HEDG

HEALTH, ECONOMETRICS AND DATA GROUP

WP 19/18

Specification and testing of hierarchical ordered response models with anchoring vignettes

William H. Greene; Mark N. Harris;
Rachel Knott and Nigel Rice

August 2019

http://www.york.ac.uk/economics/postgrad/herc/hedg/wps/

Specification and testing of hierarchical ordered response models with anchoring vignettes

William H. Greene
New York University New York U.S.A.
Mark N. Harris
Curtin University Perth Australia.
Rachel Knott
Monash University Melbourne Australia.
Nigel Rice†
University of York York U.K.

Summary.
Anchoring vignettes have been proposed as a way to correct for differential item functioning when individuals self-assess their health, or other aspects of their circumstances on an ordered categorical scale. The model relies on two key underlying assumptions of response consistency and vignette equivalence. Adopting a modified specification of the boundary equations in the compound hierarchical ordered probit model this paper develops joint and separate tests of these assumptions based on a score approach. Monte Carlo simulations show that the tests have good size and power properties in finite samples. We provide an application of the test to data from the Survey of Health, Aging and Retirement in Europe (*SHARE*), using self-reported data on pain. The tests are easy to implement, only requiring estimation of the restricted model under the null hypothesis.

Keywords: Ordered response models, Anchoring vignettes, Differential item functioning, Self-assessments, score test, *CHOPIT*.

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

†*Address for correspondence:* Nigel Rice, Department of Economics and Related Studies, University of York, York, U.K.
E-mail: nigel.rice@york.ac.uk

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 may 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 five boxes ranging from *very bad* through to *average* 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 (DIF) (Holland & Weiner 1993, Murray et al. 2002).

Anchoring vignettes have been proposed as a method to overcome DIF (King et al. 2004) and have received wide attention in the applied literature - for example, in self-reported data on health status (Soloman et al. 2004, Bago d'Uva et al. 2008, Peracchi & Rossetti 2012, Grol-Prokopczyk et al. 2011, Vonkova & Hullegie 2011); healthy behaviours (Van Soest et al. 2011); satisfaction with health systems performance (Sirven et al. 2012, Rice et al. 2012); work disability (Kapteyn et al. 2007, 2011, Angelini et al. 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. Responses to the vignettes are then used to anchor, or adjust for bias introduced by DIF, such that inter-personal comparisons can be appropriately examined. This is often achieved using the compound hierarchical ordered probit model ($CHOPIT$: see Section 2.3).

Two key identification assumptions underlie the use of the vignette approach: 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.

In practice, modelling DIF is undertaken by the allowing the bound-

ary parameters determining the cut-points between outcome categories of the standard ordered probit model to differ across individuals. To ensure the requisite ordering, these are typically specified as exponential functions of individual characteristics within the *CHOPIT* model. This paper presents a simple modification to the standard exponential functional form for the boundary equations which yields improvements in its identification properties. We show that the proposed modification is innocuous in terms of parameter estimates of the mean function of the (*CHOPIT*) model which links characteristics of respondents to outcome on the latent scale. Importantly, however, the modification lends itself to score tests for the joint assumptions of RC and VE together with separate tests for the two assumptions. Given the widespread use of the vignette approach it is surprising that there are few formal tests of the underlying assumptions behind the *CHOPIT* approach and many empirical applications simply assume these hold. We set out the score test and investigate its properties.

We show that the test is correctly sized and has appropriate power properties as one moves further from the null hypothesis of RC and VE. We also illustrate the use of the test using data from the Survey of Health, Aging and Retirement in Europe ($SHARE$), and compare our results to those of the test developed by Peracchi & Rossetti (2013). We also provide separate tests for RC and VE. These are informative when faced with a joint failure of VE and RC with regard to the underlying cause of misspecification (for example, by selecting which vignettes to use when multiple are available). As is typical with specification tests, it relies on standard parametric assumptions underlying the (*CHOPIT*) model, that the model is correctly specified, with no omitted variables, endogeneity and so on.

2. Ordered response models

2.1. Ordered probit model

Self-reports of outcomes of interest collected via survey instruments usually contain responses on a categorical (*Likert*) scale, which are often analysed using ordered response (e.g. probit or logit) models (Greene & Hensher 2010). Underlying the ordered probit (OP) model is represented by a latent variable, y^*, which is a linear function, in unknown parameters, $\tilde{\beta}$, of observed characteristics $\tilde{\mathbf{x}}$ with no constant term (throughout we denote a no constant sub-vector/matrix by use of "$\sim$"). The term ε_y represents a standard normal disturbance, such that,

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

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

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

where $\mu_{-1} = -\infty$ and $\mu_{J-1} = +\infty$. To ensure well-defined probabilities; $\mu_{j-1} < \mu_j, \forall j$. Under the assumption of normality the probabilities for the ordered outcomes are, for $y = 0$, $y = j$ and $y = J-1$, respectively,

$$\Phi\left(\mu_0 - \tilde{\mathbf{x}}'\tilde{\beta}\right); \left[\Phi\left(\mu_j - \tilde{\mathbf{x}}'\tilde{\beta}\right) - \Phi\left(\mu_{j-1} - \tilde{\mathbf{x}}'\tilde{\beta}\right)\right]; \text{ and } \Phi\left(\tilde{\mathbf{x}}'\tilde{\beta} - \mu_{J-2}\right), \tag{3}$$

where $\Phi(.)$ denotes the standard normal distribution function. The (log) density, $\ln L_{OP}(\theta)$, for this model for a $i = 1, \ldots, N$ random sample of individuals is:

$$\ln L_{OP}(\theta) = \sum_{i=1}^{N} \sum_{j=0}^{J-1} d_{ij} \ln\left[\Pr\left(y_i = j \,|\tilde{\mathbf{x}}\right)\right],$$

where d_{ij} is an indicator function returning one if individual i responded with outcome j, and zero otherwise, and θ are the parameters of the model.

2.2. Hierarchical Ordered Probit Model (HOPIT)

The standard ordered probit model assumes the use of homogeneous reporting scales across individuals such that the boundary parameters, μ_j, are common to all respondents. However, this assumption is unlikely to hold, particularly when undertaking cross-national comparisons where differences in norms and customs lead to distinct styles of reporting behaviour, or DIF (a graphical illustration of DIF is provided in Appendix A).

Differences in reporting scales across individuals can be accommodated by specifying individual-specific boundary parameters, $\mu_{i,j}$ (see, for example, Terza (1985), Pudney & Shields (2000), Boes & Winkelmann (2006), Greene & Hensher (2010), Greene et al. (2014)). This can be achieved by allowing the boundaries to depend on a set of observed characteristics $\mathbf{z}_i$ such that $\mu_{i,j} = \mathbf{z}_i'\gamma_j$, where to secure identification $\mathbf{z}_i \notin \mathbf{x}_i$. To ensure coherent probabilities most authors (see, for example, Greene & Hensher (2010)) adopt the hierarchical ordered probit ($HOPIT$) approach by specifying the boundaries as

$$\begin{aligned} \mu_{i,0} &= \mathbf{z}_i'\gamma_0, \\ \mu_{i,j} &= \mu_{i,j-1} + \exp\left(\mathbf{z}_i'\gamma_j\right), \qquad j = 1, \ldots, J-2. \end{aligned} \tag{4}$$

This model can be estimated by maximum likelihood techniques, where the μ_j in equation (2) are replaced by those of equation (4).

