Received: 25 June 2018 | Accepted: 19 June 2020

DOI: 10.1111/rssa.12612

ROYAL STATISTICAL SOCIETY
Series A Statistics in Society

ORIGINAL ARTICLE

Specification and testing of hierarchical ordered response models with anchoring vignettes

William H. Greene[1] | **Mark N. Harris**[2] | **Rachel J. Knott**[3] | **Nigel Rice**[4]

[1]New York University, New York, USA

[2]Curtin University, Perth, Australia

[3]Centre for Health Economics, Monash Business School, Monash University, Melbourne, Australia

[4]University of York, York, UK

Correspondence
Nigel Rice, Department of Economics and Related Studies, University of York, York, UK.
Email: nigel.rice@york.ac.uk

Abstract

Collection and analysis of self-reported information on an ordered *Likert* scale is ubiquitous across the social sciences. Inference from such analyses is valid where the response scale employed means the same thing to all individuals. That is, if there is no *differential item functioning (DIF)* present in the data. A priori this is unlikely to hold across all individuals and cohorts in any sample of data. For this reason, anchoring vignettes have been proposed as a way to correct for *DIF* when individuals self-assess their health (or well-being, or satisfaction levels, or disability levels, etc.) on an ordered categorical scale. Using an example of self-assessed pain, we illustrate the use of vignettes to adjust for *DIF* using the compound hierarchical ordered probit model (*CHOPIT*). The validity of this approach relies on the two underlying assumptions of response consistency (*RC*) and vignette equivalence (*VE*). Using a minor amendment to the specification of the standard *CHOPIT* model, we develop easy-to-implement score tests of the null hypothesis of *RC* and *VE* both separately and jointly. Monte Carlo simulations show that the tests have good size and power properties in finite samples. We illustrate the use of the tests by applying them to our empirical example. The tests should aid more robust analyses of self-reported survey outcomes collected alongside anchoring vignettes.

KEYWORDS

anchoring vignettes, *CHOPIT*, differential item functioning, self-assessments, score test, ordered response models

1 | INTRODUCTION

It is common in social surveys to use subjective categorical scales to elicit information in the form of self-reports; for example, levels of health, work disability or subjective well-being. Responses to such questions are often used to study differences across countries or social or demographic groups. A problem with relying on subjective responses is that individuals are likely to place different interpretations on the response scale. Information on health status might, for example, be obtained using the question: *Overall, how would you rate your health?* Respondents are asked to tick one of (typically) five boxes ranging from *very bad* through to *good* to *excellent*. Variation in responses will be due, in part, to genuine health differences, but may also be due to respondents applying different meanings to the available response categories. This type of reporting behaviour is commonly referred to as *differential item functioning*, or *DIF* (Holland & Weiner, 1993; Murray *et al*., 2002).

Figure 1 illustrates *DIF* using an example of a self-reported question about pain. Assume we have two respondents who are asked the question 'Overall in the last 30 days, how much of bodily aches or pains did you have?' and are instructed to respond by selecting one of the following: 'None', 'Mild', 'Moderate', 'Severe' or 'Extreme'. In the diagram, the vertical line represents the underlying latent scale for pain. *DIF* is depicted by the different locations of the individual-specific boundary parameters along the latent scale, μ_0 to μ_3. Although respondents have identical levels of latent pain (indicated by the bold arrows), respondent B reports mild pain, while respondent A reports no pain. Without knowing the locations of the boundary parameters, researchers would typically conclude that B has worse pain than A.

A number of approaches have been proposed to test for *DIF*. In the educational literature, where *DIF* is used to refer to test questions (items) in which individuals with the same underlying ability have differing probabilities of answering a question correctly; popular approaches include the Mantel–Haenszel procedure, item response theory and logistic regression-based methods (Holland & Thayer, 1988; Shepard *et al*., 1981; Swaminathan & Rogers, 1990). The basic idea is to compare the probability of answering a question correctly across different groups of individuals, while conditioning on underlying ability. A well-known issue is the difficulty in measuring the underlying ability of interest. For example, ability is commonly measured in terms of other test items (e.g. overall test scores) which

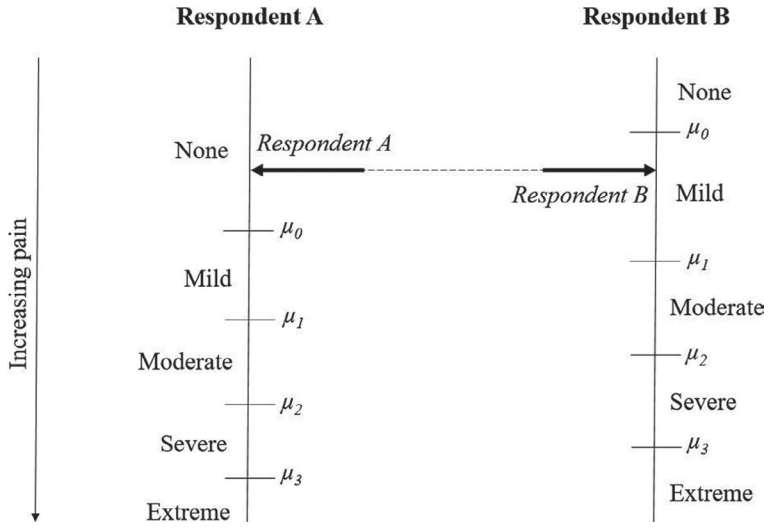


FIGURE 1 Example of DIF in self-assessed pain

themselves may be subject to *DIF*. Moreover, the item in question may depend on other forms of ability which are unobserved to the researcher (Clauser & Mazor, 1998).

Other methods include anchoring-based approaches, where common anchors are used to fix the responses of different individuals to the same response scale (Aldrich & McKelvey, 1977; Groseclose *et al*., 1999; Tay *et al*., 2013). The main challenge here is in selecting appropriate anchors that are completely free of *DIF*. The anchoring vignette method (King *et al*., 2004) addresses this issue by exploiting *DIF* in the responses to the vignette questions to adjust for *DIF* in the response of interest.

Anchoring vignettes have received wide attention in the applied literature—for example, in self-reported data on health status (Bago d'Uva *et al*., 2008; Grol-Prokopczyk *et al*., 2011; Peracchi & Rossetti, 2012; Soloman *et al*., 2004; Vonkova & Hullegie, 2011); healthy behaviours (Van Soest *et al*., 2011); satisfaction with health system performance (Rice *et al*., 2012; Sirven *et al*., 2012); work disability (Angelini *et al*., 2011; Kapteyn *et al*., 2007, 2011; Paccagnella, 2011); political efficacy (King *et al*., 2004); job satisfaction (Kristensen & Johansson, 2008); life satisfaction (Angelini *et al*., 2014); satisfaction with income (Kapteyn *et al*., 2013) and consumer satisfaction with products and services (Rossi *et al*., 2001). Together with their own situation, respondents are asked to evaluate one or more vignettes describing situations of hypothetical individuals with a given level of the domain of interest (e.g. pain). Responses to the vignettes are then used to anchor, or adjust for bias introduced by *DIF*, such that interpersonal comparisons of the self-reported outcome can be appropriately examined. This is often achieved using the compound hierarchical ordered probit model (*CHOPIT*: see Section 4.2).

Adjusting for *DIF* using the vignettes approach is valid under the two identifying assumptions of response consistency (*RC*) and vignette equivalence (*VE*). *RC* assumes that individuals use the same mapping from the underlying latent scale to the available response categories when assessing the self-assessment as they use when assessing the corresponding vignettes. This assumption allows the relationship between reporting behaviour and the characteristics of respondents identified via the vignettes to anchor responses to self-reports. *VE* assumes that 'the level of the variable represented by any one vignette is perceived by all respondents in the same way and on the same unidimensional scale' (King *et al*., 2004, p. 194). This implies that respondents agree on the underlying latent level of the concept under scrutiny—depicted by the hypothetical situation described by the vignette—except for random error.

We contribute to the literature by suggesting an amended specification to the usual vignette-based approach, which lends itself to score-based tests of the assumptions of *RC* and *VE*. The proposed score tests are informative in guiding model specification when modelling self-assessed ordered categorical outcomes using the *CHOPIT* approach. For example, if joint failure of both *RC* and *VE* is due to a failure of *VE* rather than *RC*, then this suggests that the use of alternative, or different subsets of vignettes, might be appropriate. It may also suggest that the vignette questions require refinement to better aid survey respondents' interpretation. Failure of *RC*, in contrast, potentially suggests a re-specification of the thresholds in the *CHOPIT* model. As is typical with specification tests, the score test relies on standard parametric assumptions underlying the *CHOPIT* model: that the model is correctly specified, with no omitted variables, endogeneity bias and so on.

The empirical literature that relies on vignettes to adjust for *DIF* in self-reported outcomes rarely conducts comprehensive validity checks of the approach. This is likely to be due to a previous lack of readily implementable statistical tests of *RC* and *VE*. We examine, via Monte Carlo experiments, the finite sample properties of our proposed test(s), and find that they are correctly sized and have appropriate power properties as one moves further from the relevant null hypothesis. We illustrate the use of the tests in an application to *SHARE* data. In addition, we compare our score test to a minimum

distance estimator developed by Peracchi and Rossetti (2013) and show that the score test appears to have greater power in detecting departures from the null of *RC* and *VE*.

The paper is organised as follows. Section 2 introduces the *SHARE* data, the self-reported health variable and corresponding vignettes, and illustrates the presence of *DIF*. Section 3 sets out the modelling approaches for ordered categorical outcomes in the absence of *DIF* and Section 4 in the presence of *DIF*. The latter relies on information contained within the responses to the vignettes. Our contributions are developed in Section 5, where we propose an amendment to the usual statistical approach to account for *DIF* which lends itself to simple score tests of both *VE* and *RC* individually and jointly. The tests are also derived in this Section. We apply the amended specification to *SHARE* data in Section 6 and implement the score test. Section 7 sets out the finite sample properties of the amended specification and test procedures. Section 8 provides concluding remarks.

2 | AN EMPIRICAL APPLICATION TO SELF-REPORTED PAIN FROM SHARE

This section introduces *SHARE* data including the categorical health outcomes to measure pain. Using the set of corresponding vignettes for pain, we show prima facie evidence of *DIF* in these data. *SHARE* is a multidisciplinary and cross-national panel dataset of individuals aged 50 or above and over time has expanded to covering 28 countries. The survey collects information on health, socio-economic status, and social and family networks. A particular virtue of *SHARE* is that information on self-reported health together with vignettes are included within the survey.

In the context of a diverse continent like Europe, differences in language and cultural and social norms are likely to lead to differences in the way individuals respond to survey instruments. The application of anchoring vignettes is, therefore, important for enhancing cross-country comparability. Together with self-assessments, vignettes on health were collected on subsamples of respondents in the first two waves of *SHARE*. The first wave contained three vignette questions for each domain of health and the second wave contained a single vignette only. Due to the increased number of vignettes available, which is helpful to illustrate how the score test might be applied in practical applications, we use data only from the first wave. Data from Belgium, France, Germany, Greece, Italy, the Netherlands, Spain and Sweden were included in the subsample responding to the self-assessed health questions and associated vignettes. *SHARE* data has been popular for studies investigating differences in reporting behaviour and more generally the method of anchoring vignettes (e.g. see Bago d'Uva *et al*. (2008), Angelini *et al*. (2012), Paccagnella (2013), Peracchi and Rossetti (2013), an Van Soest and Vonkova (2014), Jones *et al*. (2018)).

We consider data for the health domain representing pain and restrict our analysis to respondents aged 50–80 years. In addition to a self-assessment component, respondents were also asked to rate three vignettes for pain, representing different levels of severity, using the same response categories ('None', 'Mild', 'Moderate', 'Severe' and 'Extreme'). Appendix A contains the self-assessment question together with the vignettes, and Table 1 reports the frequencies for the responses observed in the data. The level of pain described in each vignette is increasing from vignette 1 (least pain) to vignette 3 (most pain). Due to the low prevalence of responses in the 'Extreme' category for the self-assessment and the first vignette, the responses for 'Severe' and 'Extreme' have been collapsed.

For modelling the self-assessed pain outcome, we employ the set of covariates presented in Table 2. These represent plausible determinants and indicators of pain and also feature in Peracchi and Rossetti (2013) who also used the *SHARE* data to illustrate their minimum distance estimator of the underlying assumptions of the *CHOPIT* model. As we use the data as an illustration of modelling

TABLE 1 SHARE: Self-assessed pain and corresponding vignettes

	SAH	Vignette 1 ($m1$)	Vignette 2 ($m2$)	Vignette 3 ($m3$)
None	33.64	15.91	2.26	1.13
Mild	35.95	56.60	17.96	4.60
Moderate	22.30	21.99	50.08	25.67
Severe/extreme	8.10	5.50	29.69	68.60

TABLE 2 SHARE: Descriptive statistics[a]

	Mean	Std Dev	Min	Max
Pain	1.049	0.939	0	3
Male	0.468	0.499	0	1
AnyCond	0.712	0.453	0	1
Grip35	0.531	0.499	0	1
EducPS	0.209	0.407	0	1
Age 50–65	0.643	0.479	0	1
Age 66–75	0.279	0.448	0	1
Age > 75	0.078	0.268	0	1

[a]Sample size, $N = 3802$.

self-reported outcomes in the presence of *DIF*, and the proposed score test for *RC* and *VE* in the *CHOPIT* model, rather than the substantive focus of the paper, we choose to keep the model parsimonious. The *CHOPIT* approach (see Sections 3 and 4 for more details), requires two sets of covariates; those which affect the underlying latent scale of the construct of interest, $\boldsymbol{x}$, and those which shift the inherent boundary parameters of the model, $\boldsymbol{z}$. In the absence of persuasive information on appropriate exclusion restrictions, we set $\boldsymbol{x} = \boldsymbol{z}$ (this is commonplace in the literature). Thus, the specification includes binary variables for males (*Male*: 47% of our sample); respondents aged 66–75 years (*Age* 66−75: 28%) and aged 76 and over (*Age* > 75: 8%); post-school education (*EducPS*: 21%); and the presence of health conditions (*AnyCond*: 71%). An indicator variable representing below average hand grip strength is also included (*Grip*35: 53%), which is based on up to four measurements conducted by a trained interviewer. Our working sample is 3802 individuals.

