Master Thesis

Endpoint and Midpoint Response Styles: Two Sides of the Same Coin?

Author: Martijn Schoenmakers

SNR: u1273559

ANR: 440738

Supervisor: Jesper Tijmstra

13593 words

Abstract

Response styles, the tendency of participants to respond to items regardless of the item content, have frequently been found to decrease the validity of Likert-type questionnaire results. While many models have been proposed to compensate for and model these response styles, it is still not entirely clear how these response styles relate to each other. Specifically, it is not always clear whether endpoint responding (the tendency to endorse only the extreme responses in a questionnaire) is the opposite of midpoint responding (the tendency to endorse only the middle responses in a questionnaire), or whether these response styles are two separate dimensions in a given dataset. How these response styles are modelled influences the estimation complexity, parameter estimates, and detection of and correction for response styles in IRT models. In this paper, we thus examine if it is possible to distinguish endpoint and midpoint responding as being either two separate dimensions, or being two opposite sides of a single dimension in any given dataset using the AIC and BIC. Furthermore, we assess under what circumstances (factors sample size, test length, number of substantive dimensions, response style strength, and response

style correlation) this assessment is possible and what the degree of error of this assessment is. Results indicate good performance for the AIC and BIC in a null condition and with extreme and midpoint responding as a single dimension, but worse performance for extreme and midpoint responding as two dimensions, depending on the considered factors. Additionally, interactions between factors are found.

Keywords: Response styles, endpoint response style, midpoint response style, item response theory, multidimensional nominal response model, simulation study

---

The use of Likert scales (e.g., answer scales from 1 to 5) throughout questionnaires in social science is widespread (Croasmun & Ostrom, 2011; Sullivan et al., 2013; Van Vaerenbergh & Thomas, 2013). It is however uncertain whether all participants utilize Likert response options in the same manner (Plieninger & Meiser, 2014). For example, some people may prefer to answer questions using only the ends of the scale, while others tend to answer with the middle response. These tendencies of participants to answer items a certain way regardless of content are referred to as response styles (Plieninger & Meiser, 2014; Van Vaerenbergh & Thomas, 2013). A variety of response styles exist, for example tending to use endpoints of the scale (extreme response style; ERS), often using the midpoint of the scale (midpoint response style; MRS), often agreeing with items (acquiescent response style; ARS), and many more. Participants exhibiting these response styles have repeatedly been found to reduce the validity of results obtained from Likert questionnaires in two ways (Van Vaerenbergh & Thomas, 2013). First, means and variances of results may be affected. If participants for example show ARS, agreeing with most items, means will be increased and variance will be decreased. Second, and perhaps more importantly, systematic error can be introduced into results if some participant

groups differ in their response style. For instance, a study by Moors (2012) found that a higher amount of ERS in women than men led to a spurious relation between gender and leadership styles. It is thus important to identify and correct for response styles to prevent unwarranted conclusions from being drawn. The following sections covers two often studied response styles, characteristics of participants and questionnaires that influence their occurrence, and research into their consequences in more detail (Batchelor et al., 2013; Greenleaf, 1992; He et al., 2014; Hui & Triandis, 1989; Morren et al., 2011; Naemi et al., 2009; Wetzel et al., 2013).

ERS

ERS is defined as the tendency of participants to favour answering extremely on rating scales, independent of the item content (Greenleaf, 1992). Since ERS causes participants to respond in a certain way regardless of item content, it has received much attention as a possible source of bias in questionnaire data (Greenleaf, 1992; Hui & Triandis, 1989; Morren et al., 2011; Naemi et al., 2009; Wetzel et al., 2013). ERS can bias results in three important ways. First, it may result in bias concerning measurement and relations to other variables (Batchelor et al., 2013), as exemplified earlier by the spurious correlation between gender and leadership styles (Moors, 2012). Second, ERS adds construct-irrelevant variance to the results. By adding construct-irrelevant variance, within group variance is increased and power decreases. Third, the addition of construct-irrelevant variance also reduces the magnitude of associations between constructs, analogous to the attenuation of correlations that occurs when introducing random measurement error. For example, one study found that ERS reduced the explained variance from 69.5% to 53.5% (Lau, 2007). Note that ERS can thus both lead to spurious correlations as found in Moors (2012), and to attenuation of correlations as found in Lau (2007).

Due to the various negative effects of ERS on questionnaire results, many studies have

attempted to link ERS to various person characteristics, such as personality (Naemi et al., 2009), culture (Hui & Triandis, 1989), race, gender, education, and intelligence (Batchelor et al., 2013). In addition, properties of the questionnaire have been found to affect ERS, such as response format (Lau, 2007), and remarkably even the visual distance between response options and whether the questionnaire options were presented vertically or horizontally (Weijters et al., 2021). While much research has been done on the predictors of extreme responding, the question of how well extreme responding can be detected in a dataset remains relatively unanswered. Simulation studies done on extreme responding instead tend to focus more on the impact of extreme responding on outcomes and correcting for the extreme response styles (Plieninger, 2017; Wetzel et al., 2016 for examples). One notable exception is a study by Jin and Wang (2014), which examines the accuracy of the DIC for detecting ERS as a secondary question. While this study offers encouraging results in establishing the presence of ERS, the amount of replications (20) should be increased before definitive conclusions are drawn (Jin & Wang, 2014). The following section covers the current research for midpoint response styles.

MRS

MRS is defined as the tendency of participants to favour the middle response option of a scale, regardless of item content (Van Vaerenbergh & Thomas, 2013). Just as ERS, MRS has received attention as a possible source of bias. Similar to ERS, MRS may result in invalid measurement of participants, and lead to artificially lowered or heightened correlations between variables. The effect of MRS on the variance is however quite different from ERS. Where ERS generally adds construct irrelevant variance, MRS generally deflates variance by making participants more often choose the midpoint of the scale. The deflated variance that may result from this can lead to an increased magnitude of relationships between variables, increasing the

risk of spurious findings (Van Vaerenbergh & Thomas, 2013).

Several studies have attempted to explain causes for MRS. Person characteristics that have been linked to MRS include evasiveness (not wanting to reveal one's true opinion), indecision, and indifference (Baumgartner & Steenkamp, 2001). In addition, some research has established the relation between MRS and socioeconomic development, religious denomination, Hofstede values, Schwartz values and traditionalism at the country level (He et al., 2014). Again, while research has explored the impact of MRS on obtained results, and the causes of MRS have been studied, the detection of MRS has received surprisingly little research. The following section describes various viewpoints and models present in the literature of how ERS and MRS relate to each other.

ERS and MRS

While the influences of ERS and MRS response styles on questionnaire data are relatively clear, the relation of ERS and MRS to each other is not. ERS and MRS are sometimes modelled as opposite ends of a single dimension, and other times modelled as two separate, independent dimensions (Falk & Cai, 2016). The decision to model ERS and MRS as two separate or one singular dimension has various consequences. First of all, this modelling choice will result in a more or less complex model, with different item and person parameters. Second, the ease of identifying the response style may be influenced. Finally, the correction for the response style will take a different form depending on which operationalization is used. Since the choice between these operationalizations has various important consequences, this paper aims to answer the following research question: Is it possible to empirically distinguish endpoint responding and midpoint responding as separate dimensions or as opposite points of a single dimension in a given dataset, and if so, under what conditions and with what degree of error? In

order to answer this question, a simulation study will be conducted in R. Since various approaches to modelling response styles exist, the following section gives an overview of these approaches and explains the approach used here.

Modelling response styles

A variety of approaches to modelling, measuring, and correcting for response styles exists. For example, ERS can be modelled using extensions of the rating scale, partial credit, generalized partial credit, and graded response model (Jin & Wang, 2014; Johnson, 2003; Wang et al., 2006; Wang & Wu, 2011). In addition, both mixed Rasch models and unfolding models have been proposed to model ERS and ERS + ARS respectively (Javaras & Ripley, 2007; Wetzel et al., 2013). Besides these models, various approaches using the multidimensional nominal response model exist (Takane & de Leeuw, 1987). In the multidimensional nominal response model, the probability of a participant endorsing an item category is calculated using Equation 1:

$$P(Y = k|\boldsymbol{x}, \tilde{\boldsymbol{a}}, \boldsymbol{c}) = \frac{\exp(\tilde{\boldsymbol{\alpha}}_k'\boldsymbol{x} + \boldsymbol{c}_k)}{\Sigma_{\mathrm{m=1}}^{\mathrm{k}}\exp(\tilde{\boldsymbol{\alpha}}_m'\boldsymbol{x} + \boldsymbol{c}_m)} \qquad (1)$$

Where $\tilde{\boldsymbol{\alpha}}_k'$ is the *k*'th (with subscript k denoting the item category) element of $\tilde{\boldsymbol{\alpha}}$, the vector of slopes (1 per dimension), $\boldsymbol{x}$ is the vector of participant scores on the dimension(s) (1 per dimension), and $\boldsymbol{c}_k$ is the *k*'th element of $\boldsymbol{c}$, the vector of item intercepts (1 per item category). Some of the approaches utilizing the multidimensional nominal response model add discrete latent traits to represent ERS (Moors, 2003; Morren et al., 2011) and ERS and ARS (Kieruj & Moors, 2013). Response style classes (Van Rosmalen et al., 2010) and continuous latent traits for ERS are also used (Bolt & Johnson, 2009).

