REVERSED ITEM BIAS: AN INTEGRATIVE MODEL

Bert Weijters, Hans Baumgartner, and Niels Schillewaert

- *Accepted for publication in Psychological Methods* -

Bert Weijters, Vlerick Business School and Ghent University, Belgium.

Hans Baumgartner, Smeal College of Business at The Pennsylvania State University.

Niels Schillewaert, InSites Consulting, New York.

Correspondence concerning this article should be addressed to

Bert Weijters, FPPW - Ghent University, Dunantlaan 2, B-9000 Ghent, Belgium; E-mail:

bert.weijters@ugent.be

REVERSED ITEM BIAS: AN INTEGRATIVE MODEL

ABSTRACT

In the recent methodological literature, various models have been proposed to account for the phenomenon that reversed items (defined as items for which respondents' scores have to be recoded in order to make the direction of keying consistent across all items) tend to lead to problematic responses. In this paper we propose an integrative conceptualization of three important sources of reversed item method bias (acquiescence, careless responding, and confirmation bias) and specify a multi-sample confirmatory factor analysis model with two method factors to empirically test the hypothesized mechanisms, using explicit measures of acquiescence and carelessness and experimentally manipulated versions of a questionnaire that varies three item arrangements and the keying direction of the first item measuring the focal construct. We explain the mechanisms, review prior attempts to model reversed item bias, present our new model, and apply it to responses to a four-item self-esteem scale (N = 306) and the six-item Revised Life Orientation Test (N = 595). Based on the literature review and the empirical results, we formulate recommendations on how to use reversed items in questionnaires.

Key words: reverse-keyed items, method effects, response styles, survey research, structural equation modeling.

Reverse-keyed items are items for which respondents' scores have to be recoded (i.e., reflected about the midpoint of the rating scale) in order for all the items in a multi-item scale to have the same directional relationship with the underlying construct of interest. The use of reverse-keyed items (also called oppositely-keyed, reversed-polarity, reverse-worded, negatively worded, negatively-keyed, keyed-false, or simply reversed items) is sometimes recommended to disrupt non-substantive responding and to enable the detection and control of aberrant response behavior when it occurs (e.g., Nunnally, 1978, Chapter 15; Paulhus, 1991).

However, research has shown that reversed items often lead to problems, particularly poor model fit of factor models (e.g., Marsh, 1986). In some cases, the problem is not simply that the model based on the originally hypothesized substantive factor structure is found to be inadequate, but that the lack of fit stimulates the revision of a more parsimonious conceptualization and the specification of additional substantive factors. For example, a unidimensional model in which high and low self-esteem are considered to be opposite poles of a single underlying continuum might be rejected in favor of a model in which separate, correlated factors are posited for positive and negative self-esteem corresponding to regular and reversed items (cf. Horan, DiStefano, & Motl, 2003; Motl & DiStefano, 2002).

A variety of models have been proposed to take into account differences in responding to regular and reverse-keyed items and to avoid the mistaken specification of additional substantive factors, including models with method factors or correlated uniquenesses for either the regular or the reversed items or both (e.g., Tomás & Oliver, 1999). Although these models generally achieve a better fit to the data than a model with only substantive factors, it is often difficult to distinguish between them based on statistical criteria, and the preferred model may differ in different applications (even when the same scale is analyzed). Furthermore, the conceptual

meaning of the various methodological specifications from the perspective of the psychology of survey response typically remains unclear. Often, researchers simply state that they are modeling method effects, mention several plausible mechanisms that could give rise to method effects in the discussion section (e.g., acquiescence, carelessness), or interpret the method effect in terms of a particular mechanism (e.g., acquiescence) without providing direct support that the proposed mechanism is actually at play. Although some recent papers have tried to establish relationships between method effects and personality variables (e.g., DiStefano & Motl, 2009), research investigating the specific mechanisms that have been implicated in problematic responses to reversed items is rare.

The goal of the present paper is to simultaneously consider three distinct mechanisms that could lead to method effects in response to items keyed in different directions and to propose an integrative model that takes into account and corrects for these extraneous influences. Two of the mechanisms, (net) acquiescent responding and carelessness, both encourage response inconsistencies between regular and reversed items and can therefore result in correlated errors or the emergence of spurious factors. The third mechanism, confirmation bias, leads to an upward or downward bias in respondents' scores, depending on the keying direction of the first item measuring the focal construct.

A multi-sample confirmatory factor-analytic specification combining experimental manipulation with response style measurement is proposed to model both (net) acquiescence and carelessness on the one hand and confirmation bias on the other hand as separate factors. Explicit measures of acquiescence (based on responses to a set of heterogeneous items different from the target items) and carelessness (in the second of two empirical studies) are used to separate acquiescence from carelessness. Confirmation bias is modeled via a manipulation of

two item orders in the questionnaire, depending on the keying direction of the first item measuring the target construct. The model is developed for a balanced scale with an equal number of regular and reverse-keyed items. Importantly, three different arrangements of the items in the questionnaire are considered, which enables the consideration of moderator effects of item positioning. In the grouped-alternated condition, related items are grouped together and regular and reverse-keyed items are alternated. In the grouped-massed condition, the items are also grouped together, but the reverse-keyed items follow a block of regular items (or vice versa). In the dispersed scale, the items are spread throughout the questionnaire, with unrelated buffer items spaced between the target items. These item arrangements are commonly observed in surveys. Hypotheses are developed to evaluate the effects of item positioning on response behavior under the three mechanisms. The proposed model is tested in two empirical studies in which 306 U.K. adults responded to a balanced four-item self-esteem scale and 595 U.S. adults responded to a balanced six-item optimism scale. Based on the findings and prior results, recommendations are formulated on how to use reversed items in questionnaires.

