Published in final edited form as:
J R Stat Soc Ser A Stat Soc. 2011 July ; 174(3): 575–595. doi:10.1111/j.1467-985X.2011.00694.x.

Validating the Use of Anchoring Vignettes for the Correction of Response Scale Differences in Subjctive Questions[1]

Arthur Van Soest[a], **Liam Delaney**[b], **Colm Harmon**[b], **Arie Kapteyn**[c], and **James P. Smith**[c]
Arthur Van Soest: avas@uvt.nl

[a]Tilburg University and RAND Corporation, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, Netherlands [b]Geary Institute University College, Dublin [c]RAND Corporation

Abstract

Comparing self-assessed indicators of subjective outcomes such as health, work disability, political efficacy, job satisfaction, etc. across countries or socio-economic groups is often hampered by the fact that different groups use different response scales. This paper develops an integrated framework in which objective measurements are used to validate vignette-based corrections. The framework is applied to objective and subjective self-assessments of drinking behavior by students in Ireland. Model comparisons using the Akaike information criterion favor a specification with response consistency and vignette corrected response scales. Put differently, vignette based corrections appear quite effective in bringing objective and subjective measures closer together.

Keywords

anchoring vignettes; reporting bias; hopit model

1. Introduction

In many important substantive areas, the most widely used data available to analyze individuals' behaviors and attitudes are inherently qualitative and subjective. In such data, people are typically asked to rank themselves on a subjective scale. One common example is people's ratings of their health on the traditional five-point scale from excellent to poor. Such subjective scales are pervasive in the health field and would include, in addition to general health status, measures of the ability to function in daily activities (i.e. Do you have any difficulty doing x?), work disability (Do you have a health problem that limits the kind or amount of work you can do?), and psycho-social measures (Do you feel that things in your life are beyond your control?). The widespread use of subjective scales is not limited to health. The placement of poverty thresholds, attitudes toward inequality and the effectiveness of political and governmental institutions would be just some other salient examples (cf. King et al, 2004).

These subjective scales all involve individuals' evaluation of some domain of their own objective reality (such as their true health) compared to their own subjective view of what it means to be above or below a given threshold (such as excellent, very good, etc). How

[1]This research was funded by NIA.

[2]Corresponding author: Arie Kapteyn, RAND Corporation, 1776 Main Street, Santa Monica, CA 90407; Phone (310) 393-0411, Fax (310) 451-7084.

someone situates oneself within these scales clearly depends both on the objective reality of one's situation and on one's unique subjective threshold. Since both the objective reality and the subjective thresholds can vary across individuals, it is not possible, using answers to the subjective scale questions alone, to know how much of the eventual rating of individuals on these scales reflects true objective differences among people and how much reflects variation across people in their subjective thresholds (e.g. Sen 2002, Lindeboom and van Doorslaer 2004, Christensen et al 2006).

One important new research tool that has been advanced to deal with this problem involves the use of anchoring vignettes. Respondents are first asked to evaluate their position on a scale in a given domain. Vignettes are essentially short descriptions of the positions of hypothetical persons in the same domain. Respondents are then asked to evaluate the vignette on the same scale they used to rate their own position. Because the objective situation of the person described in the vignette is the same for all respondents, anchoring vignettes have the potential to identify individual variation in subjective thresholds. The critical assumption on which identification rests is called 'response consistency'- that is respondents used the same subjective thresholds in rating the vignette persons as they use when rating themselves.

Research using anchoring vignettes has grown rapidly in recent years. Vignette questions have been applied in work on international comparisons of health (King et al, 2004, Salomon et al, 2004, and d'Uva et al, 2006), political efficacy (King et al, 2004), work disability (Kapteyn et al, 2007), job satisfaction (Kristensen and Johansson, 2006), and life satisfaction (Christensen et al, 2006). In all these applications, subjective scales were used and significant differences emerged across groups or countries in the subjective outcomes measured. Anchoring vignettes were employed to assess whether these groups also differed in their subjective thresholds.

Despite the rapidly growing use of anchoring vignettes, there has been little attempt to test the basic identifying assumption of response consistency. Anchoring vignettes will often change the adjusted distribution of responses on the subjective scale, and sometimes change them by a large amount. But how do we know that the vignette-adjusted scales are any better than the unadjusted scales? The best way to do so is having objective data to which the unadjusted and vignette adjusted distributions of qualitative responses can be compared and then test response consistency directly.

The only example we are aware of is the comparison of visual acuity between Chinese and Slovakian respondents reported by King et al (2004), who were asked self-reports of vision and shown a number of vignettes. In addition, a randomly chosen half of the respondents were administered the Snellen eye chart test. While self-reports of visual acuity show no appreciable difference between Slovakian and Chinese respondents, the eye chart test suggests that the vision of the Chinese respondents is considerably worse. Once the self-reports are corrected by using vignettes they concur with the eye chart tests in that now the Chinese respondents are shown to have considerably worse eyesight.

In this paper, we provide a more formal analysis by combining objective and subjective measures, and vignettes in a survey that we designed and conducted of drinking behavior among students at a major university in Ireland. Problem drinking of adolescents, and in particular college and university students, is a significant public health problem in many countries (see, e.g., Ham and Hope, 2003, for the US and Gill, 2002, for the UK) and Ireland is among the countries where alcohol abuse by adolescents is most prevalent (Mayor, 2001, and Ramstedt and Hope, 2005). It is well documented in the literature that excessive drinking is associated with other negative behavior (Ham and Hope, 2003; Gibson et al,

2004; Wechsler and Wuethrich, 2002) and with major causes of adolescent mortality (Schmid et al, 2003).

The specific example examined involves a subjective assessment of self-rated drinking problems. An advantage of this application is that the actual construct the question is trying to elicit is readily accessible by a simple objective behavioral measure i.e. by asking the respondent how much he or she drinks. The students were also given a set of drinking vignettes and asked to evaluate the drinking behavior of students in the vignettes. Drinking is a useful case study given the potential for social desirability and self-serving biases when rating one's own behavior and thus provides a challenging test of the vignette methodology.

The remainder of this paper is organized into four sections. The next section briefly describes the data used, including the types of vignettes we will use. Section 3 first outlines the intuition behind the use of vignettes and then presents a formal statistical model that we will we use to determine whether the critical assumption of 'response consistency' is rejected by the data. The 4th section summarizes the empirical estimates of this model and the final section highlights our main conclusions.

