

---

# WHAT CAN WE LEARN FROM (AND ABOUT) GLOBAL AGING?\*

ARIE KAPTEYN

*Although aging is a global phenomenon, there are large differences across countries in both the speed of aging and the current age composition. Furthermore, countries adopt vastly different policies. This creates a natural laboratory that scientists can use to understand how policies affect outcomes. I discuss under what circumstances data from different countries can be used for inference about policy effects. Although comparable health and retirement data are currently being collected in some 25 countries, the use of such data requires careful modeling of differences in institutions and in response styles across countries.*

**A**lthough aging is happening worldwide, there are vast differences across countries in both the speed of aging and the current age composition. In 2000, the percentage elderly in major parts of Latin America and almost all of Africa and Asia was below 7.5%. In contrast, Western Europe was already above 15%, with Italy on top at 18.4% (United Nations 2009), while in Japan, the percentage was 17.2. The percentage of citizens 65 years old or older in the United States in 2000 was 12.4, while in Africa it was 3.3. In 2050, Japan is expected to have 37.8% of its population aged 65 years and older (United Nations 2009, medium variant), while in Italy, the percentage 65 and older is then expected to be 33.3. In contrast, the percentage 65 and older in the United States in 2050 is expected to equal 21.6, while in Africa, it is projected to be 7.1.

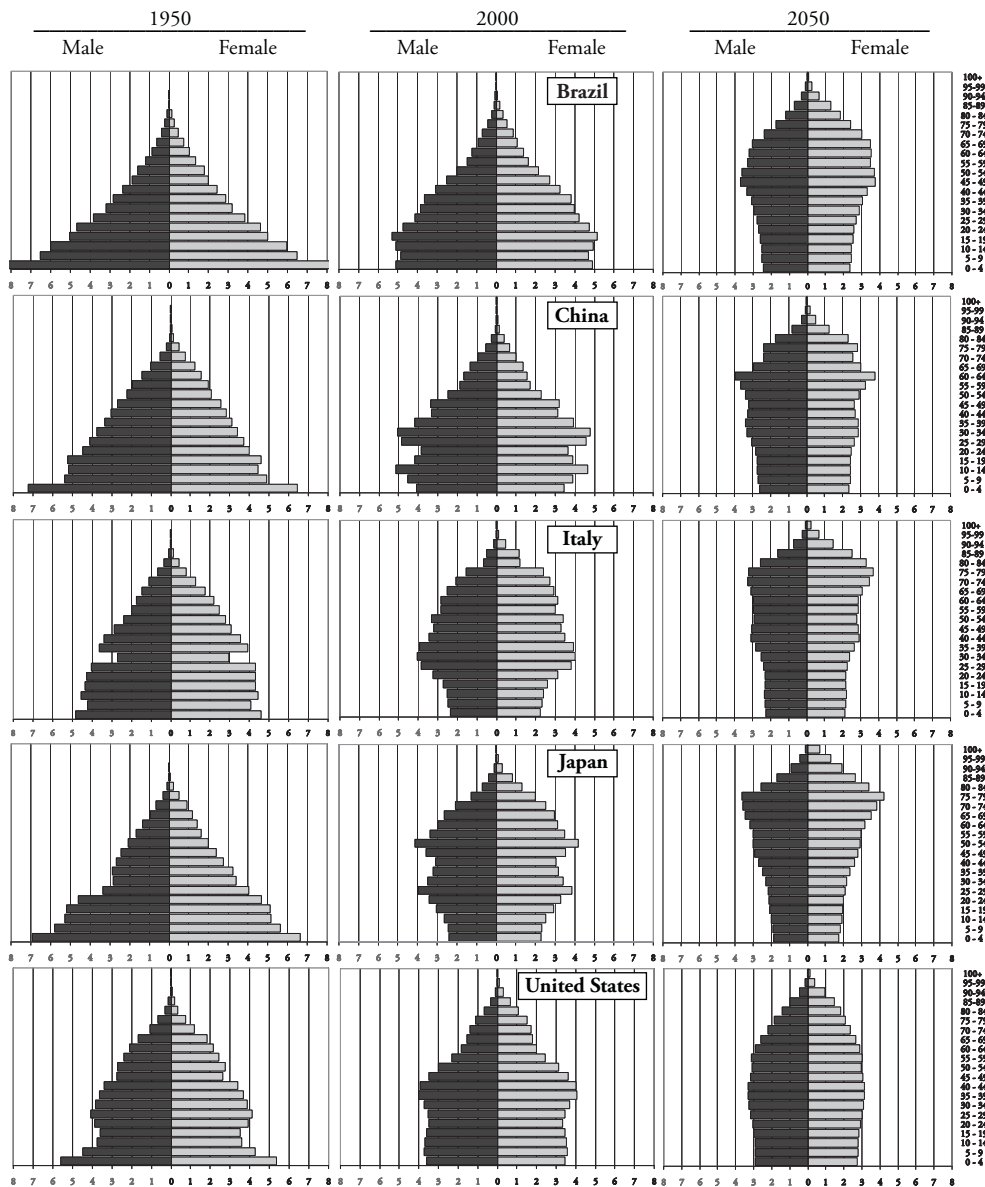
It is well known that falling birth rates are a major force behind the aging of societies. This is illustrated in Figure 1 for five countries from four different continents: Brazil, China, Italy, Japan, and the United States. All five countries show dramatic demographic transitions over the 100-year period 1950–2050, although they are currently at somewhat different stages. For Japan and Italy, the “pyramid” already has a base that is narrower than the middle, while for Brazil, China, and the United States, that pattern is less visible in 2000. In 2050, all pyramids are to some extent inverted, least so in the United States, where the different generations are expected to be of about equal size.

The change in population composition has dramatic consequences for the dependency ratio (defined as the number of individuals younger than 15 or older than 64 divided by the number of individuals aged 15–64) in various countries. In the short run, low fertility rates lead to a reduction in the dependency ratio as the working population swells relative to the population younger than 15. In the longer run, the percentage 65 or older increases and thus increases the dependency ratio. Figure 2 illustrates this phenomenon. Perhaps the most remarkable aspect of Figure 2 is the fact that we are currently living in an era with a historically low dependency ratio. In the past, the working population had to take care of the young; in the future, the workers will have to take care of the elderly. This is further underlined by Figure 3, which shows dependency ratios for both the young and the old. Notice the sharply declining dependency ratios for the young and sharply increasing dependency ratios for the old. The lines typically cross somewhere between 2000 and 2020.

---

\*Arie Kapteyn, Labor and Population, RAND Corporation, 1776 Main Street, Santa Monica, CA 90407; e-mail: kapteyn@rand.org. This research was supported by Grant 5P01AG008291 from the National Institute on Aging. I thank Lisa Klautzer for her expert research assistance.

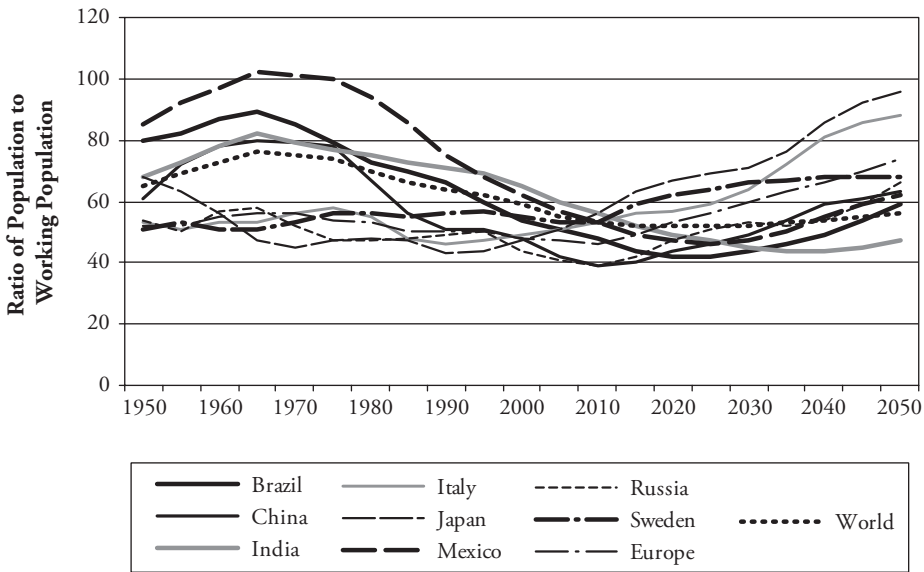
Figure 1. Population Pyramids by Gender and Five-Year Age Bands (in % of population)



Source: United Nations (2009), medium variant.