As the responses to the survey self-reports of pain are ordinal, they can be modelled as a function of covariates using a ordered probit (*OP*) model, as set out in Section 3. This approach assumes that individuals are using a given fixed reporting scale that does not differ across respondents, that is, that *DIF* does not exist in the data. We can illustrate the likely extent of *DIF* in the self-reports by simply considering responses to the set of vignettes. Since the vignettes describe fixed levels of a given domain that are provided to all respondents, variation in reporting on the vignettes by characteristics of individuals is indicative of systematic reporting behaviour. Table 3 shows reporting differences by covariates for each of the three vignettes. For each characteristic, the table reports the proportion of respondents classifying the vignette as either no or mild difficulties. For gender, Pearson *chi*-squared statistics and associated *p*-values are provided, while the corresponding *chi*-squared statistic from Kendall's τ and associated *p*-values are provided for the remaining characteristics.

The results in Table 3 indicate the likely presence of *DIF* in the levels of all covariates considered in response to at least one of the three vignettes. For example, women are more likely to rate vignette 2

TABLE 3 Vignette classification by respondent characteristics

	Vignette 1 (*m*1)	Vignette 2 (*m*2)	Vignette 3 (*m*3)
Male	0.73	0.18	0.052
Female	0.72	0.22	0.064
χ^2_1 (*p*-value)	1.24 (0.265)	5.27 (0.022)	2.39 (0.122)
AnyCond = 0	0.75	0.22	0.072
AnyCond = 1	0.71	0.19	0.051
χ^2_1 (p-value)	5.62 (0.018)	1.72 (0.190)	6.60 (0.010)
Grip35 = 0	0.74	0.19	0.058
Grip35 = 1	0.71	0.21	0.057
χ^2_1 (*p*-value)	4.82 (0.028)	0.73 (0.393)	0.013 (0.910)
EducPS = 0	0.72	0.22	0.063
EducPS = 1	0.74	0.15	0.038
χ^2_1 (*p*-value)	0.62 (0.431)	20.08 (0.000007)	7.66 (0.006)
Age 50–65	0.74	0.19	0.056
Age 66–75	0.71	0.22	0.062
Age 75+	0.70	0.24	0.054
χ^2_1 (*p*-value)	3.72 (0.054)	5.69 (0.017)	0.267 (0.605)

as no or mild pain compared to men; individuals reporting no health conditions are more likely to rate vignette 1 (least severe vignette) as no or mild pain than counterparts with health conditions; the more educated are less likely than the less educated to rate vignette 3 (most severe vignette) as no or mild pain. Younger respondents are more likely than older respondents to report vignette 1 as no or mild pain and less likely to rate vignette 2 as no or mild pain. These results provide prima facie evidence of the use of different reporting scales, or *DIF*, in respondents assessments which is likely to also exist in the self-assessments of the same health construct.

3 | MODELLING ORDERED OUTCOMES IN THE ABSENCE OF *DIF*

Our measures of pain are responses on a categorical (*Likert*) scale which can be estimated using ordered (probit or logit) response models (Greene & Hensher 2010). Underlying the *OP* model (indeed, both) is a latent variable, y^*, which is a linear (in unknown parameters, $\tilde{\boldsymbol{\beta}}$) function of observed characteristics $\tilde{\boldsymbol{x}}$ with no constant term (throughout we denote a no-constant subvector/matrix by use of '$\sim$', and denote a subvector matrix containing a constant by the absence of '$\sim$'). The term ε_y represents a standard normal disturbance term, such that

$$y^* = \tilde{\boldsymbol{x}}'\tilde{\boldsymbol{\beta}} + \varepsilon_y, \tag{1}$$

where y^* is mapped into observed $j = 0, \ldots, J-1$ outcomes via the usual mapping

$$y = j \text{ if } \mu_{j-1} \leq y^* < \mu_j \quad \text{for } j = 0, \ldots, J-1, \tag{2}$$

where $\mu_{-1} = -\infty$ and $\mu_{J-1} = +\infty$, and where to ensure well-defined probabilities; $\mu_{j-1} < \mu_j, \forall j$. The expressions for the resulting probabilities and likelihood functions are well-known (e.g. see Greene and Hensher (2010)). Applying the *OP* model to the self-reported outcomes for pain yields the set of estimates presented in column (1) of Table 4. In general, levels of pain are lower for males compared to females, and for respondents who have a post-school qualification. Respondents reporting the presence of health conditions experience greater levels of pain, as do those with below average grip strength. Pain also increases with age (test of joint significance: $\chi^2_2 = 6.74; p = 0.034$). However, for the *OP* coefficients to be unbiased, we need to assume that all respondents use the same reporting scales such that the boundary parameters, μ_j, are common to all respondents. This implies an absence of *DIF*. As we have seen in Table 3, this is unlikely to be the case.

4 | MODELLING ORDERED OUTCOMES IN THE PRESENCE OF *DIF*

We now consider extensions to the *OP* model in the presence of *DIF*. We first describe an approach that does not rely on the use of vignettes, but which imposes strong assumptions. We then consider an approach that incorporates information from vignette responses to identify the model. We conclude the Section with a discussion of approaches used in the literature to investigate the identifying assumptions of the vignette approach.

4.1 | Hierarchical ordered probit model (*HOPIT*)

Differences in reporting scales across individuals can be accommodated by specifying individual-specific boundary parameters, $\mu_{i,j}$ (see, e.g. Terza (1985), Pudney and Shields (2000), Boes and Winkelmann (2006), Greene and Hensher (2010), Greene *et al*. (2014)). This can be achieved by allowing the boundaries to depend on a set of observed characteristics $\boldsymbol{z}_i$ such that $\mu_{i,j} = \boldsymbol{z}_i'\gamma_j$. Note, however, to secure identification the approach imposes the restriction that $\boldsymbol{z}_i \notin \boldsymbol{x}_i$.

TABLE 4 Ordered response models of pain

	Ordered probit		***CHOPIT***	
	(1)		**(2)**	
***N* = 3802**	**Coef**	**SE**	**Coef**	**SE**
Male	−0.164	0.049	−0.244	0.061
AnyCond	0.627	0.042	0.588	0.051
Grip35	0.236	0.050	0.173	0.062
EducPS	−0.129	0.045	−0.168	0.055
Age 66–75	0.064	0.041	0.077	0.051
Age > 75	0.162	0.068	0.149	0.084
Boundaries				
μ_1	0.057	0.057		
μ_2	1.060	0.058		
μ_3	2.006	0.064		
Log-likelihood	−4624.80		−8824.15	

To ensure coherent probabilities most authors (see, e.g. Greene and Hensher (2010)) adopt the *Hierarchical* Ordered Probit (*HOPIT*) approach by specifying the boundaries as

$$\begin{aligned} \mu_{i,0} &= \boldsymbol{z}_i'\boldsymbol{\gamma}_0, \\ \mu_{i,j} &= \mu_{i,j-1} + \exp(\boldsymbol{z}_i'\boldsymbol{\gamma}_j), j = 1, \ldots, J-2. \end{aligned} \tag{3}$$

This model can be estimated by maximum likelihood techniques, where the μ_j in Equation (2) are simply replaced by those of Equation (3).

4.2 | The compound hierarchical ordered probit model

Empirically, it is often difficult to justify exclusion restrictions between $\boldsymbol{x}$ and $\boldsymbol{z}$. This can be seen in the above example, where from Table 3 we infer that, for example experiencing health conditions is associated with *DIF*, but also from Table 4 that health conditions are a significant predictor of pain. However, for any variable that appears in both $\boldsymbol{x}$ and $\boldsymbol{z}$, since the first threshold in Equation (3) is specified linearly, the corresponding elements of $\boldsymbol{\gamma}_0$ and $\tilde{\boldsymbol{\beta}}$ are not separately identified in the absence of further information. Identification can be resolved by the availability of (anchoring) vignettes, which are used in conjunction with the main self-report of interest. The following is an example of a vignette for pain taken from the *SHARE* (vignette $m1$ in Appendix A.1):

> "Karen has a headache once a month that is relieved after taking a pill. During the headache she can carry on with her day-to-day affairs. Overall in the last 30 days, how much of bodily aches or pains did Karen have?"

The categories (and scale) available to respondents are the same as those used to self-assess levels of pain, namely, in our example, *None*, *Mild*, *Moderate*, *Severe* and *Extreme*.

Assume that for a randomly chosen individual, the response to the self report on the latent scale, y^*, is given as model (1) and the corresponding response to the k^{th} vignette, v_k^*, as

$$v_k^* = \alpha_k + \varepsilon_k, \quad k = 1, \ldots, K, \tag{4}$$

where $\varepsilon_k \sim N(0, \sigma_k^2)$. Note that the number of vignettes available (K) will vary across surveys used, but in general is likely to be small (typically ≤ 3). When more than one are available ($K > 1$), there is a trade-off between improved model identification due to using more vignettes, and potentially increased bias due to the heightened probability that one may violate the requisite assumptions (described below). Indeed, the testing procedures developed in this paper would appear to be fundamental in the choice of vignettes used, and hence K, where there are multiple available in a given dataset.

The observed response to the self-report, y, and to each vignette, v_k, is determined as in Equation (2) before, by considering their relationship with the boundary equations. Heterogeneity across these response scales is once more accommodated by specifying the boundaries as a function of variables, $\boldsymbol{z}$ (see Equation (3)). In this set-up, we do not need to impose exclusion restrictions between $\boldsymbol{x}$ and $\boldsymbol{z}$ and it is common to assume $\boldsymbol{x} = \boldsymbol{z}$. However, to aid exposition, we retain the labelling $\boldsymbol{x}$ and $\boldsymbol{z}$ throughout.

We refer to the *HOPIT* model with vignettes as the *CHOPIT* model (see, e.g. Vonkova and Hullegie (2011), Paccagnella (2013), Van Soest and Vonkova (2014)). Identification of the model follows from

the assumptions of *RC* and *VE* (King *et al*., 2004). In practice, *RC* implies that the boundary parameters are the same across the self-report of interest and all K vignettes. Formally, *RC* imposes the following restriction

$$\boldsymbol{\gamma}_{j,k} \equiv \boldsymbol{\gamma}_{j,0}, \quad j = 0, \ldots, J-2;\ k = 1, \ldots, K, \tag{5}$$

where $k = 0$ indexes boundary equations for the self-report of interest ($\mu_{0,0}, \ldots, \mu_{J-2,0}$) and $k = 1, \ldots, K$, the corresponding boundary equations for the vignettes, ($\mu_{0,k}, \ldots, \mu_{J-2,k}$). Note that this equivalence of boundary parameters across the self-report of interest and vignette equations necessitates that all are measured on the same scale (they all have the same set of possible responses and use the same response categories).

VE, in contrast, implies that the underlying level of the construct of interest described by a vignette is perceived by all respondents in the same way and on the same unidimensional scale, except for random error (Equation (4)). The alternative is to consider the more general specification where the latent response is a function of respondent characteristics, such that

$$v_k^* = \alpha_k + \tilde{\boldsymbol{x}}' \tilde{\boldsymbol{\alpha}}_k + \varepsilon_k, \quad k = 1, \ldots, K. \tag{6}$$

VE therefore imposes the linear restriction(s) that $\tilde{\boldsymbol{\alpha}}_k = 0,\ \forall k$. In practice therefore, the usual *CHOPIT* approach simply omits the term $\tilde{\boldsymbol{x}}' \tilde{\boldsymbol{\alpha}}_k$ in estimation.

With all these elements in place, the log-likelihood function for the *CHOPIT* model consists of two distinct parts: one relating to the self-report of interest ($\ln L_{HOPIT}$) and the other to the vignette component of the model ($\ln L_V$). When there are several vignettes, $\ln L_V$ is the sum over the K of these. The first term, $\ln L_{HOPIT}$, is a function of $\boldsymbol{\beta}$ and $\mu_{j,k=0}(\boldsymbol{\gamma}_{j,k=0})$, and the second term(s), $\ln L_V$, is a function of α_k , σ and $\mu_{j,k}(\boldsymbol{\gamma}_{j,k})$, where $k > 0$. These two components of the likelihood are then linked by the common boundary parameters. The log-likelihood therefore can be written

$$\ln L = \sum_{i=1}^{N} \ln L_{i,HOPIT} + \sum_{i=1}^{N} \ln L_{i,V}.$$

Column (2) of Table 4 presents *CHOPIT* estimates using the vignette, $m1$, for pain described above. Assuming *RC* and *VE* hold, the use of vignette responses should adjust for *DIF* to produce unbiased estimates of the parameters in the outcome equation. The scaling of the primary equation of the *CHOPIT* model is the same as the *OP* ($\sigma^2_{\varepsilon_y} = 1$) and hence the parameter estimates are directly comparable. While the broad effect of covariates on outcomes remains the same across the two models—for example, levels of pain are generally lower for males compared to females, and for respondents who have a post-school qualification, the coefficients are notably changed. The coefficient on *male* is −0.163 in the *OP* results and −0.244 for the *CHOPIT* results. In absolute terms, this represents an increase of approximately 1.6 standard errors on the *OP* estimate. The estimated effect of any condition and grip strength on pain reduce by approximately 1 and 1.3 standard errors, respectively. Clearly, controlling for *DIF* appears to be important in these data. Note that the set of covariates used in the boundary equations of the *CHOPIT* model, $\boldsymbol{z}$, is the same as the set of covariates in the mean function, $\boldsymbol{x}$. The sets of boundary coefficients are presented in Appendix C. As noted above, the validity of the *CHOPIT* approach, however, rests on the assumption of *RC* and *VE*.

4.3 | Investigating the identifying assumptions of *RC* and *VE*

The empirical literature has attempted to investigate the assumptions of *RC* and *VE* in applications of the *CHOPIT* model. However, much of this literature is based on exploratory tests of the assumptions rather than a direct parametric test. For example, tests for *VE* have largely relied on indirect methods based on the relative rankings of vignettes by respondents to inform whether they are perceived in a consistent way across all survey participants. Results have tended to be ambiguous, for example, while Murray *et al*. (2003), King *et al*. (2004), Kristensen and Johansson (2008), Rice *et al*. (2011) and Hudson (2011) provide evidence in support of the assumption of *VE*, Datta Gupta *et al*. (2010), Peracchi and Rossetti (2012) and Bago d'Uva *et al*. (2011) find evidence against it.

Empirical tests for *RC* have tended to rely on the availability of *objective measures* of the concept of interest to which vignette-adjusted responses can be compared (e.g. objective measures of health). However, in practice, where objective measures exist these would offer a more plausible outcome to undertake comparison. When considering *RC*, Kapteyn *et al*. (2011) and Van Soest *et al*. (2011) provide supporting evidence, whereas Bago d'Uva *et al*. (2011) and Peracchi and Rossetti (2012) reject the null hypothesis. Van Soest and Vonkova (2014) illustrate how *RC* and *VE* be tested in the absence of objective measures. Using data from *SHARE*, they consider the ranking of a respondent's self-evaluation among the respondent's evaluations of vignettes and how these vary across socio-economic groups. These are then compared to the rankings obtained following an application of the *CHOPIT* approach. This leads to a test of the parametric assumptions inherent in the *CHOPIT* model when compared to a non-parametric alternative.