2. Survey data and the drinking vignettes

The sample for this study was recruited from a web-based survey of students attending a large Irish university, University College Dublin. In total, 4450 students started the web-based survey, from March to May 2006 of which 3,500 completed the survey. The mean age of respondents is 21.5 years, and 90 per cent of the sample is below age 25. The gender breakdown of the sample is 45 per cent male and 55 per cent female. The sample of 4,500 students represents approximately 20 per cent of the total body of 20,000 students and 50% of those who use the college email system.

How representative this sample is of the total student body is not critical for this application. Instead, selection bias would arise with this Internet panel if conditional on observables, respondents provide different vignette evaluations. Kapteyn et al. (2007) were able to test this assumption in a Dutch Internet sample where all respondents without Internet access were given a free set-top box. For this sample, it was known whether or not respondents had Internet access before they joined the panel. Kapteyn et al (2007) re-estimated a model for vignette evaluations with dummies for whether a respondent had prior access or not. These dummies were insignificant suggesting that at least for that application prior Internet access was not selective on these responses to vignette questions.

Web-based surveys may heighten fears of data privacy. Therefore, the confidence of potential respondents must be gained by assuring them about survey-confidentiality. As well as adopting strict controls on data-protection, we included an explicit assurance of survey confidentiality in our web-based questionnaire. Furthermore, all respondents were given an anonymous password that they could use to re-enter the survey at any time.

Respondents were asked several demographic, personal and family background questions. These include age; nationality; accommodation during term; relationship status; year of study; the number, age, drinking and smoking behavior of siblings; parental variables including maternal and paternal education, marital status, drinking and smoking, occupational status and gross income; individual financial information including average monthly income, income sources and average monthly expenditure.

All respondents were first asked the following basic question in relation to their drinking; "When did you last have a drink (that is more than just a few sips)?" and given five response options; "I have never had a drink"; "Not in the past year"; "More than 30 days ago but less

than a year ago"; "More than a week ago but less than 30 days ago"; "Within the last week". Of a total of 4,058 people who answered this question, 6.7 per cent were abstainers, 6.3 per cent claimed to have consumed alcohol more than 30 days but less than a year ago; 22.5 per cent consumed alcohol more than a week but less than 30 days ago, and 64.5 per cent consumed alcohol within the last week.

The 93% of the sample who did consume some alcohol during the last year were eligible to be asked the specific questions on their alcohol consumption, subjective assessments of their own drinking problems, and the vignette drinking questions.

Respondents were asked two types of questions about their own drinking behavior. The first objective variant asked them to quantify the actual amount they drank. Given that they drank at all, they were asked two subjective measures of the extent of their drinking- frequency of consumption and volume of consumption per occasion. In terms of frequency, 12 per cent of respondents drink "less than once a month"; 25 per cent drink "less than once a week"; 30 per cent of respondents drink "once a week"; 33 per cent of respondents drink "more than once a week"; and 0.66 per cent of respondents drinks daily.

The second objective measure concerns the volume of drinking per occasion.

> "How many drinks containing alcohol do you have on a typical day when you are drinking?"

with the permissible answers being less than 1, 1–2, 3–4, 5–6, 7–9, or 10 or more. In terms of volume consumed; 2% drank less than one drink; 10% drank "1 or 2"; 25% drank "3 or 4"; 32% drank "5 or 6"; 22% drank "7–9" and 9% drank "10 or more" drinks. Before this question was asked, a random half of the students were shown a screen informing them that a drink is ten grams of alcohol and were also given examples of types of drinks with a translation into grams. For example, a half pint of beer would be 9.8 grams and a pint would be 19.5 grams. As we demonstrate below, there were no statistically significant differences between the sample of students given this information and those not given this information in terms of their description of their subjective and objective drinking behavior as well as their description of the people in the drinking vignettes.

Student respondents were also asked to rate their own drinking on an ordered qualitative scale using the question:

(2) How would you describe your own drinking patterns over the course of the last year?" Mild, Moderate, Some Cause for Concern, Excessive, Extreme

26.9 per cent describe their drinking as mild; 43.9 per cent describe their drinking as moderate; 18.5 per cent describe their drinking as some cause for concern. 9.6 per cent describe their drinking as excessive; 1.5 per cent describe their drinking as extreme.

Finally, vignette questions were asked about the drinking behavior of hypothetical peers. The use of the web-surveying format allows for a complete experimental design to test the importance of various dimensions. In particular, we randomly assigned levels of severity according to frequency of drink, and the male or female names in the vignettes. The vignette drinking questions are of the form

(3) [John/Mary] is out on a given night and has [1 or 2, 3 or 4, 5 or 6, 6 or 7, 10 or more] drinks containing alcohol. Is [John/Mary]'s drinking habit-Mild, Moderate, Some Cause for Concern, Excessive, Extreme

Vignettes in (3) clearly use the same scale as in (2) for respondents' own drinking.

Table 1 lists the distribution of responses to drinking vignettes. The responses are stratified by the number of drinks mentioned in the drinking vignettes. We have sometimes combined groups to improve comparability with the categories of self-reported quantities of drinks. Not surprisingly the percent of students who thought the drinking behavior described in the vignette was either excessive or extreme rises rapidly with the number of drinks. For example, only 0.1 percent of students thought that 2–3 drinks merited the description of excessive or extreme while percent saying it was excessive or extreme for the other drinking amounts in the vignettes were 4.2% (5–6 drinks), 38.7% (7–10 drinks), and 70.3% (10 or more).

Responses to the drinking vignette questions are also presented in the right hand panel of Table 1 separately by the amount of own drinking behavior of the students. In each case we show the distribution of response categories for the person described in the vignette alongside the distribution of response categories for the students' own drinking. Since in each one of these situations the amount of drinking is approximately the same in the vignette and by the student evaluating the vignette, response consistency would imply a similar distribution of responses whether the vignette or student respondent is being described. No response consistency at all would imply that the evaluation of the drinking behavior of the vignette person would be independent of the drinking behavior of the student.

The data in Table 1 appear to strongly support response consistency. Consider for example, students who had 7–9 drinks. 19.6% of these students describe their own drinking as either excessive or extreme while 19.2% of them describe the drinking behavior of the vignette person (who has 7–10 drinks) as excessive or extreme. For this case, the qualitative subjective evaluation of own drinking problems and that of the vignette person are basically identical. Moreover, both these distributions are very different from the distribution in the first column, representing the responses of all students in the sample. Most students at this university drink less than 7–10 drinks and their assessment of the drinking behavior of these vignette persons is much harsher- 38.7% of all students describe having 7–10 drinks as excessive or extreme. The general finding that students appear to characterize vignette persons similar to the way they characterize their own drinking tends to hold for all drinking categories included in Table 1, with the possible exception of respondents who say they drink 5–6 drinks.

