tr>
<td>A &gt; B &gt; C = D</td>
<td>188</td>
<td>.07</td>
<td>.56</td>
<td>0</td>
</tr>
<tr>
<td>A &gt; C &gt; B &gt; D</td>
<td>151</td>
<td>.05</td>
<td>.61</td>
<td>1</td>
</tr>
<tr>
<td>A &gt; B = C &gt; D</td>
<td>145</td>
<td>.05</td>
<td>.66</td>
<td>0</td>
</tr>
<tr>
<td>A &gt; B &gt; D &gt; C</td>
<td>122</td>
<td>.04</td>
<td>.70</td>
<td>1</td>
</tr>
<tr>
<td>A = B &gt; C &gt; D</td>
<td>85</td>
<td>.03</td>
<td>.73</td>
<td>0</td>
</tr>
<tr>
<td>B &gt; A &gt; C &gt; D</td>
<td>73</td>
<td>.03</td>
<td>.76</td>
<td>1</td>
</tr>
<tr>
<td>A = B &gt; C = D</td>
<td>51</td>
<td>.02</td>
<td>.78</td>
<td>0</td>
</tr>
<tr>
<td>A &gt; B = D &gt; C</td>
<td>32</td>
<td>.01</td>
<td>.79</td>
<td>1</td>
</tr>
<tr>
<td>A &gt; C &gt; B = D</td>
<td>28</td>
<td>.01</td>
<td>.80</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: A > B > C > D is the expected ordering; only top 10 orderings are reported. Total proportion of cases with no order violation is 68% (some of the orders contributing to this total are not included in the top 10 list above).

Table 3: Proportion of Cases a Vignette is Less than Another

<table>
<thead>
<tr>
<th>Vignette A</th>
<th>Vignette B</th>
<th>Vignette C</th>
<th>Vignette D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vignette A</td>
<td>–</td>
<td>.145</td>
<td>.084</td>
</tr>
<tr>
<td>Vignette B</td>
<td>.727</td>
<td>–</td>
<td>.135</td>
</tr>
<tr>
<td>Vignette C</td>
<td>.854</td>
<td>.770</td>
<td>–</td>
</tr>
<tr>
<td>Vignette D</td>
<td>.887</td>
<td>.884</td>
<td>.801</td>
</tr>
</tbody>
</table>

4.3.1 Do individuals in different countries differ systematically in where they place the vignettes on our political interest scale?

If DIF across countries is an important problem for a measure and if vignettes have been constructed to reveal this (i.e., respondents place them on the same unidimensional scale, like ours did, as shown in the previous section), then we should see evidence of this in differences in the distribution of vignette placements across countries. For example, in Figure 5, we give the distribution over respondents of vignette placements for the five vignettes used in King et al.’s (2004) political efficacy study, in which DIF had very large effects, switching the ordering of Mexico and China from what it was in the DIF corrected measures.

Figure 5 reveals a great deal of disagreement between Mexican and Chinese respon-
dents in the numerical score assigned to each hypothetical individual described in King et al.’s vignettes. Indeed, there is actually more between country variation in the ratings of the identical vignettes than there is within country variation across different vignettes, which were, of course, designed to differ.

We can see this even more clearly in Figure 6, where we compare the full empirical distributions of the Mexican and Chinese respondents for each vignette. Clearly, these different groups of respondents disagreed dramatically about the placements of the individuals portrayed in the vignettes.

These examples – taken from a study in which DIF was pronounced enough to cause a mistaken ordering of countries using the unadjusted political efficacy scale – provide a good benchmark for evaluating the existence of cross-nationally consequential DIF in our raw political interest scale (and by implication the raw scales used in many other studies).
Figures 7 and 8 give the graphics that correspond to Figures 5 and 6 from the King et al.’s study – the differences could not be starker.

While the mean responses for each vignette are statistically different from one another across countries, substantively they are not very far apart and certainly they are orders of magnitude closer together than in the King et al.’s study. As above, one way to see this is to ask whether the distances between countries for a particular vignette are shorter than the differences between means across vignettes – or to put it another way: is the total variance depicted in Figure 7 composed of within vignette variance or between vignette variance?

We can answer this formally by simply estimating a random intercept model on the 12 values in Figure 7, with a random intercept for vignettes A to D. This quantifies the qualitative impression one gets from a glance at the figure: 93% of the variance in Figure 7 is explained by variation between different vignettes and only seven percent by between
countries for the same vignette. This suggests strongly that though the differences in means are statistically significant (given the very small standard errors one can get on an estimate of a mean in sample sizes of about 1000 per country), these differences are not substantively large.

