DISCUSSION PAPER SERIES

IZA DP No. 1118

International Comparisons of Work Disability

James Banks
Arie Kapteyn
James P. Smith
Arthur van Soest

April 2004

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

International Comparisons of Work Disability

James Banks
Institute of Fiscal Studies
and University College London

Arie Kapteyn
RAND and IZA Bonn

James P. Smith
RAND and IZA Bonn

Arthur van Soest
RAND and IZA Bonn

Discussion Paper No. 1118
April 2004

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
Email: iza@iza.org

This paper can be downloaded without charge at:
http://ssrn.com/abstract=533807

An index to IZA Discussion Papers is located at:
http://www.iza.org/publications/dps/

IZA Discussion Paper No. 1118
April 2004

ABSTRACT

International Comparisons of Work Disability*

Self-reported work disability is analyzed in the US, the UK and the Netherlands. Different wordings of the questions lead to different work disability rates. But even if identical questions are asked, cross-country differences remain substantial. Respondent evaluations of work limitations of hypothetical persons described in vignettes are used to identify the extent to which differences in self-reports between countries or socio-economic groups are due to systematic variation in the response scales. Results suggest that more than half of the difference between the rates of self-reported work disability in the US and the Netherlands can be explained by response scale differences. A similar methodology is used to analyze the reporting bias that arises if respondents justify being on disability benefits by overstating their work limiting disabilities.

JEL Classification: J28, I12, C81

Keywords: work limiting disability, vignettes, reporting bias, justification bias

Corresponding author:

Arthur van Soest
RAND
1700 Main Street
P.O. Box 2138
Santa Monica, CA 90407-2138
USA
Email: VanSoest@rand.org

* This research was supported by grants from the National Institute on Aging and the Lasker Foundation to the National Bureau for Economic Research, by a grant from the National Institute on Aging to RAND.

1. Introduction

Reducing work disability among the working population and particularly among older workers is an important issue on the scientific and policy agenda in many industrialized countries. The fraction of workers on disability insurance (DI) is vastly different across countries with similar levels of economic development and comparable access to modern medical technology and treatment. Institutional differences in eligibility rules or generosity of benefits no doubt contribute to explaining the differences in disability rolls. Recent survey data show that significant differences between countries are found in self-reports of work limiting disabilities and in general health. In comparing such self-reports, account should be taken of measurement issues such as differences in question wordings, justification bias and other reporting biases, as well as differences between and within countries that may exist in the scales that are used in answering questions about work disability.

The paper is a progress report on our project that evaluates and compares different approaches to the measurement of work disability. A unique aspect of the project is that it has a distinct multi-national component by using data from three countries: US, U.K., and the Netherlands. These three countries differ in several relevant dimensions—observed rates of self-reported work disability, the generosity and eligibility for government programs that provide income support for people with disability issues, and perhaps national norms about the appropriateness of not working when one is or one claims one is work disabled. However, given their similar levels of economic development and access to modern medical technology and treatment, we suspect that these countries differ less in the 'objectively' measured health status of the population. For this reason, we believe that international comparisons may be particularly useful in understanding some of the most salient research issues that have dominated the scientific literature on work disability.

A second quite unique aspect of this project is that we have been able to address several salient issues in a classic random experimental form. This is because we have had access to two reasonably large Internet samples in two of our countries allowing us to experiment along several dimensions. These samples are the Dutch CentERpanel for the Netherlands and the RAND HRS Internet panel for the United States. For example, we randomly placed experimental disability modules (with alternative forms of disability questions, etc.) into these panels. Since these are recurring panels, we also conducted test-retest reliability of several key measures. This allows us to test in a random experimental form whether the different forms of these questions in prominent US and non-US surveys lead to very different measures of disability using the same population of respondents. Moreover, the reasons for these differences can be explored using the rich information available from the core HRS interviews and the past CentERpanel interviews.

The third unique aspect of this project is the eventual use of the recently fielded English Longitudinal Survey of Aging (ELSA) panel in the U.K. ELSA will expand our work by allowing us to compare objective performance tests with the more standard subjective types of disability questions that are traditionally asked. We will be able to do this is in a large data set of over 10,000 people, many of whom lie in a relevant age range in terms of vulnerability to work disability.

A fourth unique aspect of this project is that we have utilized the vignette methodology to evaluate—once again in an experimental setting—how people within the same country as well as across different countries set thresholds that result in labeling some people work disabled while other people are not so described. Vignette questions have been applied successfully in recent work on international comparisons of health and political efficacy (King et al., 2004). In this project, we will use vignettes to identify systematic errors in self-reported work disability and to correct for justification bias. Our project proposes to apply their method to work related health limitations and extends their model in several directions, with the goal of developing a model that simultaneously explains labor market state, the actual degree of work limiting disabilities, and the bias in reported work limiting disabilities in several domains. We suspect that there may be large and possibly systematic differences in how these work disability thresholds are set.

In addition to these new data sets, we will also use some existing surveys as well (in particular HRS, PSID, and the British Household Panel Survey (BHPS). Our efforts here focus largely on the panel and time series properties of these data. For example, the HRS and BHPS both asked disability type questions in every round allowing us to investigate the amount and reasons for instability in work disability reports across rounds. Similarly, we will eventually explore what happened to disability prevalence rates when there are significant programmatic changes in eligibility (which have occurred in Great Britain in the last few decades).

The remainder of this paper is organized as follows. In the next section, we summarize our experimental results about the impact of different forms of work disability questions that have been asked in major surveys. Section 3 presents some preliminary estimates of the determinants of work disability estimated across our three countries of interest—the US, UK, and the Netherlands. Section 4 summarizes the results we have obtained to date from our research using work disability vignettes in the Dutch CentERpanel. The final section outlines some of the components of our research agenda that are still in process.

2. Does the Form of the Question Matter?

It is an understatement that there is no agreed upon standard format for asking about work disability. Thus, it is not surprising that the format and wording of questions on work disability vary not only internationally but also across the major social science surveys within a country. For example, in the United States quite different questions are asked in the principal yearly government labor force survey—The Current Population Survey or CPS ;and the principal yearly health survey—National Health Interview Survey or NHIS (see Burkhauser et al. 2002). To illustrate, the CPS question is

(a) "Does anyone in the household have a health problem or disability which prevents them from working or which limits the kind or amount or work they can do? [If so,] who is that? (Anyone else?)"

while the NHIS asks instead two questions

(b) "Does any impairment or health problem now keep you from working at a job or business?

(c) " Are you limited in the kind of amount of work you can do because of any impairment?"

To add to the potential domestic confusion, the work disability question in the HRS is

(d) " Do you have any impairment or health problem that limits the kind or amount of paid work you can do?"

and for PSID it is

(e) "Do you have any physical or nervous condition that limits the type of work or the amount of work you can do?"

In all cases, the answers permitted are yes, no, don't know, or refuse so that essentially a dichotomous disability scale can be created.

Some differences between the ways these questions are asked involve language. NHIS and HRS use the term 'impairment'; NHIS, HRS, and CPS use 'health problem'; PSID contains only the phase 'physical or nervous' condition; while the word 'disability' is only used explicitly in CPS. Another potentially important difference is that CPS first asks about anyone in the household and then in a follow-up inquires about whom that might be.

Not surprisingly, survey differences in the manner in which work disability questions are asked are not limited to the United States. For example, the basic work disability question in CentERpanel is

(f) "Do you have an impairment or health problem that limits you in the amount or kind of work you can do?"

While this sounds very similar to the HRS question format, the possible answers are now arrayed on the following 5-point scale

(1) no, not at all, (2) yes, I am somewhat limited, (3) yes, I am rather limited, (4) yes, I am severely limited, and (5) yes, I am very severely limited—I am not able to work.

Finally, in England the disability question used in the BHPS is very similar but not identical to the HRS variant—"Does your health limit the type of work or the amount of work you can do?" While ELSA did not have a work disability question in wave 1, the designers placed the following question into the first follow-up: "Do you have any health problem or disability that limits the kind or amount of work you can do?"[2]

There is a concern that variation across surveys both in the US and internationally in responses to questions about work disability may partly reflect such variation in question format and wording. For example, Burkhauser et al. (2002) report that between 1983 and 1995 NHIS work disability rates were consistently higher than those in the CPS. For example, for men the CPS rate was 8.1% compared to 10.3% in NHIS; among women the CPS rate was 7.4% compared to 10.4% in the NHIS. A possible explanation for this discrepancy is that different questions are used in the two surveys.

2.1. Reports of Disability Prevalence

In this project, we conducted several experiments to evaluate the impact of differences in question wording on reporting of disability prevalence. First, we placed the disability questions summarized above from the HRS, CPS, and NHIS into the RAND HRS Internet panel. This panel is based on a sample of about 2,700 respondents in HRS 2002 wave who had Internet access and who expressed a willingness to participate in an experimental survey on the Internet. This panel allows us to test in a random experimental setting whether the alternative forms of these questions in these three prominent surveys lead to very different measures of disability prevalence using the same population of respondents. Moreover, the reasons for any differences that emerge can be subsequently explored using the rich information available from the core HRS interviews.

In the RAND HRS Internet panel, we conducted the following experiments—half of the sample was randomly assigned the NHIS form of the disability question while the other half received the CPS variant. To test for mode differences (the internet vs. the telephone in the prior wave), the full RAND HRS Internet sample received the normal HRS question. The principal results are contained in Table 2.1.

[2] If the answer to this question is yes, ELSA follows the HRS format by asking "Is this a health problem or disability that you expect to last at least three months?" Two other British surveys ask work disability questions. For example, the Labor Force Survey (LFS) first asks, 'Do you have any health problems or disabilities that you expect will last for more than a year?' If the answer is yes, then respondents are asked in sequence "Does this health problem affect the KIND of paid work that you might do?" and then "or the AMOUNT of paid work that you might do?" The other survey is called the Family Resource Survey (FRS), which asks "Some people are restricted in the amount or type of work they can do, because they have an injury, illness or disability. Which of these statements comes closet to your own position at the moment?" 1. Unable to work at the moment; 2. Restricted in amount or type of work I can do; 3. Not restricted in amount or type of work I can do. In spite of the difference in the manner in which these questions are asked, prevalance rates from the BHPS, LFS, and FRS are remarkably close.

Table 2.1
Disability Prevalence
(% of cases who report disability)

NHIS	18.0
HRS	17.4
CPS	24.6
HRS non-married	23.5
CPS non-married	24.1
NHIS non-married	21.4

Sample is from RAND HRS Internet sample.

Contrary to the speculation in the literature, there does not appear to be any difference in estimates of disability prevalence induced by the wordings of these alternative questions. The NHIS and HRS variants produce bang-on estimates. One current complication in making these comparisons is that HRS staff has not yet coded the specific people affected in the CPS question. Fortunately, a temporary fix is available by limiting the comparisons to non-married respondents. Table 2.1 shows that in this sample HRS, CPS, and NHIS produce remarkably similar sets of estimates about disability prevalence.

While the PSID disability was not included in these experiments, one can compare PSID estimates of work disability prevalence with those obtained in the HRS for the same age group. In that case the PSID estimate of work disability was 28.7 percent while it was 26.8 percent in the HRS, about a two-percentage point difference. This also does not seem to us to be a large difference, but this conclusion must be qualified by the fact that unlike the numbers in Table 2.1 this comparison this is not a strict comparison of question wording only as other factors such as sampling frames may differ between the surveys.

Thus in our view any conflicts that emerge amongst these surveys in estimates of the prevalence of the work disabled population appear not to be due to the form of the disability questions. One possibility is that the greater concentration on health content in the NHIS alerts their respondents to health issues and results in higher reporting of disability, although differences in sampling frames may be a more likely explanation.[3]

Our next set of experiments was conducted using the Dutch CentERpanel, which includes about 2,000 households who have agreed to respond to a set of questions every weekend over the Internet. Unlike the RAND HRS Internet panel, this Dutch sample is not restricted to households with their own Internet access. If they agree to participate and do not currently have Internet access, they are provided Internet access (and if necessary, a set-top box). One advantage of the Dutch Internet panel is that these respondents had already answered many questions about their lives, including questions about their health, demographics and labor force activity. In this project, we carried out a number of experiments over about a six-month period. These included the vignette experiments, which are reported on below, test-retest experiments, and experiments with question wording.

For example, in the second round of the CentERpanel vignette disability experiments, we conducted another experiment about question wording. Randomly, half of CentERpanel respondents in the second wave of our vignette experiments were given the HRS disability question whereby one answered on a yes no basis to the disability question. Given that the first and second waves of our experiments were only a few months apart so that disability reports should not change that much, for

[3] Some evidence is available from ELSA which experimented with placing the general health status questions before and after the detailed set of questions that inquired about a long list of possible health problems. There was some tendency for general health status to be on average better when the questions were placed at the end but the principal difference was that there were fewer respondents at either tail of the five point general health scale when the questions were at the end.

these respondents one can compare the answers to this question to that given on the 5-point scale a few months earlier.

The results are presented in Table 2.2. For all but one row in the 5-point scale, the correspondence is remarkably close. Ninety-six percent of those who answered they were not at all disabled on the 5-point scale also said that they were not when using the HRS dichotomous scale. Similarly, more than 90% of Dutch respondents who said that they were more than somewhat limited replied that they had a work disability on the American 2-point scale.

The ambiguity occurs within the somewhat limited category, which splits about 50/50 when offered an opportunity to simply respond yes or no about their work disability. These are people who are clearly on the margin in terms of their work disability problems. When offered a stark yes or no choice, some will resist disability labeling. But if given a more nuanced set of alternatives, they report some degree of disability.