MECHANISMS LEADING TO REVERSED ITEM MISRESPONSE

In this section we discuss three mechanisms that can result in reversed item misresponse. The term reversed item misresponse is used in a general sense to refer to systematic differences in response to regular and reverse-keyed items (cf. Swain, Weathers, & Niedrich, 2008). Assuming that regular and reversed items are roughly equivalent indicators of the construct of interest except for the direction of keying, keying-related differences in response reflect a method effect (Podsakoff et al., 2003). As discussed below, this method effect may be due to (net) acquiescence, carelessness, or confirmation bias, and it will result in response

inconsistencies between regular and reversed items (e.g., double agreement or disagreement to an item and its opposite for net acquiescence and carelessness) or differences in mean response depending on the keying direction of the first item (for confirmation bias).

Acquiescence, disacquiescence and net acquiescence. We define acquiescence (disacquiescence) as a preference for the positive (negative) side of the rating scale. When a Likert format is used for the response scale, this means that a respondent has a tendency to agree (disagree) with items regardless of content (Paulhus, 1991). In the model of the survey response process developed by Tourangeau, Rips, and Rasinski (2000), which distinguishes between the stages of comprehension (attending to the question and interpreting it), retrieval (generating a retrieval strategy and then retrieving relevant beliefs from memory), judgment (integrating the information into a judgment), and response (mapping the judgment onto the response categories and answering the question), acquiescence affects the final stage (response), as it reflects individual differences in scale usage unrelated to content.

In prior research, acquiescence (disacquiescence) has been assessed in two distinct ways (Martin, 1964). One method is based on simultaneous (dis)agreement with regular and reversed items (i.e., endorsement of contradictory statements). The other method is based on consistent agreement with many heterogeneous items (i.e., items that vary widely in content and therefore are unlikely to share common content). Although acquiescence, defined as a preference for the agree positions on a Likert rating scale, may lead to self-contradictory responses, other mechanisms can have the same effect. Therefore, in order to distinguish between different processes, we will measure acquiescence using the second approach. Net acquiescence combines acquiescence and disacquiescence into a single index, and it can be assessed as a

person's mean response across many heterogeneous items. Defined in this way, it is basically a measure of directional (positive or negative) bias (Hui & Triandis, 1985).

Since acquiescence is a form of responding not based on content, which influences the final stage of the response process when respondents select one of the response options provided, different arrangements of the items in the questionnaire should have no differential effect on acquiescent responding. In particular, acquiescence should be evident regardless of whether items are grouped together (in either a grouped-alternated or grouped-massed sequence) or dispersed throughout the questionnaire.

Careless responding to reverse-keyed items. The term careless responding is sometimes used very broadly to refer to any type of random or nonrandom response to survey questions that is not based on content (see Meade & Craig, 2012, for a recent example). From a psychological perspective, such a broad conception of careless responding is problematic because many different mechanisms could lead to responding that is not sufficiently sensitive to content. Following Schmitt and Stults (1985) and Woods (2006), we restrict the term careless responding to situations in which respondents are inattentive to reverse-keyed items. Specifically, respondents may form expectations about what is being measured in a questionnaire and respond to individual items based on their overall position concerning the focal issue, rather than specific item content. This can occur in two ways. First, the instructions may tell respondents what the researcher is interested in, or a heading or label might be used to organize the items in the questionnaire. Second, related items may be grouped together and a respondent might infer what construct is being measured, even if the construct is not identified explicitly. If respondents answer the survey based on their expectations about what the questionnaire is trying to assess,

rather than specific item content, they may not notice that some of the items are reverse-keyed and thus respond inconsistently to regular and reversed items.

Although careless (or inattentive) responding to reverse-keyed items has the same effect as acquiescence (i.e., it leads to a response inconsistency between regular and reversed items), the problem occurs at the initial (comprehension) stage of the survey response process, not at the final (response) stage. Another difference is that, since the response is based on expectations cued by the instructions or previous items, the degree of response inconsistency should depend on the arrangement of the items in the questionnaire. Careless responding should be more pronounced when related items are grouped together, particularly if a block of items keyed in the same direction precedes an item keyed in the opposite direction (i.e., in the grouped-massed condition), because top-down processing is encouraged by this type of item arrangement. In contrast, when multiple measures of the same construct are dispersed throughout the questionnaire and mixed with other, unrelated items, respondents are less likely to form an expectation that all items are alike and careless responding should be reduced. It is not entirely clear what will happen in the grouped-alternated condition. Even though item grouping is in principle conducive to top-down processing, a single item may not be sufficient to induce the expectation that subsequent items are keyed in the same direction, especially if the keying direction of the second item is opposite to that of the first item.