If response consistency holds and people who drink more are less harsh in their evaluation of their own drinking than people who drink relatively little, this also implies that distributions of self-reported problem drinking understate the tails of the true distribution of drinking problems. For example, if the response thresholds of the median drinking were used to evaluate drinking behavior of the full population, there would be more people who would be seen as having no problem at all and more who would be designated as problem drinkers.

3. Anchoring Vignettes

In this section, we first provide an intuitive description of the use of vignettes for identifying response scale differences and then sketch our formal statistical approach. Suppose one wants to characterize the drinking behavior of two groups of individuals who may vary in their actual drinking behavior. Figure 1 presents the distribution of the density of the true but unobserved continuous drinking behavior so that group A is to the left of that in group B, implying that on average, people in B drink more than in A.

The people in these two groups, also use very different response scales if asked whether or not they have drinking problems on a five-point scale (*Mild, Moderate, Some Cause for Concern, Excessive, Extreme*). In this example, people in group B are more tolerant of

drinking than people in group A. The frequency distribution of self-reports in the two groups suggests that people in A have more of a problem with drinking than those in B—the opposite of the true drinking distribution. Correcting for the differences in the response scales (DIF, "differential item functioning," in the terminology of King et al., 2004) is essential to compare the actual drinking problems in the two groups.

Vignettes can be used to do the correction. The vignette persons given to both groups drink the same amount. For example, respondents can be asked to evaluate the drinking of a vignette person given by the dashed line. In A, this will be evaluated as "some concern." In country B, the evaluation would be "moderate." Since the actual drinking behavior of the vignette person is the same, the difference in the evaluations by the two groups must be due to DIF. Vignette evaluations thus help to identify differences between the response scales. Using the scales in one of the two groups as the benchmark, the distribution of evaluations in the other group can be adjusted by evaluating them on the benchmark scale. The corrected distribution of the evaluations can then be compared since they are now on the same scale. The underlying assumption necessary to make this adjustment is *response consistency:* a given respondent uses the same scale for the self-reports and the vignette evaluations.

We present a formal statistical model explaining both subjective qualitative self-assessments and an objective self-reported quantitative measure of drinking behavior, as well as vignette evaluations of hypothetical people with possible drinking problems. The objective measure is obtained from respondents' self-reports on the number of drinks they consume, with categorical answers on an explicitly given quantitative scale. The subjective measure has categorical answers on a subjective scale, which may be interpreted differently by different individuals, so that subjective self-assessments may be affected by DIF. Thus the subjective measure will be modeled as a function of an underlying latent index reflecting actual drinking behavior, but also of individual specific thresholds, as in the Hopit model of King et al. (2004). Vignette evaluations use the same categorical answers as the subjective self-assessment.

We entertain two alternative assumptions. The assumption of *response consistency* (RC) means that respondents use the same thresholds when they evaluate themselves as when they evaluate vignettes. The *one factor* assumption (OF) means that a common factor drives the objective measure and the subjective measure, once the latter is purged of DIF. In the most general model, we impose neither of these assumptions. We will see that we need one of the assumptions for identification. Maintaining one of the assumptions, the other one can be tested.

The other assumption emphasized by King et al (2004) is *vignette equivalence*: other than through the thresholds, the way in which respondents interpret and evaluate the vignettes must be independent of respondent characteristics, which could be violated in situations where the respondents refer to their own situation to impute missing information in the vignette descriptions. Given the straightforward nature of the issue, we do not think this is a problem in our case, so that this assumption seems much less controversial here than response consistency. We maintain it throughout the paper.

Subjective self-assessments

As mentioned before, the subjective self-assessment (Y_{si} for respondent i) is the answer to the question below, on a five point scale:

"How would you describe your own drinking patterns over the course of the last year?"

Mild, Moderate, Some Cause for Concern, Excessive, Extreme

In the empirical work we will combine the categories *Excessive* and *Extreme* because the latter does not have many observations. The self-reports are assumed to be driven by an underlying latent index Y^*_{si} reflecting actual drinking behavior, an error term reflecting the arbitrary part in each self-evaluation, and individual specific thresholds:

$$Y^*_{si}=X_i\beta_s+\pi_s D_i+\xi_{si} \quad (1)$$

$$Y_{si}=j \text{ if } \tau^{j-1}_{si}<Y^*_{si}\le\tau^{j}_{si},\ j=1,\dots 4 \quad (2)$$

Here D_i is a dummy indicating whether ($D_i = 1$) or not ($D_i = 0$) the respondent was shown a screen presenting the definition of a drink before answering the questions on drinking behavior. X_i is a set of observed respondent characteristics ξ_{si} and can be can be interpreted as unobserved heterogenety in drinking behavior combined with an idiosyncratic noise term affecting the subjective self-report but nothing else. We will assume that ξ_{si} is normally distributed with mean zero and variance normalized to $\sigma^2_\xi=1$, independent of .X_i.

The thresholds τ^j_{si} between the categories are given by

$$\begin{gathered}\tau^0_{si}=-\infty,\ \tau^4_{si}=\infty,\ \tau^1_{si}=\gamma^1_s X_i+u_i,\ \tau^j_{si}=\tau^{j-1}_{si}+\exp(\gamma^j_s X_i),\ j=2,3\\ u_i\sim N(0,\sigma^2_u),\ u_i \text{ independent of } X_i \text{ and the other error terms in the model}\end{gathered} \quad (3)$$

The fact that different respondents use different response scales τ^j_{si} represents DIF. The term u_i introduces an unobserved heterogeneity term (modeled as a random individual effect) in the response scale.

Using subjective self-reports on own drinking behavior only, parameters β and γ^1_s are not separately identified; only their difference is identified. (The γ^j_s for *j>1* will still be identified.) For example, consider nationals of different countries who may engage in different drinking behaviors. If the scales on which they report their drinking behavior can vary across countries, qualitative self-reports on drinking are not enough to identify the difference in the distribution of drinking problems across nationalities, as was illustrated in Figure 1.