Table 2.2
Correspondence between 5 and 2-point scale in Dutch panel

5-point work limitations	% in category	marginal % disabled in 2-point scale
not at all	61.8	4.3
somewhat limited	22.5	56.1
rather limited	9.9	91.2
severely limited	2.2	93.1
very severely limited	3.6	92.1

Source: Dutch CentERpanel.

Since this somewhat limited group are about a forth of Dutch respondents, the implication is that reports of disability prevalence are considerably lower if the 2-point scale is used in place of the 5-point scale. Table 2.3 shows reported US disability rates by age (from the PSID) alongside those in the UK (from the Labor Force Survey) and the Dutch disability rates using the 5 and 2-point scale obtained from CentERpanel. Especially during middle age, the Dutch have the highest rates of self-reported work disability, followed by the British, with the Americans having the lowest rates. While estimates of Dutch disability prevalence using the dichotomous scale are still much higher than that observed in the United States, a significant fraction of the disparity could be explained by the format of the disability scale. However, especially for middle age workers—say those between ages 45-64—Dutch rates of reported work disability are still about 15 percentage points higher than those in the United States even when the same question is asked in both countries. We will turn to other explanations for this difference in the next section.

2.2. Test Re-test Results

Aggregate reports of disability prevalence may be similar but specific individuals may change their responses over time even when the question wording is identical. Some of these revisions may reflect real health recovery or decay or changes in the work or family circumstances that affect the self-disability labeling of people. However, individual responses may vary over time even if no real change in the objective health or work circumstances has taken place. This raises the issue of test-retest error in disability reports.

Table 2.3
% With Work Disability by Age—US and Netherlands

	Age Group				
	25-34	35-44	45-54	55-64	65+
US	7.4	11.3	17.6	25.9	38.8
UK	9.1	12.4	19.4	30.8	NA
Netherlands					
5-point scale	25.7	30.3	42.7	44.2	53.6
US 2-point scale	17.2	23.6	38.7	37.4	38.8

US data are from PSID. UK data is from 2001 Labor Force Survey. Due to question routing, the 55-64 group contains women ages 55-59 and men ages 55-64. Netherlands data are from CenTERpanel. Netherlands 5-point scale is based on report of any limitation. All data are weighted.

The following table provides an initial perspective on this issue by dividing HRS respondents who were present in the first five survey waves into four groups. The first group, representing almost 60% of the sample, is those who never reported having a work disability in any of the first five waves. The final group—who constitutes only 8% of respondents is the mirror opposite—those who reported a work disability in all five waves. They could be thought of as the permanently disabled. If such a characterization were accurate, it would imply that the permanent disability rate is about one-third of the yearly disability rate. Note as well that there are very sharp health disability gradients in the first and final row of Table 2.4, a subject to which we will return below.

Table 2.4
Report of Disability by Education in First Five Waves of HRS

	0-11	12	13-15	16 plus	All
Never reported disability	41.7	57.1	64.2	72.4	58.1
Consistent report of new onset	16.1	11.9	10.6	9.1	12.0
Inconsistent report of disability	25.8	23.1	19.5	15.4	21.4
Always reported disability	16.4	8.0	5.6	3.1	8.4

First five waves of HRS.

Given the ages of HRS respondents, disability rates should be expected to increase across the waves and they do. Between the first and fifth HRS wave the percent with a work disability increased from 18% to 27%, or alternatively by 50%. These new onsets do not represent a reporting problem. Some of them are captured in another form in the second row of Table 2.4, which represents those HRS respondents who reported a new disability onset between the HRS waves and who did not negate that report in a subsequent wave. One in eight HRS respondents are found in this group, where once again incidence rates of new disability are also higher among the less educated.

The more problematic group lies in the third row of Table 2.4; those who reported have a work disability in one wave but who subsequently said that they had no work disability. This group represents a quite large fraction of all respondents—one in every five—and an even larger fraction if those who never reported a disability are excluded from the denominator—one out of every three. Of course, some types of work disability are only temporary and actual recovery even for more severe problems is possible. But even when we excluded respondents who stated that their work disabilities were temporary (three months or less), a large fraction subsequently changed their original position on the presence of a work disability.

Some additional insight on this issue can be obtained from the data in Table 2.5, that lists reported work disability transition rates from various surveys in our three countries. These transition rates vary in the length of periodicity between waves, running from a low of only four months in the Dutch panel to as long as two years in the normal HRS survey. Depending on the survey source, transition rates are provided for the full working age distribution and for an approximation to the HRS pre-retirement age distribution. Two types of transitions are listed- the new incidence of disability among those not previously reporting a work disability and the recovery rate (the fraction of those who previously reported a work disability who now say they are not work disabled).

Table 2.5
Transition Disability Rates

	Ages	Periodicity (years)	Disability Prevalence	New Incidence Rate	Recovery Rate	
BHPS	20 - SPA		1	13.9	6.1	28.1
BHPS	50 - SPA		1	24.1	9.4	24.2
PSID	25-64		1	14.1	5.4	28.4
PSID	50-64		1	21.8	10.4	21.8
HRS	51-61		2	17.3	10.9	22.4
HRS Internet	59-69		1		13.8	28.9
Center Panel	25-64		.33	39.6	10.9	20.1

Note: State Pension Age (SPA) in Britain is 60 for women and 65 for men.

Despite these rather large differences in periodicity, these transition rates tell a remarkably similar story. A reasonably large fraction (around a forth) of those who reported having a work disability in one wave subsequently said that they were not work disabled the next wave. Disability recovery rates are slightly lower in the older samples, but even then there appears to be a non-trivial amount of recovery.

In addition, this recovery appears to be quite rapid. For example, the RAND HRS Internet panel repeated the identical work disability question to the same HRS respondents one year later and among those respondents who claimed a work disability in the regular HRS survey one year before, 29% said they were not work disabled during the HRS internet survey. This rate is remarkably similar to the 2-year transition observed during the normal surveys. For example, when work disability reports 2 years apart in the normal survey between the first and second round of HRS saw that 29.8% of those who report disability in round 1 said that they were not disabled in round 2. This similarity in test-retest transitions for one year and two-year mode suggest that some people might be quite unsure of their work disability status and could revaluate it over very short periods of time.

The Dutch panel provides another important perspective on these reported transitions by listing transition rates using the five point scale on the same respondents 4 months apart. We can readily see that among those who moved from a disability to a non-disability category, they were concentrated in the somewhat limited category where 30% of those who claimed that they were somewhat limited in august what not at all disabled in December. Even though these questions were asked only four months apart, one in five respondents who called themselves disabled at some level in August claimed that they were not at all disabled by December. This table suggests that there are a non-trivial group of respondents who are very close to the margin of thinking of themselves as work disabled or not. Their work disabilities do not appear to be severe and when asked work disability questions their answers vary even over relatively short periods of time- measured in months.

Table 2.6
Correspondence between 5-scale in Dutch Panel Four Months Apart

5-point work limitations	not at all	somewhat limited	rather limited	severely limited	very limited
not at all	89.1	9.9	0.7	0.2	0.1
somewhat limited	29.5	54.0	14.4	1.9	0.2
rather limited	4.9	37.8	50.3	5.6	1.4
severely limited	6.2	5.3	45.1	36.2	7.3
very severely limited	3.4	.7	21.3	7.7	64.0

Source: Dutch CentERpanel, August and December waves.

3. Comparisons of Work Disability Models Across Countries

In this section, we present and discuss some simple empirical models of the predictors of work disability in three countries—the U.K., the Netherlands, and the US. To the extent possible, we limit the models to variables that are in common in all countries and focus on health and demographic predictors. These models then become the stylized facts for the vignette analysis presented below in section 4.

Disability is an important program in many countries, and one that until recently was growing rapidly over time. The number of people on disability programs is substantial, particularly among men and women in the age groups 45-64. For the US, Autor and Duggan (2003) find that the numbers of disability insurance (DI) recipients per 1000 men and women in the age group 55-64 have increased from 96 to 108 (men) and from 43 to 72 (women) between 1984 and 1999. Bound and Burkhauser (1999) report that in 1995, the number of DI recipients per 1000 workers was 103 in the age group 45-59 and 314 in the age group 60-64. Both numbers have grown substantially in the early nineties. There are also substantial differences between OECD countries. For example, the numbers of DI recipients per 1000 workers in the age category 45-59 were 87 for Germany and 271 in The Netherlands. According to Eurostat (2001), the number of 16-64 year olds receiving disability and sickness benefits is less than 3% in Italy and Greece, but almost 10% in Denmark and more than 12% in the UK.

The principal question that we ask in this project is how much of the reported differences among these countries reflect differences in some type of reporting bias and how much reflects actual differences in true work disability. Our first step in that inquiry is to estimate simple but standard types of models predicting work disability in three countries as a function of demographics and health. The models for the United States are based on two surveys—the PSID and the original HRS cohort of those born between 1931 and 1941. The PSID has the advantage of covering the complete age distribution while the HRS contains a much richer set of health variables related to work disability. The Dutch models are all based on the Dutch CentERpanel, but they are estimated over two samples—the first corresponds to the age distribution in the PSID sample (those 25 and over) while the second comes as close as possible to the HRS age distribution. Obviously, sample sizes in the already smaller Dutch sample become more of a concern in its HRS look alike sample. The English models are also estimated using two samples from the BHPS—the complete age distribution available in the BHPS and a narrower one that corresponds to the HRS sample.

One issue that arises in attempting to estimate the ‘same’ model on all three samples is how to deal with respondents’ reports of general health status, which are available in all three countries. The nature of the problem is illustrated in Table 3.1, which lists respondents’ evaluation of their health along the familiar 5-point scale—excellent, very good, good, fair, and poor. Since this comparison involves three populations where as a first approximation their ‘true’ health status is unlikely to be that much different, it is apparent that the Dutch, British, and Americans use very different criteria to place themselves within these five thresholds.

Table 3.1
Comparisons of Self Reported Health Status

	Netherlands	US	UK
Excellent	5.8	24.7	15.1
Very Good	23.9	36.0	31.9
Good	56.2	28.1	31.8
Fair	11.8	8.9	14.0
Poor	1.1	2.3	3.9

US data are from PSID. Netherlands data are from CenTERpanel. UK data from the BHPS. Ages 25-64 in all countries. All data are weighted.

The circumspect Dutch appear to run to the center not willing to make health claims at either the top or bottom while the ever optimistic Americans are four times more likely to state that they are in excellent health. Prudent as always, the British lie between these extremes. Given the size of the differences between the three countries in how health circumstances are translated into this 5-point scale, it is not all apparent that one would want to control for self-reported health status when doing international comparisons. This is especially the case when the overall objective of the research is to eventually model international differences in reported rates of work disability. Therefore in the empirical estimates summarized in this section, we model work disability without general health status as a predictor.

All models are based on probit estimates of the probability that a respondent reported having a work disability. Due to the impact of the 5-point and 2-point scale documented in the previous section, the Dutch models are estimated using both the 5-point scale (in Appendix Table 1) with any level of disability is equated with having a work disability and the more comparable 2-point HRS scale.

The Dutch models using the two point scale are listed in Tables 3.3, the corresponding American estimates based on the PSID and HRS are listed in Table 3.4, and the British models in Tables 3.5. The A panels in these tables summarize the models for the sample ages 25 and more while the B panels contain estimates that approximate the age group in the American HRS sample. All tables list estimated coefficients, derivatives, standard errors (of the derivatives), and means of variables in the columns. The principal difference between the age-restricted HRS style models and those for the entire age range is that an additional variable is added in the HRS style samples—whether the respondent suffered from pain. Unfortunately, this quite central measure is not available in the PSID.

Before discussing our estimated coefficients, it is useful to highlight first prevalence rates of the more objective chronic health conditions (hypertension, diabetes, cancer, diseases of the lung, heart problem, stroke, and arthritis) in the three countries. These prevalence rates are provided in Table 3.2. Even reporting of such conditions may be different across nations due to differential physician contact or because the precise criteria for thresholds for medical diagnosis may not be the same. Of course as before, the specific survey questions may and do vary, which may induce another layer of international non-comparability.

With those concerns as an important caveat, whether we examine the full age range or the more narrow age range in Table 3.2, in general the prevalence rates of all these health conditions actually appears higher in the US than in either the Netherlands or the UK. This is especially the case in the more age limited sample—but the point still applies with somewhat less force in the full age range especially when one recognizes that the Dutch sample is older on average. It is not central to our argument that the Dutch or British sample is much healthier than the American ones. It is enough for now to state that differential levels of these objective health measures seems unlikely to account for the much higher work disability rates observed among the Dutch and the English compared to the Americans.

Table 3.2
Prevalence of Work Disability - weighted

Variables	Dutch	US-PSID	English
	All Ages		
hypertension	.202	.227	.211
diabetes	.048	.079	.041
cancer	.037	.055	.017
disease of lung	.058	.051	NA
heart problem	.066	.088	.047
stroke	.013	.035	.044
arthritis	.109	.227	.151
emotion	.110	.060	.091
pain	.259	NA	
Variables	Dutch	US-HRS	English
	Ages HRS approximate		
hypertension	.248	.360	.246
diabetes	.044	.092	.040
cancer	.042	.053	.015
disease of lung	.063	.068	NA
heart problem	.074	.117	.046
stroke	.015	.024	.044
arthritis	.134	.364	.188
emotion	.124	.111	.102
pain	.316	.241	.278

Second, whatever the international differences, prevalence rates for many of these conditions are sufficiently small in both countries that they are unlikely to contribute in a major way towards explaining differential rates of work disability. In this group, we would include at least cancer, diabetes, strokes, and heart problems. Moreover, Tables 3.2-3.5 suggest that one disease with relatively high prevalence—hypertension—does not by itself translate at a sufficiently high rate into work disability to be likely to account for much of the cross-country difference.