While these models are valuable additions to the literature, they are often somewhat

limited in the ability to compare continuous ERS and MRS response styles in a single model. The models utilizing discrete latent traits or response style classes do not model response styles as continuous traits (Kieruj & Moors, 2013; Moors, 2003; Morren et al., 2011; Van Rosmalen et al., 2010). While some other approaches do model response styles as continuous traits, they either do not model both ERS and MRS at the same time, or do not model ERS and MRS as both separate dimensions and opposite points of a single dimension (Bolt & Johnson, 2009; Jin & Wang, 2014; Thissen-Roe & Thissen, 2013). In order to model ERS and MRS both as separate and combined continuous dimensions, an extension of the multidimensional nominal response model is used in this study (Falk & Cai, 2016). We chose this model due to its flexibility in modelling various response styles. The use of a single, flexible model to generate the data facilitates a straightforward and fair test for the presence of different possible response styles. In the extension of the multidimensional nominal model, the probability of a participant endorsing a item category is calculated using Equation 2 (Falk & Cai, 2016):

$$P(Y = k|\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{S}, \boldsymbol{c}) = \frac{\exp([\boldsymbol{a} \odot \boldsymbol{s}_k]'\boldsymbol{x} + \boldsymbol{c}_k)}{\Sigma_{\mathrm{m}=1}^{\mathrm{k}} \exp([\boldsymbol{a} \odot \boldsymbol{s}_m]'\boldsymbol{x} + \boldsymbol{c}_m)} \qquad (2)$$

where $\boldsymbol{a}$ is vector of slope parameters (1 for each dimension), $\boldsymbol{s}_k$ is the *k*'th column (with subscript k denoting the item category) of $\boldsymbol{s}$, the matrix containing the scoring functions (1 row per dimension, 1 column per item), $\odot$ denotes Schur/Habermard component wise multiplication, $\boldsymbol{x}$ is a vector of participant score(s) on the latent variable (1 per latent variable), and $\boldsymbol{c}_k$ is a the *k*'th column of $\boldsymbol{c}$, the vector of item intercepts (1 per item category). The present model gives a great deal of flexibility in modelling data, since any response style or combination of response styles can be modelled using the scoring function matrix $\boldsymbol{s}$. For instance, if we want to create a

model where no response style is present (null model), there are five answer options, and there is only one substantive dimension, we could use the following scoring function (Falk & Cai, 2016):

$$\boldsymbol{s}_m = [0 \quad 1 \quad 2 \quad 3 \quad 4] \tag{3}$$

Note that the multidimensional nominal response model given in Equation 2 reduces to the partial credit or graded partial credit model when we use the scoring function defined in Equation 3 (Falk & Cai, 2016; Masters & Wright, 1997; Muraki, 1992). In the same model with ERS and MRS added as a single dimension with ERS and MRS being modelled as opposites (ERSMRS), we could use the scoring function:

$$\boldsymbol{s}_m = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \\ 2 & 1 & 0 & 1 & 2 \end{bmatrix} \tag{4}$$

with the first row representing the scoring function of the first dimension (i.e., the dimension that relates to the latent trait of interest), and the second row representing the scoring function for the ERSMRS dimension (Falk & Cai, 2016). With this scoring function, participants with a positive score on the second dimension will show ERS and anti-MRS (Falk & Cai, 2016). Participants with a negative score on the second dimension will show the reverse pattern.

Finally, with ERS and MRS added as two separate dimensions (ERS + MRS), we would use the scoring function:

$$\boldsymbol{s}_m = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \tag{5}$$

with the first row representing the scoring function for the content factor, the second row

representing the scoring function for ERS, and the third row representing the scoring function of MRS (Falk & Cai, 2016). Due to each response style having its own row in the scoring function, a negative score on ERS does not necessarily entail a positive score on MRS, unlike the previous scoring function presented in Equation 4. The following section illustrates how these scoring functions and the extension of the multidimensional nominal response model will be implemented to answer the research question in this study.

Methods

The present study is conducted using R 4.0.4 (R Core Team, 2020). First, five-option Likert-scale data for participants are generated according to the extension of the multidimensional nominal response model (Falk & Cai, 2016), as given in Equation 2. To start, participant scores on the latent trait of interest are drawn from a normal distribution with a mean of zero and a standard deviation of one. For the condition where ERSMRS is modelled, participant's response style trait scores are drawn from a normal distribution with a mean of zero and a variable standard deviation. For the condition where ERS + MRS are modelled, response style trait scores are drawn from a bivariate normal distribution with a vector of zeros as the mean and a variable covariance matrix. This enables a correlation between response styles to be modelled. Note that the response style dimensions are always modelled as independent from the substantive dimensions (i.e., the dimension(s) corresponding to the trait(s) of interest). Depending on the model, one of the scoring functions described in Equations 3, 4 or 5 is used. The slope vector $\boldsymbol{\alpha}$ is set to a vector of ones, while the item intercepts $\boldsymbol{c}$ are set to -2, -1, 0, -1, -2. These values were chosen to ensure a symmetrical distribution of the answer probabilities (i.e., the distribution is mirrored, with for example the probability of endorsing category 1 with a trait score of -1 being equal to the probability of endorsing category 5 with a trait score of 1), with a

participant having a maximal probability of answering in the middle category when they have a score of zero on the substantive trait of interest. The answer probabilities resulting from entering the person parameters, scoring functions, slope vectors and intercept vectors into Equation 2 are then converted into answers.

To examine the research question in this study, several conditions will be examined. First, null models with no response styles present will be generated. Second, models with ERSMRS as a single dimension will be generated. Finally, models with ERS and MRS as two separate dimensions will be generated. All of these models will be evaluated using the mirt R package (Chalmers, 2012) as originating from one of three models; a model with no response styles present, a model with ERS and MRS as two dimensions, or a model with ERS and MRS as a single dimension. The fitted model with the lowest AIC (Akaike, 1998) or BIC (Schwarz, 1978) is chosen as the preferred model for each respective criterion. In this way, we obtain an overview of how often each criterion chooses the correct model under various conditions. The following section describes how the data are generated and evaluated under the null condition.

Null condition

This condition examines whether we can establish the absence of a response style in the data, and our degree of error in doing so under varying conditions. Since we are interested in a model with no response styles present, we use the scoring function from Equation 3 in combination with the person and item parameters discussed in the previous section to generate data. In addition, several factors are varied. First of all, the sample size and the test length are known to influence the accuracy of IRT model estimates (Akour & AL-Omari, 2013; Şahin & Anıl, 2017; Stone & Yumoto, 2004). Based on these previous studies, we expect higher sample sizes and higher test lengths to lead to greater accuracy. Sample sizes of 250, 500, and 1000 will

be used, with test lengths of 10 and 20 items. Second, the number of substantive dimensions (i.e., dimensions concerning a substantive trait rather than a response style) may influence the ease of detecting response styles such as ERS and MRS (Plieninger, 2017). We thus expect a higher number of substantive dimensions to lead to greater model classification accuracy. For this reason, the model will be run with 1 and 2 substantive dimensions. Note that substantive traits are independent of each other in conditions where two are modelled. The number of items per dimension will be equal to the test length divided by the number of dimensions. For the null condition, this leads to a total of $3 * 2 * 2 = 12$ data generating conditions.

After generating data with the scoring function from Equation 3 using the procedure and conditions described earlier, the mirt R package (Chalmers, 2012) is used to obtain parameter estimates for the data generated. The mirt package requires a model, an itemtype, and expected scoring functions as input. If these are provided in the following format (model = expected dimensions and which items they load on, itemtype = "gpcm", gpcm_mats = expected scoring matrix), mirt is able to use the extended multidimensional nominal response model to obtain parameter estimates (Falk & Ju, 2020). Since the gpcm_mats argument enables us to specify an expected scoring matrix, we can utilize the scoring matrices specified in Equation 3, 4, and 5. This allows mirt to evaluate all three models of interest. Evaluating the models with mirt results in multiple fit indices, including the AIC and BIC. Using these model fit indices, the model with no response styles as the expected scoring matrix should (in the vast majority of cases) be preferred over a model with ERS and MRS as two dimensions, or ERS and MRS in a single dimension as expected scoring matrices. Every data-generating condition is evaluated by mirt for all three possible scoring functions, resulting in 3 BIC's and 3 AIC's per iteration. For both criteria, the model with the lowest estimated criterion is selected as the preferred model. The

following section describes this process for the condition with ERS and MRS as a single dimension.

ERSMRS

The data-generating scoring function for ERSMRS is given in Equation 4. The approach for this scenario is similar to the approach for the null model. The only difference is that because of the inclusion of a response style dimension one condition is added. Specifically, the strength of the response styles present is likely to affect results, with more extreme response styles being more easily detectable (Cho, 2013). The standard deviation of the response styles will thus vary between values 0.6, 1 and 1.5. These values were specifically chosen in order to create an influence of response styles that is weak, normal and strong respectively, while still maintaining realistic influences (considering the influences of the response style for being +1/-1 SD from the mean, while keeping the substantive dimension constant). We expect higher values of the response style standard deviation to lead to higher model classification accuracy. This results in $3 * 2 * 2 * 3 = 36$ data-generating conditions. As in the null condition, we expect the model that generated the data (ERSMRS) to be preferred in the majority of cases over the other models, especially for higher sample sizes, test lengths and number of substantive dimensions. In addition, we expect that it will be easier to identify models in conditions with a more extreme response style, leading to a lower degree of model selection error for higher response style standard deviations. The following section describes this process for the scenario with ERS and MRS as two separate dimensions.

ERS + MRS

The data-generating scoring function for ERS and MRS being two separate dimensions is given in Equation 5. This scenario differs from the ERSMRS condition in two ways. First, we now model two response style dimensions instead of one. Second, a correlation between response styles is modelled. This is done by drawing the response style scores from a bivariate normal distribution as opposed to a univariate normal distribution. Correlations between response styles vary between -.5, 0 and .5. We expect a negative correlation between response styles to lead to a lower probability of the ERS + MRS response style being detected, while a positive correlation leads to a higher probability of the ERS + MRS response style being detected. All other hypotheses remain identical. The following section describes the results of the simulation procedure described here.

Results

Null condition

First, we describe the results of the null condition. To compare the performance of the BIC to the AIC, we utilized McNemar's tests (McNemar, 1947). McNemar's test is a test applied to 2x2 contingency tables to determine whether row and column marginal frequencies are equal (i.e., whether the AIC correct classification rate is equal to the BIC correct classification rate). These tests were conducted to compare both the overall performance of the AIC compared to the BIC (i.e., by aggregating all AIC results and all BIC results), and the performance of the AIC and BIC in each condition separately. In Table 1, the percentage of cases the BIC selects the correct model (i.e., the data are generated as having no response styles, and the BIC of the model with no response styles is lowest) is shown for each combination of the

factors test length (Number of items), sample size (N) and number of substantive dimensions (θ_N). Note that all percentages are based on 500 observations per factor combination. Conditions where the BIC significantly outperforms the AIC are marked with a *. These results for the AIC are displayed in Table 2.