2.3. Compound Hierarchical Ordered Probit Model (CHOPIT)

Empirically it may be difficult to justify exclusion restrictions between $\mathbf{x}$ and $\mathbf{z}$. However, for any variable that appears in both $\mathbf{x}$ and $\mathbf{z}$, since the first threshold in equation (4) is specified linearly, the corresponding elements of γ_0 and $\tilde{\beta}$ are not separately identified in the absence of further information. This 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 *SHARE*:

> "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 the respondent are the same as those used to respond to a self-assessment for pain, namely *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 in model (1) and the corresponding response to the k^{th} vignette, v_k^*, as,

$$v_k^* = \alpha_k + \varepsilon_k, \quad k = 1, \ldots, K, \tag{5}$$

where $\varepsilon_k \sim N\left(0, \sigma_k^2\right)$. It is common in the literature to allow σ_k^2 to be unrestricted; for example, see King et al. (2004). However, we adopt the normalisation; $\sigma_k^2 = 1, \forall k$ (we compare our key results with a specification that relaxes this restriction later in the paper). The variance parameters are generally not identified in ordered choice models (see Greene (2018) pp. 730-731.) The parameters are unidentified under the alternative hypothesis (failure of RC and VE) in the $CHOPIT$ model. A scale parameter in each vignette of equation (5) becomes identified under the null hypothesis through information about the cell probabilities and the externally imposed thresholds $\mu_{0,j}$ in equation (4). See Kapteyn et al. (2011), footnote 7 for discussion.

The observed response to the self-report, y, and to each vignette, v_k, is determined as before by considering their relationship with the boundary equations, such that

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

and

$$v_k = j \text{ if } \mu_{j-1,k} \leq v_k^* < \mu_{j,k}, \quad j = 0, \ldots, J-1;\ k = 1, \ldots, K,$$

and where $\mu_{-1,0} = \mu_{-1,k} = -\infty$ and $\mu_{J-1,0} = \mu_{J-1,k} = +\infty$. Heterogeneity across these response scales is once more accommodated by specifying the boundaries as a function of variables, $\mathbf{z}$,

$$\begin{aligned} \mu_{0,k} &= \mathbf{z}'\gamma_{0,k}, \\ \mu_{j,k} &= \mu_{j-1,k} + \exp\left(\mathbf{z}'\gamma_{j,k}\right), \qquad j = 1, \ldots, J-2;\ k = 0, 1, \ldots, K \end{aligned} \tag{6}$$

Throughout, $k = 0$ 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, $(\mu_{0,k}, \ldots, \mu_{J-2,k})$. Kapteyn et al. (2007) introduced individual-specific unobserved heterogeneity, akin to a random effects, into the specification of the boundary equations. It would be straightforward to extend our approach to incorporate unobserved heterogeneity. However, to simplify the exposition, we do not follow this approach.

Identification of the full model follows from the assumptions of RC and VE (King et al. 2004). RC assumes that individuals apply the same mapping from the latent scale to the available response categories when rating the self-reports and when rating the vignettes. In practice this means that the boundary parameters are the same across the self-report of interest and all vignettes; $k = 0, \ldots, K$. Formally RC imposes the following restriction,

$$\gamma_{j,k} = \gamma_{j,0}, \quad j = 0, \ldots, J-2;\ k = 1, \ldots, K. \tag{7}$$

VE, on the other hand, 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 uni-dimensional scale except for random error (equation (5)). The alternative is to consider the more general specification where the latent response is a function of respondent characteristics

$$v_k^* = \alpha_k + \tilde{\mathbf{x}}'\tilde{\alpha}_k + \varepsilon_k, \qquad k = 1, \ldots, K. \tag{8}$$

VE therefore imposes the linear restriction(s) that: $\tilde{\alpha}_k = 0, \forall k$.

With all these elements in place the log-likelihood function will consist of two distinct parts: one relating to the self-report of interest ($\ln L_{HOPIT}$), and a second to the vignette component of the model ($\ln L_V$) such that: $\ln L = \ln L_{HOPIT} + \ln L_V$. 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 β and $\mu_{j,k}(\gamma_{j,k})$, and the second term, $\ln L_V$, is a function of α_k, σ and $\mu_{j,k}(\gamma_{j,k})$. These two components of the likelihood are linked by the common boundary parameters. The likelihood therefore can be written

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

We refer to the *HOPIT* model with vignettes as the compound hierarchical ordered probit model (*CHOPIT*) model (e.g. Vonkova & Hullegie (2011), Paccagnella (2013), Van Soest & Vonkova (2014)).

2.4. Modified CHOPIT model

The exponential form for the boundaries in model (6) 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). We suggest a modification to the specification of this first boundary parameter. Consider the following modified form for the first boundary threshold,

$$\mu_{0,k} = \gamma_{0,k} + \exp\left(\tilde{\mathbf{z}}'\tilde{\gamma}_{0,k}\right), \quad k = 0, 1, \ldots, K. \tag{9}$$

This removes the asymmetric use of the exp(.) function in the first ($j = 0$) and subsequent boundary equations, $j = 1, \ldots J-2$. Due to the presence of the leading term $\gamma_{0,k}$, $\mu_{0,k}$ is free to lie anywhere on the real line (that is, there is no restriction that $\mu_{0,k} > 0$). The remaining $(J-2)$ boundaries follow the specification set out in equation (6).

We parameterise the model such that the linear constant term of $\gamma_{0,0}$ enters in the main effects equation for y^*, and not in this first boundary equation. This follows from location normalisations in ordered probit-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). Under the maintained assumptions of RC and VE probabilities for the outcomes of the self-report of interest can be defined in as similar manner to equation (3) (full expressions are available in Appendix A).

The amended $CHOPIT$ model re-specifies the usual boundary specification (with respect to the first boundary parameter only). While the three possible forms of specification for the boundaries (linear, standard exponential and amended exponential) are not simple reparameterisations of one another, we show empirically in Section 6.3 that what emerges is that the respective boundaries and the mean function(s) adjust in such

a way as to keep the difference between them, which drives the OP-type probabilities, and hence essentially the underlying model, the same. While leaving the underlying model unchanged, the amended specification improves model identification (by removing linearity in the first boundary) and has the advantage of being amenable to a score test of its assumptions of RC and VE. Indeed, such an approach - of reparameterising the model to facilitate a score test - has precedents in the literature (for example, see Greene & McKenzie (2015) with regard to a score approach to testing for a zero variance in nonlinear panel data models).

3. Identifying assumptions of response consistency and vignette equivalence

The empirical literature has attempted to investigated the two assumptions of RC and VE. However, much of this literature is based on exploratory tests of the assumptions rather than a direct parametric test (an important exception is Peracchi & Rossetti (2013)). 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 thave tended to be ambiguous, for example, while Murray et al. (2003), King et al. (2004), Kristensen & Johansson (2008), Rice et al. (2011) and Hudson (2011) provide evidence in support of the assumption of VE, Datta Gupta et al. (2010), Peracchi & Rossetti (2012) and Bago d'Uva et al. (2011) find evidence against. When comparing outcomes across countries, Corrado et al. (2010) are sceptical about the comparability of survey responses and develop a test that allows the identification of subsets of countries where the assumption of VE holds. On the other hand, 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 (for example, 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 & Rossetti (2012) reject the null hypothesis.