The major exception to this line of reasoning is arthritis where not only are prevalence rates reasonably large, but it is also a significant predictor of work disability. Moreover, diagnosis of arthritis is particularly imprecise as all sorts of muscular aches and pains whether severe or minor can be called arthritis. This is especially the case when self-diagnosis is permitted, as is the case in the HRS variant of the question and which accounts for the quite high prevalence in that sample. But rates of reported arthritis are actually much higher in the US than in the Netherlands, so it seems again an unlikely explanation for why Dutch work disability rates are so much higher than those in the US.

This brings us to two other also difficult to diagnose conditions which seem much more promising candidates for why disability rates differ between the two countries. These are having emotional problems and being bothered by pain. In contrast to the other more objective health conditions, these two more subjective health conditions actually have slightly higher prevalence in the two European countries compared to the US.

Pain and emotion not only have slightly higher prevalence in Europe, but in all three countries our probit estimates indicate that pain and emotion are among the strongest predictors of work disability. Since these two conditions are also among the more subjective and the more difficult to diagnose, this may indicate that the source of the international differences in reports of work disability may rest in these two conditions. It may be that for the same level of pain, that the Dutch are more likely to say that it constitutes a work disability than are the British, who in turn are more likely to claim a work disability due to pain than are the Americans. This speculation about these possible international differences in reporting leads us to try to test these ideas. Our tests will take the form of vignettes on work disability.

Before turning to the use of vignettes, consider Tables 3.6-3.8, which use the estimated probit models to summarize the effect of each explanatory variable on self-reported work disability. The procedure followed to construct the tables is as follows. We use the estimated models to predict the prevalence of work disability. (This corresponds to the bottom rows in Tables 3.3-3.5.) Next we set the coefficient of an explanatory variable equal to zero and predict once again work disability prevalence. We interpret the resulting change in predicted work disability prevalence as the effect of the corresponding variable on overall work disability. The procedure is repeated for all explanatory variables.

Furthermore, we can decompose the "total effect" of an explanatory variable in two factors. The first factor is the prevalence of that variable (all explanatory variables are dummies). The second factor is the effect on prevalence of work disability for observations for which the dummy is non-zero. The product of the two factors gives the total effect.

Tables 3.6-3.8 confirm the previous discussion. We observe that in all three countries pain is the biggest contributor to self-reported disability, with by far the biggest effect in The Netherlands. Arthritis is the second most important determinant in the US and the UK. In the older age group in The Netherlands emotional problems are the second leading cause for work disability followed by arthritis.

Table 3.3A
Dutch Probit for Work Disability—All ages
(using 2-point scale)

Variables	Coef.	DF/DX	Robust SE	Means
hypertension	-.047	-.014	.026	.190
diabetes	.636	.223	.063	.042
cancer	.161	.050	.060	.036
disease of lung	.768	.274	.053	.056
heart problem	.997	.362	.050	.066
stroke	.846	.309	.114	.011
arthritis	1.130	.410	.042	.099
emotion	.915	.327	.041	.096
female	.198	.059	.021	.451
age 35-44	.063	.019	.034	.226
age 45-54	.368	.117	.035	.240
age 54-64	.400	.129	.040	.180
age 65+	.235	.074	.042	.164
ed med	.043	.013	.025	.324
ed high	-.178	-.052	.024	.366
constant	-1.352			
observed p		.249		

Table 3.3B
Ages 45-64

Variables	Coef.	DF/DX	Robust SE	Means
hypertension	-.162	-.052	.041	.233
diabetes	.426	.154	.094	.044
cancer	.223	.078	.107	.044
disease of lung	.691	.258	.125	.048
heart problem	1.012	.382	.090	.058
stroke	2.136	.673	.082	.011
arthritis	.610	.223	.074	.109
emotion	1.099	.409	063	.128
pain	1.550	.546	.041	.265
female	.033	.011	.039	.438
ed med	.139	.046	.046	.281
ed high	-.085	-.028	.044	.350
constant	-1.374			
observed p		.310		

Table 3.4A
US-PSID Probit for Work Disability—All Ages PSID
(using 2-point scale)

Variables	Coef.	DF/DX	Robust SE	Means
hypertension	.239	.057	.011	.230
diabetes	.371	.096	.020	.072
cancer	.470	.128	.027	.043
disease of lung	.670	.196	.032	.040
heart problem	.620	.176	.024	.067
stroke	.757	.229	.038	.029
arthritis	.765	.211	.015	.189
emotion	1.009	.321	.029	.053
female	-.058	-.013	.009	.561
age 35-44	.175	.040	.015	.330
age 45-54	.291	.069	.017	.248
age 54-64	.389	.100	.023	.100
age 65+	.376	.095	.021	.154
ed med	-.196	-.044	.011	.621
ed high	-.313	-.062	.011	.205
constant	-1.399			
observed p		.181		

Table 3.4B
Ages 51-61, HRS

Variables	Coef.	DF/DX	Robust SE	Means
hypertension	.169	.042	.008	.366
diabetes	.312	085	.014	.103
cancer	.345	.096	.019	.052
disease of lung	.533	.157	.019	.068
heart problem	.645	.192	.015	.121
stroke	.887	.293	.035	.029
arthritis	.286	.072	.008	.363
emotion	.527	.153	.015	.108
pain	.984	.290	.011	.250
female	-.178	-.043	.008	.537
ed med	-.158	-.039	.008	533
ed high	-.336	-.073	.009	.175
constant	-1.324			
observed p	.215			

Table 3.5A
British Probit for Work Disability—25+
(using 2-point scale)

Variable	**coefficient**	**DF/dX**	**Robust s.e**	**Mean**
Hypertension	0.307	0.088	0.016	0.186
Diabetes	0.606	0.197	0.036	0.035
Cancer	1.017	0.359	0.073	0.015
Heart	0.614	0.199	0.035	0.047
Stroke	0.750	0.251	0.032	0.044
Arthritis	0.937	0.311	0.020	0.133
Emotion	0.917	0.310	0.023	0.088
Female	0.032	0.008	0.010	0.535
Age 35-44	0.189	0.052	0.018	0.238
Age 45-54	0.195	0.054	0.018	0.192
Age 55-64	0.370	0.109	0.021	0.147
Age 65+	0.700	0.216	0.023	0.205
Ed_med	-0.288	-0.074	0.010	0.399
Ed_High	-0.415	-0.097	0.011	0.199
Constant	-1.370			
Observed p	0.227			

Table 3.5B
British Probit for Work Disability— Adults aged 50-64
(using 2-point scale)

Variable	coefficient	DF/dX	Robust s.e	Mean
Hypertension	0.281	0.084	0.028	0.248
Diabetes	0.646	0.221	0.077	0.039
Cancer	1.782	0.627	0.097	0.014
Heart	0.835	0.295	0.075	0.049
Stroke	0.694	0.240	0.067	0.044
Arthritis	0.756	0.250	0.035	0.186
Emotion	0.725	0.247	0.047	0.101
Pain	0.972	0.314	0.029	0.281
Female	-0.027	-0.008	0.022	0.543
Ed_med	-0.242	-0.067	0.022	0.368
Ed_High	-0.411	-0.103	0.026	0.161
Constant	-1.332			
Observed p		0.252		

Table 3.6A
Decomposition of Dutch Disability—All ages
(using 2-point scale)

Variables	Total effect (%)	Prevalence (%)	Effect among individuals with characteristic (%)
hypertension	0.24	19.11	-1.26
diabetes	0.91	4.82	18.78
cancer	0.15	3.38	4.46
disease of lung	1.37	6.14	22.36
heart problem	2.22	7.27	30.53
stroke	0.44	1.90	23.33
arthritis	4.05	11.36	35.63
emotion	2.86	10.31	27.74
female	2.51	48.76	5.15
age 35-44	0.33	18.81	1.73
age 45-54	1.85	16.05	11.52
age55-64	1.78	12.83	13.86
Age 65+	1.25	15.58	8.05
ed med	0.43	31.45	1.36
ed high	-0.98	17.58	-5.55

Table 3.6B
Ages 45-64

Variables	Total effect (%)	Prevalence (%)	Effect among individuals with characteristic (%)
hypertension	-0.94	24.94	-3.79
diabetes	0.41	3.63	11.32
cancer	0.23	4.57	4.98
disease of lung	0.77	5.82	13.27
heart problem	1.86	8.30	22.37
stroke	0.58	1.09	53.34
arthritis	2.15	13.78	15.59
emotion	3.49	12.72	27.47
pain	15.79	33.56	47.05
female	0.37	46.66	0.79
ed med	0.99	32.70	3.03
ed high	-0.37	20.01	-1.86

Table 3.7A
Decomposition of US Work Disability—All Ages PSID
(using 2-point scale)

Variables	Total effect (%)	Prevalence (%)	Effect among individuals with characteristic (%)
hypertension	1.62	24.89	6.51
diabetes	0.82	7.99	10.26
cancer	0.71	5.50	12.87
disease of lung	0.95	5.06	18.85
heart problem	1.62	8.82	18.34
stroke	0.75	3.52	21.38
arthritis	5.12	22.71	22.53
emotion	1.71	6.04	28.36
female	-0.69	55.33	-1.25
age 35-44	0.64	24.73	2.57
age 45-54	1.31	24.85	5.26
age55-64	1.23	14.09	8.71
Age 65+	2.31	22.52	10.25
ed med	-2.62	60.06	-4.36
ed high	-1.44	22.73	-6.34

Table 3.7B
Ages 51-61, HRS

Variables	Total effect (%)	Prevalence (%)	Effect among individuals with characteristic (%)
hypertension	1.39	36.04	3.85
diabetes	0.75	9.16	8.16
cancer	0.44	5.25	8.44
disease of lung	1.00	6.84	14.61
heart problem	2.05	11.69	17.51
stroke	0.56	2.38	23.62
arthritis	2.61	36.41	7.17
emotion	1.62	11.14	14.58
pain	6.99	24.07	29.05
female	-1.95	52.21	-3.74
ed med	-1.86	55.86	-3.33
ed high	-1.14	19.38	-5.89

Table 3.8A
Decomposition of British Work Disability—25+
(using 2-point scale)

Variables	Total effect (%)	Prevalence (%)	Effect among individuals with characteristic (%)
hypertension	1.87	21.14	8.85
diabetes	0.72	4.07	17.65
cancer	0.45	1.73	26.18
heart problem	0.99	5.54	17.84
stroke	0.97	4.67	20.88
arthritis	4.44	15.06	29.48
emotion	2.34	9.08	25.75
female	0.41	54.97	0.75
age 35-44	0.68	21.41	3.19
age 45-54	0.70	18.86	3.73
age55-64	1.41	15.90	8.86
Age 65+	5.32	25.91	20.53
ed med	-2.39	37.58	-6.35
ed high	-1.55	18.38	-8.43

Table 3.8B
Adults aged 50-64

Variables	Total effect (%)	Prevalence (%)	Effect among individuals with characteristic (%)
hypertension	1.64	25.16	6.53
diabetes	0.61	4.22	14.36
cancer	0.51	1.50	34.02
heart problem	0.95	4.87	19.55
stroke	0.75	4.46	16.85
arthritis	4.30	19.03	22.57
emotion	1.94	10.28	18.89
pain	8.13	27.86	29.19
female	-0.33	55.24	-0.60
ed med	-1.91	36.34	-5.27
ed high	-1.29	15.39	-8.39

4. Vignettes

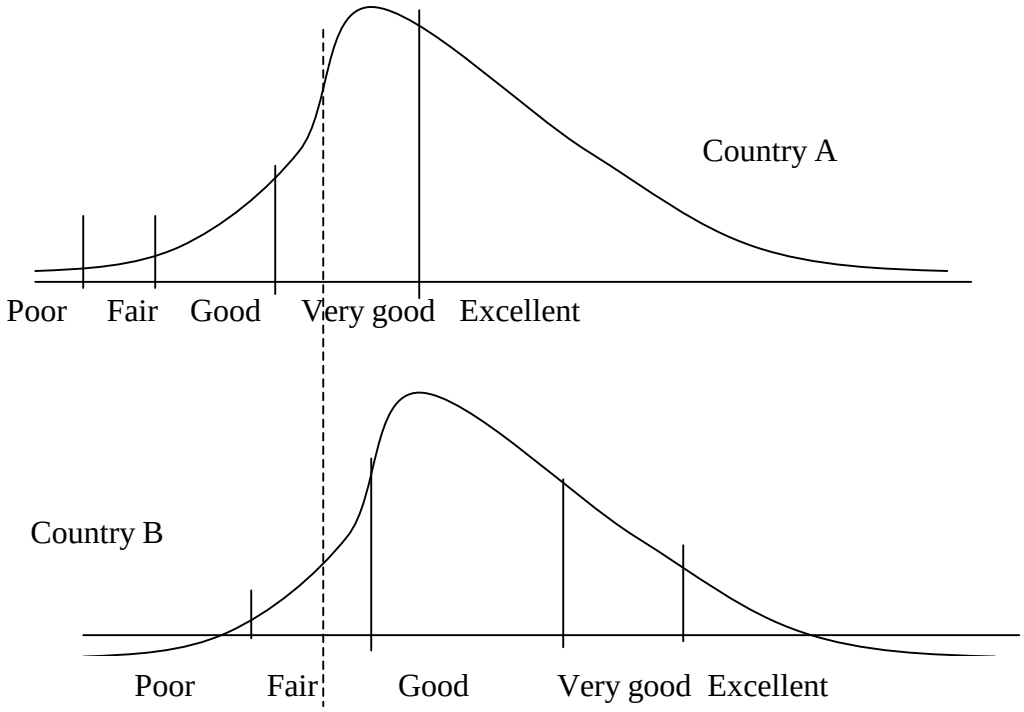
We first provide an intuitive description of the use of vignettes for identifying reporting biases, following King, Murray, Salomon and Tandon (2004). The original King et al. (2004) model shows how vignettes can help to identify systematic differences in response scales between groups (or countries), making it possible to decompose observed differences in, for example, self-reported health in a specific domain into differences due to response scale variation and genuine differences in health. Our work applies this model to work limiting disability rather than health, and extends it to incorporate justification bias.