Thus, we are in a unique era with a low dependency ratio and a future in which the economic burden of aging is likely to increase dramatically. This fact has not gone unnoticed, and policy makers around the world are grappling with the consequences of an aging world. Since different countries are at different stages of the aging process, there is considerable scope for learning from the experiences of others when designing one's own policies.

Figure 2. Dependency Ratios Between 1950 and 2050



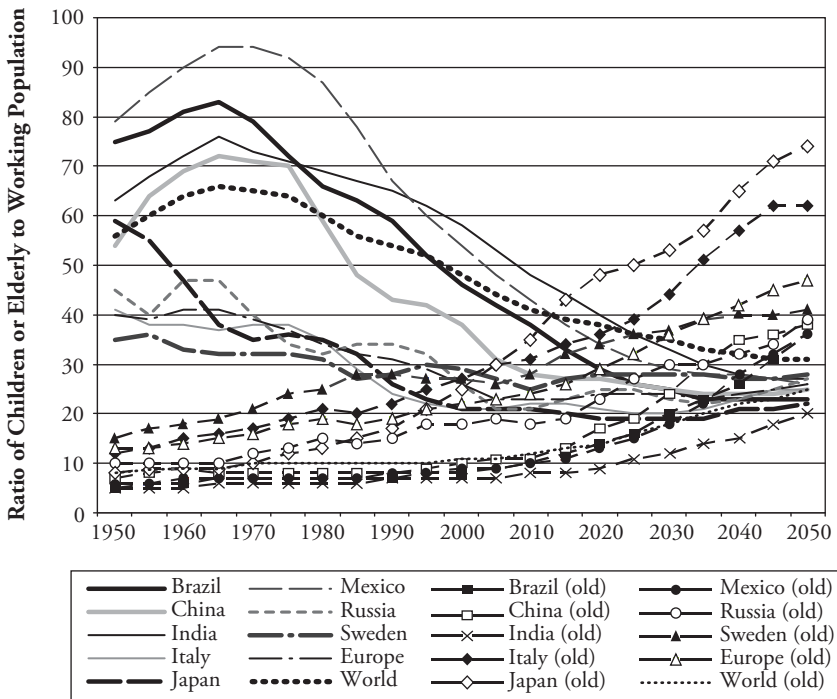
Source: United Nations (2009), medium variant.

Fortunately, due to the farsightedness of some agencies and institutions (the U.S. National Institute on Aging in particular), researchers and policy makers can increasingly rely on the availability of similar or harmonized microdata from different countries to study the relation between policies and outcomes in countries at different stages of the aging process.

Although data may be similar, there are various issues with the use of such data to learn about the effect of policies. In this article, I will concentrate on two main aspects. First, I discuss the assumptions necessary to draw conclusions about the effect of policies on outcomes. In this part, I will assume that data on outcomes are indeed comparable and that these data are available in longitudinal form. The second part of the article addresses inherent measurement issues that arise when using responses from different cultures or different countries. I will discuss *anchoring vignettes* and how they can help make measurements comparable. Anchoring vignettes are not a panacea, and they rest on two crucial identifying assumptions. I will summarize the evidence that has been collected until now to shed light on the validity of these assumptions.

### COMPARING OUTCOMES AND POLICIES

There are tremendous differences in both policies and outcomes across countries. Figure 4 shows the average male and female retirement age, life expectancies at age 65, and legal retirement ages in a number of countries. There appears to be very little relation between the legal retirement ages (however defined) and actual retirement ages. Even more striking is the contrast between life expectancy at age 65 and average retirement age. Men in European countries enjoy a high life expectancy, while at the same time retiring early. France is a case in point where males at age 65 can expect to live until about age 82.7, while the average retirement age is about 59. In Italy, life expectancy at age 65 is about the same, but males retire somewhat later, about age 61. In both countries, the effective

**Figure 3.** Dependency Ratios of Old and Young Between 1950 and 2050

Source: United Nations (2009), medium variant.

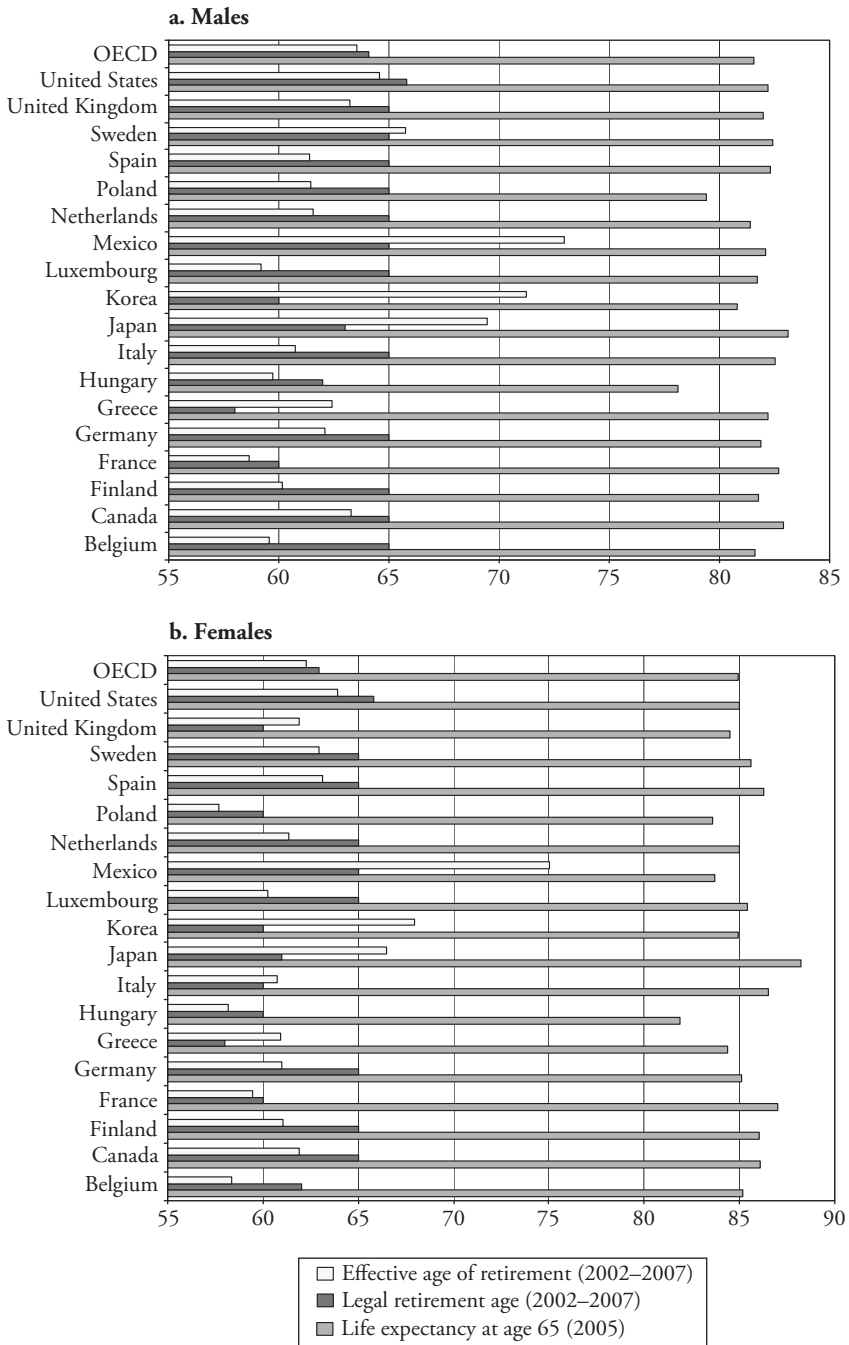
retirement age is below the legal retirement age of 60 and 65, respectively. At the other extreme, in Mexico, the effective retirement age is 73, well above the legal retirement age of 65, and less than 10 years away from the expected age of death. The picture for women is even more pronounced. At age 65, they can expect to live longer than men, while they retire earlier. Here, Belgium is a striking case: women retire on average at age 58, while at age 65 they can expect to live until 85.