Of particular relevance to the current paper, Peracchi and Rossetti (2013) provide a *direct* test of the assumptions of *RC* and *VE* by exploiting the fact that under the two assumptions, the *CHOPIT* model is over-identified. The test, applied to health domains in the *SHARE*, rejects the joint assumptions of *RC* and *VE*. They show that in the absence of the restrictions implied by the joint test for *VE* and *RC* only reduced form parameters can be estimated. These are obtained from a set of hierarchical ordered response models estimated in the spirit of Pudney and Shields (2000); see Section 4.1.

Applying the restrictions imposed by *RC* and *VE* together with the reduced form estimates, a minimum distance estimator is used to recover the underlying parameters. For example, for a model with a dependent variable containing J ordered outcomes, l regressors and K vignettes, imposing the assumption of *RC* and *VE* together with the usual required location and scale normalisation restrictions imposed in *OP* models, leads to $s = \{J(l + 1) + 1\}(K + 1)$ parameters to be estimated. Note that we adopt a different notation to Peracchi and Rossetti (2013) to be consistent with the exposition set out in Section 3 (Peracchi and Rossetti (2013), assume $R + 1$ ordered outcomes ($J = R + 1$ in the above), J vignettes ($K = J$ in the above) and k regressors ($l = k$ in the above)). Fitting $K + 1$ (K vignettes plus the self-assessment) generalised ordered probit models leads to $q = (J - 1)(l + 1)(K + 1)$ reduced form parameters. These are composite parameters, since the coefficients in the thresholds and the mean function are not separately identifiable (Peracchi and Rossetti (2013) assume linear specifications of the boundary equations). Assuming RC and VE imposes $\{(J - 1)(l + 1) + l\}K + 2$ restrictions, implying there are $p = l + (J - 1)(l + 1) + 2K$ free parameters that can be recovered through a minimum distance approach. With one or more vignettes, the *CHOPIT* model is over-identified such that under the null hypothesis that *RC* and *VE* hold; $nQ_n(\widehat{\psi}) \Rightarrow \chi^2_{q-p}$, as $n \to \infty$. $Q_n(\widehat{\psi})$ is the minimum distance criterion evaluated at the solution $\widehat{\psi}$, with $q-p$ the number of over-identifying restrictions; see Peracchi and Rossetti (2013) for further details.

The mixed findings in support, or otherwise, for *RC* and *VE* clearly indicate that whether these two assumptions hold or not, will vary across surveys, the subgroups under comparison, the instruments of interest and the particular vignettes (wording and meaning) used. We set out below an simple to

implement test statistic of the assumptions of the *CHOPIT* model that can be readily used in applications of the approach.

5 | A SCORE TEST OF THE SPECIFICATION OF THE *CHOPIT* MODEL

This section develops score tests of the assumptions of *RC* and *VC*, both jointly and independently. The score test is appealing since only the model under the null requires estimation (i.e. the *CHOPIT* model). For such an approach to be valid though, the model under the alternative must be theoretically identified, which is not the case for the standard *CHOPIT* model. However, we can achieve identification with two amendments to the model. First, we restrict the variances σ_k^2 in Equation (4) to be unity, and second, we re-specify the first boundary equation (see Equation (3)) to be an exponential function of the boundary covariates. The approach of re-parameterising a model to facilitate a score test has precedents in the literature. For example, see Greene and McKenzie (2015) with regard to testing for a zero variance in nonlinear panel data models. The amendments, why they are required and their implications, are described in more detail below.

5.1 | A modified *CHOPIT* model

5.1.1 | Restriction on the variance of the vignette equations

It is common in the literature to allow σ_k^2 to be unrestricted or to be equivalent across all K vignettes; for example, see King *et al*. (2004). We adopt the normalisation; $\sigma_k^2 = 1, \forall k$. The variance parameters are generally not identified in ordered choice models (Greene (2018), pp. 730–731). Indeed, these parameters are numerically unidentified under the alternative hypothesis (i.e. failure of *RC* and *VE*) in the *CHOPIT* model. A scale parameter in each vignette of Equation (4) becomes identified under the null hypothesis through information about the cell probabilities and the externally imposed thresholds $\mu_{0,j}$ in equation (3); see Kapteyn *et al*. (2011), footnote 7 for discussion on this point.

The suggested score test (to be described in detail below), will essentially consist of an alternative 'model' comprising a series of independent *HOPIT* models for all of the $k = 0, \ldots, K$ constructs of interest: that is, the self-assessment under scrutiny ($k = 0$) as well as the available vignette outcomes ($k > 0$). As noted the scale of these models, in a case-by-case scenario, cannot be separately identified from the structural parameters of the model (without the use of extraneous information, as afforded by the vignettes in the usual *CHOPIT* set-up). As the score test requires the alternative model to be numerically identified, this requires that the variances in the separate vignettes equation(s) are all restricted to unity.

There are two points of note here. First, in Section 7.2, we consider relaxing this restriction in a Monte Carlo experiment to assess the size of the score test; and the results suggest that the test(s) perform well regardless of whether this restriction is imposed or not. Second, testing the assumption of *RC* requires that the boundary parameters are equivalent across the self-assessment and vignettes. Taking the first boundary equation and a single vignette as an example, then from Equations (4) and (5) imposing *RC* (due to the different treatment of the scale effects in both constructs) would require that

$$\gamma_{0,0}/(\sigma_0 = 1) \equiv \gamma_{0,1}/\sigma_1. \tag{7}$$

There are two obvious implications of Equation (7): the asymmetric treatment of the scale variables across constructs appears somewhat arbitrary; and what we actually estimate in practice is $\gamma_{0,0}/\sigma_0$ and one simply, again arbitrarily, sets $\sigma_0 = 1$. Thus for these reasons, and also to facilitate explicit testing of the *RC* assumption, we simply set all scale variables throughout equal to unity (although notwithstanding the identification issues raised above, this is likely to be inconsequential, and implicitly any scale effects where σ is not directly estimated, will be absorbed into the estimation of the relevant boundary parameters).

5.1.2 | Specification of the first boundary equations

The exponential form for the boundaries in model (3) is useful as it ensures the necessary ordering of the resulting boundary parameters. However, the implementation of this approach treats the first boundary parameters ($\mu_{0,k}$) asymmetrically with respect to the other boundary parameters (which enter in a linear, and non-linear fashion, respectively). Moreover, as with the treatment of the scale effects of the vignettes as described above, our alternative/generalised model (required for the score test, see below) requires that all separate *HOPIT* models for all constructs, be numerically identified. Clearly with $\boldsymbol{x} \equiv z$ and the first boundary equation specified as $\mu_{i,0} = z_i'\boldsymbol{\gamma}_0$, this will not be the case.

To yield a model numerically identified under the alternative (where *RC* and/or *VE* do not hold), we suggest the following modification to the specification of the first boundary parameter

$$\mu_{0,k} = \gamma_{0,k} + \exp(\tilde{z}'\tilde{\boldsymbol{\gamma}}_{0,k}), \quad k = 0, 1, \ldots, K, \tag{8}$$

where, again, $k = 0$ indexes boundary equations for the self-report of interest ($\mu_{0,0}, \ldots, \mu_{J-2,0}$) and $k = 1, \ldots, K$, the corresponding boundary equations for the vignettes ($\mu_{0,k}, \ldots, \mu_{J-2,k}$).

Due to the presence of the leading term, $\gamma_{0,k}$, $\mu_{0,k}$ is free to lie anywhere on the real number line (that is, there is no restriction that $\mu_{0,k} > 0$). The remaining $(J-2)$ boundaries follow an analogous specification to that set out in Equation (3).

The simple non-linear transformation of the first boundary equation (along with the scale restrictions described above) therefore numerically identifies a *HOPIT* model of the form described in Section 4.1 *without the need for exclusion restrictions* for all of the models/constructs in the system (that is, the self-report of interest as well as all vignettes).

Note that we parameterise the model such that the linear constant term, $\gamma_{0,0}$, enters in the main effects equation for y^*, and not in this first boundary equation. This follows from location normalisations in *OP*-type models which typically restrict the constant in the main equation to zero. Alternatively, one does not constrain this parameter in the main equation, but instead restrict the constant in the first boundary equation to zero. These approaches are numerically identical (Greene & Hensher, 2010).

5.1.3 | A generalised alternative model and score test

While leaving the underlying model essentially unchanged, the amended specification described above both improves model identification (by removing the linearity in the first boundary equation and restricting scale effects) while lending itself to a score test of the explicit assumptions of *RC* and *VE* in the usual *CHOPIT* set-up. That is, they allow for a generalised model to be considered (being numerically identified), consisting of a system of independent *HOPIT* models for all constructs, that collapse to the usual *CHOPIT* model under the set of parameter restrictions implied by both *RC* and *VE*.

More formally, following the amended specification we have the usual underlying index function for the self-report of interest, of the form

$$y^* = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon_y, \qquad \varepsilon_y \sim N(0,1), \tag{9}$$

together with the generalised form for the vignette equation(s) given by Equation (6). Under *VE*, that is, $\tilde{\boldsymbol{\alpha}}_k = 0, \ \forall k$, this form collapses to Equation (4) with the exception that now $\varepsilon_k \sim N(0,1)$, as described above.

To allow us to test for *RC* in this modified set-up, we have for all of the $k = 0,1, \ldots, K$ constructs, boundary equations of the form

$$\begin{aligned} \mu_{0,k} &= \exp(\%\boldsymbol{z}'\%\boldsymbol{\gamma}_{0,k}), \\ \mu_{j,k} &= \mu_{j-1,k} + \exp(\boldsymbol{z}'\boldsymbol{\gamma}_{j,k}), j = 1, \ldots, J-1. \end{aligned} \tag{10}$$

Note that in Equation (10) the treatment of $\mu_{0,k}$ differs from that in Equation (8) in that the constant term has (equivalently) been moved into the mean Equation (9). Finally, as noted above, *RC* implies equivalence of parameters $\boldsymbol{\gamma}_{0,k}$ and $\boldsymbol{\gamma}_{j,k}$ across all boundary equations for $k = 0,1, \ldots, K$. This then provides us with a simple parameter restriction test of *RC*. So here, under *RC*, Equation (10) collapses to simply

$$\begin{aligned} \mu_0 &= \exp(\%\boldsymbol{z}'\%\boldsymbol{\gamma}_0), \\ \mu_j &= \mu_{j-1} + \exp(\boldsymbol{z}'\boldsymbol{\gamma}_j), j = 1, \ldots, J-1. \end{aligned} \tag{11}$$

This amended specification of the standard *CHOPIT* model identifies separate *HOPIT* models for all of the $k = 0,1, \ldots, K$ constructs, defined by Equations (6), (9) and (10). Under the null of *RC* and *VE*, the set of generalised *HOPIT* models collapse to the (boundary-amended) *CHOPIT* model. As all of these restrictions have been shown to be simple linear ones, they can be tested both individually and jointly by using standard score tests based on the likelihoods of the respective unrestricted model evaluated at parameter values under the null; that the restricted *CHOPIT* model is correctly specified (Greene, 2018). Not only does the score test lends itself to separate and joint tests for the assumptions of *RC* and *VE*, it does not require estimation of the more complex alternative models. Full analytical derivatives of the appropriate score vector(s), the formal null and alternative hypotheses and the corresponding form of the score test, are all presented in Appendix B (although one could also use numerical derivatives). Gauss code to undertake the tests are available at http://github.com/aptech/chopitlib.

6 | THE AMENDED *CHOPIT* MODEL AND SCORE TEST APPLIED TO SELF-REPORTS OF PAIN

In this section, we consider the practical implications of the suggested amendments to the boundary equation(s) by comparing to commonly used specifications. We then apply our suggested score tests to our empirical example of modelling pain in the *SHARE* data.

6.1 | Amended specification

Before applying our score test to the modelling of self-reported pain, we investigate the implications of our suggested amendments to model estimates. Table 5 reports the results of applying the amended

TABLE 5 Comparison of *CHOPIT* estimated parameters with different boundary specifications

	Boundary equations					
	(1)		(2)		(3)	
	Amended exponentials		Standard exponentials		Linear	
$N = 3802$	Coef	SE	Coef	SE	Coef	SE
Structural parameters ($\widehat{\beta}$)						
Male	−0.272	0.070	−0.271	0.070	−0.274	0.070
AnyCond	0.658	0.059	0.657	0.058	0.656	0.058
Grip35	0.185	0.072	0.189	0.071	0.188	0.071
EducPS	−0.194	0.063	−0.193	0.063	−0.193	0.063
Age 66–75	0.090	0.058	0.085	0.058	0.085	0.058
Age > 75	0.169	0.097	0.164	0.097	0.164	0.097
Log-likelihood	−8939.60		−8939.63		−8940.31	

CHOPIT model to the self-reported data on pain from *SHARE* (column 1). The Table also compares these results to those obtained by the standard *CHOPIT* model (column 2) and a model where the boundary equations are all specified as linear functions of the covariates (column 3). While a linear specification fails to ensure the correct ordering of the boundaries, $\mu_{i,j},\ j = 0, 1, \ldots, J-2$, it has often been applied in empirical applications (see, for example, Bago d'Uva *et al*., 2008). We include this specification for completeness.

The results illustrate the difference in model estimates from changing the specification of the boundaries. To ensure that the estimated coefficients are comparable, all models restrict the variance of the error term in the vignette equation σ_k^2 to unity as per the amended *CHOPIT* model. Accordingly, the scale of the estimates differ from the standard *CHOPIT* model (for which σ_k^2 is freely estimated) and hence are not directly comparable to the estimates provided in Table 4. However, a comparison of the relative effects of coefficients (to remove the scaling of parameter estimates) reveals very similar results. For example, the estimated coefficient for male relative to any conditions (*AnyCond*) in the *CHOPIT* model is $\frac{-0.244}{0.588} = -0.41$ (Table 4). For the amended specification in Table 5, the corresponding relativity is $\frac{-0.271}{0.657} = -0.41$. Similar relative estimates are apparent for the other covariates. While restricting the vignette variance to unity changes the scaling of the estimates, their relative interpretation remains the same as in the standard *CHOPIT* model. Accordingly, marginal effects will be unaffected by the scaling. Note that full results for the three specifications, including the boundary equations, are reported in Appendix C, Table C.