Van Soest & Vonkova (2014) and Peracchi & Rossetti (2013) provide important contributions by demonstrating how RC and VE can be tested in the absence of objective measures. Using data across a number of health domains in the Survey of Health, Ageing and Retirement in Europe ($SHARE$), Van Soest & Vonkova (2014) consider the rankings 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 using the $CHOPIT$ approach. This

leads to a test of the parametric assumptions inherent in the $CHOPIT$ model when compared to a non-parametric alternative. While both approaches maintain the assumption of RC and VE, they show that an extended $CHOPIT$ model that includes individual unobserved heterogeneity in the specification of the boundary equations performs better than the standard $CHOPIT$ model.

Peracchi & Rossetti (2013) provide a direct test of the two identifying 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 *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 generalized ordered response models estimated in the spirit of the model proposed by Pudney & Shields (2000). 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 normalization restrictions imposed in ordered probit models, leads to $s = \{J(l+1)+1\}(K+1)$ parameters to be estimated. Note that we adopt a different notation to Peracchi & Rossetti (2013) to be consistent with the exposition set out in Section 2.1 (Peracchi & 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) generalized 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 & 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; $nQ_n\left(\hat{\psi}\right) \Rightarrow \chi^2_{q-p}$, as $n \to \infty$. $Q_n\left(\hat{\psi}\right)$ is the minimum distance criterion evaluated at the solution $\hat{\psi}$, with $q-p$ the number of over-identifying restrictions. See Peracchi & 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 sub-groups under comparison, the instruments of interest and the particular vignettes (wording and meaning) used. A simple to

implement test statistic of the underlying assumptions of the $CHOPIT$ model that does not rely on additional (external to the model) information is of value.

4. A score test of model specification

Before deriving the score statistic, we spell out the model description in full. As indicated above it is difficult to justify exclusion restrictions between $\mathbf{x}$ and $\mathbf{z}$ and hence it is common to assume $\mathbf{x} = \mathbf{z}$. This is the case in the Monte Carlo experiments of Section 5 and the empirical models that follow in Section 6. To aid exposition, however, we retain the labelling $\mathbf{x}$ and $\mathbf{z}$ throughout.

We have the usual underlying index function of the form

$$y^* = \mathbf{x}'\beta + \varepsilon_y, \qquad \varepsilon_y \sim N(0,1), \tag{10}$$

and the form for the vignette equation(s) as

$$v_k^* = \alpha_k + \tilde{\mathbf{x}}'\tilde{\alpha}_k + \varepsilon_k, \tag{11}$$

which, under VE collapse to

$$v_k^* = \alpha_k + \varepsilon_k, \text{ for } k = 1, \ldots, K, \quad \varepsilon_k \sim N\left(0, \sigma^2\right). \tag{12}$$

For all of the $k = 0, 1, \ldots, K$ constructs ($k = 0$ indexes the boundary equations for the self-report), we have boundary equations

$$\begin{aligned} \mu_{0,k} &= \exp\left(\tilde{\mathbf{z}}'\tilde{\gamma}_{0,k}\right), \\ \mu_{j,k} &= \mu_{j-1,k} + \exp\left(\mathbf{z}'\gamma_{j,k}\right), \quad j = 1, \ldots, J-1. \end{aligned} \tag{13}$$

In equation (13) the treatment of $\mu_{0,k}$ differs from that in equation (9) in that the constant term has (equivalently) moved into the mean equation (10). RC implies equivalence of parameters $\gamma_{0,k}$ and $\gamma_{j,k}$ across the boundary equations for $k = 0, 1 \ldots, k$, such that (13) collapses to

$$\begin{aligned} \mu_0 &= \exp\left(\tilde{\mathbf{z}}'\tilde{\gamma}_0\right), \\ \mu_j &= \mu_{j-1} + \exp\left(\mathbf{z}'\gamma_j\right), \quad j = 1, \ldots, J-1. \end{aligned} \tag{14}$$

The score statistics derived below are based upon the restricted model facilitated by the amended specification for the first boundary equation. This amendment identifies separate $HOPIT$ models for all of the $k = 0, 1, \ldots, K$ constructs. That is, unlike the standard exponential specification, due to the nonlinear (i.e., exponential) transformation of

all boundaries, all parameters of this unrestricted general model are numerically identified. The unrestricted model consists of separate $HOPIT$ models defined by equations (10), (11) and (13). By enforcing the noted restrictions of equations (12) and (14), this generalised model collapses to the $CHOPIT$ model. As the restrictions of equations (12) and (14) are simple linear parameter restrictions, they can be tested both individually and jointly by score tests based on the likelihood of the unrestricted model, but, as usual, evaluated at parameter values under the null, which in all instances is that the (restricted) $CHOPIT$ model is correctly specified. To secure identification under both null and alternative hypotheses, we have normalized σ^2 to one throughout (see footnote 6 for discussion).

Accordingly, the score test can be applied in the usual way. To generalise, assume that the set of parameters of interest for the unrestricted model defined by equations (10), (11) and (13) is represented by θ_u. The maximum likelihood estimator, MLE $\hat{\theta}_u$ sets $s(\hat{\theta}_u) = 0$, where $s(\theta) = \partial \ln L(\theta) / \partial \theta$ is the score function. The score test is based on the closeness of $s(\tilde{\theta}_r)$ to zero, where evaluation takes place at $\tilde{\theta}_r$ which is the alternative restricted maximum likelihood estimator of the parameters, θ_r defined under the set of h restrictions imposed by assumption RC and VE defined by equations (12) and (14). Since $s(\tilde{\theta}_r\ ) \sim N(0, Var(\tilde{\theta}_r))$, the usual quadratic form of the score test is:

$$score = s\left(\tilde{\theta}_r\right)' \left[\hat{V}\left\{s\left(\tilde{\theta}_r\right)\right\}\right]^{-1} s\left(\tilde{\theta}_r\right) \sim \chi^2_h \text{ under } H_0. \qquad (15)$$

The outer product of the gradients is used to estimate the variance of the score vector - see for example, Greene (2018, p. 558). The use of the score test here is appealing as it does not require estimation of the more complex model under the alternative hypothesis. That is, we do not need to estimate MLE $\hat{\theta}_u$ and only require the scores for this function together with estimates for θ_r obtained from estimation of the amended $CHOPIT$ model. Moreover, the test lends itself to separate tests for the assumptions of RC and VE. Since each of these assumptions places restrictions on the model, we construct a score test for RC under the null hypothesis that RC holds by assuming VE is valid; and a test under the null hypothesis that VE holds under the assumption that RC is valid as follows.