Naturally, these vast differences in the length of life during which one enjoys retirement have marked consequences for the share of gross domestic product (GDP) devoted to supporting retirees. Figure 5 illustrates this by presenting the percentage of GDP spent on old-age and survivors benefits. Italy, France, and Austria spend the highest percentage of GDP on retirees. It is worth noting that these numbers do not include private pensions, so that the total resources available for retirees may be different from what is reflected in Figure 5.

Of course, not only does the level of expenditures on retirees matter, but also how fast these expenditures grow. Most OECD (Organisation for Economic Cooperation and Development) countries show an increase in expenditures. Some of the highest growth rates are in countries that start from a relatively low base, like Mexico, South Korea, and Turkey.

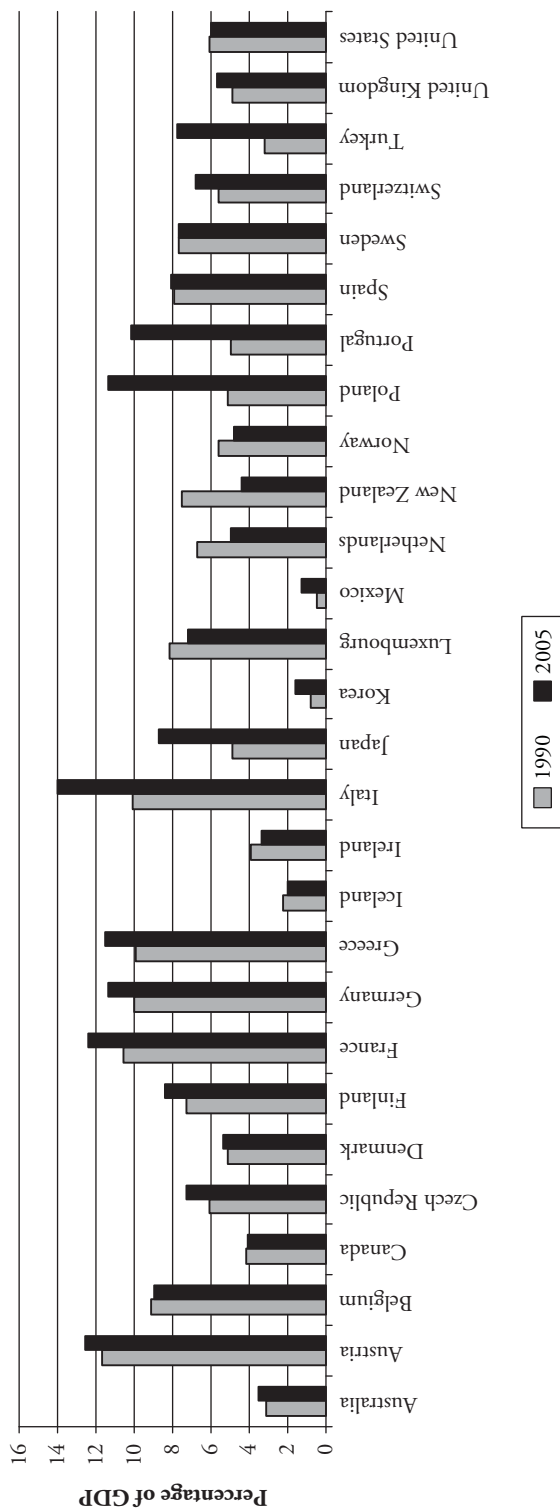
Pensions and social security are not the only spending categories associated with an aging population. Health care costs and long-term care are expected to grow fast as well. Roughly speaking, pensions make up half of the increase in total expenditures, while health and long-term care make up the other half.

**Figure 4. Effective Age of Retirement Versus Legal Retirement Age and Life Expectancy at Age 65 (in years): Males and Females**



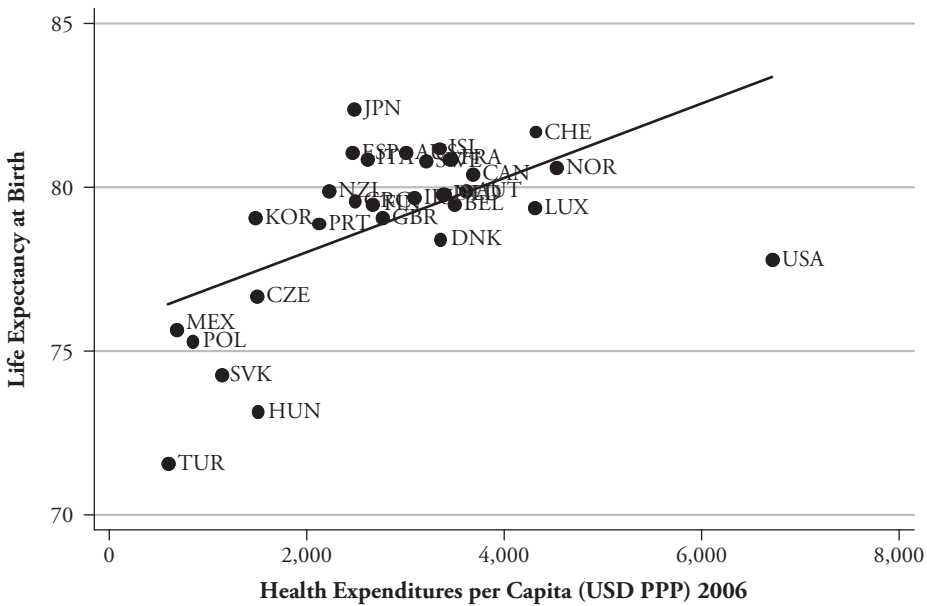
Source: OECD (2009b).

Figure 5. Old-Age and Survivor's Benefits as a Percentage of GDP in 1990 and 2005



Source: OECD (2009a).

Figure 6. Health Expenditures and Life Expectancy at Birth (2006)



Source: OECD (2009b).

It is well known that health expenditures vary tremendously across countries. Figure 6 illustrates this for a selected number of countries. The United States is a lone outlier when it comes to health care expenditures per capita. For instance, the Euro area spends about half per capita than the United States.

It is tempting to ask how much health is bought by the expenditure on health care. Figure 6 suggests that the answer is “not very much,” at least if we consider life expectancy at birth. Of course, life expectancy is determined by many factors, of which the health care system is only one. Nevertheless, the figure suggests that if increasing life expectancy is a major policy goal, one may want to explore many policies other than spending on health care.

### Learning From Differences

The previous examples show tremendous variation in outcomes, and a natural question is to what extent these differences can be ascribed to differences in policies across countries. Potentially, the variation in outcomes and in policies provides powerful opportunities to learn about the effects (intended or unintended) of policies and hence should be of considerable interest.

When using microdata from different countries with the objective of learning about the effects of different policies on outcomes for individuals or households,<sup>1</sup> one should consider several statistical and econometric issues. I will discuss some of these by using a simple illustrative econometric model and then pointing out the various assumptions needed for valid inference.

1. For simplicity, my terminology will refer to individuals, but the framework is also applicable to other units of observation, such as businesses or some other entity.

Consider the following simple panel data model to analyze the effects of policy differences across countries:

$$y_{ict} = \mu_{ic} + v_{ct} + \gamma' \mathbf{x}_{ict} + \delta_{ic} P_{ct} + \varepsilon_{ict}, \quad (1)$$

where  $y_{ict}$  is an outcome variable of interest for household (or individual, or business)  $i$  in country  $c$  at time  $t$ ;  $\mu_{ic}$  is an individual effect for household  $i$  in country  $c$ ;  $v_{ct}$  is a time effect in country  $c$ , representing aggregate effects unique to that country and that time period, and often denoted by economists as *macroshocks*;  $\mathbf{x}_{ict}$  is a vector of covariates influencing the outcomes for household  $i$  in country  $c$  at time  $t$ ;  $P_{ct}$  is a policy variable in country  $c$  and may be a vector but, for simplicity, will be treated as a scalar; and  $\varepsilon_{ict}$  is an independently and identically distributed (iid) error term.