Confirmation bias. Confirmation bias refers to the phenomenon that when respondents answer a question, they tend to activate beliefs that are consistent with the way in which the item is stated (Davies, 2003; Kunda et al., 1993). For example, if respondents are asked whether they are extraverted, they will tend to think of situations in which they were extraverted, whereas when they are asked whether they are introverted, they will tend to bring to mind situations in

which they were introverted. As a consequence, they will rate themselves as higher in extraversion when the question asks about extraversion. Prior research has shown that confirmation bias is due to a positive test strategy, where respondents primarily retrieve information supporting the question, although the generation of confirming evidence may also inhibit the production of disconfirming evidence (Davies, 2003).

In a survey context, split-ballot studies, in which respondents are randomly assigned to one of two versions of an item that are exact opposites of each other, show that responses tend to be biased in the direction in which the item is worded (e.g., McClendon, 1991; Schuman & Presser, 1981). Although this finding has been explained in terms of acquiescence, it could also be due to confirmation bias and a positive test strategy. If respondents answer a single question, it is not possible to distinguish between these two possibilities. However, let us assume that one group of respondents is exposed to a statement such as "I tend to be talkative" followed by the opposite statement "I tend to be quiet", whereas another group of respondents indicates their agreement or disagreement with the two statements in the reverse order. A within-person inconsistency in response (simultaneous agreement or disagreement with both statements) would indicate acquiescence or maybe careless responding. However, it is possible that participants respond consistently to the two items, but those in the first order indicate greater extraversion than those in the second order. Such a response pattern would reflect confirmation bias. In contrast to acquiescence and careless responding, confirmation bias occurs at the retrieval stage of the response process.

Confirmation bias has usually been studied for single items and evidence in support of the effect is obtained by comparing responses across participants, so little is known about how it will affect participants' responses to multiple items (both regular and reversed) measuring the

same construct. We assume that when related items are grouped together, respondents are unlikely to retrieve information consistent with the way in which the item is stated for each item separately. If the survey starts out with a question about being talkative, the response will depend on the extent to which respondents can retrieve situations in which they were talkative. An immediately following question about being quiet will not lead to the biased retrieval of instances in which the respondent was quiet because of satisficing (i.e., minimizing the effort expended on answering the questionnaire; Krosnick, 1991). Instead, the belief sample available based on the previous question will carry over to the second item, and responses will be biased in the direction of the first item (extraversion). On the other hand, if the first item is about being quiet, information related to being quiet will be retrieved and this information will carry over to subsequent items, even if they are keyed in the opposite direction. Therefore, responses will be biased in the direction of introversion. Overall, we expect that when the first item is reverse-keyed, the mean response to all items (after recoding reversed items) will be lower than when the first item is non-reversed.

Although confirmation bias should occur in both the grouped-alternated and grouped-massed conditions, it might be stronger in the latter case (i.e., when a reversed item follows a block of items keyed in the non-reversed direction) because the tendency to use a positive test strategy may be reinforced in this situation. On the other hand, when related items are dispersed throughout the questionnaire and separated by buffer items, which tend to activate unrelated content, carryover effects will be reduced. When respondents eventually encounter a reversed item, they may have to retrieve relevant information again, which is then biased in the direction in which the question is stated. Therefore, the differential biases induced by regular and reversed items should offset each other (i.e., upward bias for regular items and downward bias for

reversed items after recoding reversed items), and there should be no effect of the keying of the first item on the mean response to all items in the dispersed condition.

Summary. Our framework distinguishes three sources of reversed item bias: (net) acquiescence, careless responding to reversed items, and confirmation bias. Acquiescence and careless responding are expected to increase response inconsistency between regular and reversed items, in the sense that respondents' scores on observed measures (prior to recoding) will be uniformly elevated or depressed regardless of keying differences between the items. Confirmation bias will not lead to response inconsistency, but scores on regular items should be lower and scores on reversed items should be higher (prior to recoding) if a reversed rather than a regular item appears first in the questionnaire, because the first item biases memory retrieval in the direction in which the initial item is worded.

The manipulation of different ways in which the items are positioned in a questionnaire also allows us to consider the moderator effects of item arrangement. For net acquiescence, we do not expect differences in misresponse as a function of item positioning. However, the effect of careless responding on inconsistency bias and the effect of the manipulation of the keying direction of the first item on confirmation bias should be strongest in the grouped-massed condition and weakest in the dispersed condition.

MODELING METHOD EFFECTS IN THE PRESENCE OF REVERSE-KEYED ITEMS

There are many potential sources of method bias in survey research (for a review see Podsakoff et al., 2003). Our focus here is on one such source, item keying. Several models have been proposed to control for the biasing influence of item keying, and we will present a brief review of prior research before developing our own integrative model. We structure our review

along the following dimensions. We first distinguish between models using method factors and models using correlated uniquenesses to account for keying differences between items. Within the method factor approaches, we further distinguish between (a) method factors affecting all items or specific subsets of items, (b) method factors with freely estimated or constrained loadings, and (c) models with or without specifically designed indicators or antecedents of the method factor(s). We also describe the major findings and some limitations of previous research.