Vignette evaluations were collected in the Netherlands in 2003, and more recently also in the US. For the US, only 346 observations are available as yet. Most of the empirical work in this section uses only the Dutch data. Only the final part uses both US and Dutch vignettes, allowing for an international comparison. Work disability vignettes for the UK will not be available in the short run.

4.1. The King et al. Model

The basic idea of the model is sketched in Figure 1. It presents the distribution of health (in a specific domain, such as vision or emotional well-being) in two countries. The density of the continuous health variable in country A is to the left of that in country B, implying that on average, people in country A are less healthy than in country B. The people in the two countries, however, use very different response scales if asked to report their health on a five-point scale (poor-fair-good-very good-excellent, say). In the example in the figure, people in country A have a much more positive view on a given health status than people in country B.

Figure 1. Comparing self-reported health across two countries in case of DIF



For example, someone in country A with the health indicated by the dashed line would report to be in very good health, while a person in country B with the same actual health would report "fair." The frequency distribution of the self-reports in the two countries would suggest that people in country A are healthier than those in country B—the opposite of the actual health distribution. Correcting for the differences in the response scales (DIF, "differential item functioning," in the terminology of King et al.) is essential to compare the actual health distributions in the two countries.

Vignettes can be used to do the correction. A vignette question describes the health of a hypothetical person and then asks the respondent to evaluate the health of that person on the same five-point scale that was used for the self-report. The vignette descriptions are the same in the two countries, so that the vignette persons in the two countries have the same health conditions. For example, respondents can be asked to evaluate the health of a person whose health is given by the dashed line. In country A, this will be evaluated as "very good." In country B, the evaluation would be "fair." Since the actual health description of the vignette person is the same in the two countries, the difference in the country evaluations must be due to DIF. Vignette evaluations thus help to identify the differences between the response scales in the two countries. 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 underlying assumption is *response consistency:* a given respondent uses the same scale for the self-reports and the vignette evaluations.

The corrected distribution of the evaluations can then be compared to that in the benchmark country—they are now on the same scale. In the example in the figure, this will lead to the correct conclusion that people in country B are healthier than those in country A, on average. King et al. (2004) develop parametric and nonparametric models that make it possible to perform the correction. They apply their method to, for example, political efficacy and visual acuity. Their results strongly support the ability of the vignettes to correct for DIF. For example, in a comparative study of political efficacy of Chinese

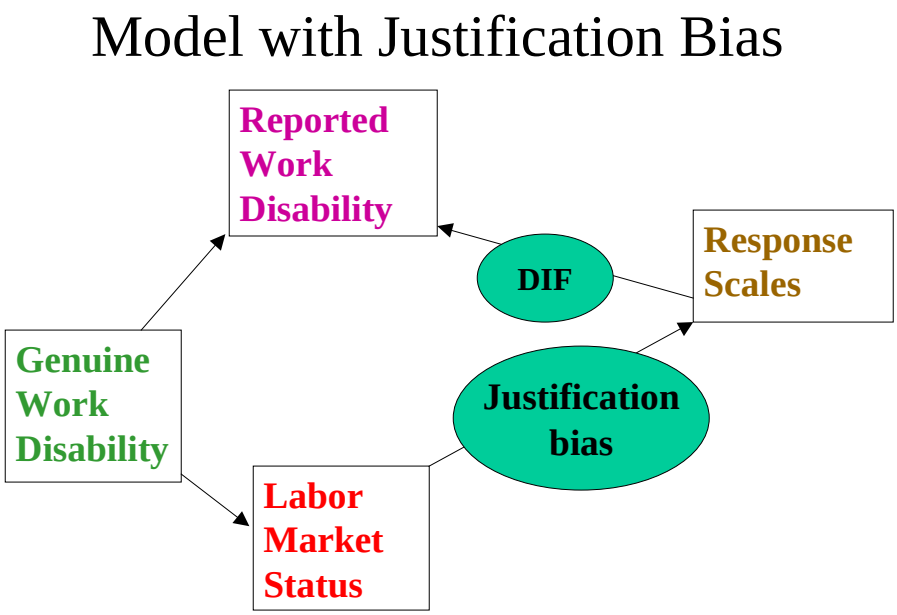
and Mexican citizens, they find that without correction the Chinese seem to have more political influence than the Mexicans. The conclusion reverses if the correction is applied.[4]

4.2. Vignettes for Work Limiting Disability

We will apply ideas in King et al. (2004) to work limiting disability, using vignettes not only to obtain international comparisons corrected for DIF, but also for comparisons of different groups within a given country. For example, *justification bias* (Bound, 1991; Kerkhofs and Lindeboom, 1995; Kreider, 1999; Currie and Madrian, 1999) can be seen as a form of DIF, with people on disability programs or other non-workers giving systematically different evaluations of their own work limiting disabilities than people who work. Comparing evaluations of vignette persons with disabilities in a certain domain (such as back pain, depression, or breathing problems, etc.) given by workers and non-workers will show whether workers and non-workers give systematically different evaluations that could reflect justification bias. This requires extending the King et al. model, since the respondent's labor market state will depend on whether the respondent has some work limiting disability or not. This feedback mechanism makes labor market state endogenous to the respondent's work limiting disability. This will be taken into account by constructing simultaneous models for labor market state, work limiting disability, reporting bias, and vignette evaluations.

Figure 2 illustrates how this model works. We take actual work limiting disabilities as exogenous to labor market status, i.e., we do not allow for feedback from labor market position to actual work

Figure 2. Model with justification bias



limiting disability. Justification bias implies that there is an effect of labor market status on the way respondents answer the question on work limiting disabilities, i.e., on the response scales they use to

[4] More applications to health are discussed in Sadana et al. (2002) and in Salomon, Tandon and Murray (2004).

distinguish between mild and moderate limitations, moderate and severe limitations, etc. Response scales and actual work limiting disabilities jointly determine reported work limiting disability. The effect of response scale differences on reported work disability is referred to as DIF. Justification bias is one source of DIF. The difference with other sources of DIF is that labor market status depends on actual work disability : people with a serious work limiting disability will have a larger probability to be on disability transfers rather than at work.

4.3. Econometric Models

We first discuss the benchmark model of King et al. (2004) and then its extension incorporating justification bias. Both models explain respondents' self-reports on work limitations and their reports on work limitations of hypothetical vignette persons. The first is the answer (Y_{ri}, i indicates respondent i) to the question

> *"Do you have any impairment or health problem that limits the type or amount of work that you can do?"*

In our data, the answers are given on a five points scale,[5] with answers "no, not at all" (Y_{ri} =1), "yes, I am somewhat limited" (Y_{ri} =2), "yes, I am moderately limited" (Y_{ri} =3), "yes, I am very limited" (Y_{ri} =4) and "yes, I am so seriously limited that I am not able to work" (Y_{ri} =5).
The questions on work limitations of the vignette persons have the same answering categories and are formulated in the same way ("Does Mr/Mrs X have any impairment or health problem that limits the type or amount of work that he or she can do?"). The answers will be denoted by Y_{li} where each respondent i evaluates a number of vignettes $l=1,\ldots,L$.

Benchmark Model

We follow King et al. (2004) with somewhat different notation. Self-reports are modeled by the following ordered response equation:

$$Y_{ri}^{*} = X_i\beta + \varepsilon_{ri};\ \ \varepsilon_{ri} \sim N(0,\sigma_r^2),\ \varepsilon_{ri} \text{ independent of } X_i,\ V_i$$
$$Y_{ri} = j \text{ if } \tau_i^{j-1} < Y_{ri}^{*} \le \tau_i^{j},\ \ j = 1,\ldots 5$$

The thresholds τ_j^i between the categories are given by

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

The fact that different respondents can use different response scales is called "differential item functioning" (DIF). In the King et al. model, response scales can vary only with observed characteristics V_i.

In King et al. (2004), both X_i and V_i include country dummies (among other variables). Using the self-reports only, the coefficients on the country dummies in β and γ^1 cannot be separately identified; the reported outcome only depends on these parameters through their difference. In other words: if two people (with the same characteristics) in two different countries can have systematically different work disability, but if the scales on which they report their work disability can also differ across countries, then the self-reports are not enough to identify the work disability difference between the countries. Vignettes are useful because they solve this identification problem.

The same applies to within country comparison of different groups. Suppose, for example, that a given health condition not only drives actual work limitations, but also has an effect on the response

[5] The HRS and PSID have self-report questions on work limiting disabilities on a two-points scale.

scale. For example, people who themselves have diabetes may have different views on who can and who cannot work than people without diabetes, resulting in a systematic difference between response scales of those who do and do not have diabetes. If people with and without diabetes report different levels of work limitations, then we do not know whether this reflects genuine differences in work limitations (β) or differences in response scales (γ^1).[6] We may be able to identify $\beta - \gamma^1$ but not β and γ^1 themselves. Ignoring that health conditions change response scales will imply that estimates of $\beta - \gamma^1$ are interpreted as estimates of β and will thus be biased if γ^1 is not equal to zero.

The evaluations of vignettes $l=1,\ldots,L$ are modeled using a similar ordered response model:

$$Y_{li}^* = \theta_l + \theta \,\text{Female}_{li} + \varepsilon_{li}$$
$$Y_{li} = j \text{ if } \tau_i^{j-1} < Y_{li}^* \le \tau_i^j,\ j = 1,\ldots 5$$
$$\varepsilon_{li} \sim N(0, \sigma^2), \text{ independent of each other, of } \varepsilon_{ri} \text{ and of } X_i,\ V_i$$

One crucial assumption of King et al. (2004) is that the thresholds τ_i^j are the same for the self-reports and the vignettes *("response consistency").* This assumption will be maintained in our work and is the basis for why vignettes help to identify DIF and help to correct for reporting differences.

The second assumption of King et al. (2004) is that Y_{li}^* does not vary with respondent characteristics in any systematic way, it only varies with vignette characteristics given in the descriptions of the vignettes (captured by a vignette specific constant θ_l and a dummy for the gender of the vignette person).

Given these assumptions, it is clear how the vignette evaluations can be used to identify β and γ $(=\gamma^1,\ldots\gamma^5)$: From the vignette evaluations alone, γ, θ, $\theta_1,\ldots\theta_5$ can be identified (up to the usual normalization of scale and location). From the self-reports, β can then be identified in addition. 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 will be estimated simultaneously by maximum likelihood.[7] This will be more efficient than the two-step procedure. Since all error terms are independent, the likelihood contribution will be a product of univariate normal probabilities over all vignette evaluations and the self-report, which is relatively easy to compute.

Correcting for DIF is straightforward in this model once the parameters are estimated. Define a benchmark respondent with characteristics $V_i = V(B)$. (For example, choose one of the countries as the benchmark country.) The DIF correction would now involve comparing Y_{ri}^* to the thresholds τ_B^j rather than τ_i^j, where τ_B^j is obtained in the same way as τ_i^j but using $V(B)$ instead of V_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 a corrected score for each individual respondent (since Y_{ri}^* is not observed) but it can be used to simulate corrected *distributions* of Y_{ri} for the whole population or conditional upon some of the characteristics in V_i and or X_i. Of course the corrected distribution will depend upon the chosen benchmark.

Labor Market State and Justification Bias

Justification bias arises if respondents who are on disability use different scales than respondents who work. It has been a central issue in the literature on the effects of health and other variables such as

[6] The γ^j for $j>1$ will still be identified.

[7] The software for ML estimation of the benchmark model is available on Gary King's home page (http://gking.harvard.edu/vign/). In extensions, the likelihood will often be approximated using simulations (Hajivassiliou and Ruud, 1994).

financial incentives on entrance into a disability program (Bound, 1991; Kerkhofs and Lindeboom, 1995; Kreider, 1999; Lindeboom and Kerkhofs, 2002). It can be incorporated in the model by including labor market state dummies in the determinants of the thresholds V_i. Since labor market status will depend on work limitations, there is feedback from Y_{ri}^* to V_i. It is essential to take this into account to obtain consistent estimates and an appropriate correction for reporting (justification) bias. We do not aim at modeling the exact way in which work limitations affect labor market status but formulate a reduced form model for labor market status, including all the exogenous variables that may affect labor market status directly or through work limitations.

We will distinguish J labor market states $j=1,\ldots,J$. In our empirical work J is equal to 5: at work for pay ($j=1$), homemaker ($j=2$), retired ($j=3$), on disability ($j=4$), and "other" ($j=5$; including students, unemployed, volunteer workers; these states have too few observations to be considered separately).

We use a standard multinomial logit model to explain labor market status from all exogenous characteristics on the respondents that we have, i.e., $Z_{i=}(X_i,V_i)$. Then the model for labor market status can be written as follows:

$$s_{ji} = Z_i^{'}\pi_j + \eta_{ji},\ j = 1,\ldots J$$
$$s_{ji} \geq s_{mi},\ m = 1,\ldots,J \Rightarrow S_i = j$$

where the error terms η_{ji} all follow a Generalized Extreme Value Type I distribution (with $P(\eta_{ji} \leq t) = e^{-e^{-t}}$), independent of each other and of Z_i. For normalization, π_1 is set to a vector of zeros. This gives the familiar multinomial logit probabilities (the index i is suppressed):

$$P[S_i = j \mid Z_i] = \exp(Z_i^{'}\pi_j) / \sum_m \exp(Z_i^{'}\pi_m),\ j = 1,\ldots,J$$

The equations for self-reported work limitations and for vignette evaluations are similar to their previous specifications except for incorporating labor market status. We assume that there is no direct feedback from labor market status to true work limitations (see Figure 2); the only effect of labor market status is that it changes the evaluation scale. This is the idea of justification bias: keeping health etc. constant, people in different labor market positions will answer the questions differently but this does not reflect genuine differences in work ability.