How do cross-country data help us learn about policy? First, consider analysis within a country. A typical panel data analysis would consider

$$\Delta y_{ict} \equiv y_{ict} - y_{ic,t-1} = \Delta v_{ct} + \gamma' \Delta \mathbf{x}_{ict} + \delta_{ic} \Delta P_{ct} + \Delta \varepsilon_{ict}, \quad (2)$$

where  $\Delta$  denotes first differences. Although taking first differences has the merit of eliminating the fixed effects, we can see immediately that the effect of a policy change, represented by the parameter  $\delta_{ic}$ , cannot be identified, since the terms  $\Delta v_{ct}$  and  $\delta_{ic} \Delta P_{ct}$  cannot be disentangled: one cannot determine if observed changes in the outcome variable  $y$  are due to a policy change  $\Delta P_{ct}$  or to other changes at the aggregate level represented by  $\Delta v_{ct}$ .

Next, consider the possibility of comparing across two countries  $c$  and  $c'$ . Then, one might consider

$$y_{ict} - y_{jc't} = \mu_{ic} - \mu_{jc'} + v_{ct} - v_{c't} + \gamma' [\mathbf{x}_{ict} - \mathbf{x}_{jc't}] + \delta_{ic} P_{ct} - \delta_{jc'} P_{c't} + \varepsilon_{ict} - \varepsilon_{jc't}. \quad (3)$$

Thus, one compares differences across countries at a given time.

There are two problems with this cross-section comparison. First of all, the way the equation is written, the effects of policies may work out differently in different countries, and comparisons across countries by themselves are not informative. The difference  $\delta_{ic} P_{ct} - \delta_{jc'} P_{c't}$  does not, of course, identify the separate parameters  $\delta_{ic}$  and  $\delta_{jc'}$ . In other words, comparing across countries does not tell us anything about how household  $i$  in country  $c$  and household  $j$  in country  $c'$  are affected by the policies of their governments. This changes if one is willing to restrict the variation of the parameters  $\delta_{ic}$ . The simplest possibility is to assume that  $\delta_{ic}$  is the same for everyone:  $\delta_{ic} = \delta$ . In other words, we assume that all households in both countries are affected in the same way by a given policy.

This simplifies Eq. (3) to

$$y_{ict} - y_{jc't} = \mu_{ic} - \mu_{jc'} + v_{ct} - v_{c't} + \gamma' [\mathbf{x}_{ict} - \mathbf{x}_{jc't}] + \delta (P_{ct} - P_{c't}) + \varepsilon_{ict} - \varepsilon_{jc't}. \quad (4)$$

If it weren't for the individual effects  $\mu_{ic}$  and  $\mu_{jc'}$ , we could now identify  $\delta$  if we had observations on a number of countries at a given time  $t$ . We could decide to follow the path of least resistance: assume that the individual effects are uncorrelated with everything else on the right side and thus absorb them into the error terms. One can then estimate an equation like (4) by ordinary least squares (OLS), preferably taking into account cluster effects (see, e.g., Moulton 1990).

The assumption  $\delta_{ic} = \delta$  can be relaxed by allowing the effect of a policy to depend on observable individual characteristics. One can write, for instance,  $\delta_{ic} = z_{ic}' \phi$  and obtain

$$y_{ict} - y_{jc't} = \mu_{ic} - \mu_{jc'} + v_{ct} - v_{c't} + \gamma' [\mathbf{x}_{ict} - \mathbf{x}_{jc't}] + [P_{ct} z_{ic}' - P_{c't} z_{jc'}'] \phi + \varepsilon_{ict} - \varepsilon_{jc't}. \quad (5)$$

Without further assumptions on the individual effects  $\mu_{ic}$ , this equation still cannot be estimated. Once again, a possibility is to assume that the individual effects are uncorrelated with the other variables on the right side of the equation and relegate them to the error term. Whether such a procedure can be justified will depend on the context in which a model like Eq. (5) is applied. If, for instance, the outcome variables  $y_{ict}$  and  $y_{jc't}$  would represent labor



force participation and  $\mathbf{x}_{ict}$  and  $\mathbf{x}_{jct}$  would contain the wage rate as one of their variables, one could easily imagine that the individual effects correlate with  $\mathbf{x}_{ict}$  and  $\mathbf{x}_{jct}$ . This is a classical case of selection bias: individuals who are more motivated to work have higher wages and are more likely to keep working, also at higher ages.

With a panel of individual observations for a number of different countries, one can exploit both the temporal and cross-country variation. Consider, for instance,

$$\Delta y_{ict} - \Delta y_{jct} = \Delta v_{ct} - \Delta v_{c't} + \gamma'(\Delta \mathbf{x}_{ict} - \Delta \mathbf{x}_{jct}) + [\Delta P_{ct} z'_{ic} - \Delta P_{c't} z'_{jct}] \varphi + \Delta \varepsilon_{ict} - \Delta \varepsilon_{jct}. \quad (6)$$

One still needs to make assumptions about the changes in time effects  $\Delta v_{ct}$  to be able to identify the parameter of interest  $\varphi$ . One option is to parameterize the aggregate time effects as functions of macro variables, like GDP per capita or the unemployment rate (see, e.g., Heckman and Robb 1985; or Kapteyn, Alessie, and Lusardi 2005).

### Endogenous Policies

So far, the assumption has been that policies are exogenous and that we can use policy variation across countries and across time to identify their effects on outcomes of interest. It is often argued, however, that policies are endogenous. For instance, in a country with a strong taste for leisure, policies that encourage short working weeks or early retirement are more likely to be implemented than in countries with a strong taste for spending on consumption goods. Such an objection need not be fatal for identification of policy effects. A simple way to reflect endogenous policies is to allow for the possibility that the individual effects  $\mu_{ic}$  partly reflect such taste differences. This would imply that the policies  $P_{ct}$  and individual effects  $\mu_{ic}$  are correlated. A fixed-effects estimator takes such correlations into account, and the parameter of interest  $\varphi$  is still identified. Clearly, this does require panel data, so if one suspects that policies are endogenous, then a cross-country type regression would most likely lead to biased estimates of policy effects.

### Comparable Microdata on Aging Populations

The point of the discussion so far is that the richer the data available, the weaker the assumptions one needs to make while still being able to learn about the effect of policies on individual outcomes. In the field of aging, the various health and retirement studies that are being conducted in many countries, often with strong support from the National Institute on Aging, provide a unique source of panel data that allow us to study a plethora of issues by comparisons across countries. The U.S. Health and Retirement Study (HRS), started in 1992, interviews some 20,000 individuals over age 50 every two years (see e.g., Juster and Suzman 1995). Discussions are continuing to add other countries, while considerable effort is expended on harmonization of these data and the design of user-friendly interfaces to facilitate easy access of the data (see, e.g., [https://mmicdata.rand.org/wiki/index.php/Main\\_Page](https://mmicdata.rand.org/wiki/index.php/Main_Page)).

The initial HRS wave took place in 1992, the first AHEAD wave in 1993–1994. To be “age eligible,” HRS respondents had to be born between 1931 and 1941 (aged 51–61 in 1992), and AHEAD respondents in 1923 or earlier (aged 70 or older in 1993). Initial sample sizes were 12,652 for HRS and 8,222 for AHEAD, but not all respondents were age-eligible because spouses were also interviewed. Subsequent waves have shown moderate loss to follow-up other than due to mortality. Newly married spouses of original respondents were added as new respondents. The HRS/AHEAD oversampled Mexican Americans, African Americans, and residents of Florida.

There have been nine HRS waves (1992 through 2008) and two AHEAD waves (1993 and 1995) to date, after which AHEAD was merged with HRS. Periodically, new cohorts are added so that the HRS remains representative of the U.S. population aged 50 and older.

HRS questionnaires include demographics, family structure, health conditions, cognition, health care utilization and costs, health status, housing, job status, expectations,

income, and net worth. In addition to these “core” questions asked of the entire sample, supplementary topical modules are administered to randomly assigned subsamples. Recently, HRS has started to collect biomarkers.