As we did above, we can also compare histograms for ratings of each vignette across countries. These are depicted in Figure 8, where we plot the empirical distribution (summarized with a kernel density) for each vignette for each country. Clearly, unlike Mexican and Chinese respondents in the political efficacy study, respondents in our three countries largely agree about where to place each vignette on the 10 point political interest scale they were given. This is strong evidence against the idea that DIF is systematically driving the kind of dramatic differences in political interest across countries that we saw in the introduction.
4.3.2 Are (all) the vignettes informative and do they discriminate?

In the last section, we saw that our respondents in different countries largely agree on the placement of our four vignettes. The next question is whether those vignettes discriminate well among different respondents. As discussed above, a set of vignettes that discriminates well, will result (after recoding) in about equal numbers of respondents in each recoded category. King et al. (2004) propose a straightforward measure of this criterion (which they call entropy) that is quite similar to familiar measures like the “effective number of parties” that measures the dispersion of votes over a set of parties. In the specific measure of entropy King et al. (2004) propose, a larger number indicates that respondents are spread more evenly over (recoded) response categories, and so has better discrimination, while a
smaller number indicates less discrimination. In general, we expect entropy to increase with more vignettes, but that need not always be the case.

Based on each possible set of vignettes we used in our data, we compute two measures of entropy (i.e., minimum entropy and estimated entropy) and plot them in Figure 9. Each vignette (A to D) corresponds to its expected order and sets of vignettes are indicated by grouped letters. For instance, “A” indicates the entropy score if one were only to use vignette A to recode the raw responses, while “AD” indicates the entropy one would get if one used vignettes A and D. Among the single vignette sets, for example, vignette B does the best job discriminating, while “BC” is the best two vignette set and “ABC” is the best three vignette set. It is encouraging that in our data the best discriminating set (by a significant margin) is the set using all four vignettes. This need not be the case and indeed the literature contains a number of examples in which few vignettes discriminate better than more. The fact that this does not happen here suggests that each of our vignettes is important to cutting up the space of the scale.

The minimum entropy score uses data recoded using the “minimum entropy” method of handling ties (see King & Wand (2007) for a description of this method), while the estimated entropy uses the ordered probit method. The fact that these differ little here, suggests the variables in the ordered probit are not having much impact on discrimination.
5 How Much Do Cross-National Differences in the Distribution of Self-Ratings of Political Interest Change When We Correct Them for DIF?

The final test of the cross-national comparability of our measure of political interest simply asks whether our comparisons between countries change when we recode them, correcting for any DIF (as illustrated in Figure 2). The idea is to produce histograms of corrected and uncorrected responses for each country and compare these. If there are few differences, or if the ordering of the countries on the scale does not change, we can conclude that the uncorrected measure does not suffer from cross-nationally consequential DIF. Thus, in this section we go into somewhat more detail about how one can use the information from our vignettes to produce corrected self-ratings and then construct a set of histograms of corrected political interest scores to compare to the original, uncorrected scores.
For respondents who have ordered the vignettes correctly, with no ties, correcting the raw scores for DIF is simple: it simply requires us to directly apply the transformation formula for $C_i$ that was given above.\footnote{Since we have four vignettes, this produces nine possible categories (i.e., “1” if the respondent’s self-rating is less than vignette D, “2” if it equals D, “3” if it is between D and C, etc.).} We do this for the 50% of cases in our data to which this condition applies. The question remains, of course, what to do with cases in which vignettes are rated by the respondent as the same category (tied) or mis-ordered? In Figure 10 below, we present corrected scores by using three different approaches. We also add the uncorrected score in the first panel in Figure 10 as a comparison to these corrected scores.

To deal with the tied and mis-ordered categories, the easiest answer is to simply ignore such cases. The second panel in Figure 10 gives the corrected histograms for this approach.\footnote{These histograms include all cases in the histogram of uncorrected self-ratings and only those with no ties or mis-orderings in the histogram of corrected self-ratings.} While this approach is not advisable,\footnote{Due to the potential bias stemming from the fact that respondents who did not place vignettes in their expected order (and so were dropped) might not be a random sample of all respondents.} in our case it leads to no real differences in the pattern of responses across our three countries. However, if we do not want to omit half the sample from the comparison, then we need to allocate the tied and mis-ordered cases to categories, correcting as best we can for each respondent’s DIF. As King et al. (2004) point out, this starts by understanding that when an individual assigns two vignettes to the same category, this means we sometimes do not know exactly how to do the assignment in equation 1. For example, if a respondent’s self-assessment fell between the value of her first ranked vignette and the value of her second and third vignette (which she had assigned the same category), we would not know for sure whether to assign her corrected self-assessment to the second or third category. In their original work, King et al. (2004) used a uniform allocation in situations like this and simply assigned the corrected rank as 2 or 3 with equal probability. In later work King & Wand (2007) improved on this by using a censored ordered probit model to estimate the probability of allocation to one of the possible ranks.\footnote{Specifically, using all the respondents who gave a unique rank to the relevant vignettes, they estimate a censored ordered probit model of choice among the possible ranks. This can be done with a cut point only model (so essentially the chance of being allocated to a possible rank is proportional to how often it was selected in the sample) or made conditional on respondent characteristics – so the probability among ranks is allocated according to these characteristics.} In these two
approaches, incorrectly ordered cases are handled by treating them as ties. The results are depicted in the third and fourth panels in Figure 10.