Vignette Evaluations

As described in Section 2, in the survey each respondent answered vignette questions on the drinking behavior of hypothetical people, using the same qualitative five point response scale that was used for the self-reports (*Mild, Moderate, Some Cause for Concern, Excessive, Extreme)*. The evaluations Y_{li} of vignettes *l=1,…,L (L=4)* are modeled using similar ordered response equations:

$$\begin{gathered}Y^*_{li}=\theta_l+\theta_f \text{ Female}_{li}+\pi_l D_i+\varepsilon_{li}\\ Y_{li}=j \text{ if } \tau^{j-1}_{vi}<Y^*_{li}\le\tau^{j}_{vi},\ j=1,\dots 5\\ \varepsilon_{li}\sim N(0,\sigma^2),\ \text{independent of each other, of } \xi_{si}, \text{ and of } X_i\end{gathered} \quad (4)$$

Apart from dummies indicating the vignettes, the only explanatory variables in the vignette evaluation equation are the dummy for having been shown the screen explaining what is a meant by a drink, and a dummy for the gender of the vignette person. The latter is included

because preliminary analysis suggested that respondents react differently to drinking vignettes with a female name than with a male name. Respondent characteristics X_i are not included in (4) - this is the maintained assumption of *vignette equivalence* discussed above.

The thresholds are modeled in a similar way as those in the self-report equation, but with different parameters:[2]

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

The standard Hopit model (see, e.g., King et al, 2004) assumes *response consistency:* $\tau_{vi}^j=\tau_{si}^j$, $j=1,\ldots 3$; $i=1,\ldots, N$. In terms of the parameters in (3) and (5), this hypothesis can be formulated as:

$$\text{RC:}\gamma_s^j=\gamma_v^j; j=1,2,3 \quad (6)$$

With this assumption, it is clear how vignette evaluations can be used to separately identify β_s and $\gamma_v(=\gamma_v^1,\ldots,\gamma_v^3)=\gamma_s(=\gamma_s^1,\ldots,\gamma_s^3)$: From the vignette evaluations alone, γ_V can be identified (up to the usual normalization of scale and location); β_s can then be identified from the self-assessments. Thus the vignettes can be used to solve the identification problem due to DIF *under the assumption of response consistency*.

In this paper, we want to consider the plausibility of assuming response consistency. In order to identify separate thresholds in the subjective self-reports and the vignette evaluations, we need more information - with the subjective self-reports and the vignette evaluations alone, identification requires the maintained assumption of response consistency. The additional information comes from an objective measure of drinking.

Objective self-assessment

The objective measure will be modeled as an ordered probit model:

$$\begin{aligned} Y_{oi}^*&=X_i\beta_o+\pi_o D_i+\xi_{oi} \\ Y_{oi}&=j \text{ if } \tau_o^{j-1}<Y_{oi}^*\le\tau_o^j,\ j=1,\ldots,6 \end{aligned} \quad (7)$$

Here we treat the category thresholds as unknown constants (with $\tau_o^0=-\infty$ and $\tau_o^6=\infty$), based upon the plausible assumption that they do not vary across individuals, in line with the literature that differential item functioning is an issue for subjective scales, not for objective scales. (We could also treat this as a grouped regression model and impose the actual values used in the question; this is somewhat more restrictive - see Appendix B.)

If no restrictions are imposed on the relation between the objective and the self-assessed measures of drinking behavior, observing the objective measure does not help for identification. A natural assumption for a perfect objective measure would be the *one factor* assumption (OF). It states that subjective and objective self-assessments are driven by the same underlying latent index for drinking behavior, i.e.:[3]

[2]The unobserved heterogeneity term is assumed to be the same in the thresholds for vignettes and subjective self-reports. This is needed for identification and without loss of generality – in the subjective self-reports, one cannot distinguish between the unobserved heterogeneity term u and the error term ξ_{si} in (1). This implies that our test for response consistency has no power in the direction of response inconsistencies not associated with the observed explanatory variables.

$$\text{OF:}\beta_s=\beta_o \quad (8)$$

We assume that ξ_{oi} is independent of X_i, u_i and ε_{li}, $l = 1,..., 4$, but can be correlated with ξ_{si}. This is because both will be affected by a common unobserved factor driving drinking behavior. The correlation will not be perfect since both measures will be affected by idiosyncratic reporting noise, and these idiosyncratic error terms will be part of ξ_{oi} and ξ_{si}. We also assume (ξ_{oi},ξ_{si}) is bivariate normally distributed.

A formal test of RC can be developed if OF is taken as a maintained assumption, thus comparing the model imposing OF and RC with a model imposing OF only. To see why in the latter model the main parameters are identified, note that the vignettes can be used to estimate γ_v^j, while the objective measure can be used to estimate $\beta = \beta_s = \beta_o$. The subjective self-reports make it possible to identify β–$\gamma_s^1, \gamma_s^2, \gamma_s^3$. With the estimates of β obtained from the objective measure equation, this means that β_s and γ_s^1 are both identified separately.

Each identified version of the model can be estimated by maximum likelihood. The likelihood contribution conditional on u_i is a product of a bivariate normal probability (for the self-report and the objective measure) and four univariate normal probabilities (for the vignettes). The unconditional likelihood contribution of respondent i can be computed numerically as an expectation over u_i. Likelihood Ratio tests can be used to formally test the assumptions of No DIF, OF, or RC, as long as there is a maintained assumption that guarantees identification. In addition to carrying out formal tests, we will also compare models using Akaike's Information Criterion (AIC; Akaike, 1974).

Checking whether Vignettes Help

An informal check for the usefulness of correcting for DIF with vignettes can be based upon the correlation between the indexes Y_{oi}^* and Y_{si}^*. This is only useful in models not imposing the one factor assumption, since imposing this assumption leads to a perfect correlation in the systematic parts Y_{oi}^* and Y_{si}^*. If the correction for DIF works well, we expect DIF corrected predicted systematic parts or simulated values of Y_{si}^* to be similar to predicted systematic parts or simulated values of Y_{oi}^* - differences due to DIF are then corrected for. Remaining differences can then be caused by 1) an imperfect correspondence between what the self-assessments measure and what the objective measure does, 2) finite sample estimation errors and, for the simulated values, 3) idiosyncratic errors in both Y_{oi}^* and Y_{si}^*. On the other hand, predicted or simulated values of Y_{oi}^* and Y_{si}^* based upon a model not allowing for DIF should be less similar to each other, since in that case, the predictions of Y_{si}^* will be affected by DIF while those of Y_{oi}^* will not. We therefore will look at the correlation between predicted as well as simulated values of Y_{si}^* and Y_{oi}^* for each model.

4. Results

Table 2 gives an overview of the models that have been estimated, imposing different subsets of the three assumptions discussed above: *No DIF* (thresholds are the same for all respondents), *OF* (one factor driving Y_{oi}^* and Y_{si}^*) and *RC* (response consistency - each respondent uses the same thresholds for vignettes and self-reports).