The identifying assumption of RC is equivalent to saying that the effect of any covariates in the boundary parameters - equation (13) - 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

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

Imposing VE is equivalent to assuming that the effect of any covariates, $\tilde{\mathbf{x}}$, entered into the model for the vignettes (see equation (11)) are zero. Accordingly, the null hypothesis of VE can be tested by

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

The full derivatives of the appropriate score vector(s) and the corresponding form of the score test, are presented in Appendix B. These are provided separately for tests of RC and VE together with a joint test for both RC and VE.

5. Monte Carlo evidence

This section considers Monte Carlo evidence for the amended $CHOPIT$ model. We first consider the amendment to the boundary specification (equation (13)) and compare this to the more common functional forms (all boundaries specified as linear functions of parameters and standard exponential specifications (equation (6))). We then implement the score test and consider its finite sample properties. Throughout we simulate data by drawing from a set of covariates within $SHARE$. $SHARE$ data are described in Section 6.1, and the set of covariates used are those in the empirical example described in Section 6.

5.1. Boundary specification

We first illustrate the difference that the amended specification has on the estimated vector of coefficients, $\tilde{\beta}$, when compared to standard exponential or linear specifications. Typically these are the parameters of most interest in empirical applications. Table 1 presents the results of the Monte Carlo exercise. Data are generated assuming standard exponential specification of the boundaries and estimated separately assuming amended exponential, standard exponential, and linear specifications. Generation of the data is undertaken assuming standard exponential functions since this is the norm in the literature on anchoring vignettes. The Monte Carlo experiment simulates data as follows: (i) use all $N = 3802$ observations and their corresponding covariates $\mathbf{x}_i$ from the $SHARE$ data; (ii) construct the latent outcome $y^* = \tilde{\mathbf{x}}'\tilde{\beta} + \varepsilon_y$ using the parameter estimates

Table 1. Specification of boundary equations†

	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
Age66-75	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
Age66-75	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
Age66-75	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

† Based on $M = 2000$ Monte Carlo repetitions from SHARE data ($N = 3802$). Simulations are generated assuming set of covariates $\tilde{\mathbf{x}}$ and parameters from column 2 of Table (C1), i.e. 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.

from the empirical example presented in Section 6.3 as column (2) of Table 6 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 6 (full model estimates including boundary parameters are provided in Table C1); (iv) the corresponding observed outcomes, $y_i, v_{i,1}$, are then constructed from their latent counterparts together with knowledge of the boundary parameters $(\tilde{\gamma}_{i,0},\ldots,\tilde{\gamma}_{i,2})$ estimated from the model reported in column (2), Table 6. *CHOPIT* estimation of the simulated y_i and $v_{i,1}$ on the set of covariates $\mathbf{x}_i$ is then undertaken. This is repeated for $M = 2000$ simulations and results for models for which convergence was achieved (S) summarised in Table 1 (convergence was deemed to have failed after 500 iterations).

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.

5.2. Finite sample performance of the score tests

We evaluate the performance of the score test 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 6. All experiments are based upon $M = 2000$ repetitions. In practice, one would not know how the thresholds were truly generated, so we consider all three variants that have been suggested above, but then conduct the tests as if the boundaries were of the amended exponential form. Thus we consider 3 sets of test size experiments, where we generate under the null hypothesis with linear boundaries (column (3), Table 6); with standard exponential boundary thresholds (column (2)); and amended exponential boundaries (column (1)). Results are reported in Table 2. When the data generating process is as the test assumes, (i.e. amended exponential boundaries), the tests are correctly sized for the score$_{joint}$ and score$_{RC}$ variants. The score$_{VE}$ variant appears to be slightly undersized (at 4.10% for a nominal 5%), however this discrepancy is small. When the true data generating process consists of linear thresholds or standard exponentials, the score$_{joint}$ and score$_{RC}$ tests appear to be marginally oversized (at 5.85% for each of score$_{joint}$ and score$_{RC}$ for the linear thresholds; and 5.60% for

Table 2. Size results, at 0.05 nominal size†

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

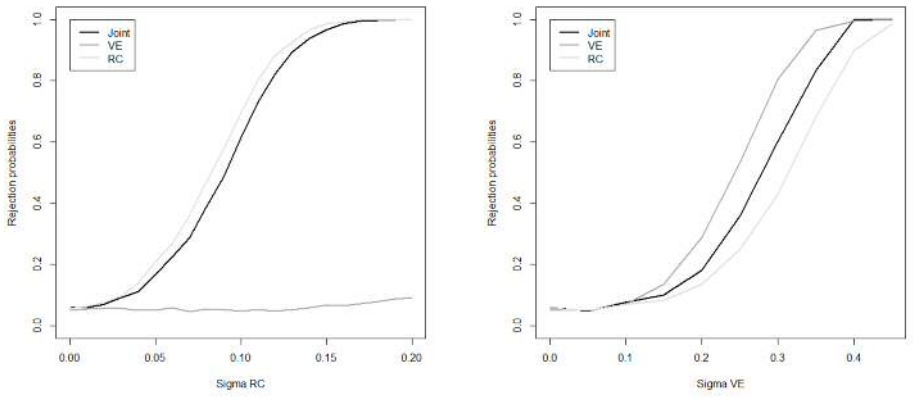
† Based on $M = 2,000$ repetitions

the standard exponential thresholds); however overall, the tests remain within acceptable range. Note that relaxing the assumption of $\sigma_k^2 = 1, \forall k$ does not materially affect size results, as evidenced in Appendix C Table C6.

We next consider power experiments using the same 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 (that is, $\gamma_{\mathbf{j1}}$ in equation (7)), perturbing at increasing values away from zero. These are undertaken for a model generated assuming amended exponential boundaries (column (1), Table 6). This is achieved by first generating a vector of standard normal random variates of the same dimension as $\mathbf{z}\,(=\mathbf{x})$. 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 this by a scalar, s_{rc} in the range $0.0 \geq s_{rc} \leq 0.20$. This ensures greater departures from the null for increasing values of s_{rc}. A similar procedure was used for violation of VE. In this case a vector of random variates of dimension $\mathbf{x}$ is first drawn. We then perturb the corresponding implicit vector of zero coefficients, $\tilde{\alpha}$ (under the null), on the covariates $\mathbf{x}$ (see equation (8)) by multiplying the random draws by a scalar, s_{ve}, and substituting these as parameters for $\tilde{\alpha}$. This process is repeated for successively larger values of s_{ve} such that $0.0 \geq s_{ve} \leq 0.50$. For the joint experiments, we simultaneously employ both approaches.

We perform $M = 2000$ experiments and record the rejection rate under the null. These are then summarised as power curves which plot the rejection probabilities against the size of the perturbation from the null of zero. Three curves are shown: a joint test for RC and VE; a test for RC alone and a test for VE alone.

The left-hand side of Figure 1 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



Departures from RC Departures from VE

Fig. 1. Power curves for rejection probabilities for departures from RC and VE

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 by construction.

The right-hand side of Figure 1 presents the power curves for the three tests when the model is subjected to violations of 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.