To further investigate the model implications of the amended specification, Table C3 of the Appendix presents averaged estimated boundaries for the three approaches. The standard exponential and linear specifications provide similar estimates of μ_0, μ_1 and μ_2. Those of the amended exponential approach are substantially larger, but *by a constant amount* relative to those of the standard (0.999) and linear (approximately 1.008) approaches. The following two panels of the table consider the location of the boundaries with respect to the estimated linear index, $\boldsymbol{x}'\widehat{\boldsymbol{\beta}}$, and separately the estimated vignette constant term (α_1 in Equation (4)). As with standard *OP*-type models, it is not just the value of the index function defining y^* that is of relevance, but the position of this index in relation to the boundaries that are essential for generating predictions from the model. Across the three specifications, we see that these quantities are essentially identical indicating (at least approximately) equivalence of the three approaches.

Further evidence of these findings are reported in Tables C4–C6. Table C4 contains the sample correlations of estimated boundary values and y^* values. These are clearly all very highly correlated, and in all cases close to one. Table C5 considers estimated probabilities. The averages of these are identical across specifications, and the correlation across the individual estimates are 1, or very close to 1, in all cases. Finally, Table C6 contains the implied partial effects for each specification. Again these are essentially equivalent across model specifications.

In summary, while individual parameter estimates may vary across the different boundary specifications, for each essentially the same model results. Importantly, the amended specification of the boundaries does not unduly enforce any implicit/explicit restriction(s) on the model that might adversely affect results and tests statistics.

6.2 | Tests of *RC* and *VE* for self-reports of pain

An application of the score tests to *SHARE* data is presented in Table 6. The data and specification follow that used in column (1) of Table 5. However, we make use of all possible permutations of the three available vignettes. The joint test of the null of both *RC* and *VE* is rejected at conventional levels for all vignettes used singularly or in combination. The test for *VE* alone (assuming *RC* holds) fails to reject the null when vignettes *V*1 or *V*3 are used singularly and when vignettes *V*2 and *V*3 are used in combination. However, *VE* is rejected in all other combinations. When we consider only *RC* (assuming *VE* holds) the score test rejects the null for all vignettes and their combinations, as does the joint test. The results emphasise the importance of testing for the identifying assumptions of *RC* and *VE* in applications of the *CHOPIT* model when attempting to correct for *DIF*.

7 | MONTE CARLO EVIDENCE

To fully explore both the general implications of the proposed change in boundary specification and the score tests, we consider a series of Monte Carlo experiments. Throughout we simulate data by drawing from *SHARE* data the set of covariates used in the empirical example described in Section 6.1.

The Monte Carlo experiment simulates data as follows: (i) use all $N = 3802$ observations and their corresponding covariates $\boldsymbol{x}_i$ from the *SHARE* data, (ii) construct the latent outcome $y^* = \tilde{\boldsymbol{x}}'\tilde{\boldsymbol{\beta}} + \varepsilon_y$ using the parameter estimates from the empirical example presented in Section 6.2 as column (2)

TABLE 6 Score tests for combinations of vignettes ($J = 4$)

	score_{joint}		score_{VE}		score_{RC}	
Vignette (V)	χ^2 (*df*)	*p−val*	χ^2 (*df*)	*p−val*	χ^2 (*df*)	*p−val*
V1	310.3 (26)	0.000	8.843 (6)	0.183	297.8 (20)	0.000
V2	195.6 (26)	0.000	14.26 (6)	0.027	190.1 (20)	0.000
V3	88.64 (26)	0.000	10.48 (6)	0.106	76.92 (20)	0.000
V1 & V2	501.9 (52)	0.000	91.07 (12)	0.000	488.5 (40)	0.000
V1 & V3	426.8 (52)	0.000	45.36 (12)	0.000	411.5 (40)	0.000
V2 & V3	328.7 (52)	0.000	15.54 (12)	0.213	311.2 (40)	0.000
V1, V2 & V3	571.5 (78)	0.000	97.42 (18)	0.000	553.2 (60)	0.000

of Table 5 together with a randomly generated standard normal error, $N(0,1)$, (iii) the latent vignette outcome, $v^*_{i,1}$, is constructed by random normal draws from the distribution $N(\alpha,1)$, with α set to the value obtained by estimation of the model in column (2) of Table 5 (full model estimates including boundary parameters are provided in Table C), (iv) the corresponding observed outcomes, $y_i, v_{i,1}$, are then constructed from their latent counterparts together with knowledge of the boundary parameters $(\tilde{\boldsymbol{\gamma}}_{i,0}, \ldots, \tilde{\boldsymbol{\gamma}}_{i,2})$ estimated from the model reported in column (2), Table 5. *CHOPIT* estimation of the simulated y_i and $v_{i,1}$ on the set of covariates $\boldsymbol{x}_i$ is then undertaken. This is repeated for $M = 2000$ simulations (as are all other Monte Carlo subsequent experiments) and results for models for which convergence was achieved (S) summarised in Table C8 (convergence was deemed to have failed after 500 maximum likelihood iterations).

7.1 | Boundary specification

We first illustrate the difference that the amended specification has on the estimated vector of coefficients, $\tilde{\boldsymbol{\beta}}$, when compared to standard exponential or linear specifications. Typically, these are the parameters of most interest in empirical applications. Table C8 presents the results. Data are generated assuming standard exponential specification of the boundaries and estimated separately assuming amended exponential, standard exponential and linear specifications.

Monte Carlo coefficients are close to their 'true' values across the different specifications of the boundaries. This can be seen by the small values reported for mean bias. The 5% coverage rate is also within expected range across all parameter estimates. However, while the standard exponential and amended exponential specifications display high convergence rates with $S/M = 0.998$ for both, the convergence rate for the linear specification of the boundaries is low ($S/M = 0.289$) illustrating the fragility of that specification. This reflects the lack of identification through not imposing non-linearity in the boundaries.

7.2 | Finite sample performance of the score tests

We evaluate the performance of the score tests by generating data under the null in a similar way to that described above again based on the estimated coefficients from the empirical models presented in Table 5. We consider three sets of test size experiments, where we generate under the null hypothesis with linear boundaries (column (3), Table 5); with standard exponential boundary thresholds (2); and amended exponential boundaries (1). We then conduct the tests as if the boundaries were of the amended exponential form.

When the data generating process (*DGP*) is as the test assumes, (amended exponential boundaries), the tests are correctly sized for the score_{joint} and score_{RC} variants (Table 7). The score_{VE} variant appears to be marginally undersized (at 4.10% for a nominal 5%). When the true *DGP*consists of linear thresholds or standard exponentials, the score_{joint} and score_{RC} tests appear to be marginally oversized, however, overall the tests remain within an acceptable range. Note that relaxing the assumption of $\sigma^2_k = 1, \ \forall k$ does not materially affect size results, as evidenced in Appendix C Table C7.

We next consider power experiments using a similar Monte Carlo experimental set-up as above, but where the assumptions of *RC* and *VE* are violated. In the experiments for departures from *RC*, we perturb the parameter vector corresponding to the boundary equations for the vignettes (i.e. $\boldsymbol{\gamma}_{j1}$ in Equation (5)), perturbing at increasing values away from zero. These are undertaken for a model generated assuming amended exponential boundaries. This is achieved by first generating a vector of standard normal random variates of the same dimension as $\boldsymbol{z}(= \boldsymbol{x})$. These draws are held fixed.

We then move away from the null of *RC* by perturbing γ_{j1} in the vignette equation only, by adding successively larger quantities to the value under the null. These quantities are dictated by the set of (fixed) random normal variates with increases achieved by multiplying by a scalar, s_{rc} in the range $0.0 \leq s_{rc} \leq 0.20$. This ensures greater departures from the null for increasing values of s_{rc}. For violations of *VE*, a vector of random variates of dimension $\boldsymbol{x}$ is first drawn. We then perturb the corresponding implicit vector of zero coefficients, $\tilde{\boldsymbol{\alpha}}$ (under the null), on the covariates $\boldsymbol{x}$ (Equation (6)) by multiplying the random draws by a scalar, s_{ve}, and substituting these as parameters for $\tilde{\boldsymbol{\alpha}}$. This process is repeated for successively larger values of s_{ve} such that $0.0 \leq s_{ve} \leq 0.50$. For the joint experiments, we simultaneously employ both approaches.

The results are then summarised as power curves, plotting rejection probabilities against the size of the perturbation from the null of zero. Three curves are shown: a joint test for *RC* and *VE*; and separate ones for *RC* and *VE*.

The left-hand side of Figure 2 displays the power curves for all three tests when we violate *RC* only. The curves are well behaved. Departures from *RC* results in *S*-shaped power curves for the test of *RC* alone and for the joint test (*RC* and *VE*). As expected, the test for *RC* uniformly dominates that for the joint test. This is due to the test maximising power in the single direction while the joint test is also testing for *VE*. In comparison, the test for *VE* remains fairly flat over the range of values for which *RC* is violated. This is encouraging as clearly *VE* is exhibiting some power as a general specification test, when *VE* is not failing but *RC* is. The right-hand side of Figure 2 presents the power curves when we violate *VE* alone. While the power curve for the test of *VE* adopts an approximate *S*-shape, it appears relatively sensitive (that is powerful) to small departures from *VE* and increases fairly rapidly across relatively small increments. Moreover, departures from *VE* are also reflected in the test for *RC*. The joint test also adopts the *S*-shaped curve, but rejects less than the test for *VE* alone, again due to the latter only testing for departures from the null in that particular direction.

A priori one would not expect increasing departures from the null with respect to *VE* (*RC*) to affect the power properties when testing for *RC* (*VE*). However, it has been well-known that it is possible to reject a false model against an alternative model, even if that alternative model is not correct (Davidson & MacKinnon 1987). In this sense, such tests that tend to reject a false model in favour of a similarly false alternative model, are often referred to as general specification tests. In this sense, the test for *RC* can be considered a useful general specification test, as it tends to similarly pick-up departures in the direction of *VE*. However, the same cannot be said of the test for *VE*, and power only marginally increases with departures from the null with respect to *RC*. It is unclear what the specific reasons for these results are, and also whether the results will hold more generally.

A probable reason for the strong performance of the *RC* test in identifying departures from the null of *VE*, is that in misspecifying the assumed outcome function in the vignette equation(s), may result in the boundary parameters having to adjust to ensure their relationship with this outcome function. In as such, by imposing non-*VE*, this may also manifest itself as a form of *RC* violation. In contrast, there appears to be a less persuasive argument for the reverse situation. By moving the boundary parameters further away from the null, the outcome function in the vignettes equation has much more limited ability to move itself to maintain the assumed relationship with the boundary parameters as implied by the null model. Prior

TABLE 7 Size results, at 5% nominal size; $M = 2000$

Boundary equations	score_{joint}	score_{VE}	score_{RC}
Linear	0.0585	0.0555	0.0580
Standard exponential	0.0560	0.0485	0.0560
Amended exponential	0.0495	0.0410	0.0495

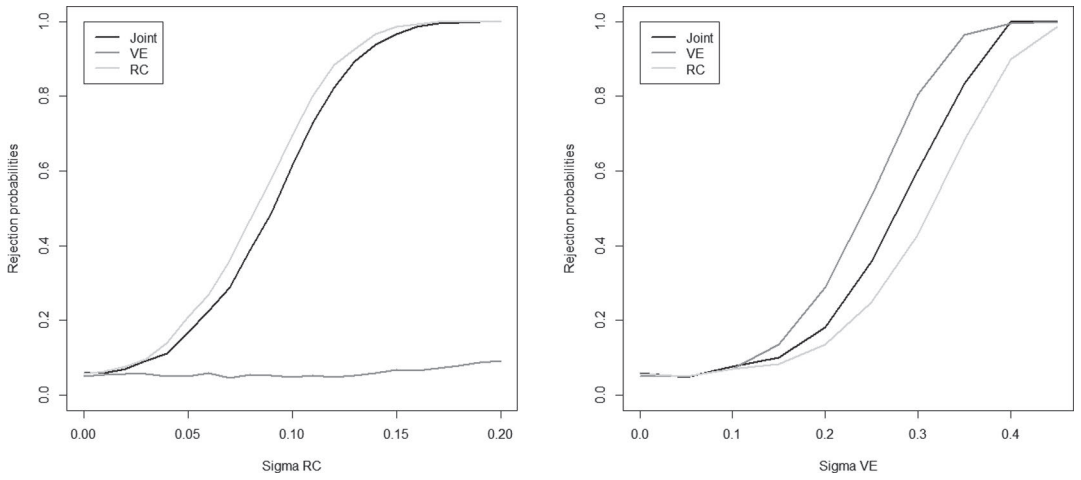


FIGURE 2 Power curves for rejection probabilities for departures from *RC* and *VE*

to undertaking individual tests, it would appear appropriate to undertake a joint test of both *RC* and *VE*; if this fails, then the individual tests for *RC* and *VE* may be informative of the reason for model failure.

Three-dimensional planes of rejection rates against simultaneous departures from both *RC* and *VE* are shown in Figure 3. The joint test and the test for *RC* perform in a similar fashion, and reflective of the results of the power curves, the joint test appears to be dominated by the test for *RC*. While the test for *VE* appears to respond to departures from the null of response consistency ($\tilde{\boldsymbol{\alpha}}_k \neq 0$), the test remains fairly flat over the range of values for which *RC* is violated.

In summary, the experiments show: (i) the individual tests have greatest power in their particular direction, (ii) both individual tests have increasing power in distance from the null with respect to the alternative violation, (iii) power increases with distance away from the null in all cases; and (iv) the joint test has increasing power in all directions, and is maximised with simultaneous deviations from the null in both directions. Note that allowing σ_k^2 to be unrestricted does not impact the power of the test substantially, as shown in Appendix C, Figures C1 and C2.

7.3 | Comparison of the proposed Score test with the minimum distance estimator approach of Peracchi and Rossetti (2013)

Section 3 of Peracchi and Rossetti (2013) investigates the finite sample performance of their minimum distance estimator using a Monte Carlo experiment. We undertake the same Monte Carlo exercise to compare their results to our score tests (full details of the Monte Carlo experiment can be found on p712, Peracchi and Rossetti (2013)). Table 8 presents the results for the situation where there are $J = 2$ threshold boundaries, $k = 1{,}2$ vignettes and a single covariate: $x = z$ (note that Peracchi and Rossetti (2013) use a different notation by indexing boundaries as $r = 1, \ldots, R$, vignettes as $j = 1, \ldots, J$ and exogenous regressors $k = 1, \ldots, K$). The sample size for the draws is $N = 250$, and each Monte Carlo exercise consists of $M = 1{,}000$ runs (as per Peracchi and Rossetti (2013)).