For completeness, we present the extended equations. Here D_i is a vector of four labor market state dummies for respondent i (at work for pay is taken as the benchmark category).

Self-reported work limitations:

$$Y_{ri}^* = X_i\beta + \varepsilon_{ri};\ \ Y_{ri} = j \text{ if } \tau_i^{j-1} < Y_{ri}^* \leq \tau_i^j,\ j = 1,\ldots 5$$
$$\tau_i^0 = -\infty,\ \ \tau_i^K = \infty,\ \ \tau_i^1 = \gamma^1 V_i + \psi D_i,\ \ \tau_i^j = \tau_i^{j-1} + \exp(\gamma^j V_i),\ j = 2,3,4$$
$$\varepsilon_{ri} \sim N(0,\sigma_r^2), \text{ independent of } X_i, V_i \text{ (but not of } D_i).$$

Vignette evaluations:

$$Y_{li}^* = \theta_i + \varepsilon_{li}$$
$$Y_{li} = j \text{ if } \tau_i^{j-1} < Y_{li}^* \leq \tau_i^j,\ \ j = 1,\ldots 5$$
$$\varepsilon_{li} \sim N(0,\sigma^2), \text{ independent of each other, of } \varepsilon_{ri}, \text{ and of } X_i, V_i \text{ and } D_i$$

There are two reasons why we cannot simply include D_i in X_i. The first is that we have given it common coefficients in self-reports and vignette evaluations, since it only enters through the thresholds. The second is the endogeneity problem, i.e., the error terms in the multinomial logit part are correlated with the error term driving actual work limitations of the respondent. Thus we want to allow for correlation between ε_{ri} and $\eta_1,...\eta_J$. We do not allow for correlation between $\eta_1,...\eta_J$ and the errors in the vignette evaluations.

To build in the correlation between the normal error term ε_{ri} and the GEV type I error terms $\eta_1,...\eta_J$, we follow the approach of Lee (1983). He does not explicitly specify the joint distribution of ε_{ri} and $\eta_1,...\eta_J$, but gives the bivariate distributions of transformations of these errors that are needed in the likelihood. Lee shows that the probability of the multinomial logit outcome *j* (given Z_i) can be written in terms of one random variable which is a function of $\eta_1,...\eta_J$ and Z_i. This random variable is transformed to a standard normal random variable. Lee then assumes that the joint distribution of this transformed variable and ε_{ri} is bivariate normal, with correlation coefficient ρ_j. Applying the inverse transformation then gives the joint distribution that is needed to write down the likelihood. See appendix for more details.

4.4. Data for the Netherlands

In August 2003, we have collected work disability self-reports and vignette evaluations in the Dutch CentERpanel (see also Section 2.1), which allows researchers to include short modules of experimental questions. This feature has been used to collect our data on work disability. The Internet infrastructure makes the CentERpanel an extremely valuable tool to conduct experiments, with possibilities for randomization of content, wording, question and response order, and regular revisions of the design. Production lags are very short, with about one month between module design and data delivery. Based upon our first analysis, we have fielded a second wave in October with different wordings of the vignette questions. In this paper we use the self –reports on work disability collected in the first wave (August 2003; see Table 2.2) and we use vignette data from both waves (August and October 2003).

Table 4.1
Some Examples of Vignette Descriptions in CentERpanel

Affect vignettes:

1. [Jim] enjoys work very much. He feels that he is doing a very good job and is optimistic about the future.
2. [Tamara] has mood swings on the job. When she gets depressed, everything she does at work is an effort for her and she no longer enjoys her usual activities at work. These mood swings are not predictable and occur two or three times during a month.

Pain vignettes:

1. [Katie] occasionally feels back pain at work, but this has not happened for the last several months now. If she feels back pain, it typically lasts only for a few days.
2. [Mark] has pain in his back and legs, and the pain is present almost all the time. It gets worse while he is working. Although medication helps, he feels uncomfortable when moving around , holding and lifting things at work

Cardio-vascular disease vignettes:

1. [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. His job is not physically demanding, but sometimes it can be hectic. He does not get along with his boss very well.
2. [Norbert] has had heart problems in the past and he has been told to watch his cholesterol level. Sometimes if he feels stressed at work he feels pain in his chest and occasionally in his arms.

We have included vignettes of people with emotional problems, back pain, and cardio vascular disease. They are adjusted versions of some of the health vignettes that can be found on the vignette web page of Gary King (http://gking.harvard.edu/vign/). Some examples are presented in Table 4.1.

In each wave, each respondent answered five vignette questions for each type of problems. The questions were always of the format: "Does … have a health problem that limits the amount or type of work he/she can do?" with a five point response scale: not at all; yes, mildly limited; yes moderately limited; yes, severely limited; yes, extremely limited/cannot work. The vignettes were preceded by a self-report question with the same wording and response scale. There were about 2250 respondents. Item non-response was negligible.

Table 4.2 presents the frequency distribution for all fifteen vignettes including the six in Table 4.1 (indicated by the names of the persons) using the August survey. On average, the distributions make a lot of sense, although some noise remains. For example, the large majority agrees that nothing is wrong with Jim (the first vignette in Table 4.1) but about 3.5% indicate that he has some limitation, and five respondents (0.22%) report that he is extremely limited and cannot work. The error terms in the equations for vignette evaluations will pick up these types of outliers. In general, the vignettes cover a broad range from no or hardly any limitation to severe limitations.

Table 4.2.
Frequencies Vignette Answers (August wave)

Affect vignettes	Affect 1	Affect 2 (Jim)	Affect 3 (Tamara)	Affect 4	Affect 5
Not at all limited	32.42	96.51	7.56	12.61	1.37
Mildly limited	53.87	2.26	34.98	43.43	5.44
Moderately limited	11.68	0.66	39.98	31.31	15.30
Severely limited	1.86	0.35	15.97	11.90	42.64
Extremely limited	0.18	0.22	1.50	0.75	35.25
Total	100.00	100.00	100.00	100.00	100.00
Pain vignettes	Pain 1 (Katie)	Pain 2	Pain 3	Pain 4	Pain 5 (Mark)
Not at all limited	24.59	10.53	0.40	0.49	0.49
Mildly limited	62.76	53.56	6.63	7.30	11.90
Moderately limited	11.10	29.10	25.92	30.83	33.61
Severely limited	1.42	6.41	50.64	46.17	43.83
Extremely limited	0.13	0.40	16.41	15.21	10.17
Total	100.00	100.00	100.00	100.00	100.00
Cvd vignettes	Cvd 1	Cvd 2 (Norbert)	Cvd 3	Cvd 4 (Tom)	Cvd 5
Not at all limited	88.77	9.33	1.99	20.26	7.61
Mildly limited	9.77	48.74	18.89	43.08	36.53
Moderately limited	1.02	28.31	36.36	26.54	31.27
Severely limited	0.35	12.61	34.06	9.73	20.65
Extremely limited	0.09	1.02	8.71	0.40	3.94
Total	100.00	100.00	100.00	100.00	100.00

Table 4.3 shows how the respondents have ordered the vignettes within each of the three categories. For example, 91% of the respondents put Jim in a lower work disability category than Tamara, and almost 8% put the two in the same category. Only 1% put Jim in a worse category than Tamara. For other pairs of vignettes, the situation is more symmetric. For example, 35% put Tom in a worse (cvd) work disability category than Norbert, but 20% does the reverse. Apparently, the descriptions for Tom and Norbert leave some ambiguity on who of the two is more work disabled. In general, however,

respondents were rather consistent in the ordering of their vignette ratings, implying that vignette descriptions were distinctive enough.

Table 4.3
Order of Vignette Responses (August wave)

Affect		aff2	aff3	aff4	aff5
aff1	-1	1.24	67.80	54.80	96.20
	0	32.82	28.22	39.81	2.83
	1	65.94	3.98	5.40	0.97
aff2	-1		91.15	86.07	97.52
	0		7.74	12.74	1.81
	1		1.11	1.19	0.66
aff3	-1			19.20	85.71
	0			41.97	12.34
	1			38.83	1.95
aff4	-1				90.80
	0				7.83
	1				1.37

Pain		pain2	pain3	pain4	pain5
Pain1	-1	44.76	94.38	92.61	90.45
	0	46.66	4.73	6.46	8.80
	1	8.58	0.88	0.93	0.75
Pain2	-1		85.45	84.79	80.01
	0		13.36	13.49	17.60
	1		1.19	1.72	2.39
Pain3	-1			24.68	18.58
	0			44.32	42.86
	1			31.00	38.57
Pain4	-1				20.21
	0				46.09
	0				33.70

Cvd		Cvd2	Cvd3	Cvd4	Cvd5
Cvd1	-1	86.33	95.71	75.76	88.99
	0	12.69	3.67	23.40	10.08
	1	0.97	0.62	0.84	0.93
Cvd2	-1		63.87	20.39	38.74
	0		30.21	44.23	46.09
	1		5.93	35.38	15.17
Cvd3	-1			5.75	13.93
	0			23.40	33.44
	1			70.85	52.63
Cvd4	-1				0.84
	0				23.40
	1				
75.76					

-1: row<column; 0: row=column; 1: row>column

4.5. Estimation Results Benchmark Models for the Netherlands

We have separately estimated the models with vignettes for affect, pain and heart problems (cvd). We have combined the August and October waves and used ten vignettes on affect, ten on pain, and ten

on cvd. In all models, the scale is fixed by setting the standard deviation of the error in the work disability equation to 10 and location is fixed by setting the constant in the equation for the threshold between the first two categories (no limitation versus mild limitation) to 0. Likelihood ratio test statistics that compare the benchmark model with a restricted model in which thresholds do not vary with respondent characteristics all lead to rejecting the null, implying that allowing thresholds to vary with respondent characteristics is always a significant improvement.

Table 4.4 presents the complete results for the benchmark model with the pain vignettes, using a basic set of background characteristics: gender, educational dummies, and dummies for age categories. To illustrate the importance of the correction of the estimates of the self-reported work disability equation, we also present the ordered probit estimates using the same normalizations. Work disability falls with education level. This is borne out both by the ordered probit estimates and by the benchmark estimates in the first panel. The benchmark estimates, however, indicate somewhat smaller effects than the ordered probit estimates. The explanation is that the pain vignettes indicate that the higher educated use higher thresholds than the lower educated, i.e., tend to assign lower work disability to the same vignette person than the lower educated. This is also revealed by the estimates for the first threshold equation (γ^1); the other threshold parameters appear not to play a large role here.[8] Correcting the self-reports for this reduces the educational differences compared to the ordered probit estimates.

Work disability increases with age. The correction for response scale differences reduces the age difference between the youngest age group and the other age groups, because the youngest respondents tend to use higher thresholds than older respondents. This is similar to the finding of Salomon et al. (2004) for mobility (as a domain of general health, not work related) who explain it from expectations: older respondents may more often expect to have some work disability and adjust their scales accordingly.

Women more often report a work disability than men of the same age and education level. This difference is reduced somewhat if response scales are corrected for DIF, since women use somewhat lower thresholds. On the other hand, there is also a systematic difference between evaluating male and female vignette persons (the parameter on the dummy female in θ). For a given vignette description, a male vignette person is seen as more work disabled than a female vignette person, by both male and female respondents.[9] The coefficients on the vignette dummies are in line with the descriptive statistics in Table 4.2: the higher the coefficient, the worse the vignette person's work disability is considered, on average.

The estimated standard deviation of the vignette evaluations is much smaller than that of the self-reports. This is in line with the fact that everyone gets the same vignette descriptions (apart from the name of the person described, determining the gender). In the self-reports, heterogeneity in respondents' own work disability not explained by gender, education or age, leads to the much larger variance of the unsystematic part.

Table 4.5 presents the estimates of the work disability equation according to ordered probit and the three benchmark models using vignettes on emotional problems, pain problems, and heart problems. The estimates using pain vignettes are the same as those in Table 4.4. The emotional and cvd vignettes do not lead to the same corrections on the education coefficients, possibly due to a counterbalancing framing effect: since most vignettes do not specify the job of the person that is described, a respondent may think that these persons have a job like the respondent's own job. On average, the lower educated respondents will probably have jobs that give more limitations for people with back pain or a heart problem. This would imply that the lower educated use lower thresholds than the higher educated, i.e., thresholds would rise with education level. In the benchmark model, framing effects like this are not explicitly incorporated. They can be addressed using vignette descriptions that are more specific about the nature of the vignette person's job; we experimented with this in the October 2003 survey.

[8] A model in which all thresholds shift with respondent characteristics in a parallel manner is statistically rejected against the model presented here, but gives very similar corrections in the work disability equation.

[9] We included an interaction term of respondent gender and gender of the vignette person but this was insignificant.