Table 1

Percent of Cases Correct Model is Chosen When Using the BIC in the Null Condition

θ_N	Number of items = 10			Number of items = 20		
	N = 250	N = 500	N = 1000	N = 250	N = 500	N = 1000
θ_N = 1	100*	100*	100*	100*	100*	100*
θ_N = 2	100*	100*	100*	100*	100*	100*

* = significantly different from the AIC at the 0.05 level using McNemar's test. Percentages are based on 500 observations per table cell.

Table 2

Percent of Cases Correct Model is Chosen When Using the AIC in the Null Condition

θ_N	Number of items = 10			Number of items = 20		
	N = 250	N = 500	N = 1000	N = 250	N = 500	N = 1000
θ_N = 1	86.2*	88.2*	90.4*	90.4*	92.6*	94.6*
θ_N = 2	87.2*	88.2*	92.2*	93.2*	93.2*	96.0*

* = significantly different from the BIC at the 0.05 level using McNemar's test. Percentages are based on 500 observations per table cell.

As can be seen in Table 1, the BIC perfectly indicates the correct model in the null condition for every combination of factors. Additionally, the BIC significantly outperforms the AIC in every condition. Table 2 shows that the AIC does not select the right model in all cases. Especially for low test length and low sample size, model classification accuracy decreases. While the overall classification accuracy of the AIC is not poor in and of itself, remaining above 85% in all cases, it is clear the BIC is to be preferred in this condition. Overall, the BIC significantly outperformed the AIC in the null condition (100% vs 91.03% correct, $\chi^2 = 536, df = 1, p < .001$).

To clarify the effects of the factors on the model classification accuracy, Table 3 displays main effects of the factors on the probability of the correct model being chosen.

Table 3

Main Effects of the Factors on the Probability of the Correct Model Being Chosen for the Null Condition

Factors	Percent correct BIC	Percent correct AIC
N = 250	100	89.2
N = 500	100	90.6
N = 1000	100	93.3*
Number of items = 10	100	88.7
Number of items = 20	100	93.3*
θ_N = 1	100	90.4
θ_N = 2	100	91.7

* = significant at the 0.05 level using dummy coded logistic regression analysis. Reference categories were: N = 250, Number of items = 10, and θ_N (number of dimensions) = 1. Significance thus indicates a significant difference in model classification accuracy compared to the reference category. Percentages are based on 500 observations per table cell.

Dummy coded logistic regression was used to obtain significance tests of main effects in Table 3. To prevent estimation problems in the form of infinite logits for the logistic regression, no conditions should have only correct or only incorrect classifications. We ensured this by subtracting 0.002 from the proportion correct if all classifications in a condition were correct, and adding 0.002 to the proportion correct if all classifications in a condition were incorrect (equivalent to changing one incorrect observation to correct, or vice versa). This was done for all conditions. The adjustments are purely to facilitate estimation, and these adjusted values are thus not displayed in any of the main effect tables (the tables always display the percentages correct obtained in the data). The results of the dummy coded logistic regression analysis without any correction can be found in Appendix A. Besides main effects, possible interactions between factors were examined. Adding interactions between factors to the model did not significantly increase model fit for the BIC or the AIC in the null condition, so they were not included in the model (BIC: $Deviance = 0, df = 7, p = 1$, AIC: $Deviance = 2.371, df = 7, p = .937$.

Concerning main effects, none are detected for the BIC. This is due to the model classification accuracy always being 100%, regardless of condition. For the AIC, three trends are visible. For increasing the sample size (250 vs 1000, $z = 4.515, p < .001$) and test lengths (10 vs 20, $z = 6.183,\ p < .001$), the correct model is chosen more often. Increasing the number of substantive dimensions from 1 to 2 has no significant effect on model classification accuracy ($z = 1.725, p = .085$). The following section describes these results for the ERSMRS condition.

ERSMRS condition

For the conditions where data were generated using the ERSMRS model, Table 4 displays the percentage of correct model choice with all combinations of the factors test length (Number of items), sample size (N), response style strength (σ_{RS}) and number of substantive dimensions (θ_N) is displayed for the BIC. Table 5 depicts these results for the AIC.

Table 4

Percent of Cases Correct Model is Chosen When Using the BIC in the ERSMRS Condition

Factors		Number of items = 10			Number of items = 20		
		N = 250	N = 500	N = 1000	N = 250	N = 500	N = 1000
σ_{RS} = 0.6	θ_N = 1	95.2*	100	100	100	100	100
	θ_N = 2	97.2*	100	100	100	100	100
σ_{RS} = 1	θ_N = 1	100	100	100	100	100	100
	θ_N = 2	100	100	100	100	100	100
σ_{RS} = 1.5	θ_N = 1	100	100	100	100	100	100
	θ_N = 2	100	100	100	100	100	100

* = significantly different from the AIC at the 0.05 level using McNemar's test. Percentages are based on 500 observations per table cell.

Table 5

Percent of Cases Correct Model is Chosen When Using the AIC in the ERSMRS Condition

Factors		Number of items = 10			Number of items = 20		
		N = 250	N = 500	N = 1000	N = 250	N = 500	N = 1000
σ_{RS} = 0.6	θ_N = 1	99.2*	100	100	100	100	100
	θ_N = 2	99.8*	100	100	100	100	100
σ_{RS} = 1	θ_N = 1	100	100	100	100	100	100
	θ_N = 2	100	100	100	100	100	100
σ_{RS} = 1.5	θ_N = 1	100	100	100	100	100	100
	θ_N = 2	100	100	100	100	100	100

* = significantly different from the BIC at the 0.05 level using McNemar's test. Percentages are based on 500 observations per table cell.

As can be seen in Table 4, the BIC performs very well in this condition. Only for weak response styles in combination with low sample sizes and short test length, accuracy drops below 99%. Table 5 displays these results for the AIC. In this condition, the AIC also performs very well. Again, only for a combination of lower sample sizes, low response style strength, and low test length a very slight drop in accuracy occurs. In these conditions, the AIC significantly outperforms the BIC. In the ERSMRS condition overall, the AIC significantly outperforms the BIC (99.78% vs 99.97% correct, $\chi^2 = 24.976, df = 1,\ p < .001$). While this effect is statistically significant, the practical significance of this 0.2% difference in favour of the AIC is limited.

To give a clear overview of the factor effects, Table 6 displays the main effects of factors on accuracy of model selection.

Table 6

Main Effects of the Factors on the Probability of the Correct Model Being Chosen for the ERSMRS Condition

Factors	Percent correct BIC	Percent correct AIC
N = 250	99.4	99.9
N = 500	100*	100
N = 1000	100*	100
Number of items = 10	99.6	99.9
Number of items = 20	100*	100
θ_N = 1	99.7	100
θ_N = 2	99.8	100
σ_{RS} = 0.6	99.4	99.9
σ_{RS} = 1	100*	100
σ_{RS} = 1.5	100*	100

* = significant at the 0.05 level using dummy coded logistic regression analysis, with the model including possible interactions between factors for the BIC. Reference categories were: N = 250, Number of items = 10, θ_N (the number of substantive dimensions) = 1, and σ_{RS} (the response style standard deviation) = 0.6. Significance thus indicates a significant difference in model classification accuracy compared to the reference category. Percentages are based on 500 observations per table cell.

Again, dummy coded logistic regression analysis was used to obtain significance estimates of the main effects. First, BIC effects will be discussed. For the BIC, including

interactions between factors significantly improved model fit ($Deviance = 35.188, df = 13, p < .001$), so interactions were added to the model. Several main effects appear in the model containing interactions. Higher sample size (250 vs 500, $z = 4.125, p < .001$, and 250 vs 1000, $z = 4.125, p < .001$), higher test length (10 vs 20 items, $z = 4.231, p < .001$), and higher response style strengths (0.6 to 1, $z = 4.125, p < .001$, and 0.6 to 1.5, $z = 4.125, p < .001$) lead to higher model selection accuracy, but only slightly. In addition, an interaction emerges between the sample size and response style strength. Increasing both the sample size and response style strength at the same time leads to a somewhat lower increase in model classification accuracy than would be expected based on the main effects alone ($z = -2.107, p = .035$ for all possible interactions between the sample size and the response style standard deviation). It has to be noted that the practical significance of these results is rather limited, given that the BIC displays imperfect performance in only 3 out of 36 conditions. Generalizing trends that occur outside these three specific imperfect cells is thus difficult to do.

For the AIC, adding interactions to the model did not significantly increase model fit ($Deviance = 2.138, df = 13, p = 1$). In addition, none of the main effects are significant. This is likely due to the very high performance of the AIC in all conditions. The following section displays these results for the ERS + MRS condition.

ERS + MRS condition

For the conditions where data were generated using the ERS + MRS model, Table 7 displays the percentage of correct model choices with all combinations of the factors test length (number of items), sample size (N), response style strength (σ_{RS}), number of substantive dimensions (θ_N), and response style correlations (r_{RS}) is displayed for the BIC.