The first column of results presents rejections rates at a nominal 5% level for the minimum distance estimator. These are followed by the joint score test and the separate score tests for *VE* and *RC*. The row labelled H_0 reports observed size of the various tests at the 5% level. All rejection frequencies are close to nominal level under the null. The panel H_1 shows results for departures from *VE*, but where *RC* holds; H_2 for departures from *RC* (where *VE* holds) and the final panel, H_3, departures from both

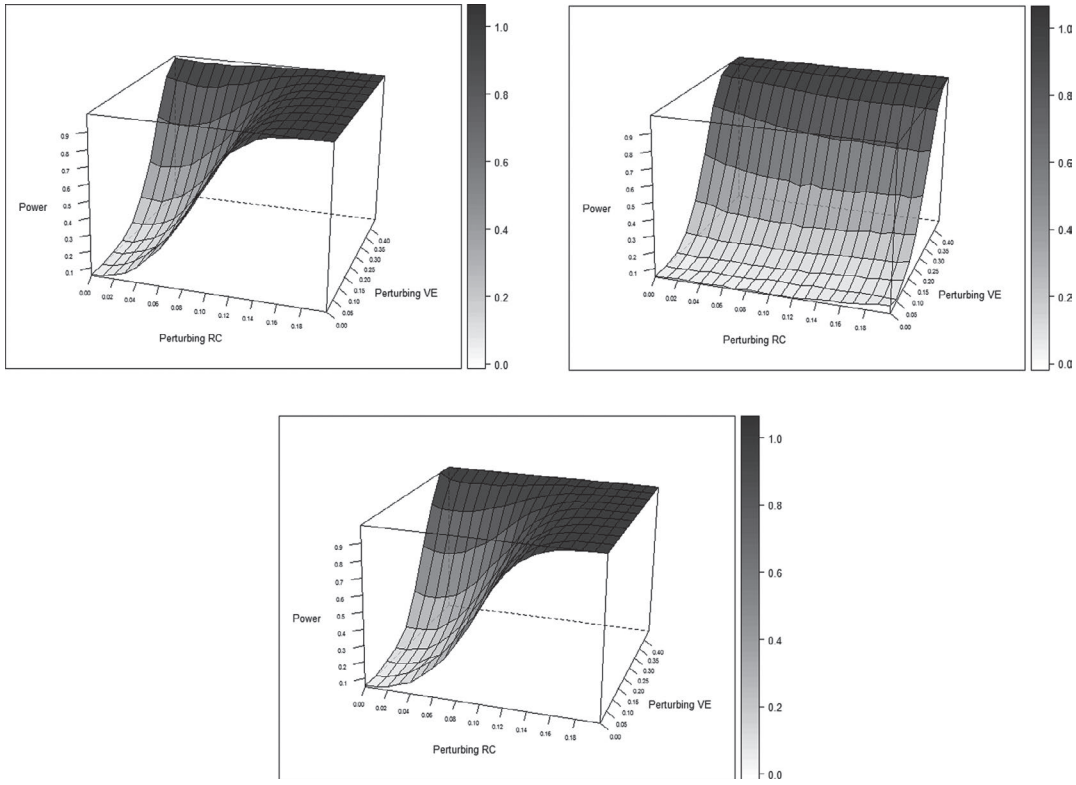


FIGURE 3 Power planes for rejection probabilities for departures from *RC* (lhs top), *VE* (rhs top) and a joint test (bottom)

VE and *RC* simultaneously. Rejection rates are reported for increasing departures from the null. The score joint test displays greater power than the test of Peracchi and Rossetti. This is the case for departures from the null for *VE* and *RC* separately and for joint departures. While the test of Peracchi and Rossetti lacks power when only a single vignette is used (but not with multiple vignettes), this is not the case for the score test which generally increases in power with increasing departure from the null even with a single vignette. The power of the test, however, also generally increases with when including a second vignette. When comparing the rejection rates across the three score tests for the different departures from the null, results closely reflect those of the Monte Carlo exercise reported and summarised in Section 7.2 above. Again, the tests have greatest power in their particular direction; power generally increases with increasing departure from the null and the score test for *VE* has the lowest power amongst the three tests when considering violations in their own respective direction.

8 | CONCLUSIONS

Inter-individual comparison of phenomena such as health status or life satisfaction that are typically self-reported on an ordered categorical scale are often subject to differential item functioning due to survey respondents' adopting different response scales. Vignettes are increasingly being collected alongside self-reports to anchor such scales and provide greater comparability across individuals. This is particularly relevant when undertaking cross-country comparisons where differences in cultural norms may lead to the use of very different response scales.

TABLE 8 Monte Carlo experiment and comparison with Peracchi and Rossetti (2013) — *p*-values of the score test

	Results for $J = 2, K = 1$ and $N = 250$				Results for $J = 2, K = 2$ and $n = 250$			
	Score test				Score test			
	P&R	**Joint**	***VE***	***RC***	**P&R**	**Joint**	***VE***	***RC***
H_0	0.057	0.055	0.063	0.056	0.056	0.055	0.055	0.052
H_1: Failure of *VE*								
$\beta_1 = 0.1$	0.062	0.066	0.065	0.068	0.056	0.059	0.051	0.057
$\beta_1 = 0.2$	0.070	0.060	0.055	0.067	0.055	0.065	0.050	0.060
$\beta_1 = 0.4$	0.057	0.075	0.072	0.063	0.080	0.068	0.068	0.053
$\beta_1 = 0.6$	0.067	0.081	0.109	0.060	0.107	0.094	0.118	0.064
$\beta_1 = 0.8$	0.055	0.147	0.120	0.084	0.172	0.119	0.180	0.067
$\beta_1 = 1.0$	0.068	0.203	0.327	0.093	0.254	0.265	0.419	0.083
H_2: Failure of *RC*								
$\alpha_{11} - \alpha_{01} = 0.1$	0.055	0.060	0.059	0.069	0.052	0.082	0.051	0.066
$\alpha_{11} - \alpha_{01} = 0.2$	0.059	0.108	0.070	0.111	0.047	0.100	0.057	0.102
$\alpha_{11} - \alpha_{01} = 0.4$	0.059	0.246	0.064	0.261	0.061	0.245	0.072	0.264
$\alpha_{11} - \alpha_{01} = 0.6$	0.074	0.472	0.071	0.506	0.103	0.482	0.082	0.522
$\alpha_{11} - \alpha_{01} = 0.8$	0.059	0.725	0.081	0.753	0.104	0.708	0.094	0.746
$\alpha_{11} - \alpha_{01} = 1.0$	0.056	0.895	0.089	0.916	0.151	0.890	0.098	0.916
H_3: Failure of *VE* & *RC*								
$\beta_1 = \alpha_{11} - \alpha_{01} = 0.1$	0.065	0.080	0.072	0.070	0.049	0.079	0.060	0.071
$\beta_1 = \alpha_{11} - \alpha_{01} = 0.2$	0.074	0.095	0.062	0.098	0.059	0.110	0.074	0.112
$\beta_1 = \alpha_{11} - \alpha_{01} = 0.4$	0.068	0.245	0.080	0.249	0.095	0.266	0.081	0.285
$\beta_1 = \alpha_{11} - \alpha_{01} = 0.6$	0.081	0.512	0.104	0.506	0.149	0.548	0.161	0.577
$\beta_1 = \alpha_{11} - \alpha_{01} = 0.8$	0.063	0.741	0.128	0.732	0.243	0.843	0.319	0.842
$\beta_1 = \alpha_{11} - \alpha_{01} = 1.0$	0.057	0.918	0.167	0.881	0.312	0.976	0.530	0.971

We illustrate the effects of correcting for *DIF* using information on vignettes in an example on self-assessments of levels of pain. We stress that the legitimate use of the vignette approach relies on the two assumptions of *RC* and *VE*. In light of this, the paper then develops single and joint tests of these assumptions based on a score approach.

Implementation of the test is within the parametric *CHOPIT* model that imposes a hierarchical structure on the boundary equations to preserve coherency of the model's probabilities. The score approach requires the model to be identified under the alternative hypothesis; the null being that *RC* and *VE* hold. This is achieved by augmenting the specification of the boundary equations by including an exponential function in the first boundary equation, together with restricting the variance of the vignette equations to unity. Such changes are innocuous in terms of parameter estimates (and marginal effects) of the coefficients in the mean function, which are typically the focus of empirical work.

An advantage of the test is its ease of implementation, requiring estimation of the restricted model under the null only. This is undertaken using the *CHOPIT* model. The test may be seen as a complement, or alternative, to Peracchi and Rossetti (2013) who also develop a joint test of *RC* and *VE*.

However, an advantage of the current approach is that separate tests of both *RC* and *VE* are available, and also that there was evidence of the current test(s) being more powerful.

We find that for the empirical example of self-assessed pain the joint null of *RC* and *VE* is rejected, such that the adjustments using vignettes may be unreliable. Monte Carlo simulations drawn from these data show that the tests have good size and power properties in finite samples, particularly for the joint test and the individual test for *RC*. Our results suggest that the assumption of *VE* may be more problematic in empirical applications than *RC*. This finding mirrors that of Peracchi and Rossetti (2013). In particular, failure of *VE* may also be picked up through rejection of *RC*. This is an area where the design of vignette questions to aid respondents' common understanding of the descriptions of the hypothetical individuals may best improve vignette equivalence. The majority of applications of the vignette approach rely on cross-sectional data. Future research might consider extensions to panel data to control more fully for individual unobserved heterogeneity in reporting behaviour. For example, extensions of such an approach to the modelling of ordered self-assessed health outcomes is provided by Bartolucci and Bacci (2014). In the context of applications to the vignette approach, where feasible, a flexible treatment of reporting heterogeneity may make reliance on the assumptions of *RC* and *VE* more plausible.

REFERENCES

Aldrich, J.H. & McKelvey, R.D. (1977) A method of scaling with applications to the 1968 and 1972 presidential elections. *American Political Science Review*, 71(1), 111–130.

Angelini, V., Cavapozzi, D. & Paccagnella, O. (2011) Dynamics of reporting work disability in Europe. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(3), 621–638.

Angelini, V., Cavapozzi, D., Corazzini, L. & Paccagnella, O. (2012) Age, health and life satisfaction among older Europeans. *Social Indicators Research*, 105, 293–308.

Angelini, V., Cavapozzi, D., Corazzini, L. & Paccagnella, O. (2014) Do Danes and Italians rate life satisfaction in the same way? Using vignettes to correct for individual-specific scale biases. *Oxford Bulletin of Economics and Statistics*, 76(5), 643–666.

Bago d'Uva, T., Van Doorslaer, E., Lindeboom, M. & O'Donnell, O. (2008) Does reporting heterogeneity bias the measurement of health disparities? *Health Economics*, 17(3), 351–375.

Bago d'Uva, T., Lindeboom, M., O'Donnell, O. & Van Doorslaer, E. (2011) Slipping anchor? Testing the vignettes approach to identification and correction of reporting heterogeneity. *Journal of Human Resources*, 46(4), 875–906.

Bartolucci, F. & Bacci, S. (2014) Longitudinal analysis of self-reported health status by mixture latent auto-regressive models. *Journal of the Royal Statistical Society. Series A*, 63(2), 267–288.

Boes, S. & Winkelmann, R. (2006) Ordered response models. *AStA Advances in Statistical Analysis*, 90(1), 167–181.

Clauser, B.E. & Mazor, K.M. (1998) Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, 17(1), 31–44.

Datta Gupta, N., Kristensen, N. & Pozzoli, D. (2010) External validation of the use of vignettes in cross-country health studies. *Economic Modelling*, 27(4), 854–865.

Davidson, R. & MacKinnon, J. (1987) Implicit alternative and the local power of test statistics. *Econometrica*, 55(6), 1305–1329.

Greene, W. (2018) *Econometric analysis*. London: Pearson. Available from: https://books.google.co.uk/books?id=xGZRvgAACAAJ

Greene, W. & Hensher, D. (2010) *Modeling ordered choices*. Cambridge: Cambridge University Press.

Greene, W. & McKenzie, C. (2015) An fLMg test based on generalized residuals for random effects in a nonlinear model. *Economics Letters*, 127, 47–50. Available from: http://www.sciencedirect.com/science/article/pii/S0165176514004923

Greene, W., Harris, M., Hollingsworth, B. & Maitra, P. (2014) A latent class model for obesity. *Economics Letters*, 123, 1–5.

Grol-Prokopczyk, H., Freese, J. & Hauser, R.M. (2011) Using anchoring vignettes to assess group differences in general self-rated health. *Journal of Health and Social Behavior*, 52(2), 246–261.

Groseclose, T., Levitt, S.D. & Snyder, J.M. (1999) Comparing interest group scores across time and chambers: adjusted ada scores for the us congress. *American Political Science Review*, 93(1), 33–50.

Holland, P.W. & Thayer, D.T. (1988) Differential item performance and the Mantel-Haenszel procedure. In Wainer, H. and Braun, H.I. (Eds.) *Test validity* (pp. 129–145). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Holland, P. & Weiner, H. (1993) *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Hudson, E. (2011) Examining the effect of socioeconomic status on child health using anchoring vignettes. Technical report, Unpublished.

Jones, A., Rice, N. & Robone, S. (2018) Anchoring vignettes and crosscountry comparability: an empirical assessment of self-reported mobility. In: Baltagi, B. and Moscone, F. (Eds.) *Health econometrics: Contributions to economic analysis* (Vol 294, pp. 145–174). Bingley, UK: Emerald Publishing, chapter 7.

Kapteyn, A., Smith, J. & Van Soest, A. (2007) Vignettes and self-reports of work disability in the United States and the Netherlands. *The American Economic Review*, 97, 461–473.

Kapteyn, A., Smith, J., van Soest, A. & Vonkova, H. (2011) Anchoring vignettes and response consistency. Technical Report WR-840, RAND.

Kapteyn, A., Smith, J. & Van Soest, A. (2013) Are Americans really less happy with their incomes? *Review of Income and Wealth*, 59(1), 44–65.

King, G., Murray, C., Salomon, J. & Tandon, A. (2004) Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, 98(1), 191–207.

Kristensen, N. & Johansson, E. (2008) New evidence on cross-country differences in job satisfaction using anchoring vignettes. *Labour Economics*, 15(1), 96–117.

Murray, C., Tandon, A., Salomon, J.A., Mathers, C. D. & Sadana, R. (2002) Cross-population comparability of evidence for health policy. Health Systems Performance Assessment: Debates, Methods and Empiricism, pp. 705–713.

Murray, C., Ozaltin, E., Tandon, A., Salomon, J., Sadana, R., Chatterji, S. et al. (2003) Empirical evaluation of the anchoring vignettes approach in health surveys. *Health Systems Performance Assessment: Debates, Methods and Empiricism*, 369, 399.