Table 4.4
Benchmark Model with Pain Vignettes & Ordered Probit

	Self-reported work disability			
	Ordered probit		Benchmark model	
	est.	s.e.	est.	s.e.
constant	-0.806	1.261	-0.452	1.233
lower voc	-0.048	1.222	0.028	1.223
sen highh	-2.411	1.345+	-1.914	1.349#
voc commun	-0.124	1.287	-0.369	1.291
voc colleg	-3.742	1.275*	-3.490	1.276*
university	-5.909	1.489*	-5.099	1.501*
age 15-24	-10.346	2.028*	-9.856	2.214*
age 25-34	-6.668	1.062*	-7.170	1.059*
age 35-44	-5.422	0.994*	-5.735	0.978*
age 45-54	-1.557	0.924+	-1.347	0.902#
age 55-64	-0.829	0.972	-0.699	0.959
woman	2.403	0.583*	2.123	0.582*

	Benchmark model, threshold parameters							
	γ^1	s.e.	γ^2	s.e.	γ^3	s.e.	γ^4	s.e.
const th 1	0.000	0.000	2.023	0.054*	1.612	0.057*	1.785	0.058*
lower voc	0.149	0.282	0.034	0.048	-0.079	0.052#	0.050	0.052
sen highh	0.775	0.297*	-0.042	0.053	-0.018	0.057	0.107	0.056+
voc commun	-0.121	0.294	0.007	0.051	-0.001	0.055	0.045	0.054
voc colleg	0.408	0.291#	0.038	0.050	-0.036	0.053	0.103	0.052*
university	1.168	0.302*	-0.049	0.054	0.011	0.057	0.164	0.057*
age 15-24	0.938	0.308*	-0.127	0.051*	-0.091	0.067#	-0.163	0.063*
age 25-34	-0.227	0.179	-0.060	0.033+	-0.005	0.039	-0.105	0.036*
age 35-44	-0.260	0.185#	-0.024	0.032	-0.023	0.038	-0.061	0.033+
age 45-54	0.232	0.167#	-0.030	0.030	-0.046	0.036	-0.189	0.034*
age 55-64	0.209	0.181	-0.079	0.034*	0.046	0.038	-0.102	0.037*
woman	-0.390	0.112*	0.030	0.019#	0.010	0.022	0.026	0.021

	Benchmark model, Vignette equation	
	θ	s.e.
dummy vig1	3.134	0.304*
dummy vig2	6.369	0.331*
dummy vig3	14.302	0.460*
dummy vig4	14.039	0.455*
dummy vig5	12.953	0.433*
vign woman	-0.170	0.054*
sig vign	4.496	0.116*

* |t-value|>1.96; +: 1.64<|t-value|<1.96; #: 1.28<|t-value|<1.64

Correcting for age related DIF leads to smaller age effects if we use the vignettes on emotional health problems. A similar result is found if vignettes on heart problems are used, and in both cases, the corrections are larger than when using the vignettes on back pain. Again, framing effects (with the older people more often in jobs that can be done with back pain but not with emotional or heart problems) might explain the difference.

Affect vignettes do not indicate DIF related to gender, while both the back pain and cvd vignettes suggest that female respondents use lower thresholds than male respondents. Correcting for this somewhat reduces the difference in work disability between women and men, although it remains significantly positive.

For both affect and cvd vignettes, we again find that there is a systematic difference between evaluating male and female vignette persons (not presented in the table: -0.707 with t-value 9.6 for affect; -0.177 with t-value 2.5 for cvd). The effect is much larger for vignettes describing emotional problems than for pain and cvd vignettes. For a given vignette description, a male vignette person with any of the three types of problems is seen as more work disabled than a female vignette person.

Table 4.5
Ordered Probit & Benchmark Models with Affect, Pain and Cvd Vignettes:
Equation for Self-reported Work-Disability

	Ordered probit		Benchmark model Affect Vignettes		Pain Vignettes		Cvd Vignettes	
	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.
constant	-0.806	1.261	-0.787	1.270	-0.452	1.233	-1.076	1.307
lower voc	-0.048	1.222	-0.518	1.230*	0.028	1.223	-0.699	1.238*
sen highh	-2.411	1.345+	-3.147	1.365**	-1.914	1.349#*	-2.362	1.372+
voc commun	-0.124	1.287	-1.331	1.302*	-0.369	1.291	-0.674	1.300*
voc colleg	-3.742	1.275*	-5.009	1.288**	-3.490	1.276*#	-3.756	1.294*
university	-5.909	1.489*	-6.657	1.515**	-5.099	1.501**	-5.647	1.527*
age 15-24	-10.346	2.028*	-8.991	2.266**	-9.856	2.214**	-9.573	2.136**
age 25-34	-6.668	1.062*	-6.235	1.083**	-7.170	1.059*	-5.848	1.089**
age 35-44	-5.422	0.994*	-4.961	1.012**	-5.735	0.978*#	-4.596	1.024**
age 45-54	-1.557	0.924+	-1.339	0.928##	-1.347	0.902##	-0.565	0.954*
age 55-64	-0.829	0.972	-0.949	0.982#	-0.699	0.959	-0.201	1.005*
woman	2.403	0.583*	2.554	0.596*	2.123	0.582**	2.188	0.595**

* |t-value|>1.96; +: 1.64<|t-value|<1.96; #: 1.28<|t-value|<1.64
First symbol: refers to |t-value| of the parameter estimate itself (β);
Second symbol: refers to |t-value| of the corresponding entry of γ^1 which drives the correction compared to ordered probit.

In Table 4.6, health conditions are added to the explanatory variables. These are answers to questions of the form "has the doctor ever told you that …." , except for pain, which is self-reported ("do you often suffer from pain? "). The same variables were already used in Section 3. They are included as exogenous background variables; we assume that these health conditions do not suffer from reporting or other measurement error.

Different health conditions have very different effects on work disability, as in the binary probits in the previous section. This does not change much after correcting for response scale differences. We find that people with emotional problems have a significantly larger tendency to judge the vignette persons as work limited.[10] This effect is strongest for the vignettes on emotional problems. Correcting for this response scale effect reduces the coefficient of emotional problems on the respondent's own work disability by about 16%. Somewhat smaller corrections on the coefficient of an emotional onset are obtained using the other vignettes. A similar effect is found for respondents with diabetes. They tend to use lower response scales and more easily report that someone is not able to work. Correcting for this reduces the effect of diabetes on work limiting disability substantially. The largest correction is obtained

[10] That is, the coefficient of the emotional health condition dummy in γ^1 is significantly positive.

using the pain vignettes, reducing the estimated effect by about one third and rendering it insignificant at the two-sided 5% (or even 10%) level.

In section 3, we found that the effect of pain on reported work disability is much larger in the Netherlands than in the US. The ordered probit results in Table 4.6 confirm this result: pain has a much larger effect than all health conditions except having had a stroke. Correcting for response scale differences between people with and without pain hardly changes this – the correction makes the coefficient even larger if the affect vignettes are used and reduces it only slightly if pain or cvd vignettes are used. Thus DIF in the Netherlands cannot explain why the effect of pain on reported work disability is so much larger in the Netherlands than in the US.

Table 4.6
Ordered Probit & Benchmark Models with Affect, Pain and Cvd vignettes – Including Health Conditions. Equation for Self-reported Work-Disability

	Ordered probit		Benchmark model Affect Vignettes		Pain Vignettes		Cvd Vignettes	
	est.	s.e.	est.	s.e.	est.	s.e.	est.	s.e.
constant	-8.085	1.580*	-8.371	1.634*	-8.145	1.627*	-9.111	1.923*
lower voc	0.201	1.424	-0.472	1.476*	0.148	1.498	-0.773	1.551*
sen highh	-1.970	1.580	-2.862	1.639+*	-1.436	1.659*	-2.177	1.711
voc commun	0.635	1.501	-0.990	1.565*	0.127	1.575	-0.325	1.624*
voc colleg	-2.671	1.499+	-4.303	1.571**	-2.373	1.583#	-2.962	1.619+
university	-3.865	1.667*	-4.737	1.744**	-2.858	1.774#*	-3.482	1.811+
age 15-24	-5.842	2.134*	-3.982	2.468#*	-5.351	2.394*+	-5.095	2.385*
age 25-34	-3.136	1.172*	-2.372	1.240+*	-3.743	1.238**	-2.293	1.261++
age 35-44	-2.121	1.083+	-1.392	1.163*	-2.557	1.141**	-1.335	1.205#
age 45-54	0.167	1.012	0.512	1.068#	0.277	1.053	0.985	1.117*
age 55-64	-0.031	1.054	-0.115	1.119	0.050	1.107	0.401	1.162
woman	0.602	0.632	0.834	0.660	0.241	0.661*	0.529	0.676
stroke	13.389	2.290*	14.235	2.600*	13.532	2.454*	13.199	2.943*
cancer	2.516	1.289+	2.549	1.362+	2.723	1.373*	2.564	1.421+
lung	4.949	1.191*	5.212	1.221*	4.365	1.274**	4.872	1.300*
heart	6.359	1.228*	6.527	1.329*#	6.114	1.271*	5.863	1.347*
high blood	0.187	0.748	0.231	0.782	-0.215	0.768*	0.353	0.800
diabetes	3.940	1.571*	3.499	1.655**	2.637	1.686#*	2.842	1.666+*
emotional	6.933	0.838*	5.822	0.897**	6.113	0.860**	6.266	0.898*#
arthritis	4.367	1.028*	4.239	1.092*	4.451	1.058*	4.754	1.094*+
vision	2.534	1.434+	3.046	1.484*	2.645	1.497+	2.197	1.629#
oft pain	11.695	0.714*	11.898	0.760**	11.352	0.750*	11.298	0.790*

* |t-value|>1.96; +: 1.64<|t-value|<1.96; #: 1.28<|t-value|<1.64
First symbol: refers to |t-value| of the parameter estimate itself (β);

Second symbol: refers to |t-value| of the corresponding entry of γ^1 which drives most of the correction for DIF.

4.6. Estimation Results Models with Labor Market States in the Netherlands

Table 4.7 summarizes the results for the model allowing for justification bias. In the current version of the paper, this model uses the August wave only and excludes the pain variable which was measured in October. Other than that, the work disability has the same specification as in Table 4.6, including age, gender, education level, and health dummies. The thresholds are allowed to depend upon the same variables. In addition, the labor market state dummies are included in γ^1, implying that labor market state shifts all thresholds in a parallel manner.

We present the results for the work disability equation and for the first threshold, as well as the correlation coefficients ρ_j that capture the correlation between the error in the work disability equation and the transformed errors in the multinomial logit model explaining labor market state. A negative value here implies a positive relation between the error in the work disability equation and the probability of state *j*. For example, those with high work disability have a much higher probability to be on disability, which seems quite plausible. Workers and respondents in retirement tend to have lower unobserved work disability components than the others. This all makes sense.

Let us focus on the effect of labor market state on the thresholds. Consider the results on the basis of the pain vignettes. Here the disability coefficient is as expected: being on disability transfers reduces the thresholds, creating a higher tendency to report work disability. This result is driven by the fact that respondents on disability transfers more often than workers tend to evaluate the pain vignettes as work limited. This is in line with justification bias: respondents on disability use lower thresholds to justify their disability status.

However, the other vignettes lead to very different conclusions. Respondents on disability do not tend to give systematically different evaluations of vignettes with heart problems from workers. They even tend to use higher thresholds for vignettes with emotional conditions. This would suggest the opposite of justification bias.

The effects for the other groups are more stable across vignette types. Both homemakers and others (mainly students) tend to use higher thresholds than workers. Thus the groups who probably have the least experience in the labor market have a tendency to report that health problems are not work limiting.

4.7. International Comparison using Vignettes

The vignette questions discussed above were also fielded in the RAND MS Internet panel, an Internet survey for US respondents aged 40 and over. At this stage, 346 observations are available for a potentially selective sample of respondents with Internet connection. A control group will be interviewed by phone at a later stage.

Table 4.8 compares the vignette evaluations in the US to the Dutch evaluations which were already presented in Table 4.2. There are some substantial differences in the evaluations between the two countries. In particular, for the first two vignettes, which describe people with relatively mild work limitations, the US respondents much more often report that the vignette persons have no limitation at all, where the Dutch respondents have a larger tendency to use the intermediate categories “mildly” and “moderately.” The same tendency towards the extremes in the US and towards the middle for the Netherlands is seen in the fourth vignette, describing a person with relatively serious work limitations. The US respondents much more often evaluate this person as severely or extremely limited, where the Dutch still tend to use the answer “moderately.” Thus the general picture seems to be the same as for the self-reports in general health in Table 3.1, with the Dutch much less often choosing the “extreme” categories.