[3]One might expect a location parameter and a scaling factor here but these are normalized to 0 and 1, respectively. As a consequence, no further normalizations on equation (7) are needed if OF is imposed.

The most restrictive model does not allow for threshold variation across individuals (No DIF), assumes that these thresholds are the same in self-reports and vignettes (RC), and assumes that objective and subjective measures are driven by the same underlying factor (OF). This model ranks lowest in terms of the AIC. The second model has different thresholds for vignettes and self-reports (i.e. does not impose response consistency), but, because of the need to normalize scale and location of the ordered response equation for the vignettes, has only one additional parameter. This model is significantly better than the first model according to a likelihood ratio test (and has a better AIC), so response consistency would be rejected under the maintained assumption of one factor and no DIF. Of course as we demonstrate below the no DIF assumption in particular will be strongly rejected by the data.

Model 1 is also rejected against a model that does not impose that objective and subjective health measures are driven by the same factor (model 3). This model leads to a correlation between predicted objective and subjective drinking indexes of 0.915 (systematic parts only). The correlation between unobservables is 0.606. The unobserved parts exhibit much more variation than the systematic parts, explaining why the correlation between the simulated values is not much larger than 0.606 (i.e. .635). Again, imposing No DIF seems particularly strong here (and will be rejected below), so we should not take rejecting OF under the maintained assumption as evidence against OF. On the other hand, OF seems unlikely to hold exactly for our data, since the objective measure refers to only one feature of drinking behavior - the number of drinks on a typical drinking day - and not, for example, to the number of drinking days.

Model 4 relaxes model 3 in the same way as model 2 relaxes model 1. Again, response consistency is rejected by a formal LR test, now under the (implausible) maintained assumption of No DIF.

Model 5 relaxes the assumption that everyone uses the same thresholds (i.e. allows for DIF), while maintaining the other two assumptions. This leads to a huge improvement of the likelihood, and consequently also of the AIC. It also leads to higher estimates of the correlation between the objective and subjective health indexes. This increase in correlation is due both to a higher correlation between the systematic parts and the error terms. Model 5 is the model we would want from a theory point of view if the objective and subjective measures were in perfect accordance with each other, i.e., if the one factor assumption is valid and the vignettes do their work, i.e., response consistency is valid. The evidence that people use different response scales is strongly supported by this data.

Model 5 is formally rejected against both more general models (6 and 7), although the likelihood difference is much smaller than between model 5 and the earlier models. As discussed above, this can be seen as evidence against either the one factor assumption, or against response consistency, or both. The identification problem implies that we cannot really distinguish between these two alternatives. Since the objective measure is certainly not perfect - reflecting only one quantitative dimension of drinking behavior (number of drinks on a typical drinking day) and not the other (number of typical drinking days), the one factor assumption does not seem very plausible in our case. Thus we should not interpret this result as strong evidence against response consistency. This view is reinforced if we consider the AIC. According to the AIC, model 6 is the preferred model. In other words, according to this criterion a model that assumes response consistency, but does not impose OF provides the best fit to the data.

Not only does the likelihood improve substantially by allowing for DIF, it also brings objective and subjective indexes much more in line with each other. The best way to see this

is by comparing models 3 and 6. The correlation between the systematic parts increases (half the gap between this correlation and its theoretical maximum 1 is bridged), and the correlation between unsystematic parts (ξ_{oi} and ξ_{si}) increases as well. Thus vignettes certainly help a great deal to reduce the problem of differential item functioning. In our example, correcting for DIF using vignettes bridges a substantial part of the gap between objective and subjective measures of drinking behavior. It does not completely bridge the gap - and this may be due to the fact that the objective and subjective measure do not exactly measure the same thing, as also suggested by the AIC.

Parameter Estimates of Selected Models

We present parameter estimates of two models imposing response consistency and not imposing the one factor assumption, the model allowing for DIF (model 6 in Table 2) and the model not allowing for DIF (model 3 in Table 2). We report the parameter estimates for self-assessed drinking behavior, for the vignette thresholds, and the objective drinking measure.

There are a number of covariates entering these models that can be separated into three classes - personal attributes of the students, family background including attributes of parents and number of siblings, and drinking behaviors of parents. Appendix A provides a detailed description of each of the covariates.

The student level variables include a quadratic in age, gender of student (female = 1), nationality (non-Irish national =1), marital status (married =1), single and dating (going out =1), and undergraduate (bachelor =1 with mainly masters and PhD students as the reference group).

Family background variables include measures of the education of father and mother into three groups - education high (Father edu high, Mother edu high; education equals higher education, university) education medium (Father edu med, Mother edu med; education equals upper secondary) with education low as the reference group (primary or lower secondary), parental income (coded 1 to 8 depending on which of eight equally spaced income intervals parents income belongs to), whether the parents are separated, the number of siblings 16 or over and the number of siblings younger than 16.

Because attitudes and drinking habits can be transmitted across generations, we include measures of how much the father and mother drink each time they are drinking. For each parent, there are two variables describing their drinking behavior. The variable alcohol is treated as cardinal and goes from 1 = “abstainer” to 6 “consumes alcohol daily”. The second variable measures the quantity of drinks when drinking and is derived from the answers to the following question “Roughly how many drinks does your father (mother) consume each time he/she is drinking?).” Indicator variables for missing values for any variables mentioned above are included in the models but not listed in the tables.

Table 3 presents the parameter estimates for our ordered probit model of self-reported drinking behavior based on answers to the question (2) cited in Section 2 on the severity of one’s drinking behavior. The first two columns present parameter estimates in the no DIF case (not adjusted for the scales obtained from the vignettes) while the last two columns list parameter estimates adjusted for the vignette differences in thresholds.

Taking first the no DIF estimates- that is the model one would estimate without any vignette correction for different thresholds among respondents- drinking problems are reported to be less severe among female students,[4] among married students and those singles who are dating, and among those students who are not Irish nationals. In contrast, drinking problems

are more severe among those with more siblings over age 16, those who are single and not dating (the reference compared to married and single "going out"), and the more that either of their parents drinks. The education of neither parent affects drinking problems of these college students, but we find that higher parental income is associated with more alcohol consumption.[5]

The last two columns of Table 3 present the parameter estimates using the vignettes to correct for differences in thresholds among these students. The estimated effects of gender and parents' drinking behavior increase compared to the model that does not take into account DIF. The largest difference between parameter estimates accounting and not accounting for DIF is the coefficient of "Non-Irish." The difference between being an Irish National and not being an Irish National is much larger than self reports of problem drinking would have one believe.