Paccagnella, O. (2011) Anchoring vignettes with sample selection due to non-response. *Journal of the Royal Statistical Society, Series A*, 3(174), 665–687.

Paccagnella, O. (2013) Modelling individual heterogeneity in ordered choice models: anchoring vignettes and the Chopit model. *QdS Journal of Methodological and Applied Statistics*, 15, 69–94.

Peracchi, F. & Rossetti, C. (2012) Heterogeneity in health responses and anchoring vignettes. *Empirical Economics*, 42(2), 513–538.

Peracchi, F. & Rossetti, C. (2013) The heterogeneous thresholds ordered response model: identification and inference. *Journal of the Royal Statistical Society Series A*, 176(3), 703–722.

Pudney, S. & Shields, M. (2000) Gender, race, pay and promotion in the British nursing profession: estimation of a generalized ordered probit model. *Journal of Applied Econometrics*, 15, 367–399.

Rice, N., Robone, S. & Smith, P. (2011) Analysis of the validity of the vignette approach to correct for heterogeneity in reporting health system responsiveness. *The European Journal of Health Economics*, 12(2), 141–162.

Rice, N., Robone, S. & Smith, P. (2012) Vignettes and health systems responsiveness in cross-country comparative analyses. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(2), 337–369.

Rossi, P., Gilula, Z. & Allenby, G. (2001) Overcoming scale usage heterogeneity: a Bayesian hierarchical approach. *Journal of the American Statistical Association*, 453(96), 20–31.

Shepard, L., Camilli, G. & Averill, M. (1981) Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics*, 6(4), 317–375.

Sirven, N., Santos-Eggimann, B. & Spagnoli, J. (2012) Comparability of health care responsiveness in Europe. *Social Indicators Research*, 2(105), 255–271.

Soloman, J., Tandon, A. & Murray, C. (2004) Comparability of selfrated health: cross-sectional multi-country survey using anchoring vignettes. *British Medical Journal*, 328, 258–260.

Swaminathan, H. & Rogers, H.J. (1990) Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27(4), 361–370.

Tay, L., Vermunt, J. & Wang, C. (2013) Assessing the item response theory with covariate (IRT-C) procedure for ascertaining differential item functioning. *International Journal of Testing*, 13, 201–222.

Terza, J. (1985) Ordered probit: a generlization. *Communications in Statistics - A. Theory and Methods*, 14, 1–11.

Van Soest, A. & Vonkova, H. (2014) Testing the specification of parametric models by using anchoring vignettes. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 177(1), 115–133.

Van Soest, A., Delaney, L., Harmon, C., Kapteyn, A. & Smith, J. (2011) Validating the use of anchoring vignettes for the correction of response scale differences in subjective questions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(3), 575–595.

Vonkova, H. & Hullegie, P. (2011) Is the anchoring vignettes method sensitive to the domain and choice of the vignette? *Journal of the Royal Statistical Society, Series A*, 174(3), 597–620.

How to cite this article: Greene WH, Harris MN, Knott RJ, Rice N. Specification and testing of hierarchical ordered response models with anchoring vignettes. *J R Stat Soc Series A*. 2020;00:1–34. https://doi.org/10.1111/rssa.12612

APPENDIX A

VIGNETTES AND THE CHOPIT MODEL

A.1. Vignette descriptions

The three vignettes available for pain within *SHARE* are:

> Vignette ($m1$): "Karen has a headache once a month that is relieved after taking a pill. During the headache she can carry on with her day-to-day affairs. Overall in the last 30 days, how much of bodily aches or pains did Karen have?"
>
> Vignette ($m2$): "Maria has pain that radiates down her right arm and wrist during her day at work. This is slightly relieved in the evenings when she is no longer working on her computer. Overall in the last 30 days, how much of bodily aches or pains did Maria have?"
>
> Vignette ($m3$): "Alice has pain in her knees, elbows, wrists and fingers, and the pain is present almost all the time. Although medication helps, she feels uncomfortable when moving around, holding and lifting things. Overall in the last 30 days, how much of bodily aches or pains did Alice have?"

A.2. CHOPIT model probabilities

Expressions for the probabilities derived for the self-report from the *CHOPIT* model.

$$
\begin{aligned}
P_{0,0} &= \Phi[\exp(\tilde{z}'\tilde{\gamma}_0) - x'\beta], \\
P_{1,0} &= \Phi[\exp(\tilde{z}'\%\tilde{\gamma}_0) + \exp(z'\gamma_1) - x'\beta] - \Phi[\exp(\tilde{z}'\tilde{\gamma}_0) - x'\beta], \\
P_{2,0} &= \Phi[\exp(\tilde{z}'\tilde{\gamma}_0) + \exp(z'\gamma_1) + \exp(z'\gamma_2) - x'\beta] \\
&\quad - \Phi[\exp(\tilde{z}'\%\tilde{\gamma}_0) + \exp(z'\gamma_1) - x'\beta], \\
&\vdots \\
P_{J-1,0} &= \Phi\left[x'\beta - \exp(\tilde{z}'\tilde{\gamma}_0) - \sum_{j=1}^{J-2} \exp(z'\gamma_j)\right].
\end{aligned}
$$

Corresponding probabilities for the vignette outcome(s), for $k = 1, \ldots, K$, are

$$
\begin{aligned}
P_{0,k} &= \Phi[(\exp(\tilde{z}'\tilde{\gamma}_0) - \alpha_k)/\sigma], \\
P_{1,k} &= \Phi[(\exp(\tilde{z}'\tilde{\gamma}_0) + \exp(z'\gamma_1) - \alpha_k)/\sigma] \\
&\quad - \Phi[(\exp(\tilde{z}'\tilde{\gamma}_0) - \alpha_k)/\sigma], \\
P_{2,k} &= \Phi[(\exp(\tilde{z}'\tilde{\gamma}_0) + \exp(z'\gamma_1) + \exp(z'\gamma_2) - \alpha_k)/\sigma] \\
&\quad - \Phi[(\exp(\tilde{z}'\tilde{\gamma}_0) + \exp(z'\gamma_1) - \alpha_k)/\sigma], \\
&\vdots \\
P_{J-1,k} &= \Phi\left[(\alpha_k - \exp(\tilde{z}'\tilde{\gamma}_0) - \sum_{j=1}^{J-2} \exp(z'\gamma_j)) \Big/ \sigma\right].
\end{aligned}
$$

APPENDIX B

SCORE VECTORS FOR THE TESTS OF RESPONSE CONSISTENCY AND VIGNETTE EQUIVALENCE

In this Appendix, we set out formally the various score vectors required for the score tests described in Section 5 and how these are combined for the separate tests of *RC* and *VE*, together with the joint test of both *RC* and *VE*. Note that although analytical expressions are given, one could also use numerical derivatives.

First we derive the score vector for the restricted *CHOPIT* model (Equations (4) and (11)), and show how particular elements can be adapted to derive the proposed score statistic(s). That is, the score is derived for the boundary-amended *CHOPIT* model under the assumption, or restrictions, that both *VE* and *RC* hold. Testing for *VE*, on the assumption that *RE* holds, in this setting requires replacing the appropriate elements of the score corresponding to the derivatives with respect to α_k, with those of the more general model—Equation (6)—which does not impose *VE*. As usual, the score test evaluates the score for the more general (here the model allowing for *VE*), but at parameter values under the null hypothesis (assuming *VE*).

Similarly, testing for *RC*, on the assumption that *VE* holds, is based on the full score of this restricted boundary-amended *CHOPIT* model, but now replacing the appropriate elements corresponding to the derivatives with respect to the boundary parameters implied by the restricted specification of Equation (11), with those implied by the generalised version of Equation (10). Again, the test is based on evaluating the generalised score at parameter values under the null hypothesis.

Finally, the joint test of both *VE* and *RC* involves replacing the elements of the restricted *CHOPIT* score with respect to both α_k and the boundary parameters, with those implied by the more general specifications. Again, the test involves evaluating the score of the generalised model at parameter values under the null hypothesis.

The score vector for this model consists of a series of partitions. The first corresponds to $\boldsymbol{\beta}$, $(\nabla\boldsymbol{\beta})$, such that

$$
\begin{aligned}
\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\bigg|_{j=0} &= \frac{-\boldsymbol{x}\phi(\mu_0 - \boldsymbol{x}'\boldsymbol{\beta})}{P_{0,0}}, \\
\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\bigg|_{j=1} &= \frac{-\boldsymbol{x}[\phi(\mu_0 - \boldsymbol{x}'\boldsymbol{\beta}) - \phi(\mu_1 - \boldsymbol{x}'\boldsymbol{\beta})]}{P_{1,0}}, \\
&\vdots \\
\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\bigg|_{j=J-1} &= \frac{\boldsymbol{x}[\phi(\boldsymbol{x}'\boldsymbol{\beta} - \mu_{J-2})]}{P_{J-1,0}},
\end{aligned}
$$

where $\phi(.)$ is the standard normal density, and $\ln L(\boldsymbol{\theta})$ is the log-likelihood function and $\boldsymbol{\theta}$ contains all of the parameters of the relevant model. The second is a partition due to $\tilde{\boldsymbol{\gamma}}_0$ from the equation for the self-report ($\nabla \tilde{\boldsymbol{\gamma}}_{0,0}$),

$$\begin{aligned}
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \% \tilde{\boldsymbol{\gamma}}_{0,0}}\right|_{j=0} &= \frac{\tilde{z}\exp(\tilde{z}'\tilde{\boldsymbol{\gamma}}_0)\phi(\mu_0 - \boldsymbol{x}'\boldsymbol{\beta})}{P_{0,0}}, \\
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \% \tilde{\boldsymbol{\gamma}}_{0,0}}\right|_{j=1} &= \frac{\tilde{z}\exp(\tilde{z}'\tilde{\boldsymbol{\gamma}}_0)[\phi(\mu_1 - \boldsymbol{x}'\boldsymbol{\beta}) - \phi(\mu_0 - \boldsymbol{x}'\boldsymbol{\beta})]}{P_{1,0}}, \\
&\vdots \\
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \% \tilde{\boldsymbol{\gamma}}_{0,0}}\right|_{j=J-1} &= \frac{\tilde{z}\exp(\tilde{z}'\tilde{\boldsymbol{\gamma}}_0)\phi(\boldsymbol{x}'\boldsymbol{\beta} - \mu_{J-2})}{P_{J-1,0}}.
\end{aligned}$$

Similarly from the vignette equation(s) ($\nabla \tilde{\boldsymbol{\gamma}}_{0,k}$),

$$\begin{aligned}
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \% \tilde{\boldsymbol{\gamma}}_{0,k}}\right|_{j=0,k} &= \frac{[\tilde{z}\exp(\tilde{z}'\tilde{\boldsymbol{\gamma}}_0)/\sigma]\phi[(\mu_0 - \alpha_k)/\sigma]}{P_{0,k}}, \\
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \% \tilde{\boldsymbol{\gamma}}_{0,k}}\right|_{j=1,k} &= \frac{[\tilde{z}\exp(\tilde{z}'\tilde{\boldsymbol{\gamma}}_0)/\sigma]\{\phi[(\mu_1 - \alpha_k)/\sigma] - \phi[(\mu_0 - \alpha_k)/\sigma]\}}{P_{1,k}}, \\
&\vdots \\
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \% \tilde{\boldsymbol{\gamma}}_{0,k}}\right|_{j=J-1,k} &= \frac{[\tilde{z}\exp(\tilde{z}'\tilde{\boldsymbol{\gamma}}_0)/\sigma]\phi[(\alpha_k - \mu_{J-2})/\sigma]}{P_{J-1,k}}.
\end{aligned}$$

Collecting the above terms together gives $\nabla \tilde{\boldsymbol{\gamma}}_0 = \nabla \tilde{\boldsymbol{\gamma}}_{0,0} + \sum_1^K \nabla \tilde{\boldsymbol{\gamma}}_{0,k}$. The score with respect to $\boldsymbol{\gamma}_1$ again consists of a quantity from the y^* equation ($\nabla \boldsymbol{\gamma}_{1,0}$)

$$\begin{aligned}
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_{1,0}}\right|_{j=1} &= \frac{z\exp(z'\boldsymbol{\gamma}_1)\phi(\mu_{1,0} - \boldsymbol{x}'\boldsymbol{\beta})}{P_{1,0}}, \\
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_{1,0}}\right|_{j=2} &= \frac{z\exp(z'\boldsymbol{\gamma}_1)[\phi(\mu_2 - \boldsymbol{x}'\boldsymbol{\beta}) - \phi(\mu_1 - \boldsymbol{x}'\boldsymbol{\beta})]}{P_{2,0}}, \\
&\vdots \\
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_{1,0}}\right|_{j=J-1} &= \frac{z\exp(z'\boldsymbol{\gamma}_1)\phi(\boldsymbol{x}'\boldsymbol{\beta} - \mu_{J-2})}{P_{J-1,0}}.
\end{aligned}$$

Similarly from the corresponding vignette components ($\nabla \boldsymbol{\gamma}_{1,k}$),

$$\begin{aligned}
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_{1,k}}\right|_{j=1,k} &= \frac{[z\exp(z'\boldsymbol{\gamma}_1)/\sigma]\phi[(\mu_1 - \alpha_k)/\sigma]}{P_{1,k}}, \\
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_{1,k}}\right|_{j=2,k} &= \frac{[z\exp(z'\boldsymbol{\gamma}_1)/\sigma]\{\phi[(\mu_2 - \alpha_k)/\sigma] - \phi[(\mu_1 - \alpha_k)/\sigma]\}}{P_{2,k}}, \\
&\vdots \\
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_{1,k}}\right|_{j=J-1,k} &= \frac{[z\exp(z'\boldsymbol{\gamma}_1)/\sigma]\phi[(\alpha_k - \mu_{J-2})/\sigma]}{P_{J-1,k}}.
\end{aligned}$$