The specification of method factors to account for keying differences

A method factor is a latent variable assumed to contribute to the variability of a set of observed response variables that share a common method. Two types of method factors have been considered in the context of responses to regular and reversed items: a general method factor that underlies all observed variables (with uniformly positive loadings for all items, provided reversed items have *not* been recoded) and a method factor (possibly multiple method factors) for subsets of items. In the latter case, researchers have specified method factors for the regular items, reverse-keyed items, or both.

Models with a general method factor underlying all items. Several authors have considered a model with a single substantive factor (with positive or negative loadings for regular and reversed items, assuming reversed items have *not* been recoded) and a single method factor which influences both the regular and reversed items. The loadings on the method factor are set to one if the variance of the method factor is freely estimated, or they are constrained to be equal if the variance of the method factor is set to one (see Welkenhuysen-Gybels et al., 2003, for an example). This model is a special case of the random intercept model proposed by Maydeu-Olivares and Coffman (2006), and it is shown in Panel A of Figure 1.

– Insert Figure 1 about here –

The model with uniform positive loadings on a single, general method factor has also been applied to regular and reversed items measuring multiple substantive constructs (e.g., Billiet & McClendon, 2000). If multiple substantive factors are included, it is not necessary to constrain the loadings on the method factor to be equal and these loadings can be freely estimated (e.g., Levin & Montag, 1989). An example is shown in panel B of Figure 1. Operationally, this model is similar to a bifactor model in which the method factor is the general factor and the substantive factors are the sub-factors, although the substantive factors are allowed to be correlated (which is feasible since the signs of the loadings for the substantive and method factors differ).

Models with method factors for subsets of items. Some authors have entertained models for a single substantive construct in which a method factor with freely estimated loadings is specified for either the regular or reversed items (but not both). Illustrative examples are contained in Tomás and Oliver (1999). This specification is a special case of the so-called correlated trait, correlated method minus one or CTC(M-1) model when only one trait is available and either the reverse-keyed items or the regular items serve as the reference factor (Geiser et al., 2008). The model can be extended to several substantive factors or situations where the same substantive factor is measured repeatedly over time, in which case the longitudinal invariance of the substantive and method factors can be investigated (e.g., Motl & DiStefano, 2002). Panels C and D in Figure 1 show examples of models with one or two substantive factors and a method factor for the reversed items.

If there is a single substantive construct and separate uncorrelated method factors are specified for the regular and reversed items (with method factor loadings freely estimated), this model is a special case of the bifactor model in which the substantive construct is the general

factor and the method factors are the sub-factors (see panel E of Figure 1). It is also possible to allow the method factors to be correlated and an example of this specification is provided in Tomás and Oliver (1999). If there are multiple correlated substantive factors and separate correlated method factors are included for the regular and reversed items, the model is an application of the general correlated trait correlated method (CTCM) model (panel F of Figure 1; see Harris & Bladen, 1994, for an example).

Explicit correlates of method factors. Usually, the method factor is not directly related to a specific variable designed to explicitly capture the particular method effect in question. Exceptions are the papers by Watson (1992) and Billiet and McClendon (2000). Watson calls the method factor an acquiescence factor, which is indicated by an explicit acquiescence index (computed as the extent of strong agreement with seven control items) and which is also related to a measure of education. In the study by Billiet and McClendon (2000), a general style factor on which all indicators of two substantive constructs load equally is related to an explicit agreement index (computed as the sum of agreements with 14 items) as well as age and education (see Welkenhuysen-Gybels et al., 2003, for a similar approach). Marsh (1986) showed that the consistency of children's responses to a self-concept instrument containing both regular and reversed items was positively related to grade level and that for fifth-graders a method factor based on the reversed items was substantially correlated with reading achievement (see also Marsh, 1996). This implies that reversed items may be especially problematic in research with children and other respondents who have relatively low verbal ability.

The specification of correlated uniquenesses to account for keying differences

Method effects have also been modeled via correlated uniquenesses. As implied by the name, in this model the unique factors (error terms) associated with items that share a common

method are allowed to be correlated. Under certain circumstances, models with method factors and correlated uniquenesses can yield identical model fits, but in general the correlated uniqueness model is both *more* restrictive (method effects corresponding to regular and reversed items are assumed to be independent) and *less* restrictive (method effects are not assumed to be unidimensional) than some method factor models. Although correlated uniqueness models have been shown to be less susceptible to convergence problems, it is difficult to relate method effects to direct measures or other variables of interest. As in the case of method factors, correlated uniquenesses can be specified for the regular items, for the reversed items, or for both (for examples see Marsh, 1996, and Tomás & Oliver, 1999; for a discussion of the consequences of including correlated uniquenesses for only some but not all of the methods see Cole, Ciesla, & Steiger, 2007).