Table 4.7
Model with Labor Market States in Thresholds

	Affect Vignettes		Pain Vignettes		Cvd Vignettes	
	est.	s.e.	est.	s.e.	est.	s.e.
Work disability:						
constant	-7.834	1.405*	-7.945	1.444*	-7.801	1.436*
primary sch	6.625	1.528*	5.443	1.554*	4.812	1.551*
lower voc	5.421	1.120*	4.611	1.147*	3.602	1.146*
sen highh	2.139	1.264+	1.668	1.276#	1.595	1.296
voc commun	5.194	1.172*	4.961	1.202*	4.401	1.206*
voc colleg	1.553	1.151#	1.630	1.171#	1.285	1.189
age 15-24	-8.299	2.198*	-9.247	2.277*	-9.218	2.213*
age 25-34	-5.389	1.364*	-6.653	1.411*	-4.842	1.402*
age 35-44	-4.487	1.347*	-5.568	1.382*	-3.930	1.392*
age 45-54	-2.620	1.261*	-2.342	1.306+	-1.268	1.317
age 55-64	-2.497	1.245*	-2.457	1.275+	-1.364	1.278
d woman	2.621	0.612*	1.605	0.622*	1.606	0.620*
stroke	12.635	2.642*	12.834	2.133*	11.912	3.023*
cancer	0.453	1.313	0.605	1.361	0.990	1.408
lung	6.531	1.124*	5.728	1.119*	7.018	1.218*
heart pr	6.310	1.153*	6.534	1.188*	5.560	1.175*
high blood	-0.174	0.724	-0.363	0.712	0.135	0.744
diabetes	2.437	1.452+	1.096	1.411	1.174	1.603
emotional	6.129	0.833*	6.309	0.797*	6.365	0.824*
arthritis	8.449	0.916*	8.360	0.902*	9.138	0.922*
vision	5.284	1.310*	4.365	1.278*	4.025	1.332*
Reporting bias:						
primary sch	1.358	0.407*	0.181	0.421	-0.413	0.455
lower voc	0.588	0.295*	-0.025	0.310	-1.085	0.306*
sen highh	0.080	0.326	-0.275	0.324	-0.185	0.318
voc commun	-0.842	0.323*	-0.834	0.341*	-1.508	0.330*
voc colleg	-0.381	0.305	-0.340	0.304	-0.573	0.296+
age 15-24	1.180	0.459*	0.220	0.444	-0.324	0.532
age 25-34	0.484	0.363#	-0.845	0.338*	0.893	0.373*
age 35-44	0.804	0.345*	-0.522	0.345#	1.033	0.358*
age 45-54	-0.065	0.341	0.184	0.314	1.436	0.340*
age 55-64	-0.298	0.337	0.291	0.310	1.163	0.336*
d woman	0.657	0.174*	-0.022	0.186	-0.111	0.186
stroke	0.750	0.750	0.950	0.678#	-0.677	0.827
cancer	-0.610	0.525	-0.503	0.531	-0.544	0.599
lung	0.211	0.377	-0.733	0.437+	0.792	0.370*
heart pr	0.687	0.335*	0.806	0.339*	-0.354	0.314
high blood	-0.166	0.207	-0.346	0.228#	0.316	0.223#
diabetes	-0.277	0.427	-1.683	0.532*	-2.029	0.556*
emotional	-1.068	0.281*	-0.903	0.339*	-0.623	0.306*
arthritis	0.371	0.265#	-0.021	0.362	0.960	0.306*
vision	1.268	0.440*	-0.048	0.471	-0.465	0.477
Labor market states dummies ψ_j :						
working	0		0		0	
homemaker	1.554	0.171*	0.901	0.121*	0.577	0.155*
retired	-0.122	0.244	0.170	0.184	0.291	0.213#
disabilit	0.798	0.249*	-0.836	0.181*	0.013	0.246
other	0.775	0.185*	0.663	0.149*	1.020	0.199*
Correlations ρ_j						
working	0.338	0.074*	0.275	0.081*	0.310	0.078*
homemaker	-0.110	0.079#	-0.107	0.080#	-0.047	0.077
retired	0.404	0.109*	0.335	0.122*	0.353	0.120*
disabilit	-0.774	0.041*	-0.751	0.045*	-0.763	0.044*
other	-0.091	0.067#	-0.112	0.064+	-0.112	0.064+

* |t-value|>1.96; +: 1.64<|t-value|<1.96; #: 1.28<|t-value|<1.64

Table 4.8
Pain Vignette Evaluations in United States and Netherlands

	Pain 1		Pain 2		Pain 3		Pain 4		Pain 5	
Limited?	NL	US	NL	US	NL	US	NL	US	NL	US
Not at all	24.89	36.09	10.52	29.20	0.35	0	0.46	0.30	0.46	0.59
Mildly	63.28	50.30	53.46	48.08	6.22	7.44	7.28	2.66	11.94	9.17
Moderately	10.47	11.24	29.44	19.76	26.56	28.57	31.11	15.38	33.79	37.87
Severely	1.32	0.59	6.27	1.47	50.89	48.21	46.28	57.99	43.90	39.35
Extremely	0.05	1.78	0.30	1.47	15.98	15.77	14.87	23.67	9.91	13.02

Sources: Netherlands: CentERpanel, August 2003, 1977 observations; US: RAND MS Internet Panel, January 2004, 346 observations.

This also implies that the Dutch seem harder on the vignette persons with a serious limitation and softer on those with a minor limitation. For the self-reports, the latter is probably more important than the former, since the large majority of all respondents categorize themselves in one of the minor limitations categories. Thus Table 4.8 suggests that the Dutch would be harder on themselves if they would use the US scales. Using the US scales would thus reduce self-reported work disability prevalence, and would thus also reduce the difference in this prevalence between the two countries.

Model for International Comparison

To estimate the model comparing work disability in the US and the Netherlands, three data sets are combined: the Dutch CentERpanel (waves 1, 2 and 3, in August, October and December 2003), the US RAND MS Internet panel, and the US HRS wave 1. They all have different age selections (all age groups in CentERpanel; 40+ in RAND MS Internet Panel; 51-61 in HRS), but since we condition on age, this should not be a problem. CentERpanel and RAND MS have exactly the same vignette questions on pain problems, emotional problems, and cardio-vascular disease. HRS has no vignettes.

CentERpanel has self-reports on work limiting disability on a five-point scale (August 2003) and on a two-point scale (October 2003 for 50% of all the observations, December 2003 for the other 50%). Both US surveys have self-reports on the two-point scale only. In order to link the US (and NL) self-reports on the two-point scale to the US (and NL) vignette evaluations on a five point scale, we expand the model discussed above with a transformation from the five-point scale to the two-point scale. Table 2.2 suggested that the cut-off point between "yes" and "no" for the two-point scale is somewhere between the cut-off points between "no" and "mildly" and "mildly" and "moderately" for the five-point scale. In line with this, we model the cut-off point $\tau_i(2)$ on the two-point scale as a weighted mean of the two first cut-off points on the five-point scale:

$$\tau_i(2) = \lambda\tau_i^1 + (1-\lambda)\tau_i^2$$

We assume that the weight λ does not vary with individual characteristics and is the same in the US and the Netherlands. Thus the thresholds on the five-point scale and the thresholds on the two-point scale can have completely different structures in the two countries, but the relation between them is the same. If the Dutch have lower thresholds on the five-point scale, they also have a lower threshold on the two-point scale, etc. This assumption is needed as long as there are no five-point scale self-reports on the five-point scale for the US. Intuitively, the parameter λ can be identified from the Dutch self-reports on both scales, and can then also be implemented for the US respondents. In practice, all parameters are estimated simultaneously by Maximum Likelihood, taking into account that for the US respondents, the five-point scale self-report is an unobserved variable.

Table 2.6 also suggests that there is some random error in the two-point and/or five-point scale evaluations that is not transferred to the other scale. To account for this, we adjust the equation for the respondent's own work limiting disability as follows, partitioning the error term in a genuine unobserved

component of work disability affecting both the two-point and the five-point scale reports, and an idiosyncratic error term affecting only one report and independent of everything else. To be precise, the two-point scale and five-point scale self-reports are modelled as:

$$Y_{ri}^{*} = X_i\beta + \varepsilon_{ri};\ \ \varepsilon_{ri} \sim N(0,\sigma_r^2),\ \varepsilon_{ri} \text{ independent of } X_i,\ V_i$$

Five-point scale:

$$Y_{ri} = j \text{ if } \tau_i^{j-1} < Y_{ri}^{*} + u_i^5 \le \tau_i^j,\ \ j = 1,\ldots 5$$

Two-point scale:

$$Y_{ri} = 0 \text{ if } Y_{ri}^{*} + u_i^2 \le \tau_i\ (2);\ Y_{ri} = 1 \text{ if } {}^{*}_{ri} + u_i^2 > \tau_i\ (2)$$

$$u_i^2, \sim N(0,\sigma_{u^2}^2); u_i^5 \sim N(0,\sigma_{u^5}^5); u_i^2, u_i^5 \text{ independent of each other and of other error terms}$$

Estimation results of the complete model are presented below. The equations for work disability and for the thresholds all include a complete set of interactions with the country dummy for the Netherlands. Vignette evaluation equations and the auxiliary parameters introduced above concerning the transformation from the two-point to the five-point scale do not include such interactions. Panel A of Table 4.9 presents the results for the work disability equation in the complete model and in a model without any form of DIF, i.e., a model in which thresholds do not vary with country or with individual characteristics or health conditions. The latter model is clearly rejected against the former by a likelihood ratio test (the log likelihood's are -32022.55 for the complete model and -32242.02 for the model without DIF).

The main differences between the two models concern the effects of education level. Education level in the US is much more important in the complete model than in the model without DIF. In the Netherlands, the correlation between education level and work disability is much weaker, both before and after correcting for DIF. Age is insignificant once health conditions are controlled for directly. (The large coefficients on the youngest age group is somewhat misleading since this group is quite small in the US data.) The role of gender is smaller in the model which controls for DIF. Due to data limitations in the RAND MS Survey, only a limited set of health conditions could be included. Health conditions play similar roles in the two countries and the results are similar to those discussed in earlier models. Both before and after correcting for DIF, the only significant (at the 5% level) interaction with the country dummy is with the variable indicating whether the respondent suffers from pain. Pain is a much more important cause for work disability in the Netherlands than in the UK, as already concluded in previous sections. Correcting for DIF still increases the difference between the effects of pain in the two countries.

Table 4.9
Estimation Results US-NL Model

Panel A	Work disability			
	Model without DIF		Complete model	
	est.	s.e.	est.	s.e.
constant	-12.394	1.807*	-12.815	2.240*
ed_med	-2.501	0.349*	-3.835	0.761*
ed_high	-4.871	0.513*	-6.240	0.986*
age 15-44	-11.418	7.766#	-9.942	12.049
age 45-54	-0.605	1.724	1.666	2.417
age 55-64	1.246	1.710	2.353	2.435
woman	-1.467	0.321*	-0.837	0.641#
high blood	2.641	0.329*	2.779	0.732*
diabetes	4.124	0.465*	2.719	0.992*
cancer	3.583	0.604*	3.057	1.053*
lung	6.408	0.543*	7.598	1.255*
heart	7.633	0.464*	9.206	1.243*
emotional	5.979	0.467*	5.231	1.061*
oft pain	11.653	0.451*	10.786	0.731*
Interactions With dummy NL:				
Constant	1.033	2.054	-0.627	2.719
ed_med	2.086	0.884*	3.411	1.135*
ed_high	1.949	0.980*	3.924	1.310*
age 15-44	9.034	7.834	6.699	12.096
age 45-54	1.597	2.027	-0.871	2.656
age 55-64	-0.331	2.043	-1.562	2.701
woman	2.420	0.758*	1.437	0.959#
high blood	-1.680	0.879+	-2.167	1.118+
diabetes	1.486	1.614	1.528	1.963
cancer	-1.072	1.525	-0.118	1.814
lung	0.416	1.357	-1.485	1.844
heart	1.154	1.288	-0.290	1.756
emotional	2.017	1.030+	1.850	1.424#
oft pain	3.846	0.862*	4.717	1.065*

Normalization: $\sigma_r^2 = 1$

Panel B presents the estimates of the threshold parameters. Most parameters are individually insignificant, but a likelihood ratio test indicates they are jointly significant. The estimates for the first threshold show that DIF related to education level is rather different in the two countries, in line with the corrections in panel A. In the US, the higher educated are harder on the people with pain vignettes, but this is not the case in the Netherlands. In the US, respondents with pain are harder on the back pain vignettes than respondents without pain, although the difference is significant at the two-sided 20% level only. Combined with the interaction of this dummy and the country dummy indicates that in the Netherlands, respondents with and without back pain give the same evaluations (ceteris paribus).

Table 4.9
Estimation Results US-NL Model, continued

Panel B	Threshold Parameters							
	γ^1	s.e.	γ^2	s.e.	γ^3	s.e.	γ^4	s.e.
constant	0.000	0.000	2.152	0.200*	1.828	0.204*	1.999	0.157*
ed_med	-1.376	0.820+	0.029	0.126	0.013	0.131	0.020	0.109
ed_high	-1.483	1.044#	0.074	0.160	-0.031	0.157	-0.056	0.127
age 15-44	1.518	1.378	0.014	0.187	0.001	0.216	-0.220	0.188
age 45-54	2.453	1.126*	-0.107	0.155	0.015	0.181	0.070	0.131
age 55-64	1.229	1.112	-0.073	0.153	0.103	0.185	0.114	0.119
woman	0.682	0.666	-0.029	0.098	-0.117	0.100	0.063	0.086
high blood	0.275	0.749	-0.087	0.111	0.159	0.128	-0.010	0.103
diabetes	-1.386	1.047#	-0.017	0.166	0.198	0.156	-0.096	0.145
cancer	-0.833	1.043	0.167	0.142	-0.113	0.185	-0.046	0.147
lung	1.750	1.252#	-0.403	0.213+	0.096	0.245	-0.158	0.173
heart	1.568	1.412	-0.008	0.238	-0.381	0.277#	0.151	0.191
emotional	-0.874	1.037	0.066	0.180	-0.139	0.227	0.171	0.122#
oft pain	-1.067	0.675#	0.103	0.105	-0.004	0.115	0.056	0.094
Interactions with dummy NL								
c th 1 NL	-2.185	1.289+	0.241	0.198	0.098	0.203	0.221	0.155#
ed_med	1.463	0.843+	-0.067	0.129	0.045	0.134	0.004	0.112
ed_high	2.128	1.063*	-0.094	0.162	0.073	0.159	0.126	0.130
age 15-44	-2.240	1.404#	-0.038	0.189	-0.032	0.219	0.117	0.191
age 45-54	-2.492	1.158*	0.101	0.158	-0.070	0.185	-0.267	0.136*
age 55-64	-1.129	1.146	0.006	0.157	-0.061	0.189	-0.221	0.125+
woman	-1.129	0.687+	0.059	0.100	0.128	0.103	-0.046	0.089
high blood	-0.886	0.779	0.155	0.114#	-0.136	0.131	0.005	0.107
diabetes	-0.517	1.146	0.110	0.173	-0.178	0.169	0.122	0.157
cancer	1.331	1.120	-0.227	0.152#	0.170	0.196	0.148	0.156
lung	-2.648	1.320*	0.427	0.219+	-0.046	0.251	0.172	0.179
heart	-1.317	1.450	-0.011	0.241	0.338	0.282	-0.231	0.196
emotional	-0.209	1.072	-0.057	0.183	0.164	0.231	-0.126	0.126
oft pain	1.138	0.702#	-0.116	0.107	-0.044	0.118	-0.113	0.097

Panel C has the estimates for the vignette equations. These results are similar to those in Table 4.4, which were based upon the Dutch vignette evaluations only. For example, we again find that female persons in the vignette descriptions are evaluated as less work disabled than men with the same vignette description.