The explanation for these differences in parameter estimates is clear from Table 4, which gives estimates of threshold parameters for the model accounting for DIF. The critical differences show up in the first threshold. To illustrate, Non-Irish students have very different (and stricter) norms on what is considered mild versus moderate drinking, and a similar shift applies to the other thresholds. What Irish students call mild drinking is often called moderate drinking by foreign students. Similarly, female students have a lower threshold for what constitutes problem drinking. In contrast, additional drinking by parents raises the threshold of what constitutes problem drinking. Given the narrow age range in this sample, one should not make much of the estimated quadratic age terms. But with that caveat, it appears that up to age 23.6 students are becoming looser on drinking standards and after that a bit stricter.

Table 5 presents our estimates of the ordered probit equation for the objective measure of drinking behavior, that is the answer to the question "How many drinks containing alcohol do you have on a typical day when you are drinking?" with answers 1 "less than 1", 2 "1 or 2", 3 "3 or 4", 4 "5 or 6", 5 "7–9" or 6 "10 or more". With the objective measure, the two models with and without DIF give very similar results. To illustrate, with the objective measure the effect of being a Non Irish student is the same in the DIF and non DIF models. This is as it should be since differential item functioning leads to thresholds variation for the subjective drinking measure but not for the objective measure of drinking. The parameter estimates on the objective measure are more in line with the DIF estimate in Table 3 than with the no-DIF estimate. See the parameter on Non-Irish, for example.

The first panel of Table 6 presents the estimates for the equation explaining the vignette evaluations. The vignette dummies are in line with vignette descriptions - they are ordered from least drinking (vignette 1) to most drinking (vignette 4). Explaining how a drink is defined on an introductory screen makes the vignette evaluations move slightly to less excessive drinking, but the difference is small and marginally significant only for one of the four vignettes. The sign suggests that for most respondents, "a drink" is less serious than what they had thought. The sign of the coefficient of the dummy for whether a female name is used in the vignette shows that the same drinking behavior is considered significantly more excessive if the vignette person is female than if the person is male.

[4]This gender differential is also found in many other studies. Rimal and Real (2005) find it can be fully explained by gender differences in perceived benefits from alcohol consumption.

[5]Weitzman and Chen (2005) find a positive association between alcohold consumption and parents' education, not controlling for parental income or drinking behavior.

Combining observed and unobserved variation shows that the correlation between objective and subjective reports is 0.636 in the model without DIF, and 0.734 in the model with DIF - a substantial improvement.

5. Conclusions

In this paper, we have investigated the validity of anchoring vignettes, which have been advanced to deal with the problem that different people may have different thresholds when answering qualititative questions on a subjective scale. We put forth a formal test of the validity of anchoring vignettes testing the key identifying assumption of response consistency. Response consistency implies that people use the same threshold in answering questions about themselves as they use in the anchoring vignettes. Using a sample of college students in Dublin, which has both objective and subjective measures of their drinking behavior as well as a set of anchoring vignettes about drinking, we find that the vignettes do a very good job in bringing self-reports on the severity of one's drinking in line with objective information about the quantity of their alcohol consumption.

This is clearly illustrated by the results in Table 1, where students who consume a certain amount of alcohol tend to exhibit very similar responses regarding their own drinking and the drinking of vignette persons who approximately consume the same amount of alcohol. According to Akaike's Information Criterion, the model maintaining response consistency, but not imposing the one factor model, provides the best fit to the data. In addition we find that relaxing DIF is extremely important for improving the fit of the model and in raising the correlation between the subjective drinking scale and the objective drinking measure.

The test applied in this paper is facilitated by the fact that there exists an objective measure that is relatively easy to observe, with a clear relation to the domain in which we are eliciting subjective responses. In cases with more ambiguity about the exact objective situation on which one is eliciting subjective responses, the use of anchoring vignettes may be less successful. In essence this would be caused by the fact that a vignette description has to be brief and therefore will tend to be incomplete. Even in the current application, the description of the vignettes is not complete. For instance, we describe a given situation and then how much the vignette person drinks at that occasion. But we do not specify how often the vignette person consumes that quantity. This in itself makes it all the more remarkable how good a job the vignettes are doing in correcting for differences in response scales.

Yet, additional tests of response consistency are necessary for other uses of vignettes especially when the correspondence between the objective and subjective measures are not as transparent as they are in the drinking application used here.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974; AC-19(6):716–723.

Christensen K, Herskind AM, Vaulpel JW. Why Danes are Smug: Comparative study of life satisfaction in the European Union. British Medical Journal. 2006; 333:1289–1291. [PubMed: 17185710]

D'Uva, TB.; Van Doorslaer, ED.; Lindeboom, M.; O'Donnell, O.; Chatterji, S. Tinbergen Institute Discussion Papers 06–033/3. 2006. Does reporting heterogeneity bias the measurement of health disparities?.

Gibson C, Schreck CJ, Miller JM. Binge drinking and negative alcohol-related behaviors: A test of self-control theory. Journal of Criminal Justice. 2004; 32:411–420.

Gill JS. Reported levels of alcohol consumption and binge drinking within the UK undergraduate student population over the last 25 years. Alcohol and Alcoholism. 2002; 37(2):109–120. [PubMed: 11912065]

Ham LS, Hope DA. College students and problematic drinking: A review of the literature. Clinical Psychology Review. 2003; 23:719–159. [PubMed: 12971907]

Kapteyn A, Smith JP, Van Soest A. Vignettes and self-reported work disability in the US and the Netherlands. American Economic Review. 2007; 97(1):461–473.

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

Kristensen, N.; Johansson, E. Aarhus School of Business Department of economics Working Paper 06–1. Labour Economics; 2006. New Evidence on Cross-Country Differences in Job Satisfaction Using Anchoring Vignettes. forthcoming

Lindeboom M, van Doorslaer E. Cut-point shift and index shift in self-reported health. Journal of Health Economics. 2004; 23:1083–1099. [PubMed: 15556237]

Mayor S. Alcohol and drug misuse sweeping world, says WHO. British Medical Journal. 2001; 322(7284):449. [PubMed: 11222413]

Ramstedt M, Hope A. The Irish drinking habits of 2002 - Drinking and drinking-related harm in a European comparative perspective. Journal of Substance Use. 2005; 10(5):273–283.

Rimal RN, Real K. How behaviors are influenced by perceived norms: A test of the theory of normative social behavior. Communication Research. 2005; 32:389–414.