Since the HRS was first conducted, several similar studies have emerged in other countries.

**ELSA** (English Longitudinal Study of Aging) began following a cohort of people aged 50 and older and their partners in 2002. The ELSA sample was drawn from respondents to a prior survey, the Health Survey for England (HSE 1998, 1999, and 2001), so that fairly rich data are already available at baseline. More than 12,000 people participated in the first round, and ELSA covers the same topics as the HRS, but its health section includes additional questions on symptoms, and its cognitional and psychosocial sections are expanded, including the nature of social networks, support, and participation. The ELSA survey instrument was constructed to be as comparable as possible to the HRS. Over the course of the proposed grant period, we expect releases of the 2008, 2010, and 2012 data waves.

**SHARE** (Survey of Health, Ageing and Retirement in Europe) is a multidisciplinary and cross-national panel interview survey on health, socioeconomic status, and social and family networks of individuals aged 50 or older in Europe. Eleven countries contributed data to the 2004 SHARE baseline survey. Participating countries are a balanced representation of the various regions in Europe: Austria, Belgium, Denmark, France, Germany, Greece, Italy, the Netherlands, Spain, Sweden, and Switzerland. For these countries, second and third waves of data collection took place in 2006 and 2009. Since its start, five additional countries have joined SHARE: in 2006, the Czech Republic, Poland, and Ireland had their first wave of data collection. Data were collected in 2005–2006 for Israel and in 2008–2009 for Slovenia. In total, SHARE includes data for more than 30,000 adults aged 50 or older in 16 countries.

**KLoSA** (Korean Longitudinal Study of Aging) is a nationally representative panel survey on aging in Korea. KLoSA is a biennial survey of over 10,000 Korean elders aged 45 and older who reside in a community. All age-eligible adults are interviewed. Sample weights are applied to ensure population representativity. The age of 45 was chosen to permit the study of early retirement due to mandatory retirement policies (and practices, such as up-or-out promotion rules) common in Korea. The baseline data were collected through face-to-face interviews in 2006, and the second wave was conducted in 2008. In addition, employment history data were collected in 2007, as an off-year study.

**MHAS** (Mexican Health and Aging Study) is a biennial panel study, modeled after the HRS, of individuals born before 1951 and their spouses in Mexico. The first wave of data was collected in 2001, and the second was collected in 2003. The data set is nationally representative and contains information on demographic and employment characteristics, firm size, health status, indicators of current and childhood health, health behaviors, parental education and longevity, health care services provisions, family transfers, migration history to the United States, income, business assets, and wealth for 9,862 households. There are plans to reinterview the respondents to the original two waves.

**CHARLS** (Chinese Health, Aging, and Retirement Longitudinal Study) is a biennial survey in 10 Chinese provinces that aims to be representative of the residents of China aged 45 and older. The age of 45 was chosen because the official retirement age for female workers is 50 (for some jobs), and many qualify for early retirement. Moreover, the health transition typically starts at an earlier age in developing countries. Similar to all other country versions of the HRS, both spouses are interviewed in the survey. A pilot survey of 1,500 respondents was conducted in 2008 in two provinces: Zhejiang, located in the developed coastal region, and Gansu, located in the less developed western region. Gansu is one of the poorest, most rural provinces in China, while Zhejiang is one of the most dynamic provinces in terms of its fast economic growth, private sector, small-scale industrialization,

and export orientation. Follow-up interviews are planned in 2011 and 2013, with a target sample size of 10,000 across ten provinces.

**LASI** (Longitudinal Aging Study India) is a multidisciplinary, internationally harmonized panel data set that aims to represent the elderly population in India. A full-scale biennial survey of 30,000 people representing Indians aged 45 and older is planned, and with this goal, a pilot study is currently being conducted with the support of the National Institute on Aging. For this pilot study, the initial sample of about 2,000 persons is drawn from four states (Karnataka, Kerala, Punjab, and Rajasthan) using the *2001 Census Primary Census Abstract*, and the data will be collected during October through November 2010 by in-person, computer-assisted personal interviews (CAPI). Full-scale surveys are planned for 2011 and 2013.

### COMPARABILITY OF OUTCOME MEASURES

Until now, we have assumed that the outcome measures are comparable across countries. There are several ways in which comparability may be compromised. For instance, the health and retirement studies tend to measure the prevalence of health conditions by asking a question of the following form: “Has a doctor ever told you that. . . .” Apart from the fact that this question is somewhat ambiguous (e.g., how to answer the question if one has been misdiagnosed with some condition in the past?), clearly a positive answer can only logically be given if one has seen a doctor. Where access to care, frequency of doctor visits, and diagnostic tools vary across countries, a straight comparison of the answers to such questions across countries may be highly misleading. Indeed, undiagnosed disease is a major problem in developing countries. But even within or across developed countries, there are reasons to be wary of such self-reports. For instance, Banks et al. (2006), in their study of health differences among the late-middle-aged population in England and the United States, used both self-reports and biological data to document the health of Americans and English. Johnston, Propper, and Shields (2009) considered self-reported and nurse-diagnosed hypertension in the Health Survey of England. They found that for self-reported hypertension, there is no evidence of a socioeconomic status (SES) health gradient, but for nurse-diagnosed hypertension, there is a strong relationship between SES and health.

In other cases, national definitions may be different. For instance, *unemployment* is defined in the United States as referring to people who do not have a job, have actively looked for work in the past four weeks, and are currently available for work. Also, people who were temporarily laid off and are waiting to be called back to that job are counted as unemployed. In the Netherlands, one has to be registered with the labor exchange in order to be counted as unemployed, and one must be looking for a job of at least 12 hours per week. Definitions are also affected by temporary policy measures. For instance, in some countries, unemployed individuals above a certain age may not have to look for a job in order to be eligible for unemployment benefits. Unemployment then effectively becomes retirement.

Similar issues arise with program participation. For instance, the percentage of the working-age population drawing disability benefits varies tremendously across countries (see, e.g., Kapteyn, Smith, Van Soest, and Banks 2010). To the extent that this variation is the result of different policies (e.g., eligibility rules or financial incentives), it does not pose a problem in principle. After all, the differences in policies and their effects on outcomes are precisely why we collect data in different countries. The situation becomes more complicated if the way citizens experience work disability is different across countries. In that case, the difference in disability rolls across countries may not just point at different policies and their effects but also at different definitions of outcomes, and estimating a model like (6) above would potentially be highly misleading.

One reason why outcomes might be defined differently in different countries is adaptation to policy. For instance, in a country with generous work disability policies, many

workers may be induced to draw disability insurance benefits. This may shift social norms so that opinions on what constitutes a work disability shift. Workers become “soft.” A straightforward way to model this is the following: Let  $y_{ict}^*$  be the true measure, while  $y_{ict}$  is observed. The observation and the true measure are related by an observation function:  $y_{ict} = h(y_{ict}^* | P_{ct}, \mathbf{x}_{ict})$ . Thus the true model is

$$y_{ict}^* = \mu_{ic} + v_{ct} + \gamma' \mathbf{x}_{ict} + \delta_{ic} P_{ct} + \varepsilon_{ict}, \quad (7)$$

If instead we estimate an equation like (1) (e.g., using cross-country panel data), then in principle, estimates will be biased. Consider, for instance, the following observation function:  $y_{ict} = h(y_{ict}^* | P_{ct}, \mathbf{x}_{ict}) \equiv y_{ict}^* + \lambda_c P_{ct} + \eta' \mathbf{x}_{ict}$ . We would then be estimating an equation of the form

$$y_{ict} = \mu_{ic} + v_{ct} + (\gamma' + \eta') \mathbf{x}_{ict} + (\delta_{ic} + \lambda_c) P_{ct} + \varepsilon_{ict}. \quad (8)$$

Rather than estimating parameters  $\delta_{ic}$ , we would be estimating  $\delta_{ic} + \lambda_c$ , that is, the direct policy effect plus an effect of the policy on the disability norms in a country.

Clearly, one would want to disentangle the two effects. One recent promising approach is the use of anchoring effects (King et al. 2004).