Summary of prior findings modeling method effects due to keying differences

The following conclusions have emerged from research on method effects caused by the use of reverse-keyed items based on the models distinguished in the previous sections. First, models in which method effects are included generally yield a much better fit to the data than models in which only substantive factors are included. Second, it is often difficult to clearly distinguish between different method effect specifications on the basis of statistical criteria, because different models of method effects fit about equally well. Third, the psychological processes causing method effects are frequently left unspecified, or researchers merely speculate about possible reasons for method effects in the discussion section. Several papers have used explicit measures of acquiescence to validate the interpretation of the method factor as an acquiescence factor (e.g., Billiet & McClendon, 2000). However, the acquiescence indices used in these studies are sometimes problematic, because due to data limitations the items on which

the acquiescence indices are based overlap to a great deal with the items used as indicators of the method factor. Furthermore, the indices do not contain a large number of regular or reversed items measuring the same construct and/or the items are not heterogeneous in content (both of which are prerequisites for a good acquiescence measure). To unambiguously validate the meaning of the method factor, it is necessary to correlate the method factor with an independent and reliable direct measure of the method effect of interest (e.g., acquiescence).

Finally, although method factors have been related to a variety of other psychological constructs, the choice of these other constructs often seems somewhat *ad hoc*. For example, in DiStefano and Motl (2009), social desirability, behavioral inhibition and activation, fear of negative evaluation, self-monitoring, and self-consciousness were studied as antecedents of a reverse-keyed item factor. It is not entirely clear why these specific constructs were studied as antecedents of method effects, and it would seem to be useful to investigate more immediate antecedents of method effects. Little or no research has studied method factors as a function of experimentally manipulated task characteristics such as different arrangements of the items in a questionnaire.

AN INTEGRATIVE MODEL OF SURVEY RESPONSE TO REGULAR AND REVERSED ITEMS

In this section we propose a model representing the response process to sets of items varying in keying direction as a function of different arrangements of these items in the questionnaire. This model has three characteristic features. First, it incorporates three distinct mechanisms – (net) acquiescence, careless responding to reversed items, and confirmation bias – that have been shown to influence responses to regular and reversed items. Second, the model

distinguishes between two types of method effects, response inconsistency between regular and reversed items and difference in mean response depending on whether the first item measuring the focal construct is a regular or reversed item. Acquiescence and careless responding are assumed to lead to the former, confirmation bias to the latter. In order to distinguish between acquiescence and careless responding, an explicit (net) acquiescence index (based on the extent of agreement with unrelated control items) and a direct proxy measure of careless responding are specified as antecedents of the response inconsistency factor. Confirmation bias is modeled as the difference in mean response to regular and reversed items by respondents who encounter either a regular or reversed item first (which requires a between-participant manipulation of whether a regular or reversed measure of the focal construct appears first in the questionnaire). Third, the model is developed as a multi-sample specification in order to represent different experimentally manipulated item arrangements in the questionnaire (grouped-alternated, grouped-massed, and dispersed), which enables the consideration of moderator effects of item positioning on misresponse in surveys.

Figure 2 depicts the proposed model graphically. For simplicity, we assume that there is a single focal construct, which is measured by 3 regular (non-reversed) and 3 reversed items. The regular items are denoted as p1, p2, and p3 (p for positive), the reversed items as n1, n2, and n3 (n for negative). Reversed items have not been recoded, which means that the loadings on the substantive factor (ξ) are positive for regular and negative for reversed items; otherwise, the substantive loadings are freely estimated. Of course, the model can be easily extended to multiple substantive constructs and more items.

– Insert Figure 2 about here –

The responses to the six items are modeled as a function of the substantive construct, denoted by ξ, one unique factor per item (ε_{p1}, ε_{p2}, ε_{p3}, ε_{n1}, ε_{n2}, ε_{n3}) and two method factors (η_1, η_2) . The substantive construct and unique factors conform to the usual confirmatory factor analysis (CFA) specification (Mulaik, 2009). In addition, we specify two method factors. The first method factor (η_1) models inconsistency in response to regular and reversed items (inconsistency bias). This is reflected in the fact that it has positive unit loadings for all items, regardless of their keying direction (remember that reversed items have *not* been recoded). To distinguish acquiescence from careless responding, the response inconsistency factor is regressed on an explicit measure of net acquiescence called NARS for net acquiescence response style (with regression weight γ_{NARS}) and a proxy measure of careless responding called IMC for Instructional Manipulation Check (with regression weight γ_{IMC}); the residual term of this regression is denoted as Z. The idea underlying IMC will be explained in the context of Study 2.

The second method factor (η_2) captures confirmation bias. Regular items have a positive unit loading on the confirmation bias factor, reversed items have a negative unit loading. In order to capture confirmation bias, respondents have to be randomly assigned to conditions in which the first item measuring the focal construct is either a regular or reversed item (so the first item in the scale is either p1, a regular item, or n1, a reverse-keyed item). A dummy variable (denoted as 'First item reversed' or FIR) is used to code the two conditions (1 for respondents who encounter a reversed item first, 0 otherwise). The confirmation bias factor is regressed on the 'First item reversed' dummy variable (with weight $\gamma_{Confbias}$) and has no residual variance (which, although not necessary under all conditions, facilitates model identification and convergence). The proposed specification is equivalent to directly regressing the observed items on the 'First item reversed' dummy variable and fixing the regression weights to be equal across

items (with a negative or positive sign for the regular and reversed items, respectively). However, the model in Figure 2 more clearly conveys the meaning of the confirmation bias factor and can be easily implemented with standard software.