Table 7

Percent of Cases Correct Model is Chosen When Using the BIC in the ERS + MRS Condition

Factors			Number of items = 10			Number of items = 20		
			N = 250	N = 500	N = 1000	N = 250	N = 500	N = 1000
$r_{RS} = -0.5$	$\sigma_{RS} = 0.6$	$\theta_N = 1$	0.0*	0.0*	0.0*	0.0*	0.0*	0.0*
		$\theta_N = 2$	0.0*	0.0*	0.2*	0.0*	0.0*	0.0*
	$\sigma_{RS} = 1$	$\theta_N = 1$	0.2*	3.0*	6.6*	1.4*	13.2*	56.8*
		$\theta_N = 2$	1.0*	3.0*	7.4*	1.4*	12.4*	52.6*
	$\sigma_{RS} = 1.5$	$\theta_N = 1$	25.6*	54.6*	86.6*	69.8*	99.2	100
		$\theta_N = 2$	22.8*	50.4*	88.8*	68.8*	97.4*	100
$r_{RS} = 0$	$\sigma_{RS} = 0.6$	$\theta_N = 1$	0.0*	1.0*	19.6*	0.0*	15.4*	91.0*
		$\theta_N = 2$	0.0*	1.2*	22.0*	0.2*	18.0*	88.4*
	$\sigma_{RS} = 1$	$\theta_N = 1$	66.2*	99.4	100	98.0*	100	100
		$\theta_N = 2$	66.4*	99.0	100	98.8*	100	100
	$\sigma_{RS} = 1.5$	$\theta_N = 1$	100	100	100	100	100	100
		$\theta_N = 2$	100	100	100	100	100	100
$r_{RS} = 0.5$	$\sigma_{RS} = 0.6$	$\theta_N = 1$	0.0*	0.8*	52.2*	0.2*	51.0*	100
		$\theta_N = 2$	0.0*	1.6*	50.8*	0.4*	39.0*	100
	$\sigma_{RS} = 1$	$\theta_N = 1$	96.6*	100	100	100	100	100
		$\theta_N = 2$	98.2*	100	100	100	100	100
	$\sigma_{RS} = 1.5$	$\theta_N = 1$	100	100	100	100	100	100
		$\theta_N = 2$	100	100	100	100	100	100

* = significantly different from the AIC at the 0.05 level using McNemar's test. Percentages are based on 500 observations per table cell.

Table 7 shows a very strong contrast between conditions, with correct model choice varying between 0% and 100%. Conditions with low response style strength, negative response style correlations, low test lengths and low sample sizes lead to especially poor performance. Table 8 displays these results for the AIC.

Table 8

Percent of Cases Correct Model is Chosen When Using the AIC in the ERS + MRS Condition

Factors			Number of items = 10			Number of items = 20		
			N = 250	N = 500	N = 1000	N = 250	N = 500	N = 1000
$r_{RS} = -0.5$	$\sigma_{RS} = 0.6$	$\theta_N = 1$	15.0*	18.8*	14.4*	11.8*	19.8*	28.0*
		$\theta_N = 2$	12.2*	13.2*	13.2*	14.8*	20.8*	24.4*
	$\sigma_{RS} = 1$	$\theta_N = 1$	35.0*	47.8*	56.2*	67.0*	87.2*	99.0*
		$\theta_N = 2$	33.0*	40.4*	58.4*	62.8*	86.0*	98.8*
	$\sigma_{RS} = 1.5$	$\theta_N = 1$	77.6*	92.2*	98.2*	99.2*	100	100
		$\theta_N = 2$	74.2*	90.8*	99.0*	99.2*	100*	100
$r_{RS} = 0$	$\sigma_{RS} = 0.6$	$\theta_N = 1$	61.8*	85.6*	99.0*	88.6*	100*	100*
		$\theta_N = 2$	60.0*	84.6*	98.6*	87.0*	100*	100*
	$\sigma_{RS} = 1$	$\theta_N = 1$	99.6*	100	100	100*	100	100
		$\theta_N = 2$	98.6*	100	100	100*	100	100
	$\sigma_{RS} = 1.5$	$\theta_N = 1$	100	100	100	100	100	100

Factors			Number of items = 10			Number of items = 20		
			N = 250	N = 500	N = 1000	N = 250	N = 500	N = 1000
		$\theta_N = 2$	100	100	100	100	100	100
$r_{RS} = 0.5$	$\sigma_{RS} = 0.6$	$\theta_N = 1$	76.4*	98.8*	100*	99.6*	100*	100
		$\theta_N = 2$	76.8*	98.6*	100*	99.6*	100*	100
	$\sigma_{RS} = 1$	$\theta_N = 1$	100*	100	100	100	100	100
		$\theta_N = 2$	100*	100	100	100	100	100
	$\sigma_{RS} = 1.5$	$\theta_N = 1$	100	100	100	100	100	100
		$\theta_N = 2$	100	100	100	100	100	100

* = significantly different from the BIC at the 0.05 level using McNemar's test. Percentages are based on 500 observations per table cell.

Overall, the results for the AIC seem to follow the same pattern as the BIC. The AIC does tend to be less variable in its model selection success, with percentage of correct choices varying between 11.8% and 100% rather than 0% and 100%. The AIC is the very clear winner in the ERS + MRS condition, significantly outperforming the BIC by a large margin (59.2% vs 84.4% correct, $\chi^2 = 13613,\ df = 1,\ p < .001$).

To enhance interpretability of the factor effects on ERS + MRS model classification accuracy, Table 9 displays main effects of the factors.

Table 9

Main Effects of the Factors on the Probability of the Correct Model Being Chosen for the ERS + MRS Condition

Factors	Percent correct BIC	Percent correct AIC
N = 250	47.7	79.2
N = 500	57.2*	85.7*
N = 1000	72.9*	88.5*
Number of items = 10	52.3	80.1
Number of items = 20	66.2*	88.8*
θ_N = 1	59.4	84.8
θ_N = 2	59.1	84.2
σ_{RS} = 0.6	18.1	67.3
σ_{RS}= 1	68.9*	88.0*
σ_{RS} = 1.5	90.7*	98.1*
r_{RS} = -0.5	28.4*	58.6*
r_{RS}= 0	74.7	97.4
r_{RS} = 0.5	77.5*	98.6*

* = significant at the 0.05 level using dummy coded logistic regression analysis, including possible interactions between factors for the BIC and AIC. Reference categories were: N = 250, Number of items = 10, θ_N (the number of substantive dimensions) = 1, σ_{RS} (the response style standard deviation) = 0.6, and r_{RS} (the correlation between response styles) = 0. Significance thus indicates a significant difference in model classification accuracy compared to the reference category. Percentages are based on 500 observations per table cell.

Clear main effects of factors emerge from Table 9 for both the BIC and AIC. First, the BIC results are discussed. Adding interactions between factors significantly increased model fit for the BIC ($Deviance = 889.83, df = 25, p < .001$). Increasing the sample size (250 to 500, $z = 9.255, p < .001$, and 250 to 1000, $z = 14.958, p < .001$), the test length (10 to 20 items, $z = 8.223\ p < .001$), the response style standard deviation (0.6 to 1, $z = 19.005, p < .001$, and 0.6 to 1.5, $z = 22.719, p < .001$), and the response style correlation (0 to .5, $z = 4.775, p < .001$) leads to more accurate model selection. Decreasing the response style correlation leads to less accurate model selection (0 to -.5, $z = -6.591, p < .001$). There is no significant effect of the number of substantive dimensions ($z = 1.210, p = .226$). Since several significant interaction effects are present, the full logistic regression model including interactions for the BIC is presented in Table 10.

Table 10

Results of the Dummy Coded Logistic Regression Model in the ERS + MRS Condition for the BIC

Factor	Estimate	std. error	z-value	p-value
Intercept	-9.286	0.530	-17.541	$< .001$
N = 500	4.685*	0.506	9.255	$< .001$
N = 1000	7.858*	0.525	14.958	$< .001$
Number of items = 20	2.003*	0.244	8.223	$< .001$
θ_N = 2	0.220	0.182	1.210	$.226$
σ_{RS} = 1	10.012*	0.527	19.005	$< .001$
σ_{RS} = 1.5	13.808*	0.608	22.719	$< .001$

Factor	Estimate	std. error	z-value	p-value
r_{RS} = -0.5	-4.221*	0.640	-6.591	$< .001$
r_{RS} = 0.5	2.200*	0.461	4.775	$< .001$
N = 500 : Number of items = 20	1.078*	0.181	5.944	$< .001$
N = 500 : θ_N = 2	-0.150	0.128	-1.170	$.242$
N = 500 : σ_{RS} = 1	-1.348*	0.474	-2.843	$.004$
N = 500 : σ_{RS} = 1.5	-1.928*	0.538	-3.582	$< .001$
N = 500 : r_{RS} = -0.5	-1.380*	0.368	-3.748	$< .001$
N = 500 : r_{RS} = 0.5	-1.451*	0.414	-3.505	$< .001$
N = 1000 : Number of items = 20	1.639*	0.226	7.248	$< .001$
N = 1000 : θ_N = 2	-0.086	0.152	-0.569	$.570$
N = 1000 : σ_{RS} = 1	-3.612*	0.588	-6.144	$< .001$
N = 1000 : σ_{RS} = 1.5	-3.816*	0.672	-5.679	$< .001$
N = 1000 : r_{RS} = -0.5	-0.949*	0.465	-2.042	$.041$
N = 1000 : r_{RS} = 0.5	-0.660	0.454	-1.454	$.146$
Number of items = 20 : θ_N = 2	-0.176	0.099	-1.784	$.074$
Number of items = 20 : σ_{RS} = 1	0.719*	0.287	2.504	$.012$
Number of items = 20 : σ_{RS} = 1.5	1.691*	0.429	3.939	$< .001$
Number of items = 20 : r_{RS} = -0.5	-1.602*	0.278	-5.765	$< .001$
Number of items = 20 : r_{RS} = 0.5	0.692*	0.220	3.151	$.002$
θ_N = 2 : σ_{RS} = 1	-0.082	0.186	-0.442	$.659$
θ_N = 2 : σ_{RS} = 1.5	-0.220	0.276	-0.795	$.427$

Factor	Estimate	std. error	z-value	p-value
θ_N = 2 : r_{RS} = -0.5	0.029	0.197	0.147	.883
θ_N = 2 : r_{RS} = 0.5	-0.205	0.122	-1.682	.093
σ_{RS} = 1 : r_{RS} = -0.5	-2.348*	0.510	-4.608	< .001
σ_{RS} = 1 : r_{RS} = 0.5	0.616	0.439	1.405	.160
σ_{RS} = 1.5 : r_{RS} = -0.5	-1.452*	0.674	-2.156	.031
σ_{RS} = 1.5 : r_{RS} = 0.5	-1.902*	0.568	-3.350	.001

Reference categories used: 250 for the sample size, with N=500 representing the 250 to 500 sample size increase, and N=1000 representing the 250 to 1000 sample size increase. 10 for Number of items, with number of items = 20 representing the increase from 10 to 20 items. 1 for dimensions, with θ_N = 2 representing the increase of 1 to 2 dimensions. 0.6 for the response style standard deviation, with σ_{RS} = 1 representing the increase from 0.6 to 1 response style standard deviation, and σ_{RS} = 1.5 representing the increase from 0.6 to 1.5 response style standard deviation. Finally, 0 for the response style correlation, with r_{RS} = -.5 representing the decrease from 0 to -.5 response style correlation, and r_{RS} = .5 representing the increase from 0 to .5 response style correlation. **:** between variables refer to interactions, and * behind estimates refer to significance at the 0.05 level.