### ANCHORING VIGNETTES

Anchoring vignettes have been designed to overcome the measurement problems mentioned above. The exposition below in large outline follows earlier work by Kapteyn, Smith, and Van Soest (2010).

For concreteness, I will consider work disability, although later on I will discuss other domains. Consider the following question: *Do you have any impairment or health problem that limits the kind or amount of work you can do?* Response options are (1) none, (2) mild, (3) moderate, (4) severe, and (5) extreme/cannot do. Quite possibly, respondents with a different background or from a different country may interpret this question differently and hence give different answers even though in some objective sense they have the same health problem. This phenomenon is known as *differential item functioning* (DIF). How do we find out if this is the case? Conceptually, one can imagine conducting extensive physical examinations in different countries and comparing answers by respondents who seem to have the same health. In practice, the costs and logistics of such an enterprise are likely to be prohibitive, at least if one wants to compare a large number of countries and interview a large number of respondents.

King et al. (2004) proposed using *anchoring vignettes*. In the context of work disability, anchoring vignettes would be introduced to a respondent as follows:<sup>2</sup>

We would now like to give you a number of examples of persons with some health problems. We would like you to indicate the extent to which you think these people would be limited in the kind or amount of work they can do. In terms of their age, their education, and their work histories, you should imagine that these men or women are similar to yourself. Other than the conditions explicitly mentioned, you should imagine the individual is in reasonably good health.

Next, a respondent gets questions like the following, with response options again ranging from (1) mild to (5) extreme/cannot do:

*[Eva] feels worried all the time. She gets depressed once a week at work for a couple of days in a row, thinking about what could go wrong and that her boss will disapprove of her*

2. The introduction given here and the vignette questions cited are from SHARE 2004.

*condition. But she is able to come out of this mood if she concentrates on something else. How much is [Eva] limited in the kind or amount of work she can do?*

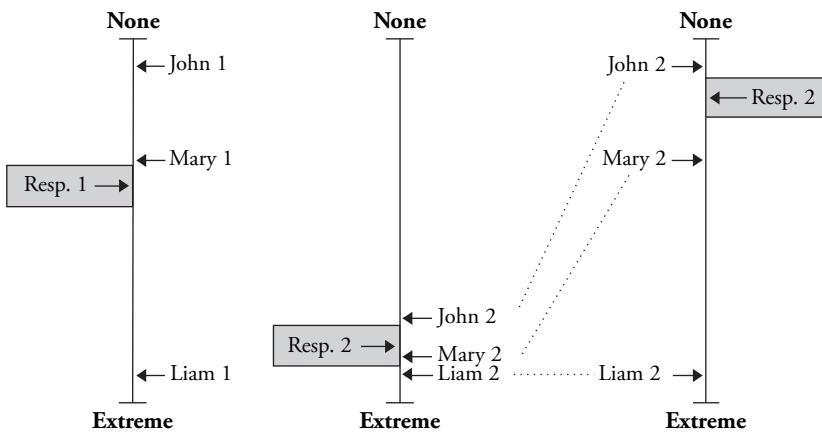
*[Tom] has been diagnosed with high blood pressure. His blood pressure goes up quickly if he feels under stress. Tom does not exercise much and is overweight. How much is [Tom] limited in the kind or amount of work he can do?*

The genders of the names associated with the vignettes are randomized. Notice that the response scale for the vignettes is the same as for the self-reports. To illustrate how vignettes help to detect differences in the use of response scales, consider the hypothetical example in Figure 7. First consider Respondent 1; this respondent places himself (let’s assume the respondent is male) in the middle of the scale between *none* and *extreme/cannot do*. The picture also shows where he locates the vignettes John, Mary, and Liam. He feels that he has more work limitations than John and Mary, but less than Liam.

Now consider Respondent 2. He places himself close to being completely work disabled. However, we also observe that he places Mary and Liam even closer to being completely work disabled. Comparing Respondent 2’s ratings of the vignettes with those of Respondent 1 suggests that Respondent 2 is much more likely to rate a health condition as work-limiting than Respondent 1. We might say that Respondent 2 is “softer.” Thus, we may be confronted with a case of DIF. Under the assumption of *response consistency* (if a respondent has the same work limitation as a vignette person, then he rates the vignette person’s limitations and his own limitations the same) and *vignette equivalence* (different respondents interpret the description of a vignette in the same way), we can make the scales of the respondents comparable. This is illustrated in the right panel in the figure. There, the positions of John, Mary, and Liam for Respondent 2 are at the same level as for Respondent 1. Since Respondent 2 placed himself between Mary and Liam, this implies that his own rating moves in the direction of less work disability, and we now conclude that Respondent 2 has less work disability than Respondent 1.

It does not matter that we took the scale of Respondent 1 as the base and adjusted the scale of Respondent 2; we could also have taken the scale of Respondent 2 as the base and

**Figure 7. Individual Response Scales and Vignette Adjustments for Differential Item Functioning (DIF)**



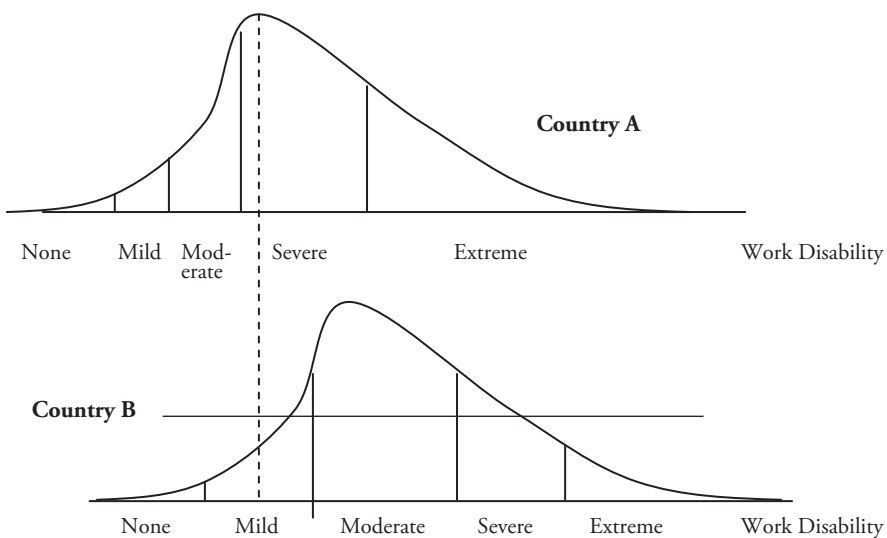
adjusted the scale of Respondent 1. The idea illustrated here can be applied in more general terms, and one can, for instance, take the response scale in one country as a base and then adjust response scales of respondents in other countries so that these become comparable with the scale in the base country.

A different way to represent the way the vignette correction works is illustrated in Figure 8, which presents the distribution of work disability in two hypothetical countries. The density of the continuous work disability variable in country A is to the left of that in country B, implying that on average, people in country A suffer less work disability than in country B. The people in the two countries, however, use very different response scales if asked to report their work limitations on a five-point scale (with response options of none, mild, moderate, severe, and extreme). These differences may be caused by cultural differences or may simply be the result of inadequate translation because, for instance, there exist no exact one-to-one translations of concepts from one language to the other. In the example in the figure, people in country A have a much more negative view of their capacity for work than people in country B. Someone in country A with the health indicated by the dashed line would report having a severe work disability, while a person in country B with the same actual work limitation would report only a mild work disability. The frequency distribution of the self-reports in the two countries would suggest that people in country A are more work-disabled than those in country B—the opposite of the true disability distribution.

Suppose that respondents in both countries are asked to evaluate the work limitation of a person whose disability is given by the dashed line. In country A, this will be evaluated as “severe.” In country B, the evaluation would be “mild.” Since the actual work disability is the same in the two countries, the difference in the country 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 distribution of evaluations in the other country can be adjusted by evaluating them on the benchmark scale. The

**Figure 8. Comparing Self-reported Health-Related Work Limitations**



corrected distribution of the evaluations can then be compared with 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 people in country B are more disabled than those in country A, on average.