No covariances are shown between the exogenous variables ξ (the substantive construct), NARS, IMC, and FIR. Since respondents are randomly assigned to the 'First item reversed' conditions, the dummy variable should be unrelated to the other exogenous variables. The remaining variables could be correlated. It is an empirical question whether these covariances are necessary, although a conceptual rationale should be provided if an association is to be introduced.

Figure 2 does not explicitly show that the model is formulated as a multi-sample specification, but this is assumed. Specifically, respondents are randomly assigned to conditions in which all items measuring the construct of interest are grouped together or dispersed throughout the questionnaire, separated by unrelated buffer items. In the grouped conditions, the substantive items are arranged in an alternating order based on keying direction (in the grouped-alternated condition) or a block of regular (reversed) items is followed by several reversed (regular) items (in the grouped-massed condition). For example, in our empirical studies, four items keyed in the same direction are followed by two or three items keyed in the opposite direction. Further details are provided in the context of the empirical studies.

In terms of the method effect specifications discussed earlier, our model is most similar to Billiet and McClendon (2000) and Welkenhuysen-Gybels et al. (2003), and because of the unit loadings on the inconsistency bias factor across all items, it can be considered as a random intercept factor model (Maydeu-Olivares & Coffman, 2006). However, the model is more general because acquiescence is distinguished from careless responding, confirmation bias is

taken into account, and an extension to a multi-sample context is considered so that the moderating effect of different item arrangements on the three response mechanisms can be investigated.

STUDY 1

In the first study, we conducted an initial test of the proposed model based on a four-item self-esteem scale with two regular and two reversed items. In this study, we do not have an explicit measure of careless responding, and we interpret the residual Z of the response inconsistency factor (after accounting for NARS) as a proxy for careless responding.

Method

We collected the data from the UK online panel of a European panel provider, for an effective sample size of $N = 306$. Respondents were randomly sampled from the panel, which is representative of the general population. The average age of the respondents was 46 (Min = 12; Max = 69; SD = 12.4), 51 percent of respondents were female, and 43 percent had a higher education, where higher education refers to formal education (college or university) beyond secondary school.

The focal construct in this study was self-esteem, which was measured with two regular items (formulated in the high self-esteem direction) and two reversed items: ‘Generally speaking, I feel pleased with myself’ (p1); ‘On the whole, I feel satisfied with myself’ (p2); ‘Generally speaking, I feel annoyed with myself’ (n1); and ‘On the whole, I feel frustrated with myself’ (n2). These items are similar to those found in the Rosenberg Self-Esteem Scale (Rosenberg, 1965). For the grouped-massed conditions, we wrote two additional regular or reversed items: ‘Overall, I take a positive attitude toward myself’ (p3) and ‘Generally speaking, I feel good

about myself' (p4), or 'Overall, I take a negative attitude toward myself' (n3) and 'Generally speaking, I feel bad about myself' (n4). These additional items were not included in the model specification; the purpose of these items was simply to simulate a situation in which several items keyed in the same direction lead respondents to expect a common keying direction. All items were rated on 7-point Likert scales with categories labeled (1) Fully disagree, (2) Disagree, (3) Slightly disagree, (4) Neutral, (5) Slightly agree, (6) Agree, (7) Fully agree.

We randomly assigned respondents to one of six conditions using a 3 (type of item arrangement: grouped-alternated, grouped-massed, or dispersed) by 2 (keying direction of the first focal item: regular or reversed) experimental design. In the grouped-alternated conditions (N = 106), the regular and reversed self-esteem items were presented in an alternating order (either p1, n1, p2, n2 or n1, p1, n2, p2, depending on the keying direction of the first item). In the grouped-massed conditions (N = 96), four regular items preceded two reversed items (p1, p2, p3, p4, n1, n2) or four reversed items preceded two regular items (n1, n2, n3, n4, p1, p2). In both the grouped-alternated and grouped-massed conditions, the focal self-esteem items (which were always shown on the same page) were preceded and followed by a total of 48 filler items; these were used to construct the NARS index. In the dispersed conditions (N = 104), only one self-esteem item was shown on a given page (which was always the first item on the page) and the individual self-esteem items were separated by 12 buffer items. As in the grouped conditions, the first focal item in the dispersed conditions was either a regular or reversed item.

The buffer items included to measure net acquiescence were deliberately very diverse in content as they were randomly sampled from existing scales and questionnaires. Examples of buffer items were 'I think quantitative information is difficult to understand', 'Fashion is irrelevant', 'I try to anticipate and avoid situations where there is a likely chance of my getting

emotionally involved', and 'When I go shopping, I find myself spending very little time checking out new products and brands.' To construct the NARS measure, we computed the average score across the 48 unrelated buffer items (Min = 1.85, Max = 5.04, M = 3.51, SD = .53, Cronbach's alpha = .83). The resulting index indicates a person's tendency to agree or disagree with items regardless of specific item content (Baumgartner & Steenkamp, 2001). In order to investigate the effect of confirmation bias, we formed a dummy variable called 'First item reversed' (FIR) based on the manipulation of the keying direction for the first item (coded 1 for respondents who encountered a reversed self-esteem item first and 0 otherwise).