Since the main effects originating from the model in Table 10 are already discussed above, this part will focus on the interactions between factors. Several interactions emerge in Table 10. First of all, a clear interaction between the test length and the sample size appears. Increasing both the test length and the sample size simultaneously leads to a bigger increase in model classification accuracy than one would expect based on the main effects alone. In addition, an interaction between the sample size and the response style correlation appears. The positive main effect of increasing the sample size from 250 to either 500 or 1000 is reduced when the response style correlation drops from zero to -.5. Similarly, the positive effect of raising the sample size from 250 to 500 is reduced if the correlation between response styles is increased from 0 to .5. This trend also occurs for increasing the sample size from 250 to 1000, but is not significant. Notably, the effect of raising the sample size from 250 to 500 in

combination with increasing the response style correlation from 0 to .5 is reversed from the interaction effect for the AIC appearing later. An explorative post hoc analyses (described in Appendix C) revealed the differing direction of this interaction for the AIC and BIC to be a consequence of the high rate of completely correct and completely incorrect model classifications occurring in the AIC and BIC data cells, where the statistical correction applied to these cells may have had an impact on the estimated effect. While this makes the effect of the sample size interaction with positive response style correlations difficult to interpret, the interaction of the sample size with negative response style correlations is clear, since the same trends occur for the AIC and BIC. A final interaction concerning the sample size emerges. Increasing the sample size from 250 to 500 or 1000 has a lower positive effect on correct model classification accuracy if the response style standard deviation is simultaneously increased from 0.6 to either 1 or 1.5. Notably, this interaction is in a different direction than the interaction for the AIC appearing later. The same post hoc analysis as described earlier revealed the direction of this interaction for the BIC to be another artefact.

Second, several interactions concerning the test length appear. A positive interaction between the test length and the response style standard deviation emerges. Raising the test length from 10 to 20 items, in combination with raising the response style standard deviation from 0.6 to either 1 or 1.5 leads to a substantially higher increase in model classification accuracy than one would expect based on the main effects of the test length and the response style standard deviation alone. Additionally, there is an interaction between the test length and the response style correlation. Increasing the test length while simultaneously decreasing the response style correlation from 0 to -.5 leads to a greatly reduced increase in model classification accuracy. For positive correlations, the reverse pattern appears, where the longer test leads to better model

classification than can be expected by main effects alone.

Finally, two interactions appear involving the response style standard deviation. Increasing the standard deviation from 0.6 to 1 or 1.5 leads to a lower increase in model classification accuracy than would be expected based on main effects if the response style correlation is also lowered from 0 to -.5. The reverse pattern appears for positive correlations, where the increase in model classification accuracy is bigger than what would be expected given the main effects, but only for raising the standard deviation from 0.6 to 1.5. The combination of raising the response style standard deviation from 0.6 to 1 and increasing the response style correlation does not result in a significant effect. The following paragraphs discuss the results for the AIC.

For the AIC, adding interactions to the model also significantly increased model fit ($Deviance = 874.53, df = 25, p < .001$). Several main effects emerged from the model containing interactions. These main effects are comparable to the BIC main effects discussed previously. Increasing the sample size (250 to 500, $z = 12.506, p < .001$, and 250 to 1000, $z = 14.643, p < .001$), the test length (10 to 20 items, $z = 13.146, p < .001$), the response style standard deviation (0.6 to 1, $z = 15.242, p < .001$, and 0.6 to 1.5, $z = 13.282, p < .001$), and the response style correlation (0 to -.5, $z = -20.325, p < .001$, and 0 to .5, $z = 6.482, p < .001$) leads to more accurate model classification. The full logistic regression for the AIC, including interactions between factors, is presented in Table 11.

Table 11

Results of the Logistic Regression in the ERS + MRS Condition for the AIC

Factor	Estimate	std. error	z-value	p-value
Intercept	0.530	0.079028	6.706	$< .001$
N = 500	1.409*	0.112651	12.506	$< .001$
N = 1000	3.360*	0.229479	14.643	$< .001$
Number of items = 20	1.514*	0.115143	13.146	$< .001$
θ_N = 2	-0.119	0.094213	-1.268	$.205$
σ_{RS} = 1	3.699*	0.242657	15.242	$< .001$
σ_{RS} = 1.5	4.050*	0.304899	13.282	$< .001$
r_{RS} = -0.5	-2.377*	0.117	-20.325	$< .001$
r_{RS} = 0.5	0.8267*	0.128	6.482	$< .001$
N = 500 : Number of items = 20	0.541*	0.102	5.281	$< .001$
N = 500 : θ_N = 2	-0.053	0.089	-0.591	$.554$
N = 500 : σ_{RS} = 1	0.296*	0.113	2.615	$.009$
N = 500 : σ_{RS} = 1.5	0.808*	0.162	4.981	$< .001$
N = 500 : r_{RS} = -0.5	-1.158*	0.137	-8.461	$< .001$
N = 500 : r_{RS} = 0.5	1.105*	0.245	4.512	$< .001$
N = 1000 : Number of items = 20	1.100*	0.123	8.908	$< .001$
N = 1000 : θ_N = 2	0.088	0.103	0.856	$.392$
N = 1000 : σ_{RS} = 1	1.138*	0.126	9.055	$< .001$
N = 1000 : σ_{RS} = 1.5	2.418*	0.245	9.854	$< .001$

Factor	Estimate	std. error	z-value	p-value
N = 1000 : r_{RS} = -0.5	−3.382*	0.247	−13.700	< .001
N = 1000 : r_{RS} = 0.5	−0.307	0.374	−0.821	.412
Number of items = 20 : θ_N = 2	0.081	0.083	0.970	.332
Number of items = 20 : σ_{RS} = 1	1.612*	0.010	16.156	< .001
Number of items = 20 : σ_{RS} = 1.5	2.750*	0.233	11.823	< .001
Number of items = 20 : r_{RS} = -0.5	−1.662*	0.137	−12.088	< .001
Number of items = 20 : r_{RS} = 0.5	1.057*	0.260	4.068	< .001
θ_N = 2 : σ_{RS} = 1	0.004	0.092	0.047	.962
θ_N = 2 : σ_{RS} = 1.5	0.032	0.139	0.232	.816
θ_N = 2 : r_{RS} = -0.5	−0.044	0.118	−0.369	.712
θ_N = 2 : r_{RS} = 0.5	0.120	0.157	0.762	.446
σ_{RS} = 1 : r_{RS} = -0.5	−2.522*	0.251	−10.051	< .001
σ_{RS} = 1 : r_{RS} = 0.5	−0.557	0.382	−1.459	.145
σ_{RS} = 1.5 : r_{RS} = -0.5	−0.844*	0.312	−2.705	.007
σ_{RS} = 1.5 : r_{RS} = 0.5	−0.966*	0.421	−2.293	.022

Reference categories used: 250 for the sample size, with N=500 representing the 250 to 500 sample size increase, and N=1000 representing the 250 to 1000 sample size increase. 10 for Number of items, with number of items = 20 representing the increase from 10 to 20 items. 1 for dimensions, with θ_N = 2 representing the increase of 1 to 2 dimensions. 0.6 for the response style standard deviation, with σ_{RS} = 1 representing the increase from 0.6 to 1 response style standard deviation, and σ_{RS} = 1.5 representing the increase from 0.6 to 1.5 response style standard deviation. Finally, 0 for the response style correlation, with r_{RS} = -.5 representing the decrease from 0 to -.5 response style correlation, and r_{RS} = .5 representing the increase from 0 to .5 response style correlation. **:** between variables refer to interactions, and * behind estimates refer to significance at the 0.05 level.

Interaction effects for the AIC are by and large similar to the BIC. Since the effects of the two indices are similar, only the AIC interaction effects that are different will be discussed here. First of all, increasing the sample size and the response style standard deviation simultaneously leads to a larger increase in model classification accuracy than would be expected based on main effects alone for the AIC, while the opposite is true for the BIC. Second, increasing the sample size from 250 to 500 while the response style correlation is raised to .5 from 0 leads to a negative interaction effect for the BIC, while the AIC shows a positive interaction effect. Note that both of these effects that differed for the AIC and BIC were discovered to be due to artefacts, as was discussed earlier in the BIC interaction results. For all other interactions involving the AIC, the direction is the same as the BIC.

Discussion

The present study set out to establish whether it is possible to empirically distinguish endpoint responding and midpoint responding as separate dimensions or opposite points of a single dimension in a given dataset using the AIC and BIC, and if so, under what conditions and with what degree of model classification error this is possible. Data were generated under three conditions to answer this question. These conditions will be discussed below in order of appearance.

In the null condition, the absence of a response style present in the data could be established well by both the AIC and BIC. The sample size and the test length were established to have a positive influence on model selection accuracy, but only for the AIC. The BIC performed perfectly in this condition. The AIC obtained a respectable 91.03% overall correct model classification. Despite this performance by the AIC, the BIC is the better fit index to establish the absence of response styles in data. This makes sense, as the BIC naturally selects

more parsimonious models, and the null model is the most parsimonious model of the three possible models considered.