The vignette approach cannot only be used to obtain international comparisons corrected for DIF, but also for comparisons of different groups within a given country. For example, it is often hypothesized that men report themselves to be in better health than objective circumstances would warrant; that as people age, they adjust their norms downward about what constitutes good health; and that some of the SES health gradient reflects different health thresholds by SES rather than true health differences. Vignettes offer the potential for systematic testing of these hypotheses.

In order to apply the basic idea in practice, one needs to formalize it mathematically. A relatively straightforward way of doing that is the conditional hierarchical ordered probit (HOPIT) model (or CHOPIT) of King et al. (2004), of which the formulation below is a slight generalization (see, e.g., Kapteyn, Smith, and Van Soest 2007).

Self-reports are modeled as a function of respondent characteristics  $X_i$  and an error term  $\varepsilon_i$  by the following ordered response equation:

$$Y_i^* = X_i\beta + \varepsilon_i; \varepsilon_i \sim N(0,1), \varepsilon_i \text{ independent of } X_i \quad (9)$$

$$Y_i = j \text{ if } \tau_i^{j-1} < Y_i^* \leq \tau_i^j, \quad j = 1, \dots, 5 \quad (10)$$

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

$$\tau_i^0 = -\infty, \tau_i^5 = \infty, \tau_i^1 = X_i\gamma^1 + u_i, \tau_i^j = \tau_i^{j-1} + \exp(X_i\gamma^j), \quad j = 2, 3, 4 \quad (11)$$

$$u_i \sim N(0, \sigma_u^2), \text{ independent of } \varepsilon_i \text{ and } X_i \quad (12)$$

The error term  $u_i$  reflects unobserved heterogeneity in the thresholds. The fact that different respondents can use different response scales  $\tau_i^j$  is called *differential item functioning* (DIF), noted earlier.

Using the self-reports on own work disabilities only, the parameters  $\beta$  and  $\gamma^1$  cannot be separately identified;<sup>3</sup> the reported outcome depends on these parameters only through their difference. For example, consider country dummy variables: two people (with the same characteristics) in two different countries can have systematically different work disabilities, but if the scales on which they report their work disability can also differ across countries, then self-reports alone are not enough to identify the work disability difference between the countries. This was illustrated in Figure 8.

Suppose that we have at our disposal a common set of  $L$  vignette questions. The evaluations of vignettes  $l = 1, \dots, L$  are modeled using similar ordered response equations:

$$Y_{li}^* = \theta_l + \varepsilon_{li} \quad (13)$$

$$Y_{li} = j \text{ if } \tau_i^{j-1} < Y_{li}^* \leq \tau_i^j, \quad j = 1, \dots, 5 \quad (14)$$

$$\varepsilon_{li} \sim N(0, \sigma_v^2), \text{ independent of each other, of } \varepsilon_i \text{ and of } X_i \quad (15)$$

The assumption of *response consistency* means that the thresholds  $\tau_i^j$  are the same for the self-reports and the vignettes. The assumption of *vignette equivalence* implies that the genuine work-related health of the vignette person  $Y_{li}^*$  does not depend on  $X_i$ ; it depends only on the vignette description ( $l$ ) and an idiosyncratic error term.<sup>4</sup>

3. The  $\gamma^j$  for  $j > 1$  will still be identified.

4. Allowing the vignette evaluations to depend on gender of the vignette person (as was done in Kapteyn et al. 2007) does not affect the results.

Given these assumptions, it is clear how the vignette evaluations can be used to separately identify  $\beta$  and  $\gamma (= \gamma^1, \dots, \gamma^4)$ : from the vignette evaluations alone,  $\gamma, \theta_1, \dots, \theta_L$  can be identified (up to the usual normalization of scale and location). From the self-reports,  $\beta$  can also then be identified. Thus, the vignettes can be used to solve the identification problem due to DIF. The two-step procedure is sketched only to make intuitively clear why the model is identified. In practice, all parameters are usually estimated jointly by maximum likelihood.<sup>5</sup>

Adjusting for DIF is straightforward in this model once the parameters are estimated. Define a benchmark respondent with characteristics  $X_i = X(B)$ . (For example, choose one of the countries as the benchmark country.) The DIF adjustment now involves comparing  $Y_i^*$  to the thresholds  $\tau_B^i$  rather than  $\tau_i^i$ , where  $\tau_B^i$  is obtained in the same way as  $\tau_i^i$  but using  $X(B)$  instead of  $X_i$ . Thus, a respondent's work ability is computed using the benchmark scale instead of the respondent's own scale. This does not lead to an adjusted score for each individual respondent (since  $Y_i^*$  is not observed), but it can be used to simulate adjusted distributions of  $Y_i$  for the whole population or conditional on some of the characteristics in  $X_i$ . Of course, the adjusted distribution will depend on the chosen benchmark.

Since the publication of the King et al. (2004) paper, the anchoring vignette methodology has been used in numerous papers, covering several domains: health (Bago d'Uva, O'Donnell, and van Doorslaer 2008; Bago d'Uva et al. 2008; Datta Gupta, Kristensen, and Pozzoli 2009; Salomon, Tandon, and Murray 2004), drinking (Van Soest et al. 2007), job satisfaction (Kristensen and Johansson 2008), and life satisfaction (Kapteyn, Smith, and Van Soest 2010), among others.

Because response consistency is a crucial assumption underlying the vignette approach, it is clearly important to consider empirical evidence that can help gauge the plausibility of the assumption. The amount of work on tests of vignette consistency is still fairly limited.

As noted above, response consistency means that the threshold parameters in Eqs. (10) and (14) are the same. That is, Eq. (11) applies to both. Without additional information, it is not possible to test the response consistency assumption. One approach is to use some extraneous objective measure that allows one to estimate the parameters in Eq. (9) independently from the vignettes. This has been dubbed the *one-factor assumption* by Van Soest et al. (2007). The name refers to the fact that both the outside measure and  $Y_i^*$  are explained by the same equation. Clearly if one "knows" the parameter  $\beta$  in Eq. (9), then one does not need the vignettes for identification and one can relax the response consistency assumption.

This approach was first used by Van Soest et al. (2007) in the context of drinking by Irish students. The extraneous measure in their case is the number of glasses students reported drinking at a given occasion, whereas  $Y_i^*$  is a latent variable indicating to what extent the student's drinking behavior is a cause for concern. These authors compared models with the one-factor assumption (and without response consistency) and models with response consistency (and without the one-factor assumption) using the Akaike information criterion (AIC). They concluded that the model imposing response consistency is to be preferred based on the AIC.

Datta Gupta et al. (2009) followed the approach taken by Van Soest et al. (2007) and applied it to a cross-country comparison of SHARE countries. In their case,  $Y_i^*$  represents self-reported work disability, while the extraneous measure is grip strength. They came to the opposite conclusion: that the model with the one-factor assumption is to be preferred to the model imposing response consistency, once again based on the AIC. However, their basic model is more restrictive than the model presented in Eq. (11), since they omitted the individual heterogeneity terms  $u_i$ . This would seem to be important because vignette evaluations appear to suffer from a fair amount of random variation (e.g., Juerges 2009).

5. This is more efficient than the two-step procedure. Since all error terms are independent, the likelihood contribution is a product of univariate normal probabilities over all vignette evaluations and the self-report, which is relatively easy to compute.



My colleagues and I take a very different approach (Kapteyn, Smith, Vonkova, and Van Soest 2010). We administered a survey of health conditions and self-rated health in a number of domains to participants of the American Life Panel<sup>6</sup> in December 2008. In the spring of 2009, the same respondents were again asked for their self-rated health, but also their ratings of a number of health vignettes in different domains. One of the vignettes in each domain was constructed based on the information provided by the respondent in the December 2008 interview. In other words, that vignette described the respondent's own health a few months earlier. By comparing the self-ratings in December 2008 with the vignette rating in the spring of 2009, one can construct a very direct test of response consistency. The analysis is still ongoing, but it looks as if response consistency is a reasonable but not perfect approximation to the way in which individuals rate their own health and the health of vignette persons in a number of health domains.