Figure 10: Non-Corrected and Corrected Self-Assessments

In making these comparisons, it is important to keep in mind that the category numbers (and their meanings) are different in the corrected and uncorrected assessments. In the former case, the 10 categories on the x-axis are the labelled answer categories that were given to respondents. In the latter case, the nine categories on the x-axis result from the possible ways that self-assessments can be recoded, given four vignettes (e.g., if there had been only two vignettes, then there would have been only five categories in the corrected self-assessments). Consequently, these have a substantive meaning that is different from similarly labelled categories on the original scale. For example, in the recoded metric, category 3 is interpreted as “the level of political interest above vignette D and below vignette C.” Given this, the relevant comparisons across graphs are only in the broad shapes of the histograms between the corrected and uncorrected self-assessments and not the specific categories. Importantly, however, one can compare the relative rankings of the countries on as other people like this respondent allocated their choices.
Table 4: Relative Ranking for Each Country

<table>
<thead>
<tr>
<th>relative ranking of countries</th>
<th>Netherlands</th>
<th>UK</th>
<th>France</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean raw self-assessments (SA)</td>
<td>Highest</td>
<td>Middle</td>
<td>Lowest</td>
</tr>
<tr>
<td>Mean corrected SA (omit tied cases)</td>
<td>Highest</td>
<td>Middle</td>
<td>Lowest</td>
</tr>
<tr>
<td>Mean corrected SA (uniform allocation)</td>
<td>Highest</td>
<td>Middle</td>
<td>Lowest</td>
</tr>
<tr>
<td>Mean corrected SA (probit allocation)</td>
<td>Highest</td>
<td>Middle</td>
<td>Lowest</td>
</tr>
<tr>
<td>Mean % interested in politics (ESS)</td>
<td>Highest</td>
<td>Middle</td>
<td>Lowest</td>
</tr>
</tbody>
</table>

various meaningful measures (e.g., the means, medians, and variances of the empirical distributions) across the corrected and uncorrected measures – looking, of course, for any differences in these rankings resulting from the DIF corrections.

In Figure 10, we provide the means of each country across graphs to help readers better see these comparisons. Comparing across columns, there is clearly not a great deal of difference between the pattern in the DIF corrected and the uncorrected histograms across countries. Table 4 makes this clear by giving the relative rankings for each country and mean self assessments for each of the histograms in Figure 10. The last row in Table 4 provides the relative ranking of our sample countries we observed in Figure 3.

To sum up, our result provides the first evidence that cross-national comparisons based on standard political interest questions may indeed reflect underlying differences in the distribution of political interest across populations rather than systematic differences in the way questions are interpreted. And most importantly, the cross-countries difference (i.e., rankings) we found in our data do not differ from what we have seen in other cross-national surveys.

6 Conclusion

In this paper, we have explored whether the substantial differences in the distribution of political interest that are observable across countries are likely to result from systematic
differences in how the concept is measured rather than from more substantive sources (like differences in political institutions). Given the simplicity of the typical political interest question, as well as the extraordinary similarity in its wording and structure across surveys, it is unlikely the differences in wording and format that do exist could account for much of the observed cross-national variation in levels of interest. Thus, we have focused our investigation on the possibility of DIF in respondents’ interpretation of the typical question’s ordinal response categories. Similar investigations of other important concepts in political science have sometimes (but not always) found this kind of DIF to be an important source of systematic differences across countries.

Our investigation took advantage of original survey work designed to apply King’s anchoring vignette methodology to the problem. Our results suggest that respondents have little trouble correctly ordering vignettes describing different levels of political interest and that the vignettes we wrote discriminated well among different levels of political interest. Most importantly, we found that conclusions about the relative distribution of interest across countries remained the same after recoding to correct for DIF (since there was not a great deal of DIF apparent).

This strongly suggests that political interest does not suffer from significant, cross-nationally relevant DIF and implies that the wealth of survey data on the concept (at least for the set of countries similar in political and economic development to the three we examine here) should be used more often to explore the substantive sources of the large cross-national differences in the distribution of political interest that we demonstrated at the beginning of this paper.
References