Both the BIC and the AIC were excellent in detecting the presence of ERSMRS in a dataset under nearly all conditions. Only when the response style strength is weak, the sample size is small, and the test length is low, slightly lower model classification accuracy occurs. The test length, the sample size and response style strength influenced model selection accuracy, but only for the BIC. While the AIC formally outperforms the BIC in this condition, the 0.2% difference between the two is of minimal if any practical significance. Both fit indices are well suited to detecting this particular response style.

The condition with ERS and MRS as two separate dimensions was the place where both the difference between the fit indices and the effects of the factors were most pronounced. For negative response style correlations and weak to middling response style strengths, the BIC was almost completely unable to correctly identify the model that generated the data, instead often preferring the simpler null and ERSMRS models. Even for the zero or positive response style correlations, weak response style strengths decreased the BIC's model selection accuracy to great extents. The AIC's performance generally followed the same trend as the BIC, although its overall performance was substantially higher. Both fit indices were strongly affected by the sample size, the test length, the response style standard deviation and the response style correlation for correct model classification. The effect of the response style correlation on the model classification accuracy can be explained by the fact that a -.5 response style correlation in the ERS + MRS condition makes the ERS + MRS model more similar to the ERSMRS model. Conversely, a .5 response style correlation in the ERS + MRS condition makes the model less similar to the ERSMRS model. It must be noted that the effect of the number of substantive

dimensions on the correct model classification was not observed in this and all other conditions. Overall, the AIC was most definitely the better fit index in the ERS + MRS condition.

In addition to the main effects of factors discussed above, the interacting effects of several factors were established. Of particular note here are the interactions involving the sample size and the test length, since these are the only factors researchers can readily influence. Generally, raising the sample size increases the positive effect of other variables on model classification accuracy, as can be seen in the interactions with the test length and the response style standard deviation. The negative response style correlations are a notable exception to this rule. Researchers must thus take note that simply increasing the sample size may have less effect than desired if negatively correlated response styles are present. Raising the test length follows the same pattern as raising the sample size.

Several limitations are present in this study. First of all, conducting a simulation requires various simplifying assumptions to be made. These assumptions are unlikely to hold in practice, and may influence the results obtained. For example, the item slopes and intercepts were the same for all items, which will not occur with a real questionnaire. In addition, all participant trait and response style scores were drawn from a normal distribution with a mean of zero and a standard deviation of one. In practice, it is likely participants will not all draw from the same distribution, especially if groups differ in age, gender, race, and other predictors of response styles. Furthermore, the response styles are thought to affect each item equally, which is also unlikely to hold in practice. Since midpoint responding has been found to be affected by participant fatigue, later items may for example be more affected by it than earlier items. Finally, only two response styles were modelled in this study, while more may be present in real-life data. The addition of other response styles to a dataset might complicate the process of

identifying which response styles exactly are present. Future research can expand on this study by experimenting with different participant trait score distributions, varying item parameters, varying effects of response styles on items, adding other response styles, and other variations on this study design.

Second, the number of replications in this study was limited to 500 to reduce the computational time. Ideally, at least 1000 replications would be conducted for such a study. While it is unlikely increasing the number of replications will greatly change the results, some of the computational problems with the logistic regression encountered here may have been avoided this way.

From the findings of this study, several recommendations for practical research arise. First of all, the use of the AIC or BIC seems to make quite a difference in some conditions when attempting to detect response styles. Which index is best to use will thus depend on research aims and hypotheses. If a researcher wants to have a high degree of accuracy in determining the absence of response styles, use of the BIC is the safest choice. In contrast, if a researcher expects a response style but wishes to distinguish between ERSMRS or ERS + MRS, the AIC performs better. Another approach that may be considered is a combined approach, where the BIC is used to establish the presence of response styles, and the AIC is used to establish which response style is present. Results of this approach are presented in Appendix B, and appear promising. Based on the results in this appendix, both the use of the AIC alone or the combination of the AIC and BIC is defensible. The AIC performs better than the combined approach in the ERS + MRS condition, but the combined approach is better at detecting the absence of any response style in the data. Which approach is used must thus be determined by weighing the cost of detecting a response style that is not present against the cost of not detecting a response style that is present.

Using the BIC alone is not recommended, as this results in inferior or equal performance in every condition compared to the combined approach.

Besides the choice between the AIC and BIC, the role of factors must also be discussed. The sample size and the test length must be sufficient, especially for the more complex multidimensional response style models. While the response style standard deviation and the response style correlation are not readily influenced, they must nevertheless not be neglected. As is visible in the ERS + MRS condition, these factors have substantial impacts on the sample size and the test length required to obtain accurate model classification, and a negative response style correlation can even reduce the effects of raising the sample size and the test length. Researchers may make use of the estimated parameters of models containing response styles to gain insight which condition they are most likely in. For illustration, a short example is provided. A researcher could a priori choose to use the AIC based on his aversion to not detecting a response style that is present. If the AIC selects the ERSMRS model as the preferred model, the researcher could additionally check the estimated parameters of the ERS + MRS model (i.e., the response style standard deviation and the response style correlation) and compare them to Table 8 to get some insight into how often the AIC would pick the ERS + MRS model if the ERS + MRS response style was present in reality. If the researcher had a sample size of 250, a test length of 10 items, a single substantive dimension and the estimated response style standard deviation and the estimated response style correlation were 0.6 and -.5 respectively, the ERS + MRS response style is difficult to detect (15% of 500 replications in this study). The researcher could use this information to note that while the AIC displays a preference for an ERSMRS model, a complex response style such as ERS + MRS is unlikely to be detected in current conditions, even if it was present.

Overall, the question of whether it is possible to distinguish ERS and MRS from ERSMRS can be answered with a yes. Only when a negative response style correlation occurs for the ERS + MRS condition in combination with a low response style standard deviation do we really run into problems distinguishing ERSMRS from ERS + MRS that cannot be solved by simply increasing the test length or the sample size. A researcher utilizing a test with twenty items and around 500 participants should thus not anticipate major problems in the vast majority of cases when attempting to distinguish ERSMRS from ERS + MRS. If a sample size of 500 participants is not possible, even sample sizes of 250 often suffice for this goal. It should again be noted that we do not recommend using a BIC-only approach here, as this can lead to poorer performance even under the conditions outlined above.

Summarizing, the current study establishes the possibility of empirically distinguishing between ERSMRS and ERS + MRS utilizing the AIC and the BIC. In addition, the current study further charts the effects of conditions and factors, including interactions, on the model classification accuracy. Finally, an alternative approach to only using the BIC is proposed, where the AIC is added as a second step.

References

Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected Papers of Hirotugu Akaike* (pp. 199–213). Springer. https://doi.org/10.1007/978-1-4612-1694-0_15

Akour, M., & AL-Omari, H. (2013). Empirical Investigation of the Stability of IRT Item-Parameters Estimation. *International Online Journal of Educational Sciences*, 11.

Batchelor, J., Miao, C., & Mcdaniel, M. (2013). *Extreme response style: A meta-analysis*.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response Styles in Marketing Research: A Cross-National Investigation. *Journal of Marketing Research*, *38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

Bolt, D. M., & Johnson, T. R. (2009). Addressing Score Bias and Differential Item Functioning Due to Individual Differences in Response Style. *Applied Psychological Measurement*, *33*(5), 335–352. https://doi.org/10.1177/0146621608329891

Chalmers, R. (2012). Mirt: A Multidimensional Item Response Theory Package for the R Environment. *JSS Journal of Statistical Software*, *48*. https://doi.org/10.18637/jss.v048.i06

Cho, Y. (2013). *The mixture distribution polytomous Rasch model used to account for response styles on rating scales: A simulation study of parameter recovery and classification accuracy*. https://drum.lib.umd.edu/handle/1903/14511

Croasmun, J. T., & Ostrom, L. (2011). Using Likert-Type Scales in the Social Sciences. *Journal of Adult Education*, *40*(1), 19–22.

Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, *21*(3), 328–347. https://doi.org/10.1037/met0000059

Falk, C. F., & Ju, U. (2020). Estimation of Response Styles Using the Multidimensional Nominal Response Model: A Tutorial and Comparison With Sum Scores. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.00072

Greenleaf, E. A. (1992). Measuring Extreme Response Style. *Public Opinion Quarterly*, *56*(3), 328. https://doi.org/10.1086/269326

He, J., Van de Vijver, F. J., Espinosa, A. D., & Mui, P. H. (2014). Toward a unification of acquiescent, extreme, and midpoint response styles: A multilevel study. *International Journal of Cross Cultural Management*, *14*(3), 306–322. https://doi.org/10.1177/1470595814541424

Hui, C. H., & Triandis, H. C. (1989). Effects of Culture and Response Format on Extreme Response Style. *Journal of Cross-Cultural Psychology*, *20*(3), 296–309. https://doi.org/10.1177/0022022189203004

Javaras, K. N., & Ripley, B. D. (2007). An "Unfolding" Latent Variable Model for Likert Attitude Data. *Journal of the American Statistical Association*, *102*(478), 454–463. https://doi.org/10.1198/016214506000000960

Jin, K.-Y., & Wang, W.-C. (2014). Generalized IRT Models for Extreme Response Style. *Educational and Psychological Measurement*, *74*(1), 116–138. https://doi.org/10.1177/0013164413498876

Johnson, T. R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika*, *68*(4), 563–583. https://doi.org/10.1007/BF02295612

Kieruj, N. D., & Moors, G. (2013). Response style behavior: Question format dependent or personal style? *Quality & Quantity*, *47*(1), 193–211. https://doi.org/10.1007/s11135-011-9511-4

Lau, M. Y. (2007). *Extreme response style: An empirical investigation of the effects of scale response format and fatigue*. https://search.proquest.com/openview/0bd10b01659c43d2d46115b7ce007d79/1?pq-origsite=gscholar&cbl=18750&diss=y

Masters, G. N., & Wright, B. D. (1997). The Partial Credit Model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 101–121). Springer. https://doi.org/10.1007/978-1-4757-2691-6_6

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157. https://doi.org/10.1007/BF02295996

Moors, G. (2003). Diagnosing Response Style Behavior by Means of a Latent-Class Factor Approach. Socio-Demographic Correlates of Gender Role Attitudes and Perceptions of Ethnic Discrimination Reexamined. *Quality and Quantity*, *37*(3), 277–302. https://doi.org/10.1023/A:1024472110002

Moors, G. (2012). The effect of response style bias on the measurement of transformational, transactional, and laissez-faire leadership. *European Journal of Work and Organizational Psychology*, *21*(2), 271–298. https://doi.org/10.1080/1359432X.2010.550680

Morren, M., Gelissen, J. P. T. M., & Vermunt, J. K. (2011). Dealing with Extreme Response Style in Cross-Cultural Research: A Restricted Latent Class Factor Analysis Approach.