Repeating for $\boldsymbol{\gamma}_2$ we have ($\nabla\boldsymbol{\gamma}_{2,0}$ and $\nabla\boldsymbol{\gamma}_{2,k}$). The score with respect to $\boldsymbol{\gamma}_2$ again consists of a quantity from the y^* equation ($\nabla\boldsymbol{\gamma}_{2,0}$)

$$\begin{aligned}
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_{2,0}}\right|_{j=2} &= \frac{z\exp(z'\boldsymbol{\gamma}_2)\phi(\mu_2 - \boldsymbol{x}'\boldsymbol{\beta})}{P_{2,0}},\\
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_{2,0}}\right|_{j=3} &= \frac{z\exp(z'\boldsymbol{\gamma}_2)[\phi(\mu_3 - \boldsymbol{x}'\boldsymbol{\beta}) - \phi(\mu_2 - \boldsymbol{x}'\boldsymbol{\beta})]}{P_{3,0}},\\
&\vdots\\
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_{2,0}}\right|_{j=J-1} &= \frac{z\exp(z'\boldsymbol{\gamma}_2)\phi(\boldsymbol{x}'\boldsymbol{\beta} - \mu_{J-2})}{P_{J-1,0}},
\end{aligned}$$

and $\nabla\boldsymbol{\gamma}_{2,k}$

$$\begin{aligned}
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_{1,k}}\right|_{j=2,k} &= \frac{[z\exp(z'\boldsymbol{\gamma}_2)/\sigma]\phi[(\mu_2 - \alpha_k)/\sigma]}{P_{2,k}},\\
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_{1,k}}\right|_{j=3,k} &= \frac{[z\exp(z'\boldsymbol{\gamma}_2)/\sigma]\{\phi[(\mu_2 - \alpha_k)/\sigma] - \phi[(\mu_1 - \alpha_k)/\sigma]\}}{P_{3,k}},\\
&\vdots\\
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_{1,k}}\right|_{j=J-1,k} &= \frac{[z\exp(z'\boldsymbol{\gamma}_2)/\sigma]\phi[(\alpha_k - \mu_{J-2})/\sigma]}{P_{J-1,k}}.
\end{aligned}$$

The progression continues for $j, j > 2$ to $j = J-1$. Under parameter equivalence implied by *RC*, the elements of the score would be the sum of the respective gradients such that

$$\nabla\boldsymbol{\gamma}_j = \nabla\boldsymbol{\gamma}_{j,0} + k^{\sim}\tilde{\sum}\nabla\boldsymbol{\gamma}_{j,k}.$$

Derivatives with respect to α_k are given by ($\nabla\alpha_k$)

$$\begin{aligned}
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \alpha_k}\right|_{j=0,k} &= \frac{-\phi[(\mu_0 - \alpha_k)/\sigma]\sigma^{-1}}{P_{0,k}},\\
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \alpha_k}\right|_{j=1,k} &= \frac{-\{\phi[(\mu_1 - \alpha_k)/\sigma] - \phi[(\mu_0 - \alpha_k)/\sigma]\}\sigma^{-1}}{P_{1,k}},\\
&\vdots\\
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \alpha_k}\right|_{j=J-1,k} &= \frac{\phi[(\alpha_k - \mu_{J-2})/\sigma]\sigma^{-1}}{P_{J-1,k}}.
\end{aligned}$$

Finally, $\nabla\sigma$ is given by

$$\begin{aligned}
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \sigma}\right|_{j=0,k} &= \frac{\phi[(\mu_0 - \alpha_k)/\sigma](\alpha_k - \mu_0)\sigma^{-2}}{P_{0,k}},\\
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \sigma}\right|_{j=1,k} &= \frac{-\{\phi[(\mu_1 - \alpha_k)/\sigma] - \phi[(\mu_0 - \alpha_k)/\sigma]\}(\mu_1 - \mu_0)\sigma^{-2}}{P_{1,k}},\\
&\vdots\\
\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \sigma}\right|_{j=J-1,k} &= \frac{-\phi[(\alpha_k - \mu_{J-2})/\sigma](\alpha_k - \mu_{J-1})\sigma^{-2}}{P_{J-1,k}}.
\end{aligned}$$

B.1. Score test for vignette equivalence (*VE*)

Imposing *VE* is equivalent to assuming that the effect of any covariates, $\tilde{\boldsymbol{x}}$, entered into the model for the vignettes (see specification (6)) are zero. Accordingly, the null hypothesis of *VE* can be tested by

$$H_0: \tilde{\boldsymbol{\alpha}}_k = \mathbf{0},$$
$$H_1: \text{at least one element is non-zero.}$$

The use of the score test here is appealing as it does not require estimation of the more complex model under H_1. Here, the appropriate partition of the score vector under the null replaces $(\nabla \boldsymbol{\alpha}_k)$ (defined above), with the generalised version $(\nabla \boldsymbol{\alpha}_{VE,k})$ such that

$$\begin{aligned}
\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}_k}\bigg|_{j=0,k} &= \frac{-\boldsymbol{x}\phi[(\mu_0 - \boldsymbol{x}'\boldsymbol{\alpha}_k)/\sigma]\sigma^{-1}}{P_{0,k}}, \\
\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}_k}\bigg|_{j=1,k} &= \frac{-\boldsymbol{x}\phi[(\mu_1 - \boldsymbol{x}'\boldsymbol{\alpha}_k)/\sigma] - \phi[(\mu_0 - \boldsymbol{x}'\boldsymbol{\alpha}_k)/\sigma]\sigma^{-1}}{P_{1,k}}, \\
&\vdots \\
\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}_k}\bigg|_{j=J-1,k} &= \frac{-\boldsymbol{x}\phi[(\boldsymbol{x}'\boldsymbol{\alpha}_k - \mu_{J-2})/\sigma]\sigma^{-1}}{P_{J-1,k}},
\end{aligned}$$

The above expressions are evaluated under the *CHOPIT* null: that is, at α_k estimated in the *CHOPIT* model, and setting $\tilde{\boldsymbol{\alpha}}_k = \mathbf{0}$ (where $\boldsymbol{x}'\boldsymbol{\alpha}_k = \alpha_k + \tilde{\boldsymbol{x}}'\tilde{\boldsymbol{\alpha}}_k$). The quadratic form of the score test is

$$S_{VE} = (\nabla \boldsymbol{\beta}, \nabla \boldsymbol{\gamma}_j, \nabla \boldsymbol{\alpha}_{VE,k}, \nabla \sigma)[\boldsymbol{I}(\widehat{\theta}_{VE})]^{-1}(\nabla \boldsymbol{\beta}, \nabla \boldsymbol{\gamma}_j, \nabla \boldsymbol{\alpha}_{VE,k}, \nabla \sigma)'.$$

Under H_0 $S_{VE} \sim \chi^2_{q_{VE}}$, where

$$q_{VE} = \dim(\nabla \boldsymbol{\beta}, \nabla \boldsymbol{\gamma}_j, \nabla \boldsymbol{\alpha}_{VE,k}, \nabla \sigma) - \dim(\nabla \boldsymbol{\beta}, \nabla \boldsymbol{\gamma}_j, \nabla \alpha_k, \nabla \sigma) = \dim(\tilde{\boldsymbol{x}})K.$$

We use the outer product of gradients to estimate the variance of the score vector—see, for example Greene (2018), p. 558). Thus all other elements of the components of the score vector remain unchanged; and the score of this generalised version allowing for a relaxation of *VE*, is evaluated under the null. That is, evaluation takes place at parameter values estimated under the null: here we simply set $\boldsymbol{\alpha}_k = (\alpha_k, \mathbf{0})$ where α_k is the scalar estimated from the *CHOPIT* model, and the dimensions of the null vector in $\tilde{\boldsymbol{\alpha}}_k$ will be determined by the number of variables in $\boldsymbol{x}$. Note that in deriving the score test for *VE* we assume the assumption of *RC* holds.

B.2. Score test for response consistency (*RC*)

The identifying assumption of *RC* is equivalent to saying that the effect of any covariates in the boundary parameters—Equation (10)—for the self-report ($k = 0$) and vignettes equations, ($k = 1, \ldots, K$) are identical. Accordingly, the null of *RC* (assuming here that *VE* holds) can be tested as

$$H_0: \tilde{\boldsymbol{\gamma}}_{0,0} = \tilde{\boldsymbol{\gamma}}_{0,k}; \boldsymbol{\gamma}_{j,0} = \boldsymbol{\gamma}_{j,k}, \quad \forall j = 1, \ldots, J-1;\ k = 1, \ldots, K,$$
$$H_1: \text{at least one element differs.}$$

Here we replace the $\nabla\boldsymbol{\gamma}_j$ elements of the restricted *CHOPIT* score, with those from the generalised version, $\nabla\boldsymbol{\gamma}_{j,RC}$. For example, the generalised partition of the score vector due to $\tilde{\gamma}_0$ from the vignette equations $(\nabla\tilde{\boldsymbol{\gamma}}_{0,k})$ is

$$\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \%\tilde{\boldsymbol{\gamma}}_{0,k}}\right|_{j=0,k} = \frac{[\tilde{z}\exp(\tilde{z}'\tilde{\boldsymbol{\gamma}}_{0,k})/\sigma]\phi[(\mu_{0,k}-\alpha_k)/\sigma]}{P_{0,k}},$$
$$\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \%\tilde{\boldsymbol{\gamma}}_{0,k}}\right|_{j=1,k} = \frac{[\tilde{z}\exp(\tilde{z}'\tilde{\boldsymbol{\gamma}}_{0,k})/\sigma]\{\phi[(\mu_{1,k}-\alpha_k)/\sigma]-\phi[(\mu_{0,k}-\alpha_k)/\sigma]\}}{P_{1,k}},$$
$$\vdots$$
$$\left.\frac{\partial \ln L(\boldsymbol{\theta})}{\partial \%\tilde{\boldsymbol{\gamma}}_{0,k}}\right|_{j=J-1,k} = \frac{[\tilde{z}\exp(\tilde{z}'\tilde{\boldsymbol{\gamma}}_{0,k})/\sigma]\phi[(\alpha_k-\mu_{J-2,k})/\sigma]}{P_{J-1,k}}.$$

Similarly, generalising other elements of $\nabla\boldsymbol{\gamma}_j$, the score test is given by

$$S_{RC} = (\nabla\boldsymbol{\beta}, \nabla\boldsymbol{\gamma}_{j,RC}, \nabla\alpha_k, \nabla\sigma)[\boldsymbol{I}(\widehat{\theta}_{RC})]^{-1}(\nabla\boldsymbol{\beta}, \nabla\boldsymbol{\gamma}_{j,RC}, \nabla\alpha_k, \nabla\sigma)'.$$

Under H_0 $S_{RC} \sim \chi^2_{q_{RC}}$, where

$$q_{RC} = \dim(\nabla\boldsymbol{\beta}, \nabla\boldsymbol{\gamma}_{j,RC}, \nabla\alpha_k, \nabla\sigma) - \dim(\nabla\boldsymbol{\beta}, \nabla\boldsymbol{\gamma}_j, \nabla\alpha_k, \nabla\sigma).$$

B.3. A joint score test for vignette equivalence and response consistency

As noted, the above test for *RC* is performed under the assumption that *VE* holds, and *vice versa*. A *joint* score test of both assumptions simultaneously holding, is defined by simply combining the above two approaches such that

$$S_{joint} = (\nabla\boldsymbol{\beta}, \nabla\boldsymbol{\gamma}_{j,RC}, \nabla\boldsymbol{\alpha}_{VE,k}, \nabla\sigma)[\boldsymbol{I}(\widehat{\theta})]^{-1}(\nabla\boldsymbol{\beta}, \nabla\boldsymbol{\gamma}_{j,RC}, \nabla\boldsymbol{\alpha}_{VE,k}, \nabla\sigma)'.$$

Under H_0 where $S_{joint} \sim \chi^2_q$, where

$$q = \dim(\nabla\boldsymbol{\beta}, \nabla\boldsymbol{\gamma}_{j,RC}, \nabla\boldsymbol{\alpha}_{VE,k}, \nabla\sigma) - \dim(\nabla\boldsymbol{\beta}, \nabla\boldsymbol{\gamma}_j, \nabla\alpha, \nabla\sigma).$$

APPENDIX C

SUPPORTING TABLES

Table C1 *CHOPIT* estimates including boundary equations

$N = 3802$	Mean function		1st boundary		2nd boundary		3rd boundary	
	Coef	SE	Coef	SE	Coef	SE	Coef	SE
Constant			−0.076	0.061	0.295	0.046	−0.178	0.082
Male	−0.244	0.061	0.041	0.043	−0.181	0.039	−0.148	0.063
AnyCond	0.588	0.051	−0.011	0.035	−0.080	0.031	−0.034	0.058
Grip35	0.173	0.062	0.015	0.044	−0.164	0.039	−0.031	0.066
EducPS	−0.168	0.055	−0.074	0.039	0.074	0.034	−0.043	0.060
Age 66–75	0.073	0.051	−0.005	0.036	−0.010	0.033	0.117	0.051
Age > 75	0.149	0.084	−0.027	0.061	−0.001	0.056	0.061	0.079
Parameters of the vignette equation								
Vig 1 constant (α_1)	0.616	0.070						
Sigma vig	0.712	0.016						
Log-likelihood	−8824.15							

Table C2 Comparison of *CHOPIT* estimated parameters with different boundary specifications

	Boundary equations					
	Amended exponentials		Standard exponentials		Linear	
$N = 3,802$	Coef	SE	Coef	SE	Coef	SE
Structural parameters ($\hat{\beta}$)						
Constant	1.177	0.062	0.177	0.062	0.172	0.062
Male	−0.272	0.070	−0.271	0.070	−0.274	0.070
AnyCond	0.658	0.059	0.657	0.058	0.656	0.058
Grip35	0.185	0.072	0.189	0.071	0.188	0.071
EducPS	−0.194	0.063	−0.193	0.063	−0.193	0.063
Age 66–75	0.090	0.058	0.085	0.058	0.085	0.058
Age > 75	0.169	0.097	0.164	0.097	0.164	0.097
1st boundary (μ_0)						
Male	0.044	0.054	0.047	0.057	0.039	0.057
AnyCond	0.008	0.047	0.008	0.047	0.004	0.047
Grip35	0.015	0.057	0.021	0.058	0.017	0.058
EducPS	−0.092	0.056	−0.090	0.052	−0.092	0.052
Age 66–75	0.004	0.046	−0.024	0.048	−0.002	0.048

(Continues)

Table C2 (Continued)