Results

As the data were approximately normally distributed (all endogenous observed variables had skewness and kurtosis values between –1 and +1) and no missing values occurred, we fit the model to the data using multi-group structural equation modeling with the ML estimator, using MPlus 6.1. The means, standard deviations and correlations of the observed variables, as well as details about the specification of the model, are provided in the supplemental notes for this paper.

We initially estimated a one-factor model for the three groups defined by item arrangement condition using only the four self-esteem items, with one substantive factor (self-esteem) underlying the four observed variables and all estimated parameters unconstrained across the three groups. This model fit very poorly (χ^2 (6) = 153.81, $p < .0001$; RMSEA = .491; CFI = .839; TLI = .518), and the sizable modification indices for the covariances among the unique factors suggested the presence of reversed item bias.

We then specified the model shown in Figure 2, except that there were only four endogenous observed variables (instead of six) and no direct measure of careless responding was

available (i.e., the model does not include IMC). The loadings on the substantive factor were freely estimated except for one unit loading for the marker item (with positive loadings for the regular items and negative loadings for the reversed items), but the loadings on the inconsistency and confirmation bias factors were set to unity (as indicated in Figure 2). Model comparison tests for the substantive loadings across the three item arrangement conditions showed that the loadings did not differ across groups, so they were set invariant across conditions. The exogenous variables of self-esteem, NARS, and FIR were initially specified to be uncorrelated, but the modification indices for the covariance between self-esteem and NARS were substantial in each group and the model was therefore revised to incorporate this covariance. The resulting negative correlation between self-esteem and NARS is consistent with findings that low status individuals and respondents with lower education and lower income (which should also have lower self-esteem) show more acquiescence (Baumgartner & Steenkamp, 2001).

Our hypotheses imply that the effect of NARS on inconsistency bias will be invariant across the three item arrangement conditions, whereas the effect of carelessness on inconsistency bias and the effect of the keying direction of the first target item (FIR) on confirmation bias should depend on the positioning of the items in the questionnaire. To test these hypotheses, we conducted invariance tests of the relevant parameters (γ_{NARS}, σ^2_Z, and $\gamma_{Confbias}$). In support of the first prediction, γ_{NARS} did not differ across groups ($\Delta\chi^2$ (2) = .87, n.s.). We do not have an explicit measure of carelessness in this study, but if the residual in the inconsistency bias factor (σ^2_Z) is interpreted as a proxy measure for careless responding, the findings support the predicted lack of invariance of the effect of careless responding across the item arrangement conditions ($\Delta\chi^2$ (2) = 16.75, $p < .001$). Contrary to predictions, $\gamma_{Confbias}$ did not differ across groups ($\Delta\chi^2$ (2) = 2.44, n.s.). The final model, in which γ_{NARS} and $\gamma_{Confbias}$ were specified to be invariant across

item arrangement conditions but the residual variance of the inconsistency bias factor was freely estimated (see the supplemental notes for details of the model specification), fit well: χ^2 (31) = 40.98, p = .11; RMSEA = .056 (90% C.I. 0.000 – 0.099); SRMR = .054; CFI = .990; TLI = .986. The estimation results (including 95 percent confidence intervals) are shown in Table 1.

– Insert Table 1 about here –

The findings indicate that NARS significantly affects the inconsistency bias factor (γ_{NARS} = .17, p < .01) and the effect does not vary by item arrangement. Both results are in line with our predictions. The variance of Z (the residual variance of the inconsistency bias factor after accounting for NARS) is significant in all conditions (p < .05), and the effect differs across groups, with substantially less variance in the dispersed condition (σ^2_Z = .10) than in the grouped-alternated (σ^2_Z = .17) and especially the grouped-massed (σ^2_Z = .37) conditions. If one interprets the residual in the inconsistency bias factor (after accounting for NARS) as a proxy for the careless responding effect, these findings support our predictions. Finally, our data show evidence of confirmation bias since respondents reported somewhat lower self-esteem if the first item was reversed ($\gamma_{Confbias}$ = -.29, p < .05). Since the estimated parameter does not differ by item arrangement, the hypothesis that confirmation bias would primarily emerge in the grouped conditions was not supported.

To assess the relative contribution of the various factors to observed item scores, we applied variance decomposition, separately for regular and reversed items (i.e., averaging across the two regular and two reversed items). Because of the negative covariance between NARS and self-esteem, the variance for the regular items contains a negative term, which creates problems for the variance partitioning, but the negative terms are small and the variance decomposition

does provide useful insights. The results are reported in Table 2 (including 95 percent confidence intervals). It is reassuring that most of the observed variance is accounted for by the construct of interest. Also, although the effect of NARS is statistically significant, NARS contributes a negligible amount to the variance in observed scores. Similarly, the effect of positioning a reversed item first (confirmation bias) is statistically significant but very small. An interesting finding is that in the grouped-massed condition a substantial portion of the variance in observed scores is accounted for by careless responding (assuming that the residual of the inconsistency bias factor can be interpreted as a proxy for careless responding). In fact, in the grouped-massed condition careless responding is the second most important contributor to item variance after the target construct, self-esteem. In contrast, in the dispersed condition careless responding is a much less important variance component. However, the unique variances of the items are much larger in that case. These findings suggest that while the item arrangement in the grouped-massed condition increases systematic error (greater inconsistency bias), dispersing items throughout the questionnaire leads to greater non-systematic (random) error.