Sociological Methodology, *41*(1), 13–47. https://doi.org/10.1111/j.1467-9531.2011.01238.x

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. https://doi.org/10.1177/014662169201600206

Naemi, B. D., Beal, D. J., & Payne, S. C. (2009). Personality Predictors of Extreme Response Style. *Journal of Personality*, *77*(1), 261–286. https://doi.org/10.1111/j.1467-6494.2008.00545.x

Plieninger, H. (2017). Mountain or Molehill? A Simulation Study on the Impact of Response Styles. *Educational and Psychological Measurement*, *77*(1), 32–53. https://doi.org/10.1177/0013164416636655

Plieninger, H., & Meiser, T. (2014). Validity of Multiprocess IRT Models for Separating Content and Response Styles. *Educational and Psychological Measurement*, *74*(5), 875–899. https://doi.org/10.1177/0013164413514998

R Core Team. (2020). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. https://www.R-project.org/

Şahin, A., & Anıl, D. (2017). The Effects of Test Length and Sample Size on Item Parameters in Item Response Theory. *Educational Sciences: Theory & Practice*. https://doi.org/10.12738/estp.2017.1.0270

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Stone, M., & Yumoto, F. (2004). The Effect of Sample Size for Estimating Rasch/IRT Parameters with Dichotomous Items. *Journal of Applied Measurement*, *5*(1), 48–61.

Sullivan, G. M., Artino, A. R., & Jr. (2013). Analyzing and Interpreting Data From Likert-Type Scales. *Journal of Graduate Medical Education*, *5*(4), 541. https://doi.org/10.4300/JGME-5-4-18

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408. https://doi.org/10.1007/BF02294363

Thissen-Roe, A., & Thissen, D. (2013). A Two-Decision Model for Responses to Likert-Type Items. *Journal of Educational and Behavioral Statistics*, *38*(5), 522–547. https://doi.org/10.3102/1076998613481500

Van Rosmalen, J., Van Herk, H., & Groenen, P. J. F. (2010). Identifying Response Styles: A Latent-Class Bilinear Multinomial Logit Model. *Journal of Marketing Research*, *47*(1), 157–172. https://doi.org/10.1509/jmkr.47.1.157

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response Styles in Survey Research: A Literature Review of Antecedents, Consequences, and Remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217. https://doi.org/10.1093/ijpor/eds021

Wang, W.-C., Wilson, M., & Shih, C.-L. (2006). Modeling Randomness in Judging Rating Scales with a Random-Effects Rating Scale Model. *Journal of Educational Measurement*, *43*(4), 335–353. https://doi.org/10.1111/j.1745-3984.2006.00020.x

Wang, W.-C., & Wu, S.-L. (2011). The Random-Effect Generalized Rating Scale Model. *Journal of Educational Measurement*, *48*(4), 441–456. https://doi.org/10.1111/j.1745-3984.2011.00154.x

Weijters, B., Millet, K., & Cabooter, E. (2021). Extremity in horizontal and vertical Likert scale format responses. Some evidence on how visual distance between response categories

influences extreme responding. *International Journal of Research in Marketing*, *38*(1), 85–103. https://doi.org/10.1016/j.ijresmar.2020.04.002

Wetzel, E., Böhnke, J. R., & Rose, N. (2016). A Simulation Study on Methods of Correcting for the Effects of Extreme Response Style. *Educational and Psychological Measurement*, *76*(2), 304–324. https://doi.org/10.1177/0013164415591848

Wetzel, E., Carstensen, C. H., & Böhnke, J. R. (2013). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality*, *47*(2), 178–189. https://doi.org/10.1016/j.jrp.2012.10.010

Appendix A

Results of the logistic regression analyses without the correction for only incorrect or only correct classifications within conditions

Null condition

Table A1

Main Effects of the Factors on the Probability of the Correct Model Being Chosen for the Null Condition

Factors	Percent correct BIC	Percent correct AIC
N = 250	100	89.2
N = 500	100	90.6
N = 1000	100	93.3*
Number of items = 10	100	88.7
Number of items = 20	100	93.3*
$\theta_N = 1$	100	90.4
$\theta_N = 2$	100	91.7

* = significant at the 0.05 level using dummy coded logistic regression analysis. Reference categories were: N = 250, Number of items = 10, and θ_N (number of dimensions) = 1. Significance thus indicates a significant difference in model classification accuracy compared to the reference category.

For the null condition, no changes in significance of effects appear.

ERSMRS condition

Table A2

Main Effects of the Factors on the Probability of the Correct Model Being Chosen for the ERSMRS Condition

Factors	Percent correct BIC	Percent correct AIC
N = 250	99.4	99.9
N = 500	100	100
N = 1000	100	100
Number of items = 10	99.6	99.9
Number of items = 20	100	100
$\theta_N = 1$	99.7	100
$\theta_N = 2$	99.8	100
$\sigma_{RS} = 0.6$	99.4	99.9
$\sigma_{RS} = 1$	100	100
$\sigma_{RS} = 1.5$	100	100

* = significant at the 0.05 level using dummy coded logistic regression analysis, with the model including possible interactions between factors for the BIC. Reference categories were: N = 250, Number of items = 10, θ_N (the number of substantive dimensions) = 1, and σ_{RS} (the response style standard deviation) = 0.6. Significance thus indicates a significant difference in model classification accuracy compared to the reference category.

Without the correction, no effects are significant for the BIC. Results for the AIC remain identical.

ERS + MRS condition

Table A3

Results of the Dummy Coded Logistic Regression Model in the ERS + MRS Condition for the BIC

Factor	Estimate	std. error	z-value	p-value
Intercept	-8.044	0.8377	-9.603	$< .001$
N = 500	3.806	0.832	4.574	$< .001$
N = 1000	6.600	0.8359	7.896	$< .001$
Number of items = 20	0.871	0.2703	3.223	$.001$
$\theta_N = 2$	0.301	0.2015	1.494	$.135$
$\sigma_{RS} = 1$	8.661	0.8404	10.306	$< .001$
$\sigma_{RS} = 1.5$	31.605	2439.114	0.013	$.990$
$r_{RS} = -0.5$	-1.380	1.4019	-0.984	$.325$
$r_{RS} = 0.5$	-0.774	0.9874	-0.784	$.433$
N = 500 : Number of items = 20	1.775	0.2182	8.134	$< .001$
N = 500 : $\theta_N = 2$	-0.194	0.140	-1.385	$.166$
N = 500 : $\sigma_{RS} = 1$	0.510	0.875	0.582	$.560$
N = 500 : $\sigma_{RS} = 1.5$	0.830	0.929	0.893	$.372$
N = 500 : $r_{RS} = -0.5$	-3.259	0.454	-7.175	$< .001$
N = 500 : $r_{RS} = 0.5$	0.299	0.943	0.317	$.751$
N = 1000 : Number of items = 20	2.805	0.256	10.940	$< .001$
N = 1000 : $\theta_N = 2$	-0.140	0.166	-0.839	$.402$
N = 1000 : $\sigma_{RS} = 1$	0.828	1.527	0.543	$.587$

Factor	Estimate	std. error	z-value	p-value
N = 1000 : σ_{RS} = 1.5	1.965	1.552	1.267	.205
N = 1000 : r_{RS} = -0.5	-5.354	1.380	-3.880	< .001
N = 1000 : r_{RS} = 0.5	2.330	0.986	2.362	.018
Number of items = 20 : θ_N = 2	-0.206	0.104	-1.985	.047
Number of items = 20 : σ_{RS} = 1	2.714	0.369	7.362	< .001
Number of items = 20 : σ_{RS} = 1.5	4.751	0.528	9.003	< .001
Number of items = 20 : r_{RS} = -0.5	-3.517	0.355	-9.912	< .001
Number of items = 20 : r_{RS} = 0.5	2.049	0.329	6.223	< .001
θ_N = 2 : σ_{RS} = 1	-0.187	0.225	-0.832	.406
θ_N = 2 : σ_{RS} = 1.5	-0.380	0.336	-1.131	.258
θ_N = 2 : r_{RS} = -0.5	0.130	0.237	0.548	.584
θ_N = 2 : r_{RS} = 0.5	-0.238	0.128	-1.859	.063
σ_{RS} = 1 : r_{RS} = -0.5	-3.950	1.378	-2.863	.004
σ_{RS} = 1 : r_{RS} = 0.5	3.860	0.996	3.875	< .001
σ_{RS} = 1.5 : r_{RS} = -0.5	-23.374	2439.114	-0.010	.992
σ_{RS} = 1.5 : r_{RS} = 0.5	1.168	3706.520	0.000	1

Reference categories used: 250 for the sample size, with N=500 representing the 250 to 500 sample size increase, and N=1000 representing the 250 to 1000 sample size increase. 10 for Number of items, with number of items = 20 representing the increase from 10 to 20 items. 1 for dimensions, with θ_N = 2 representing the increase of 1 to 2 dimensions. 0.6 for the response style standard deviation, with σ_{RS} = 1 representing the increase from 0.6 to 1 response style standard deviation, and σ_{RS} = 1.5 representing the increase from 0.6 to 1.5 response style standard deviation. Finally, 0 for the response style correlation, with r_{RS} = -.5 representing the decrease from 0 to -.5

response style correlation, and r_{RS} = .5 representing the increase from 0 to .5 response style correlation. **:** between variables refer to interactions, and * behind estimates refer to significance at the 0.05 level.