The above power results suggest the following when individual tests of RC and VE are used. Rejection of either RC or VE but not the other is straightforward to interpret; as is non-rejection of both. Rejection of RC and VE suggests either failure of the assumption of VE alone (and that this is reflected as an incorrect rejection of RC), or rejection of both RC and VE. Prior to undertaking individual tests, it would appear sensible to undertake a joint test of the null of both RC and VE. If this joint test 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 2. These are provided

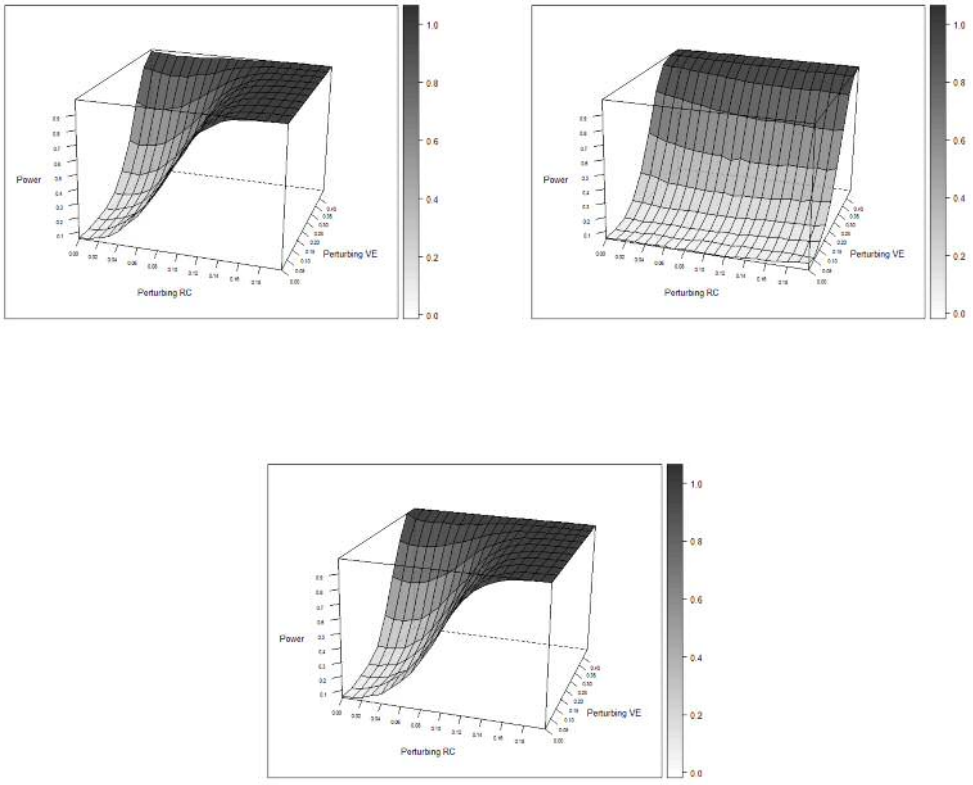


Fig. 2. Power planes for rejection probabilities for departures from RC (lhs top), VE (rhs top) and a joint test (bottom)

for the score tests for RC and VE alone (top left and top right plane) and the joint test (bottom plane). 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 ($\tilde{\alpha}_{\mathbf{k}} \neq 0$), the test remains fairly flat over the range of values for which RC is violated.

We summarise the findings of the Monte Carlo experiments as: (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

Table 3. Monte Carlo experiment and comparison with Peracchi & Rossetti (2013)

	Results for $J = 2$, $K = 1$ and $N = 250$				Results for $J = 2$, $K = 2$ and $n = 250$			
	P&R	score test			P&R	score test		
		Joint	VE	RC		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

power in all directions, and is maximised with simultaneous deviations from the null in both directions. Allowing σ_k^2 to be unrestricted does not impact the power of the test substantially, as shown Appendix C as Figures C1 and C2.

5.3. Comparison of the score test with the minimum distance estimator approach of Peracchi and Rossetti (2013)

Section 3 of Peracchi & 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 & Rossetti (2013)). Table 3 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 & 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 $M = 1000$ runs (as per Peracchi & 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 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 & 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 5.2 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.

6. Empirical example

6.1. Data

We use data from the Survey of Health, Ageing and Retirement in Europe ($SHARE$). $SHARE$ is a multidisciplinary and cross-national panel dataset of individuals aged 50 or over which has expanded over time to covering 27 European countries and Israel. A strength of the survey is allowing cross-country comparative analysis. 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 were included within the survey. In the context of a diverse continent like Europe, however, comparison of outcomes suffer from differences in the way people answer survey questions, particularly self-evaluations such as health status. Differences in language and cultural and social norms are more pronounced across countries than, for example, across socio-economic groups within a country. The application of

Table 4. SHARE: SAH and vignettes

	SAH	Vignette 1 (m_1)	Vignette 2 (m_2)	Vignette 3 (m_3)
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

anchoring vignettes is, therefore, important for enhancing cross-country comparability. Together with self-assessments, vignettes on health were collected in the first two waves of $SHARE$. The first wave of data, collected in 2004, was undertaken across 11 European countries; the second wave included and an additional three countries.

Three vignettes questions on each particular domain of health were included in the first wave and a single vignette in the second wave. In both waves data on self-assessments and vignettes were collected on a sub-sample of the overall $SHARE$ survey sample. For Belgium, France, Germany, Greece, Italy, Netherlands, Spain, and Sweden self-assessments and vignettes were included in both waves, and for the Czech Republic, Denmark and Poland, only the second wave. We focus on the first wave of data only. Given self-reports and vignettes were only collected on a subset of respondents and on only eight of the countries (together with some missing item response values for covariates), our working sample is 3,802 individuals.

$SHARE$ data is useful for illustrating the score test for the underlying assumptions of RC and VE in the $CHOPIT$ model since it's cross-country component has made it a particular focus of studies investigating differences in reporting behaviour and more generally the method of anchoring vignettes (for example, see Bago d'Uva et al. (2008), Angelini et al. (2012), Paccagnella (2013), Peracchi & Rossetti (2013), Van Soest & 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 differing levels of severity, using the same response categories (i.e., "None", "Mild", "Moderate", "Severe", and "Extreme"). Appendix A contains the self-assessment question together with the vignettes and Table 4 reports the frequencies for responses to these observed in the $SHARE$ 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"

Table 5. SHARE: Descriptive statistics†

	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
Age50-65	0.643	0.479	0	1
Age66-75	0.279	0.448	0	1
Age > 75	0.078	0.268	0	1

† Sample size, $N = 3802$.

and "Extreme" have been collapsed (hence $J = 4$).

We adopt the set of covariates presented in Table 5. These are included in both the index function (equation (10)) on the latent scale (as $\mathbf{x}$) and the boundary (as $\mathbf{z}$) components of the *CHOPIT* model. The specification includes binary variables for males (Male: 47% of our sample); respondents aged 66 to 75 years (Age66-75: 28%) and aged 76 and over (Age > 75: 8%); post-school education (EducPS: 21%); the presence of health conditions (AnyCond: 71%). An indicator variable representing below average hand grip strength is also included (Grip35: 53%), which is based on up to four measurements conducted by a trained interviewer.