Salomon JA, Tandon A, Murray CJ. Comparability of self rated health: Cross sectional multi-country survey using anchoring vignettes. British Medical Journal. 2004; 328(7434):258–260. [PubMed: 14742348]

Schmid H, Ter Bogt T, Godeau E, Hublet A, Ferreira Dias S, Fotiou A. Drunkenness among young people: A cross-national comparison. Journal of Studies on Alcohol. 2003; 64:650–661. [PubMed: 14572187]

Sen A. Health: perception versus observation. British Medical Journal. 2002; 324:860–861. [PubMed: 11950717]

Wechsler, H.; Wuethrich, B. Dying to drink: Confronting binge drinking on college campuses. Emmaus, PA: Rodale; 1999.

Weitzman ER, Chen YY. Risk modifying effect of social capital on measures of heavy alcohol consumption, alcohol abuse, harms, and secondhand effects: national survey findings. Journal of Epidemiology and Community Health. 2005; 59:303–309. [PubMed: 15767384]

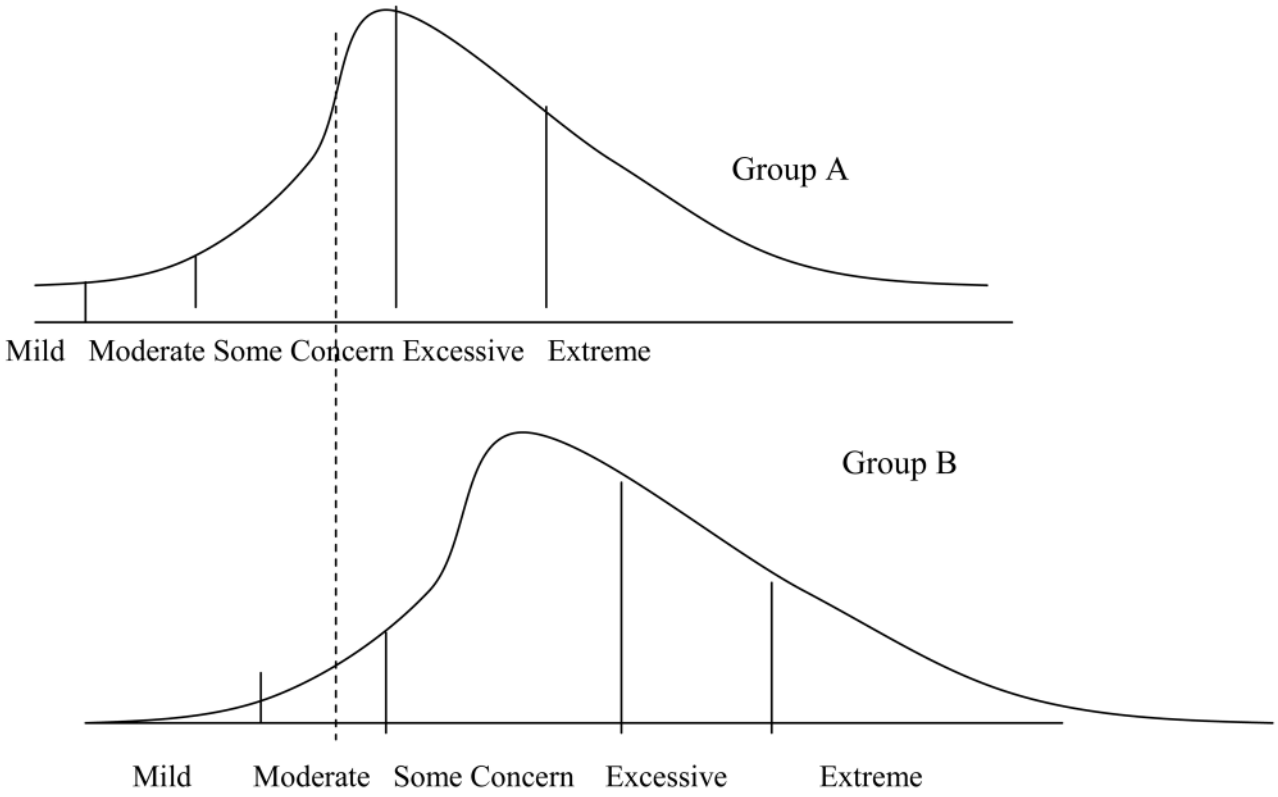

Figure 1.
Comparing self-reported drinking problems in two groups in case of DIF

Table 1

Responses to Vignettes Compared to Responses on Own Drinking Behavior

For vignettes that Describe 10 or More Drinks			
	All respondents	**Respondents who drink 10 or more**	
	Vignettes	**Vignettes**	**Self**
Mild	0.5	2.3	1.2
Moderate	7.2	24.7	21.0
Cause for concern	22.1	33.1	29.9
Excessive	36.4	30.8	37.4
Extreme	33.9	9.1	10.4
For vignettes that describe 7–10 drinks			
	All respondents	Respondents who drink 7–9 drinks	
	Vignettes	Vignettes	Self
Mild	0.8	1.0	6.6
Moderate	23.4	40.8	41.0
Cause for concern	37.1	39.0	32.9
Excessive	30.6	17.9	17.6
Extreme	8.1	1.3	2.0
For vignettes that describe 5–6 drinks			
	All respondents	Respondents who drink 5–6 drinks	
	Vignettes	Vignettes	Self
Mild	9.7	7.8	18.7
Moderate	66.3	75.9	55.1
Cause for concern	19.7	14.1	19.3
Excessive	4.0	2.2	6.2
Extreme	0.2	0.0	0.7
For vignettes that describe 2–3 drinks			
	All respondents	Respondents who drink 1–4 drinks	
	Vignettes	Vignettes	Self
Mild	75.0	59.7	49.1
Moderate	24.1	39.2	43.2
Cause for concern	0.6	0.8	6.2
Excessive	0.0	0.0	1.6
Extreme	0.1	0.2	0.0

Table 2

Models, log likelihoods, and correlations between Y^*_{oi} and Y^*_{si}

Model	Restrictions	# Par.	Log likel.	AIC (rank)*	Correlation Y^*_{oi} and Y^*_{si}	Corr. $\widehat{Y}^*_{oi}$ and $\widehat{Y}^*_{si}$	Corr. ξ_{oi} and ξ_{si}
1	No DIF, RC, OF	46	−18,522.01	37136.02 (7)	0.636	1	0.600
2	No DIF, OF	47	−18,517.92	37129.84 (6)	0.635	1	0.599
3	No DIF, RC	69	−18,488.34	37114.68 (5)	0.635	0.915	0.606
4	No DIF	70	−18,484.24	37108.48 (4)	0.635	0.915	0.605
5	OF, RC	119	−17,000.95	34239.90 (2)	0.738	1	0.690
6	RC	142	−16,972.59	34229.18 (1)	0.743	0.962	0.696
7	OF	191	−16.946.33	34274.66 (3)	0.723	1	0.665

*Indicates the rank of the model with respect to the AIC.