For the BIC, several effects differ. Most notably, the huge standard deviation of the strong response style factor makes the main effect and interaction effects of this factor nonsignificant. This incredibly large standard error is a consequence of the estimation problems that occur given only correct or only incorrect observations in a cell. In addition, the main effects of the response style correlation are no longer significant.

Table A4

Results of the Logistic Regression in the ERS + MRS Condition for the AIC

Factor	Estimate	std. error	z-value	p-value
Intercept	4.453	7.999	5.574	< .001
N = 500	1.501	1.161	12.932	< .001
N = 1000	4.030	3.017	13.34	< .001
Number of items = 20	1.677	1.206	13.914	< .001
θ_N = 2	-1.197	9.670	-1.238	.216
σ_{RS} = 1	4.481	3.431	13.062	< .001
σ_{RS} = 1.5	2.387	3.452	0.007	.994
r_{RS} = -0.5	-2.009	1.161	-17.309	< .001
r_{RS} = 0.5	7.337	1.301	5.638	< .001
N = 500 : Number of items = 20	8.328	1.071	7.779	< .001
N = 500 : θ_N = 2	-6.482	9.074	-0.714	.475
N = 500 : σ_{RS} = 1	5.107	1.144	4.464	< .001

Factor	Estimate	std. error	z-value	p-value
N = 500 : σ_{RS} = 1.5	1.307	1.687	7.748	< .001
N = 500 : r_{RS} = -0.5	-1.527	1.420	-10.752	< .001
N = 500 : r_{RS} = 0.5	1.682	3.077	5.466	< .001
N = 1000 : Number of items = 20	1.599	1.316	12.151	< .001
N = 1000 : θ_N = 2	7.987	1.074	0.743	.457
N = 1000 : σ_{RS} = 1	1.509	1.301	11.602	< .001
N = 1000 : σ_{RS} = 1.5	3.515	3.015	11.661	< .001
N = 1000 : r_{RS} = -0.5	-4.464	3.166	-14.097	< .001
N = 1000 : r_{RS} = 0.5	1.908	3.668	0.005	.996
Number of items = 20 : θ_N = 2	7.255	8.645	0.839	.401
Number of items = 20 : σ_{RS} = 1	1.881	1.028	18.296	< .001
Number of items = 20 : σ_{RS} = 1.5	4.291	3.728	11.51	< .001
Number of items = 20 : r_{RS} = -0.5	-2.199	1.448	-15.185	< .001
Number of items = 20 : r_{RS} = 0.5	2.638	5.186	5.087	< .001
θ_N = 2 : σ_{RS} = 1	-4.842	9.468	-0.051	.959
θ_N = 2 : σ_{RS} = 1.5	1.663	1.512	0.11	.912
θ_N = 2 : r_{RS} = -0.5	-2.689	1.244	-0.216	.829
θ_N = 2 : r_{RS} = 0.5	1.327	1.715	0.774	.439
σ_{RS} = 1 : r_{RS} = -0.5	-3.567	3.492	-10.215	< .001
σ_{RS} = 1 : r_{RS} = 0.5	1.910	4.520	0.004	.997
σ_{RS} = 1.5 : r_{RS} = -0.5	-2.110	3.452	-0.006	.995
σ_{RS} = 1.5 : r_{RS} = 0.5	-1.247	4.467	0	1

Reference categories used: 250 for the sample size, with N=500 representing the 250 to 500 sample size increase, and N=1000 representing the 250 to 1000 sample size increase. 10 for Number of items, with number of items = 20 representing the increase from 10 to 20 items. 1 for dimensions, with θ_N = 2 representing the increase of 1 to 2 dimensions. 0.6 for the response style standard deviation, with σ_{RS} = 1 representing the increase from 0.6 to 1 response style standard deviation, and σ_{RS} = 1.5 representing the increase from 0.6 to 1.5 response style standard deviation. Finally, 0 for the response style correlation, with r_{RS} = -.5 representing the decrease from 0 to -.5 response style correlation, and r_{RS} = .5 representing the increase from 0 to .5 response style correlation. **:** between variables refer to interactions, and * behind estimates refer to significance at the 0.05 level.

Effects for the AIC are similar to the BIC.

Appendix B

Results if BIC is used to establish presence of response styles, and AIC to establish which response style is present

Table B1

Percent of Cases the Correct Model is Chosen When Using the BIC to Establish the Presence of a Response Style, then the AIC to Establish Which Response Style is Present in the ERSMRS Condition

Factors		Number of items = 10			Number of items = 20		
		N = 250	N = 500	N = 1000	N = 250	N = 500	N = 1000
σ_{RS} = 0.6	θ_N = 1	94.6	100	100	100	100	100
	θ_N = 2	97.0	100	100	100	100	100
σ_{RS} = 1	θ_N = 1	100	100	100	100	100	100
	θ_N = 2	100	100	100	100	100	100
σ_{RS} = 1.5	θ_N = 1	100	100	100	100	100	100
	θ_N = 2	100	100	100	100	100	100

* = significantly different from the BIC at the 0.05 level using McNemar's test. Percentages are based on 500 observations per table cell.

As can be seen in Table 16, performance of the combined BIC/AIC approach leads to a slight decline in performance compared to both the AIC and BIC. Compared to using just the AIC, performance drops from 99.2% to 94.6% in the low sample size, low response style strength, low test length and one substantive dimension condition, while dropping from 99.8% to 97% in the two substantive dimensions condition. Compared to the BIC only approach, the first

cell drops from 95.2% to 94.6%, and the second from 97.2% to 97.0%. The combined approach thus performs worse than both approaches individually in this condition. It must be noted that this decrease in performance only occurs in two cells, and is not very large. The overall performance in this approach is 99.76%.

Table B2

Percent of Cases the Correct Model is Chosen When Using the BIC to Establish the Presence of a Response Style, then the AIC to Establish Which Response Style is Present in the ERS + MRS Condition

Factors			Number of items = 10			Number of items = 20		
			N = 250	N = 500	N = 1000	N = 250	N = 500	N = 1000
$r_{RS} = -0.5$	$\sigma_{RS} = 0.6$	$\theta_N = 1$	8.80	18.6	14.4	11.4	19.8	28.0
		$\theta_N = 2$	6.2	13.2	13.2	14.2	20.8	24.4
	$\sigma_{RS} = 1$	$\theta_N = 1$	35.0	47.8	56.2	67.0	87.2	99.0
		$\theta_N = 2$	33.0	40.4	58.4	62.8	86.0	98.8
	$\sigma_{RS} = 1.5$	$\theta_N = 1$	77.6	92.2	98.2	99.2	100	100
		$\theta_N = 2$	74.2	90.8	99.0	99.2	100	100
$r_{RS} = 0$	$\sigma_{RS} = 0.6$	$\theta_N = 1$	6.8	50.6	98.6	48.6	98.8	100
		$\theta_N = 2$	6.8	59.8	98.4	45.8	99.8	100
	$\sigma_{RS} = 1$	$\theta_N = 1$	99.6	100	100	100	100	100
		$\theta_N = 2$	98.6	100	100	100	100	100
	$\sigma_{RS} = 1.5$	$\theta_N = 1$	100	100	100	100	100	100

Factors			Number of items = 10			Number of items = 20		
			N = 250	N = 500	N = 1000	N = 250	N = 500	N = 1000
		$\theta_N = 2$	100	100	100	100	100	100
$r_{RS} = 0.5$	$\sigma_{RS} = 0.6$	$\theta_N = 1$	0.0	2.0	54.0	0.2	51.6	100
		$\theta_N = 2$	0.0	3.0	52.8	0.6	40.2	100
	$\sigma_{RS} = 1$	$\theta_N = 1$	96.8	100	100	100	100	100
		$\theta_N = 2$	98.8	100	100	100	100	100
	$\sigma_{RS} = 1.5$	$\theta_N = 1$	100	100	100	100	100	100
		$\theta_N = 2$	100	100	100	100	100	100

As can be seen in Table 17, the performance of the AIC/BIC combined approach is in-between that of the BIC and AIC by themselves. Especially for the low sample size, the low response style strength, and the low test length, performance decreases compared to the AIC-only approach. The approach does yield better results than the BIC only approach. Overall, this combined approach leads to 75.06% overall correct model classification, compared to 59.2% for the BIC and 84.4% for the AIC. If we arbitrarily assume that the probability of being in the null condition, ERSMRS condition, or ERS + MRS condition are equal ($\frac{1}{3}$), we would thus arrive at the following model classification accuracies:

BIC only: $1 * \frac{1}{3} + 0.9978 * \frac{1}{3} + 0.592 * \frac{1}{3} = 0.863267$

AIC only: $0.9103 * \frac{1}{3} + 0.997 * \frac{1}{3} + 0.844 * \frac{1}{3} = 0.9171$

AIC/BIC combined: $1 * \frac{1}{3} + 0.9976 * \frac{1}{3} + 0.7506 * \frac{1}{3} = 0.916067$

From these results, it is clear both the AIC and the combined approach are both defensible. The choice of which of these to use must be based on a researcher's aversion to type I or type II errors, with the conservative researcher preferring the combined approach and the liberal researcher preferring the AIC. Alternatively, one could take a more Bayesian approach and shift around their prior beliefs about the response style condition they are in to obtain the index best suited to them, with a heavier weight on the null condition resulting in a preference for the combined approach, and a heavier weight on the ERS + MRS condition resulting in a preference for the AIC. This use of priors must naturally be well reasoned and justified. The BIC only approach leads to almost identical or severely inferior performance to the AIC/BIC combined approach in all conditions, and we would thus recommend to at least not use this approach.

Appendix C

Explorative Post-hoc Analysis

The explorative post-hoc analysis described in the results section consisted of changing the data to reflect 50.000 observations rather than 500 replications. This reduced the impact of the correction for perfect and imperfect cells, where perfect cells had one correct observation replace by an error, and completely wrong cells had one error replace by a correct observation. The post-hoc analysis was conducted to make sure effects contradictory interaction effects between the AIC and BIC were not caused by the correction.