6.2. Comparison of boundary specification

Table 6 presents the estimated parameters, $\hat{\tilde{\beta}}$, for the index function adopting the three forms for the specification of the boundary equations in the *CHOPIT* model. Column (1) presents the amended exponential form (equation (9)). The standard exponential specification, equation (6) is provided in column (2) and column (3) presents estimates for a model with linear equations for all boundaries. As is evident, the parameter estimates are similar across the three model specifications in terms of both significance and magnitude, as are the log likelihoods.

Thus, in general, we find that 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, although the effects are not statistically significant at conventional levels. Full results for the three specifications, including the boundary equations, are reported in Appendix C, Table C1.

To further investigate the model implications of the amended specification Table C2 of the Appendix presents averaged estimated boundaries for the three approaches. These are based on the estimates, μ_0, μ_1 and μ_3.

As might be expected, given their similarity with regard to the specification of the first boundary parameter, the standard exponential and linear approaches are similar. 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, $\mathbf{x}'\hat{\beta}$, and separately the estimated vignette constant term (α_1 in equation (5)). As with standard ordered probit 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 of the boundary parameters we see that these quantities are essentially identical (to three decimal places), indicating (at least approximately) equivalence of the three approaches.

Further evidence of these findings are reported in Tables C3 to C5. Table C3 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 C4 considers estimated probabilities. The average 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 C5 contains the implied partial effects for each specification; again here we see that these are essentially equivalent across model specifications.

In summary, while individual parameter estimates may vary across the different boundary specifications, it is clear that in each case essentially the same model results. 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.

6.3. Application of the score test to SHARE data

An application of the score tests to *SHARE* data is presented in Table 7. The data and specification follows that used in column (1) of Table 6. However, we make use of all three available vignettes (see Appendix A for a description). 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 $V1$ or $V3$ are used singularly and when vignettes $V2$ and $V3$ are used in combination; but rejects for $V2$ when used alone and combinations of $V1$ and $V2$; $V1$ and $V3$; and $V1$, $V2$ and $V3$. Similarly to the joint test, when we consider only RC (assuming VE holds) the score test rejects the null for all vignettes and their combinations.

The results emphasize the importance of testing for the identifying assumptions of RC and VE in applications of the $CHOPIT$ model when

Table 6. Comparison of *CHOPIT* estimated parameters with different boundary specifications

$N = 3802$	Boundary equations (1) Amended exponentials		(2) Standard exponentials		(3) Linear	
	Coef	SE	Coef	SE	Coef	SE
Structural parameters $\left(\hat{\beta}\right)$						
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
Age66-75	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	

attempting to correct for DIF. This echoes the findings of Vonkova & Hullegie (2011) who using $SHARE$, consider the performance of single vignettes for the health domains cognition, breathing and mobility, for which multiple vignettes are available. By comparing whether DIF-adjusted self-assessments are closer to an objective measure of the situation than the unadjusted self-assessment, they find that the $CHOPIT$ approach is sensitive to the choice of vignette for cognition and breathing but not for mobility. Our results suggest that for pain the assumption of VE is sensitive to the choice of vignette. However, no single vignette nor combination of vignettes, fails to reject the null for RC, or the joint test.

7. Conclusions

Inter-individual comparison of phenomena such as health status or life satisfaction that are typically self-reported on an ordered categorical scale are subject to differential item functioning due to survey respondents' adopting different response scales. Anchoring 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. The legitimate use of the vignette approach relies on the two assumptions of RC and VE. This paper develops a joint test of these assumptions based on the *score* approach. We also consider an extension to individual tests for RC and VE.

Implementation of the test is within the parametric $CHOPIT$ model

Table 7. Lagrange-multiplier test 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

that imposes a hierarchical structure on the boundary equations to preserve coherency of the boundary probabilities. The score approach requires the model to be identified under the alternative. This is achieved by augmenting the specification of the hierarchical boundary equations by including an exponential function in the first boundary. We show that this is innocuous in terms of parameter estimates 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 & Rossetti (2013) who also develop a joint test of RC and VE.

We apply the test to data from the first wave of *SHARE*, using the self-reported information on respondents' ratings of pain. Monte Carlo simulations drawn from these data show that the tests have good size and power properties in finite samples, particularly for a 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 & Rossetti (2013). In particular, failure of VE may be picked up through rejection of RC together with rejection of VE itself. This is an area where the design of vignette questions to aid respondents' common understanding of the descriptions of the hypothetical individuals can improve vignette equivalence.

References

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.

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.

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.

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.

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

Corrado, L., Weeks, M. et al. (2010), *Identification strategies in survey response using vignettes*, University of Cambridge, Faculty of Economics.

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.

Greene, W. (2018), *Econometric Analysis, 8th Edition*, Pearson Education Limited.

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

Greene, W. & Hensher, D. (2010), *Modeling Ordered Choices*, Cambridge University Press.

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

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.

Holland, P. & Weiner, H. (1993), *Differential Item Functioning*, 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 cross-country comparability: An empirical assessment of self-reported mobility, *in* B. Baltagi & F. Moscone, eds, 'Health Econometrics: Contributions to Economic Analysis, Vol 294', Emerald Publishing, United Kingdom, chapter 7, pp. 145–174.

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* pp. 461–473.

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.

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

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., 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.

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.

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**, 367399.

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.

Sirven, N., B., B. S.-E. & 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 self-rated health: Cross-sectional multi-country survey using anchoring vignettes', *British Medical Journal* **328**, 258–260.

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

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.

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.

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.

Supplementary Appendices for Greene, Harris, Knott & Rice: Specification and testing of hierarchical ordered response models with anchoring vignettes

Appendix A: Differential Item Functioning and example vignettes

Figure A1, which uses the example of a self-assessed question about pain from $SHARE$. 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 differing 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.

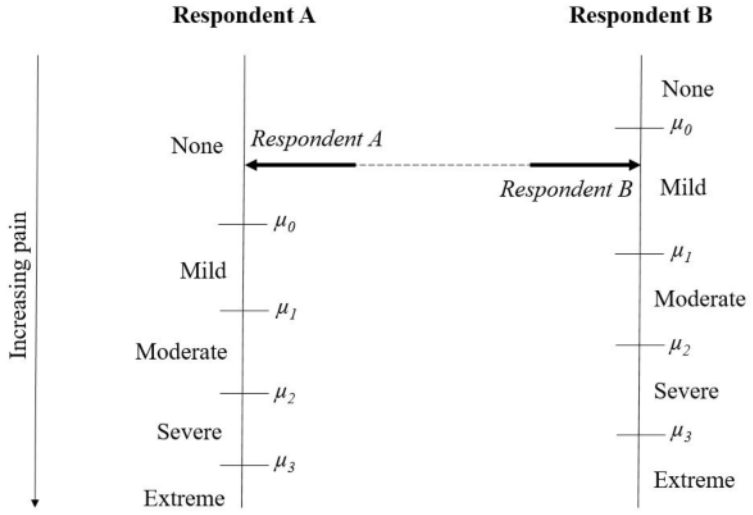


Fig. A1. Example of DIF in self-assessed pain

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?"

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

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

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

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

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 4 and how these are combined for the separate tests of RC and VE together with the joint test of both RC and VE.