Table 3

Models of Self-report on Own Drinking Behavior

	No DIF		DIF	
	par.	**s.e.**	**par.**	**s.e.**
Constant	0.399	0.837	0.253	1.224
Age/10	−0.099	0.693	1.129	0.836
(Age/10) squared	−0.025	0.040	−0.274	0.177
Female	−0.271*	0.040	−0.296*	0.045
Married	−0.460*	0.135	−0.453*	0.159
Going out	−0.222*	0.042	−0.245*	0.047
Non-Irish	−0.223*	0.078	−0.573*	0.088
Bachelor	0.059	0.059	0.077	0.067
Siblings 16+	0.036*	0.015	0.040*	0.018
Siblings 16−	0.024	0.026	0.042	0.029
Father edu med	0.035	0.062	−0.024	0.069
Father edu high	0.046	0.060	0.004	0.066
Father alcohol	0.011	0.020	0.023	0.023
Father drinks	0.021*	0.009	0.036*	0.009
Mother edu med	−0.026	0.062	−0.011	0.070
Mother edu high	−0.050	0.062	−0.047	0.071
Mother alcohol	0.011	0.020	0.009	0.023
Mother drinks	0.063*	0.015	0.098*	0.016
Parents' income	0.067*	0.012	0.066*	0.013
Parents separated	−0.028	0.069	−0.021	0.081
Screen shown	0.055	0.041	0.049	0.054

*indicates statistical significant at the 5% level and

+indicates statistical significant at the 10% level.

Table 4

Models of Vignette Thresholds with DIF

	threshold 1		threshold 2		threshold 3	
	par.	**s.e.**	**par.**	**s.e.**	**par.**	**s.e.**
constant	0.000	0.000	0.040	0.372	−0.646	0.682
Age/10	1.267*	0.398	−0.094	0.309	0.031	0.567
(Age/10) squared	−0.268*	0.084	0.023	0.065	0.014	0.119
Female	−0.087*	0.021	0.041*	0.017	0.028	0.032
Married	−0.027	0.082	−0.087	0.070	−0.083	0.112
Going out	−0.061*	0.023	0.010	0.018	0.026	0.032
Non-Irish	−0.403*	0.043	0.026	0.037	0.125*	0.061
Bachelor	0.008	0.030	0.014	0.026	−0.010	0.046
Siblings 16+	0.010	0.007	−0.001	0.007	0.007	0.013
Siblings 16−_	0.019	0.013	0.004	0.010	−0.013	0.020
Father edu med	−0.054+	0.032	−0.032	0.026	0.032	0.048
Father edu high	−0.023	0.031	−0.042+	0.025	0.040	0.047
Father alcohol	0.017	0.011	0.001	0.008	−0.025	0.016
Father drinks	0.014*	0.005	0.003	0.005	0.017*	0.008
Mother edu med	−0.011	0.034	0.047+	0.027	−0.073	0.048
Mother edu high	−0.024	0.033	0.039	0.028	−0.092+	0.050
Mother alcohol	−0.007	0.011	0.012	0.009	−0.005	0.016
Mother drinks	0.038*	0.007	0.000	0.007	−0.002	0.013
Parents' income	0.006	0.006	0.004	0.005	0.001	0.009
Parents' separated	0.043	0.036	−0.032	0.029	−0.084	0.058

*indicates statistical significance at the 5% level and

+indicates statistical significant at the 10% level.

Table 5

Models of Objective Measure of Drinking

	No DIF		DIF	
	par.	**s.e.**	**par.**	**s.e.**
const obj me	−0.366	0.840	−0.370	1.011
Age/10	1.986*	0.698	1.991*	0.851
(Age/10) squared	−0.477*	0.146	−0.478*	0.180
Female	−0.504*	0.041	−0.504*	0.047
Married	−0.554*	0.129	−0.554*	0.146
Going out	−0.155*	0.042	−0.155*	0.048
Non-Irish	−0.836*	0.080	−0.838*	0.091
Bachelor	0.191*	0.061	0.191*	0.069
Siblings 16+	0.068*	0.016	0.068*	0.018
Siblings 16−	0.043	0.026	0.043	0.030
Father edu med	−0.055	0.063	−0.056	0.071
Father edu high	−0.119+	0.061	−0.119+	0.067
Father alcohol	0.035*	0.020	0.035	0.023
Father drinks	0.037*	0.003	0.038*	0.001
Mother edu med	0.003	0.064	0.003	0.072
Mother edu high	−0.041	0.064	−0.042	0.072
Mother alcohol	−0.051*	0.020	−0.051*	0.023
Mother drinks	0.115*	0.010	0.115*	0.016
Parents' income	0.091*	0.012	0.091*	0.013
Parents' separated	−0.005	0.074	−0.006	0.082
Screen shown	0.041	0.044	0.041	0.045

* indicates statistical significance at the 5% level and

+ indicates statistical significance at the 10% level.

Table 6

Other Parameters Model with DIF

Vignette dummies & gender vignettes			
	par.	**s.e.**	**t-val**
dummy vignette 1	1.169*	0.475	2.46
dummy vignette 2	2.211*	0.476	4.65
dummy vignette 3	2.952*	0.477	6.19
dummy vignette 4	3.394*	0.478	7.11
dummy screen shown vignette 1	−0.027	0.026	1.01
dummy screen shown vignette 2	−0.043+	0.024	1.79
dummy screen shown vignette 3	−0.005	0.023	0.23
dummy screen shown vignette 4	−0.032	0.025	1.30
dummy vignette has female name	0.115*	0.010	11.71
Covariance structure errors			
	par.	s.e.	t-val
sigma error selfreport	1.000	0.000	0.00
sigma heterogeneity subj. threshold	0.400*	0.011	37.93
sigma vignette evaluation	0.333*	0.007	45.22
sigma objective report	1.040*	0.017	60.39
correlation objective and subj. reports	0.696*	0.011	61.59
threshold objective report 1	0.000	0.000	0.00
threshold objective report 2	0.917*	0.044	21.08
threshold objective report 3	1.914*	0.047	40.66
threshold objective report 4	2.862*	0.050	57.14
threshold objective report 5	3.800*	0.058	65.83