	Boundary equations					
	Amended exponentials		Standard exponentials		Linear	
$N = 3{,}802$	Coef	SE	Coef	SE	Coef	SE
Age > 75	−0.030	0.079	−0.036	0.080	−0.036	0.080
2nd boundary (μ_1)						
Constant	0.492	0.044	0.493	0.043	1.592	0.058
Male	−0.185	0.039	−0.185	0.039	−0.188	0.054
AnyCond	−0.099	0.032	−0.098	0.031	−0.121	0.045
Grip35	−0.166	0.040	−0.168	0.039	−0.196	0.055
Educ PS	0.073	0.034	0.072	0.034	0.009	0.049
Age 66–75	−0.014	0.033	−0.012	0.033	−0.018	0.045
Age > 75	0.001	0.056	0.003	0.056	−0.033	0.075
3rd boundary (μ_2)						
Constant	0.051	0.081	0.051	0.081	2.632	0.090
Male	−0.120	0.063	−0.120	0.063	−0.298	0.071
AnyCond	−0.037	0.058	−0.037	0.058	−0.155	0.064
Grip35	−0.045	0.066	−0.046	0.066	−0.229	0.074
EducPS	−0.032	0.060	−0.033	0.060	−0.028	0.066
Age 66–75	0.101	0.050	0.101	0.050	0.081	0.060
Age > 75	0.056	0.078	0.057	0.078	0.021	0.095
Parameters of the vignette equation						
Vig 1 constant (α_1)	1.829	0.065	0.830	0.065	0.821	0.064
Log-likelihood	−8939.60		−8939.63		−8940.31	

Table C3 *SHARE*: Averaged estimated boundaries and first probability index†

Boundaries	Amended exponential	Standard exponential	Linear
$\widehat{\mu}_0$	1.015	0.016	0.007
$\widehat{\mu}_1$	2.316	1.317	1.307
$\widehat{\mu}_2$	3.289	2.289	2.279
$\widehat{\mu}_0 - \boldsymbol{x}'\widehat{\boldsymbol{\beta}}$	−0.600	−0.600	−0.600
$\widehat{\mu}_1$ with $(\widehat{\mu}_0 - \boldsymbol{x}'\widehat{\boldsymbol{\beta}})$	0.701	0.701	0.701
$\widehat{\mu}_2$ with $(\widehat{\mu}_0 - \boldsymbol{x}'\widehat{\boldsymbol{\beta}})$	1.674	1.674	1.672
$\widehat{\mu}_0 - \widehat{\alpha}_1$	−0.814	−0.814	−0.814
$\widehat{\mu}_1$ with $(\widehat{\mu}_0 - \widehat{\alpha}_1)$	0.487	0.487	0.487
$\widehat{\mu}_2$ with $(\widehat{\mu}_0 - \widehat{\alpha}_1)$	1.459	1.459	1.458

†Note, $\widehat{\alpha}_1$ is the estimated constant term for the vignette (α_1 in Equation (4)). $\boldsymbol{x}'\widehat{\boldsymbol{\beta}}$ is the estimated linear index including the constant. Boundary equations are estimated hierarchically such that $\mu_j = \mu_{j-1} + exp(\boldsymbol{x}'\widehat{\boldsymbol{\beta}})$. Accordingly, subtracting the linear index, $\boldsymbol{x}'\widehat{\boldsymbol{\beta}}$, from the first boundary, μ_0, affects all subsequent boundaries. This is denoted above for μ_1 and μ_2 as '$\widehat{\mu}_1$ with $(\widehat{\mu}_0 - \boldsymbol{x}'\widehat{\boldsymbol{\beta}})$' and '$\widehat{\mu}_2$ with $(\widehat{\mu}_0 - \boldsymbol{x}'\widehat{\boldsymbol{\beta}})$', respectively. Similarly for subtracting the vignette constant, $\widehat{\alpha}_1$ from each boundary.

Table C4 Correlations across estimated latent health and boundaries

	Amended exponential	Standard exponential	Linear
	Latent index, $\hat{y}^*$		
Amended exponential	1.000	1.000	1.000
Standard exponential	1.000	1.000	1.000
Linear	1.000	1.000	1.000
	1st boundary, $\hat{\mu}_0$		
Amended exponential	1.000	0.996	0.992
Standard exponential	0.996	1.000	0.996
Linear	0.992	0.996	1.000
	2nd boundary, $\hat{\mu}_1$		
Amended exponential	1.000	1.000	0.994
Standard exponential	1.000	1.000	0.995
Linear	0.994	0.995	1.000
	3rd boundary, $\hat{\mu}_2$		
Amended exponential	1.000	1.000	0.996
Standard exponential	1.000	1.000	0.997
Linear	0.996	0.997	1.000

Table C5 *SHARE*: Average predicted probabilities, and correlations across predicted probabilities

	Amended exponential	Standard exponential	Linear
	Average probabilities		
$j = 0$	0.288	0.288	0.288
$j = 1$	0.447	0.447	0.447
$j = 2$	0.202	0.302	0.202
$j = 3$	0.063	0.063	0.063
	Correlations across individual $P(y = 0)$		
Amended exponential	1.0000	1.0000	1.0000
Standard exponential	1.0000	1.0000	1.0000
Linear	1.0000	1.0000	1.0000
	Correlations across individual $P(y = 1)$		
Amended exponential	1.0000	0.9977	0.9976
Standard exponential	0.9977	1.0000	0.9999
Linear	0.9976	0.9999	1.0000
	Correlations across individual $P(y = 2)$		
Amended exponential	1.0000	0.9998	0.9997
Standard exponential	0.9998	1.0000	1.0000
Linear	0.9997	1.0000	1.0000
	Correlations across individual $P(y = 3)$		

Table C5 (Continued)

	Amended exponential	Standard exponential	Linear
Amended exponential	1.0000	0.9998	0.9997
Standard exponential	0.9998	1.0000	1.0000
Linear	0.9997	1.0000	1.0000

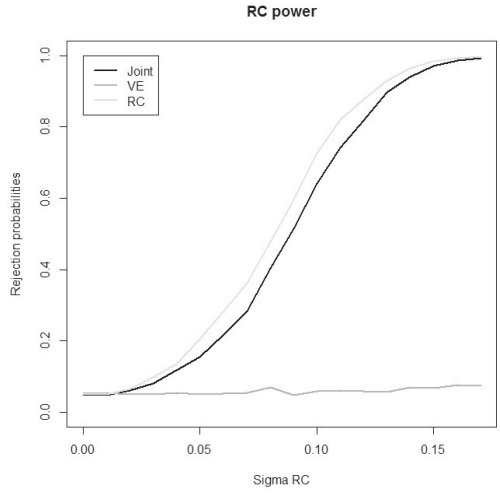
Table C6 *SHARE*: Partial effects (evaluated at sample means) and standard errors (in parentheses)

	Observed categorical outcomes							
	$j = 0$		$j = 1$		$j = 2$		$j = 3$	
Amended exponential								
Male	0.105	(0.019)	−0.081	(0.016)	−0.028	(0.015)	0.004	(0.007)
AnyCond	−0.216	(0.015)	−0.027	(0.014)	0.162	(0.013)	0.081	(0.007)
Grip35	−0.057	(0.019)	−0.064	(0.017)	0.078	(0.015)	0.043	(0.008)
EducPS	0.034	(0.017)	0.028	(0.014)	−0.045	(0.013)	0.016	(0.007)
Age 66–75	−0.028	(0.016)	−0.004	(0.014)	0.032	(0.012)	0.0006	(0.006)
Age > 75	−0.066	(0.027)	0.004	(0.024)	0.048	(0.020)	0.014	(0.009)
Standard exponential								
Male	0.106	(0.019)	−0.082	(0.016)	−0.028	(0.015)	0.004	(0.007)
AnyCond	−0.217	(0.015)	−0.027	(0.014)	0.162	(0.013)	0.081	(0.007)
Grip35	−0.056	(0.019)	−0.065	(0.017)	0.078	(0.015)	0.043	(0.008)
EducPS	0.034	(0.017)	0.027	(0.014)	−0.045	(0.014)	0.016	(0.007)
Age 66–75	−0.029	(0.016)	−0.003	(0.014)	0.032	(0.012)	0.0004	(0.006)
Age > 75	−0.067	(0.027)	0.005	(0.024)	0.048	(0.020)	0.014	(0.009)
	Linear							
Male	0.104	(0.019)	−0.077	(0.016)	−0.029	(0.015)	0.002	(0.007)
AnyCond	−0.217	(0.015)	−0.025	(0.014)	0.162	(0.013)	0.080	(0.007)
Grip35	−0.057	(0.019)	−0.063	(0.016)	0.079	(0.015)	0.041	(0.007)
EducPS	0.034	(0.017)	0.029	(0.015)	−0.047	(0.014)	0.016	(0.007)
Age 66–75	−0.029	(0.016)	−0.003	(0.013)	0.032	(0.012)	0.0003	(0.006)
Age > 75	−0.067	(0.027)	0.005	(0.023)	0.047	(0.020)	0.014	(0.009)

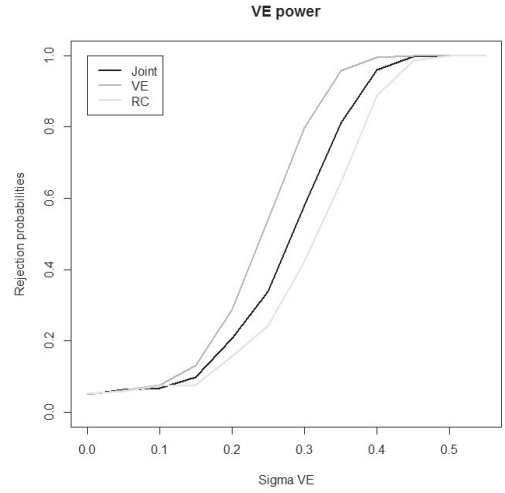
Table C7 Size results, at 0.05 nominal size; unrestricted σ_k^{2a}

Boundary equations	score_{joint}	score_{VE}	score_{RC}
Linear	0.0560	0.0550	0.0535
Standard exponential	0.0575	0.0535	0.0565
Amended exponential	0.0470	0.0525	0.0565

[a]Based on $M = 2000$ repetitions.



Departures from RC



Departures from VE

FIGURE C1 Power curves for rejection probabilities for departures from *RC* and *VE*: unrestricted σ_k^2

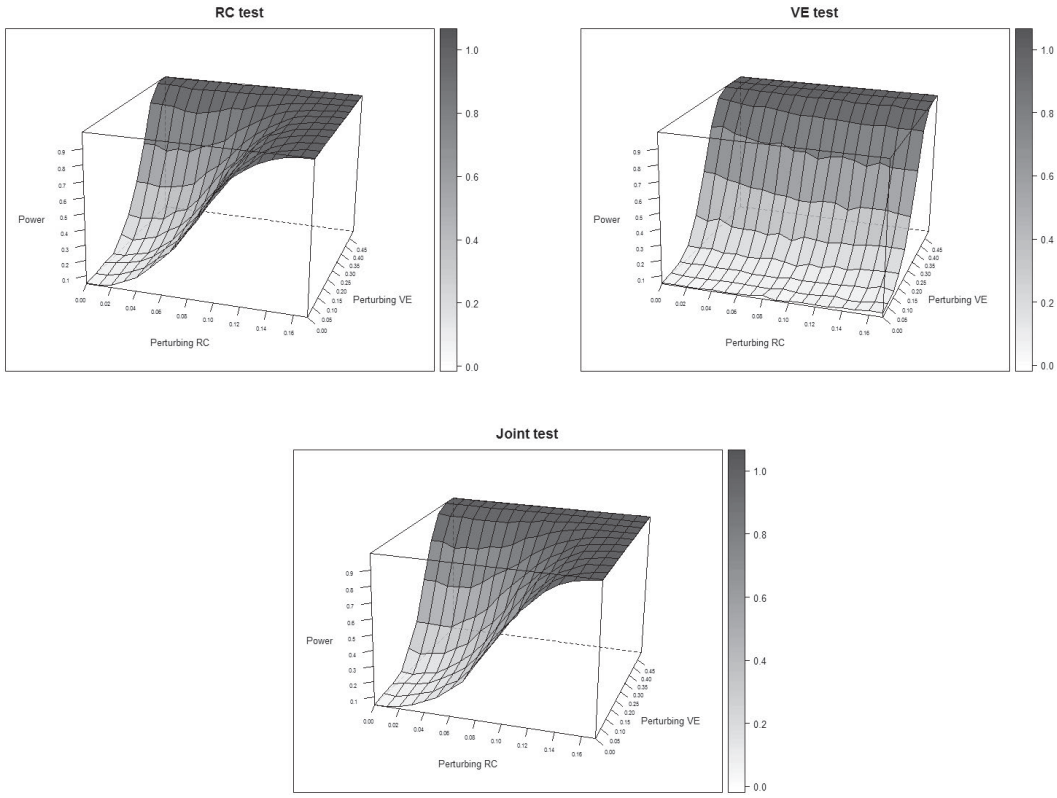


FIGURE C2 Power planes for rejection probabilities for departures from *RC* (lhs top), *VE* (rhs top) and a joint test (bottom): unrestricted σ_k^2

Table C8 Specification of boundary equations[a]

	True value	Boundary equations: amended exponential				
		$M = 2000; S = 1995$				
	Coef	Coef	SE (Coef)	MB	SE (MB)	Cov
Male	−0.271	−0.270	0.072	0.0007	0.072	0.943
AnyCond	0.657	0.659	0.059	0.002	0.059	0.946
Grip35	0.189	0.190	0.071	0.001	0.071	0.947
EducPS	−0.193	−0.194	0.065	−0.001	0.065	0.938
Age 66–75	0.085	0.084	0.057	−0.0005	0.057	0.952
Age > 75	0.164	0.165	0.097	0.001	0.097	0.955
	True value	Boundary equations: standard exponential				
		$M = 2000; S = 1996$				
Male	−0.271	−0.270	0.021	0.0009	0.072	0.944
AnyCond	0.657	0.659	0.059	0.002	0.059	0.949
Grip35	0.189	0.190	0.071	0.001	0.071	0.948
EducPS	−0.193	−0.194	0.065	−0.001	0.065	0.936
Age 66–75	0.085	0.084	0.057	−0.0006	0.057	0.953
Age > 75	0.164	0.165	0.097	0.001	0.097	0.954
	True value	Boundary equations: linear				
		$M = 2000; S = 578$				
Male	−0.271	−0.277	0.067	−0.006	0.067	0.958
AnyCond	0.657	0.658	0.059	0.0009	0.059	0.945
Grip35	0.189	0.187	0.071	−0.002	0.071	0.952
EducPS	−0.193	−0.194	0.061	−0.0005	0.061	0.955
Age 66–75	0.085	0.088	0.055	0.003	0.055	0.960
Age > 75	0.164	0.165	0.095	0.001	0.095	0.955

[a]Based on $M = 2000$ Monte Carlo repetitions from SHARE data ($N = 3802$). Simulations are generated assuming set of covariates $\tilde{\boldsymbol{x}}$ and parameters from column 2 of Table (C) that is assuming standard exponential boundaries. Models are estimated assuming (i) amended exponential boundaries, (ii) standard exponential boundaries, and (iii) linear boundaries. S represents the number of model repetitions that converged. MB is mean bias; Cov is the 5% coverage rate.