First we derive the score vector for the restricted $CHOPIT$ model (equations (12) and (14)), and show how particular elements can be adapted to derive the proposed score statistic(s). The score vector for this model consists of a series of partitions. The first corresponds to β, $(\nabla\beta)$, such that

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

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

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

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

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

Collecting the above terms together gives $\nabla\tilde{\gamma}_0 = \nabla\tilde{\gamma}_{0,0} + \sum_1^K \nabla\tilde{\gamma}_{0,k}$.

The score with respect to γ_1 again consists of a quantity from the y^* equation ($\nabla\gamma_{1,0}$)

$$\begin{aligned}
\frac{\partial \ln L(\theta)}{\partial \gamma_{1,0}}|_{j=1} &= \frac{\mathbf{z} \exp(\mathbf{z}'\gamma_1)\, \phi(\mu_{1,0} - \mathbf{x}'\beta)}{P_{1,0}}, \\
\frac{\partial \ln L(\theta)}{\partial \gamma_{1,0}}|_{j=2} &= \frac{\mathbf{z} \exp(\mathbf{z}'\gamma_1)\, [\phi(\mu_2 - \mathbf{x}'\beta) - \phi(\mu_1 - \mathbf{x}'\beta)]}{P_{2,0}}, \\
&\vdots \\
\frac{\partial \ln L(\theta)}{\partial \gamma_{1,0}}|_{j=J-1} &= \frac{\mathbf{z} \exp(\mathbf{z}'\gamma_1)\, \phi(\mathbf{x}'\beta - \mu_{J-2})}{P_{J-1,0}}.
\end{aligned}$$

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

$$\begin{aligned}
\frac{\partial \ln L(\theta)}{\partial \gamma_{1,k}}|_{j=1,k} &= \frac{[\mathbf{z} \exp(\mathbf{z}'\gamma_1)/\sigma]\, \phi[(\mu_1 - \alpha_k)/\sigma]}{P_{1,k}}, \\
\frac{\partial \ln L(\theta)}{\partial \gamma_{1,k}}|_{j=2,k} &= \frac{[\mathbf{z} \exp(\mathbf{z}'\gamma_1)/\sigma]\, \{\phi[(\mu_2 - \alpha_k)/\sigma] - \phi[(\mu_1 - \alpha_k)/\sigma]\}}{P_{2,k}}, \\
&\vdots \\
\frac{\partial \ln L(\theta)}{\partial \gamma_{1,k}}|_{j=J-1,k} &= \frac{[\mathbf{z} \exp(\mathbf{z}'\gamma_1)/\sigma]\, \phi[(\alpha_k - \mu_{J-2})/\sigma]}{P_{J-1,k}}.
\end{aligned}$$

Repeating for γ_2 we have ($\nabla\gamma_{2,0}$ and $\nabla\gamma_{2,k}$). The score with respect to γ_2 again consists of a quantity from the y^* equation ($\nabla\gamma_{2,0}$)

$$
\begin{aligned}
\frac{\partial \ln L(\theta)}{\partial \gamma_{2,0}}|_{j=2} &= \frac{\mathbf{z}\exp(\mathbf{z}'\gamma_2)\,\phi(\mu_2-\mathbf{x}'\beta)}{P_{2,0}},\\
\frac{\partial \ln L(\theta)}{\partial \gamma_{2,0}}|_{j=3} &= \frac{\mathbf{z}\exp(\mathbf{z}'\gamma_2)\,[\phi(\mu_3-\mathbf{x}'\beta)-\phi(\mu_2-\mathbf{x}'\beta)]}{P_{3,0}},\\
&\vdots\\
\frac{\partial \ln L(\theta)}{\partial \gamma_{2,0}}|_{j=J-1} &= \frac{\mathbf{z}\exp(\mathbf{z}'\gamma_2)\,\phi(\mathbf{x}'\beta-\mu_{J-2})}{P_{J-1,0}},
\end{aligned}
$$

and $\nabla\gamma_{2,k}$

$$
\begin{aligned}
\frac{\partial \ln L(\theta)}{\partial \gamma_{1,k}}|_{j=2,k} &= \frac{[\mathbf{z}\exp(\mathbf{z}'\gamma_2)/\sigma]\,\phi[(\mu_2-\alpha_k)/\sigma]}{P_{2,k}},\\
\frac{\partial \ln L(\theta)}{\partial \gamma_{1,k}}|_{j=3,k} &= \frac{[\mathbf{z}\exp(\mathbf{z}'\gamma_2)/\sigma]\,\{\phi[(\mu_2-\alpha_k)/\sigma]-\phi[(\mu_1-\alpha_k)/\sigma]\}}{P_{2,k}},\\
&\vdots\\
\frac{\partial \ln L(\theta)}{\partial \gamma_{1,k}}|_{j=J-1,k} &= \frac{[\mathbf{z}\exp(\mathbf{z}'\gamma_2)/\sigma]\,\phi[(\alpha_k-\mu_{J-2})/\sigma]}{P_{J-1,k}}.
\end{aligned}
$$

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

$$
\nabla\gamma_j = \nabla\gamma_{j,0}+\sum\nolimits_k \nabla\gamma_{j,k}.
$$

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

$$
\begin{aligned}
\frac{\partial \ln L(\theta)}{\partial \alpha_k}|_{j=0,k} &= \frac{-\phi[(\mu_0-\alpha_k)/\sigma]\,\sigma^{-1}}{P_{0,k}},\\
\frac{\partial \ln L(\theta)}{\partial \alpha_k}|_{j=1,k} &= \frac{-\{\phi[(\mu_1-\alpha_k)/\sigma]-\phi[(\mu_0-\alpha_k)/\sigma]\}\,\sigma^{-1}}{P_{1,k}},\\
&\vdots\\
\frac{\partial \ln L(\theta)}{\partial \alpha_k}|_{j=J-1,k} &= \frac{\phi[(\alpha_k-\mu_{J-2})/\sigma]\,\sigma^{-1}}{P_{J-1,k}}.
\end{aligned}
\tag{16}
$$

Finally, $\nabla\sigma$ is given by

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

B.1 score test for vignette equivalence (VE)

Imposing VE is equivalent to assuming that the effect of any covariates, $\tilde{\mathbf{x}}$, entered into the model for the vignettes (see specification (11)) are zero. Accordingly, the null hypothesis of VE can be tested by

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

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 equation (16), with the generalised version ($\nabla\alpha_{VE,k}$) such that

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

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

$$score_{VE} = (\nabla\beta, \nabla\gamma_j, \nabla\alpha_{VE,k}, \nabla\sigma)\left[\mathbf{I}\left(\hat{\theta}_{VE}\right)\right]^{-1}(\nabla\beta, \nabla\gamma_j, \nabla\alpha_{VE,k}, \nabla\sigma)' \sim \chi^2_{q_{VE}}, \tag{17}$$

where $q_{VE} = \dim(\nabla\beta, \nabla\gamma_j, \nabla\alpha_{VE,k}, \nabla\sigma) - \dim(\nabla\beta, \nabla\gamma_j, \nabla\alpha_k, \nabla\sigma) = \dim(\tilde{\mathbf{x}})\, K$.