## CONCLUSIONS

The notion that one can learn from international comparisons is not new. In particular, the pioneering work of the Gruber and Wise team (1999, 2004, 2007, 2010) has shown the tremendous power of international comparisons for the explanation of policy effects. A major challenge of the work by the Gruber and Wise team has been to deal with data sets from different countries, often containing unequal amounts of information and different variable definitions. With the advent of new harmonized data sets, the power of international comparisons can only increase.

The increased availability of internationally comparable microdata on aging populations around the world is an extremely promising development. Having such data opens up possibilities for important new research. As always, new options come with new challenges. I have described some of the challenges in this article.

In order to be able to learn from international differences, one needs variation in explanatory variables and in outcomes. I have documented some (well-known) differences across the globe in age structure of populations, fertility rates and dependency ratios, retirement ages, mortality, and health care expenditures. Many of these variables show enormous variation across countries and over time.

To take advantage of such variation, comparable longitudinal data for many countries are needed. The health and retirement studies that are now underway on at least three continents provide exactly those data. Modeling the effect of policy or other factors that differ across countries on outcome variables of interest (e.g., retirement, health, or income) requires awareness of various potential pitfalls. These include feedback between behavior and policies,<sup>7</sup> heterogeneity in individual responses to policy changes, and the exclusion of other factors (e.g., macroshocks). Even when such factors are properly taken into account, one still has to establish that seemingly comparable variables are indeed comparable across countries and cultures. I have discussed anchoring vignettes as one possible approach to gauging the comparability of empirical measures across countries. Anchoring vignettes are only one possible approach. The use of biomarkers or physical performance measures is another. These different tools for making outcomes comparable all have their own strengths and weaknesses. It is thus natural to argue for the use of the various tools in combination partly to cross-validate and partly to increase efficiency of international comparisons.

---

6. The American Life Panel is an Internet panel of respondents aged 18 and older. Respondents in the panel log on to the Internet using either their own computer or a Web TV, which allows them to access the Internet using their television and a telephone line. The technology allows respondents who did not have previous Internet access to participate in the panel and furthermore use the Web TVs for browsing the Internet or using e-mail.

7. Policies influence behavior, but policies are partly a reflection of the preferences of citizens, while at the same time, preferences may be influenced by policy to the extent that a policy sets norms for what is acceptable or adequate behavior.

A major motivation of scientific agencies and governments around the world for funding health and retirement studies lies in the notion that the information collected, and the analysis based on the data, will help improve policy decision-making in the face of the dramatic compositional change of the global population. Ultimately, this means that insights obtained from international comparisons need to be translated into policy so that buy-in from politicians is needed. How easy that will be remains to be seen. Politicians on both sides of the Atlantic, for instance, sometimes seem less than eager to learn from others. If it comes to learning from the United States, European politicians are wont to say that they want to avoid “American scenes,” while American politicians have an equally strong inclination to label any European example as “socialist.” Neither of these terms is meant as a compliment.

Thus, the potential for learning exists; how much what is learned will be used is a different matter.

## REFERENCES

- Bago d’Uva, T., O. O’Donnell, and E. van Doorslaer. 2008. “Differential Health Reporting by Educational Level and Its Impact on the Measurement of Health Inequalities Among Older Europeans.” *International Journal of Epidemiology* 37:1375–83.
- Bago d’Uva, T., E. van Doorslaer, M. Lindeboom, and O. O’Donnell. 2008. “Does Reporting Heterogeneity Bias Health Inequality Measurement?” *Health Economics* 17:351–75.
- Banks, J., M. Marmot, Z. Oldfield, and J.P. Smith. 2006. “Disease and Disadvantage in the United States and in England.” *Journal of the American Medical Association* 295:2037–45.
- Datta Gupta, N., N. Kristensen, and D. Pozzoli. 2009. “External Validation of the Use of Vignettes in Cross-Country Health Studies.” IZA Discussion Paper No. 3989. Institute for the Study of Labor, Bonn, Germany.
- Gruber, J. and D. Wise, eds. 1999. *Social Security Programs and Retirement Around the World*. Chicago: University of Chicago Press.
- . 2004. *Social Security Programs and Retirement Around the World: Micro-Estimation*. Chicago: University of Chicago Press.
- . 2007. *Social Security Programs and Retirement Around the World: Fiscal Implications of Reform*. Chicago: University of Chicago Press.
- . 2010. *Social Security Programs and Retirement Around the World: The Relationship to Youth Employment*. Chicago: University of Chicago Press.
- Heckman, J. and R. Robb. 1985. “Using Longitudinal Data to Estimate Age, Period and Cohort Effects in Earnings Equations.” Pp. 137–50 in *Cohort Analysis in Social Research. Beyond the Identification Problem*, edited by W. Mason and S. Fienberg. New York: Springer.
- Johnston, D.W., C. Propper, and M.A. Shields. 2009. “Comparing Subjective and Objective Measures of Health: Evidence From Hypertension for the Income/Health Gradient.” *Journal of Health Economics* 28:540–52.
- Juerges, H. 2009. “Test-Retest Reliability of Health Vignette Ratings: Evidence From SHARE.” Working paper. Mannheim Research Institute for the Economics of Aging (MEA), University of Mannheim. Available online at <http://www.mea.uni-mannheim.de/juerges/TestRetestNov09.pdf>.
- Juster, F.T. and R. Suzman. 1995. “An Overview of the Health and Retirement Study.” *Journal of Human Resources* 30:7–56.
- Kapteyn, A., R. Alessie, and A. Lusardi. 2005. “Explaining the Wealth Holdings of Different Cohorts: Productivity Growth and Social Security.” *European Economic Review* 49:1361–91.
- Kapteyn, A., J.P. Smith, and A. Van Soest. 2007. “Vignettes and Self-reports of Work Disability in the United States and the Netherlands.” *American Economic Review* 97:461–73.
- . 2010. “Comparing Life Satisfaction.” Pp. 70–105 in *International Differences in Well-being*, edited by E. Diener, J.F. Helliwell, and D. Kahneman. New York: Oxford University Press.
- Kapteyn, A., J.P. Smith, A. Van Soest, and J. Banks. 2010. “Labor Market Status and Transitions During the Pre-Retirement Years: Learning From International Differences.” Pp. 63–92 in *Research Findings in the Economics of Aging*, edited by D. Wise. Chicago: University of Chicago Press.

- Kapteyn, A., J.P. Smith, H. Vonkova, and A. Van Soest. 2010. "Anchoring Vignettes and Response Consistency." Working paper. RAND, Santa Monica, CA.
- King, G., C.J.L. Murray, J.A. Salomon, and A. Tandon. 2004. "Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research." *American Political Science Review* 98:567–83.
- Kristensen, N. and E. Johansson. 2008. "New Evidence on Cross-Country Differences in Job Satisfaction Using Anchoring Vignettes." *Labour Economics* 15:96–117.
- Moulton, B.R. 1990. "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units." *Review of Economics and Statistics* 72:334–38.
- Organisation for Economic Co-operation and Development (OECD). 2009a. "Pensions at a Glance 2009: Retirement-Income Systems in OECD Countries." Available online at [http://www.oecd.org/document/49/0,3343,en\\_2649\\_34757\\_42992113\\_1\\_1\\_1\\_1,00.html](http://www.oecd.org/document/49/0,3343,en_2649_34757_42992113_1_1_1_1,00.html).
- . 2009b. "Society at a Glance 2009—OECD Social Indicators." Available online at [www.oecd.org/els/social/indicators/SAG](http://www.oecd.org/els/social/indicators/SAG).
- Salomon, J.A., A. Tandon, and C.J.L. Murray. 2004. "Comparability of Self Rated Health: Cross Sectional Multi-Country Survey Study Using Anchoring Vignettes." *British Medical Journal* 328:258–61.
- United Nations. 2009. "World Population Prospectus: The 2008 Revision." New York: United Nations. UN Population Database available online at <http://esa.un.org/unpp>.
- Van Soest, A., L. Delaney, C. Harmon, A. Kapteyn, and J.P. Smith. 2007. "Validating the Use of Vignettes for Subjective Threshold Scales." RAND Working Paper WR-501. Available online at [http://www.rand.org/pubs/working\\_papers/WR501](http://www.rand.org/pubs/working_papers/WR501).