Finally, panel D presents the auxiliary parameters related to the transformation between the two-point and the five-point scale. The cut-off point for the two-point scale is a weighted mean of the first and second threshold in the five-point scale, with an estimated weight for the first threshold of 0.79. Both idiosyncratic errors in the vignette reports play a role, and are of similar order of magnitude as the unobserved heterogeneity term in “true” latent work disability, which is common in both reports.

Table 4.9
Estimation Results US-NL Model, continued

Panel C	Vignette equation		
		θ	s.e.
dummy vig1		1.459	1.264
dummy vig2		5.883	1.282*
dummy vig3		17.514	1.460*
dummy vig4		17.349	1.455*
dummy vig5		15.616	1.424*
v woman		-0.265	0.078*
sig vign		6.471	0.271*

Panel D	Two-point and Five-point scales		
		Coeff.	s.e.
λ		0.789	0.046*
σ_{u^2}		4.310	0.767*
σ_{u^5}		7.218	0.532*

Table 4.10 compares predictions of work disability on the two-point scale of the models with and without DIF (the same two models presented in the first panel of Table 4.9). The model without DIF work disability rates of 34.8% in the Netherlands and 20.7% for the US, close to the observed work disability rates on the two-point scale for this age group. For the model with DIF, the estimated thresholds for the US are used. For the US sample, this again closely reproduces the observed work disability rate. This is due to the way the prediction is computed: there is no correction for within US DIF, only for cross-country DIF. For the Netherlands, however, the result is quite different. For every Dutch respondent, the work disability probability is computed as if this respondent would use the threshold of a US respondent with the same characteristics (age, education level, gender, health conditions). The results show that, if the Dutch would use the American thresholds, the self-reported work disability rate in the Netherlands would be reduced by about 7.6 percentage points to 27.3%. Thus correcting for cross-country DIF reduces the gap between the US and the Netherlands from 14.1 percentage points to 6.6 percentage points, a reduction of about 54%.

The other rows in Table 4.10 predict how much each health condition contributes to explaining work disability according to both models, again using US response scales for the model with DIF. Work disability is recomputed after setting the dummy for the given health condition equal to zero, and the reduction in work disability compared to the first row is reported. The differences between the two models are small. Pain remains the dominating factor in both countries, and is much more important in the Netherlands than in the US. Thus we find that there is a considerable difference in response scales between Dutch and US respondents explaining a large part of the observed difference in the work disability rate, but the difference is not related to whether respondents suffer from a health condition or not. All health conditions together explain most of reported work disability according to both models. They explain more in the Netherlands than in the US, again due to the effect of pain.

Table 4.10
Predicted Work Disability and Health Conditions

	Model without DIF		Model with DIF	
	NL	US	NL	US
total work disability	34.79	20.73	27.34	20.73
work disability explained by				
hypertension	0.61	2.06	0.35	2.15
diabetes	0.73	0.94	0.51	0.63
cancers	0.28	0.44	0.30	0.38
lung diseases	0.99	1.13	0.98	1.32
heart diseases	1.97	2.35	1.93	2.77
emotional diseases	2.70	1.74	2.34	1.53
pain	15.21	7.70	14.66	7.24
all health conditions	22.49	16.35	21.08	16.02

Age group 45-64, CentERpanel and HRS; Weighted using respondent weights. First row: total work disability. Other rows: Reduction in total work disability if dummy for given health condition (or dummies for all health conditions) is always zero. In model with DIF, work disability is predicted using US response scales.

Table 4.11 gives the prevalence rates of the health conditions in the age group 45-64 and the average marginal effect of each health condition on the probability of work disability. As in Table 4.10, the estimated US response scales are used for both the Dutch and the American respondents. Table 4.11 decomposes the contributions to work disability in Table 4.10 in two components: prevalence and the marginal effect. There are some differences between the models that do and do not correct for DIF across countries, but the qualitative conclusions remain the same. Pain has both the largest prevalence rate and the largest marginal effect in both countries, explaining why it has by far the strongest contribution on work disability. In the Netherlands, both prevalence and marginal effect are substantially larger than in the US, explaining why the contribution of pain to explaining work disability is larger in the Netherlands than in the US.

Table 4.11
Prevalence and Marginal Effects

	Prevalence (in %)		Average marginal effect (%-points)			
			Model without DIF		Model with DIF	
	NL	US	NL	US	NL	US
hypertension	25.38	36.04	2.42	5.71	1.40	5.97
diabetes	4.64	9.16	15.69	10.29	11.01	6.91
cancer	4.53	5.25	6.21	8.29	6.71	7.14
lung disease	6.35	6.84	15.51	16.56	15.44	19.37
heart disease	8.42	11.69	23.40	20.07	22.95	23.67
emotional dis.	12.81	11.14	21.09	15.64	18.28	13.75
pain	32.09	24.07	47.40	32.00	45.67	30.09

Age group 45-64, CentERpanel and HRS; Weighted using respondent weights. Prevalence: fraction of the sample with the given health condition. Average marginal effect taken over all observations with given health condition.

5. Future Research Agenda

This paper is a first installment on a long run project dealing with international differences in work disability. In this final section, we sketch out some components of the project to which we will turn in the near future.

There are two important extensions that we anticipate in the UK. First, in addition to offering another national setting, the use of the ELSA panel in the U.K. expands our work by allowing us to compare objective performance tests with the more standard subjective types of disability questions that are asked. This will be done in a large data set (over 10,000 people) who all lie in the relevant age range for this project.

Second, as an alternative identification strategy we will also analyze the UK benefit reforms of 1995, where eligibility rules changed according to gender and age, and where, in addition there was a reduction in generosity and a tightening of the stringency of the work test. The 'grandfathering' of the reforms in the UK meant that those (men) arriving at age 65 before 1995 could stay on benefits to age 70, while those reaching age 65 in or after 1995 were forced off after age 65. By following all these groups over the reform period we will be able to look at the effect of the eligibility criterion independently of the potential effects of financial incentives and the work test reform, which were affecting all groups.

Despite these reforms, differences across gender still exist within the disability system in the UK. These differences will also be used to look at potential effects of the disability system on health reporting behavior (e.g. whether you are work disabled or not), by comparing men and women conditional on detailed health conditions and health measurements. The 2001 reform primarily affected the generosity of benefits, particularly for those on higher incomes, by introducing means testing. The ongoing data collection of the BHPS and LFS datasets allows us to use this reform in addition, in order to understand the potential effects of changes in financial incentives associated with disability benefits.

We will also be conducting many additional experiments with our internet panels. The most important extension is to field a parallel set of internet interviews and vignette experiments on a US sample. This will be done with the RAND MS Internet panel, which mimics the CentERpanel in the US. The Internet sample has been recruited from respondents of age 40 and older to the Monthly Survey (MS) of the University of Michigan's Survey Research Center (SRC).[11] Those who agree to participate are added to the panel of households to be interviewed regularly over the Internet. In addition MS-respondents without Internet access are recruited into a control group, which is interviewed over the phone. The MS Internet panel comprises 1000 households and the control group comprises 500 households. A first experimental module with self-reports and vignettes on work limiting disabilities in the MS panel will be fielded in early 2004.

Another goal of this project is to extend the analysis of work-limiting disability, reporting bias, and labor market status to a dynamic framework. Most existing studies investigating the relation between health and labor market status use cross-section models. But the data we have presented in this paper indicates that there is considerable short run fluctuations in thresholds over time. In light of this, we will collect several waves of vignette data and link them to existing panels in the Netherlands and the US. Panel data models for labor market state, work-limiting disability, and reporting bias will be analyzed to investigate to which extent observed transitions in reported work-limiting disabilities reflect genuine changes and to investigate the dynamic relation between work-related disability and entry into disability or other labor market transitions.

[11] The MS is the leading consumer sentiments survey that incorporates the long-standing Survey of Consumer Attitudes (SCA) and produces the widely used Index of Consumer Expectations. SRC screens MS respondents. It asks MS-respondents age 40 or older if they have Internet access and, if yes, whether they would be willing to participate in Internet surveys.

Appendix to Section 3

Table A1
Dutch Probit for Work Disability—All ages
(using 5-point scale)

Variables	Coef.	DF/DX	Robust SE	Means
hypertension	.105	.038	.029	.191
diabetes	.512	.196	.063	.041
cancer	.222	.083	.066	.037
disease of lung	.661	.255	.056	.050
heart problem	1.025	.392	.048	.064
stroke	1.223	.456	.101	.011
arthritis	1.124	.426	.039	.089
emotion	.728	.279	.038	.101
female	.180	.064	.022	.465
age 35-44	.025	.009	.035	.229
age 45-54	.272	.100	.036	.233
age 54-64	.257	.095	.040	.174
age 65+	.399	.149	.043	.162
ed med	-.141	-.050	.027	.326
ed high	-.396	-.137	.025	.362
constant	-.886		.099	
observed p		.329		

Table 3.2B
Ages 45-64

Variables	Coef.	DF/DX	Robust SE	Means
hypertension	.051	.019	.039	.242
diabetes	.692	.271	081	.048
cancer	.305	.120	.118	.045
disease of lung	.702	.274	.115	.048
heart problem	.946	.361	.081	.064
stroke	1.579	.519	.110	.014
arthritis	1.124	.298	.071	.118
emotion	.728	.360	.064	.118
pain	1.481	.541	.043	.274
female	.056	.022	.043	.442
ed med	-.110	-.042	.049	.284
ed high	-.480	.178	.047	.333
constant	-.901		.125	
observed p		.389		

CentERpanel

Appendix to Section 4: Details on Applying Lee (1983)

To be more precise on the model incorporating the choice of labor market state, let $\eta^*_{ji} = Max_{m \neq j} s_{mi} - \eta_{ji}$. Then $S_i = j$ if and only if $s_{ji} = Z_i'\pi_j + \eta_{ji} \geq Max_{m \neq j} s_{mi}$, i.e., if $\eta^*_{ji} \leq Z_i'\pi_j$. We know that the probability that $S_i = j$ is given by the multinomial logit probability given above, so we get:

$$P[\eta^*_{ji} \leq Z_i'\pi_j \mid Z_i] = \exp(Z_i'\pi_j) / \sum_m \exp(Z_i'\pi_m)$$

Since this holds for any value of $Z_i'\pi_j$, it implies that the distribution function of η^*_{ji} given Z_i is given by:

$$F_j(t \mid Z_i) = P[\eta^*_{ji} \leq t \mid Z_i] = \exp(t) / [t + \sum_{m \neq j} \exp(Z_i'\pi_m)]$$

The Lee approach is to specify the joint distribution of η^*_{ji} and ε_{ri} by taking the known marginals, transforming them to standard normals, and specifying a bivariate normal with arbitrary correlation coefficient (to be estimated) for the joint distribution of the two transformed variables. The transformation needed to obtain a standard normal from ε_{ri} is simply dividing by σ_r. The transformation needed for η^*_{ji} is $\Phi^{-1} \circ F_j(. \mid Z_i)$. Thus we assume:

$$\varepsilon_{ri} / \sigma_r, \Phi^{-1}(F_j(\eta_{ji} \mid Z_i) \mid Z_i \sim N_2(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_j \\ \rho_j & 1 \end{pmatrix})$$

In computing the likelihood contributions, we will need the probability $P(Y_{ri} = r, S_i = j \mid Z_i)$. Using the specification given above, this probability can be rewritten as a bivariate normal probability, and can be evaluated relatively easily. The actual likelihood contribution can then be determined by integrating out the product of this probability with the univariate vignette probabilities—conditional on ε_{ri} over all possible values of ε_{ri}.

References

Autor, D. and M. Duggan. 2003. The Rise in the Disability Rolls and the Decline in Unemployment, *Quarterly Journal of Economics* 118(1), 157-206.

Bound, J. 1991. Self-reported versus Objective Measures of Health in Retirement Models, *Journal of Human Resources* 26(1), 106-138.

Bound, J. and R. Burkhauser. 1999. Economic Analysis of Transfer Programs Targeted on People with Disabilities, *Handbook of Labor Economics, Vol. 3C*, O. Ashenfelter and D. Card (eds.), 3417-3528.

Burkhauser, R., M. Daly, A. Houtenville, and N. Nargis. 2002. Self-Reported Work Limitation Data—What They Can and Cannot Tell Us, *Demography* 39(3), 541-555.

Currie, J. and B. Madrian. 1999. Health, Health Insurance and the Labor Market, *Handbook of Labor Economics, Vol. 3C*, O. Ashenfelter and D. Card (eds.), 3309-3416.

Eurostat. 2001. *Disability and Social Participation in Europe*, Luxembourg: Office for Official Publications of the European Communities.

Kerkhofs, M. and M. Lindeboom. 1995. Subjective Health Measures and State Department Reporting Errors, *Health Economics* 4, 221-235.

King, G., C. Murray, J. Salomon, and A. Tandon. 2004. Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research, *American Political Science Review* 98(1), 567-583.

Kreider, B. 1999. Latent Work Disability and Reporting Bias, *Journal of Human Resource* 34, 734-769.

Lee, Lung-fei. 1983. Generalized Econometric Models with Selectivity, *Econometrica* 51(2), 507-512.

Lindeboom, M. and M. Kerkhofs. 2002. Health and Work of the Elderly, IZA Discussion Paper 457, Institute for the Future of Labor, Bonn.

Salomon, J., A. Tandon, and C. Murray. 2004. Comparability of Self rated Health: Cross Sectional Multi-country Survey Using Anchoring Vignettes, *British Medical Journal* 328 (7434), 258-260.