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. To evaluate expression (17) under the null, we simply set $\alpha_k = (\alpha_k, \mathbf{0})$ where α_k is the scalar estimated from the *CHOPIT* model, and the dimensions of the null vector in $\tilde{\alpha}_k$ will be determined by the number of variables in $\mathbf{x}$. 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 (13) - 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

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

Here we replace the $\nabla\gamma_j$ elements of the score with the generalised version, $\nabla\gamma_{j,RC}$. For example, the generalised partition of the score vector due to $\tilde{\gamma}_0$ from the vignette equations ($\nabla\tilde{\gamma}_{0,k}$) (corresponding to expressions in 16) is

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

Generalising other elements of $\nabla\gamma_j$, the score test is given by

$$score_{RC} = (\nabla\beta, \nabla\gamma_{j,RC}, \nabla\alpha_k, \nabla\sigma) \left[\mathbf{I}\left(\hat{\theta}_{RC}\right)\right]^{-1} (\nabla\beta, \nabla\gamma_{j,RC}, \nabla\alpha_k, \nabla\sigma)' \sim \chi^2_{q_{RC}},$$

where $q_{RC} = \dim(\nabla\beta, \nabla\gamma_{j,RC}, \nabla\alpha_k, \nabla\sigma) - \dim(\nabla\beta, \nabla\gamma_j, \nabla\alpha_k, \nabla\sigma)$.

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

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

$$score_{joint} = (\nabla\beta, \nabla\gamma_{j,RC}, \nabla\alpha_{VE,k}, \nabla\sigma) \left[\mathbf{I}\left(\hat{\theta}\right)\right]^{-1} (\nabla\beta, \nabla\gamma_{j,RC}, \nabla\alpha_{VE,k}, \nabla\sigma)' \sim \chi^2_q,$$

where $q = \dim(\nabla\beta, \nabla\gamma_{j,RC}, \nabla\alpha_{VE,k}, \nabla\sigma) - \dim(\nabla\beta, \nabla\gamma_j, \nabla\alpha, \nabla\sigma)$.

Appendix C: Supporting tables

Table C1. Comparison of *CHOPIT* estimated parameters with different boundary specifications

$N = 3802$	Boundary equations: Amended exponentials		Standard exponentials		Linear	
	Coef	SE	Coef	SE	Coef	SE
Structural parameters $\left(\hat{\beta}\right)$						
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
Age66-75	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
Age66-75	0.004	0.046	-0.024	0.048	-0.002	0.048
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
Age66-75	-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
Age66-75	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 C2. $SHARE$: Averaged estimated boundaries and first probability index†

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

† Note, $\hat{\alpha}_1$ is the estimated constant term for the vignette (α_1 in equation (5)). $\mathbf{x}'\hat{\beta}$ is the estimated linear index including the constant. Boundary equations are estimated hierarchically such that $\mu_j = \mu_{j-1} + exp\left(\mathbf{x}'\hat{\beta}\right)$. Accordingly, subtracting the linear index, $\mathbf{x}'\hat{\beta}$, from the first boundary, μ_0, affects all subsequent boundaries. This is denoted above for μ_1 and μ_2 as "$\hat{\mu}_1$ with $\left(\hat{\mu}_0 - \mathbf{x}'\hat{\beta}\right)$" and "$\hat{\mu}_2$ with $\left(\hat{\mu}_0 - \mathbf{x}'\hat{\beta}\right)$" respectively. Similarly for subtracting the vignette constant, $\hat{\alpha}_1$ from each boundary.

Table C3. 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 C4. $SHARE$: 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 $P(y = 3)$		
Amended exponential	1.0000	0.9998	0.9997
Standard exponential	0.9998	1.0000	1.0000
Linear	0.9997	1.0000	1.0000

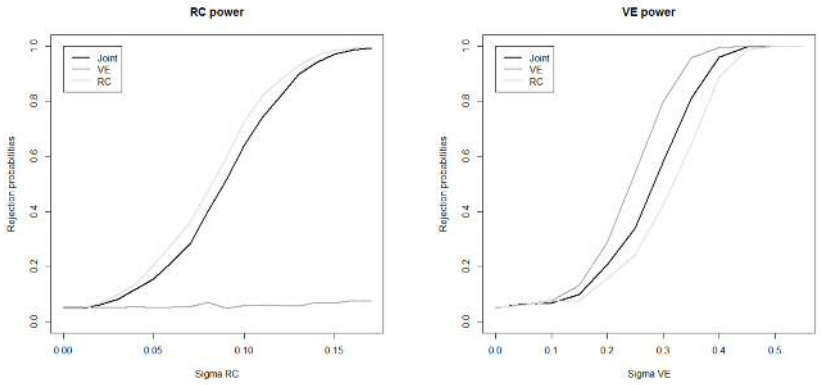
Table C5. $SHARE$: Partial effects (evaluated at sample means) and standard errors (in parentheses)

	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)
Age66-75	-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)
Age66-75	-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)
Age66-75	-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 C6. Size results, at 0.05 nominal size; unrestricted σ_k^2†

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

† Based on $M = 2,000$ repetitions.



Departures from RC Departures from VE

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

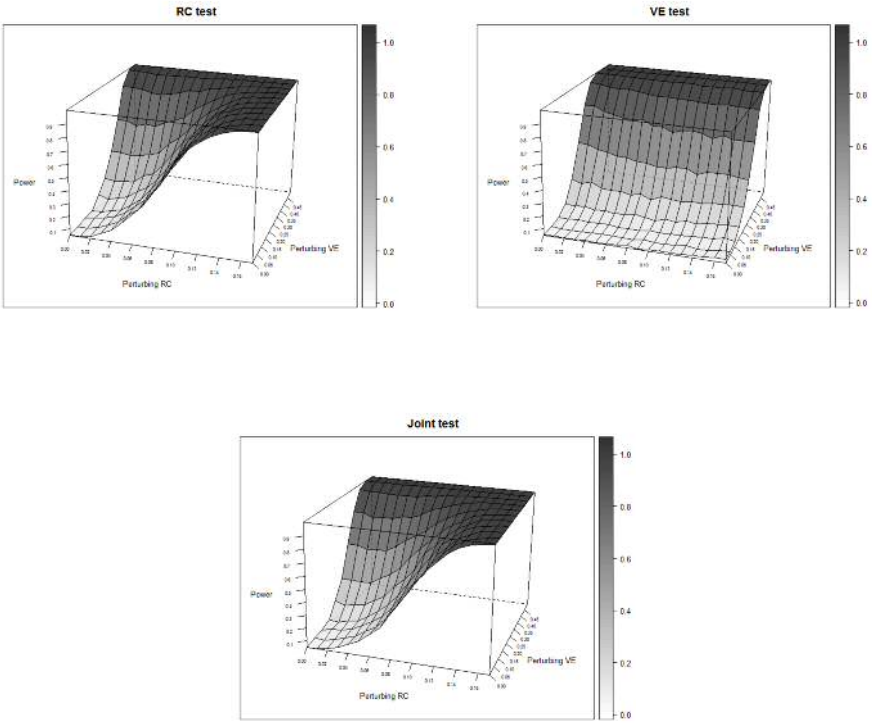


Fig. C2. Power planes for rejection probabilities for departures from RC (lhs top), VE (rhs top) and a joint test (bottom): unrestricted σ_k^2