– Insert Table 2 about here –

STUDY 2

Although we believe that the model tested in the first study provides useful insights into the phenomenon of reversed item bias, four limitations should be pointed out. First, we do not have direct evidence for the interpretation of the residual of the inconsistency bias factor as careless responding. The moderating effect of item arrangement is consistent with our hypothesis that careless responding should be strongest in the grouped-massed condition and therefore supports this interpretation, but it would be useful to have a direct measure of careless

responding. Second, the self-esteem scale used in Study 1 was not an established instrument, there were only four items, and the items were probably more similar in wording than the items found in many personality and attitude scales. Third, we used an ad hoc measure of net acquiescent responding by randomly selecting presumably unrelated items from different scales found in the literature. Finally, the sample size for the study was relatively small.

To address these limitations, we conducted another study in which we used an established measure of another substantive construct, optimism (Scheier, Carver & Bridges, 1994), based on three regular and three reversed items. We also included different filler items to construct our index of net acquiescent responding (NARS), and we collected data from a larger sample of respondents in the U.S. In addition, we extended the original model by adding a direct measure of carelessness. Specifically, we propose the use of an Instructional Manipulation Check (IMC) as an explicit proxy indicator of careless responding. As described by Oppenheimer, Meyvis, and Davidenko (2009), an IMC consists of a question embedded in the questionnaire that is similar to the other questions in terms of length and response format (e.g., Likert scale), but differs from the other questions in that participants are asked to ignore the standard response format and instead provide an indication that they paid attention to the specific wording of the question (e.g., "For this statement, please do not check a response option, but simply click on the continue button below"). Since careless responding is based on top-down processing cued by respondent expectations about what construct is being assessed, IMC is expected to capture careless or inattentive responding to reversed items.

Method

We collected the data from the Amazon Mechanical Turk panel (see Goodman, Cryder, & Cheema, forthcoming, for an evaluation of Mechanical Turk participants). We set a target response

rate of 600 completed questionnaires, and the survey remained active for 12 days until this target was reached. Because of missing data, the final effective sample size was N = 595. A small monetary incentive was used to encourage participation, and respondents spent an average of 4 minutes to complete the online survey. The average age of the respondents was 36 (Min = 16; Max = 79; SD = 13.0), 60 percent of respondents were female, and 46 percent had at least a four-year college education.

The focal construct in Study 2 was the six-item Revised Life Orientation Test or LOT (Scheier, Carver, & Bridges, 1994). Life orientation data have been used previously in studying reversed item method effects (e.g., Maydeu-Olivares & Coffman, 2006; McPherson & Mohr, 2005). Three items measured life orientation in a positive direction: 'In uncertain times, I usually expect the best'; 'I'm always optimistic about my future'; 'Overall, I expect more good things to happen to me than bad' (p1 to p3). Three items measured life orientation in a negative direction and were thus reversed: 'If something can go wrong for me, it will'; 'I hardly ever expect things to go my way'; 'I rarely count on good things happening to me' (n1 to n3). For the grouped-massed condition, we used two additional items (both from the original LOT scale; Scheier & Carver, 1985), either 'I always look on the bright side of things' for the regular item first condition (p4) or 'Things never work out the way I want them to' for the reversed item first condition (n4). All items were rated on 5-point Likert scales with endpoints of strongly disagree (1) and strongly agree (5).

To measure NARS, we included 16 items identified by Greenleaf (1992) as being highly diverse in content, which are thus well suited for capturing response style variance. Greenleaf (1992) originally proposed the items as a scale for measuring extreme response style, but the heterogeneity of the items in terms of substantive content serves our purpose as well. Based on a

preliminary screening of the items for potential content relations with life orientation, we deleted three items ('No matter how fast our income goes up, we never seem to get ahead'; 'Investing in the stock market is too risky for most families'; 'I will probably have more money to spend next year than I have now') and replaced them with three items that are sometimes included as filler items among the life orientation items (i.e., 'I enjoy my friends a lot'; 'It's important for me to keep busy'; 'It's easy for me to relax'; see Scheier et al., 1994). To construct the NARS measure, we computed the average score across the 16 unrelated buffer items (Min = 1.00, Max = 5.00, M = 3.31, SD = .85, Cronbach's alpha = .42). The reliability of the composite was relatively low, but this is not entirely surprising given that the scale contains only 16 items and the items in the scale were specifically selected to share no substantive content. Watson (1992) reported comparable reliabilities in her study.

We used six versions of the questionnaire, corresponding to the six conditions in Study 1. In the grouped-alternated conditions (N = 204), the six LOT items were shown on the same page in an alternating sequence, starting with either a regular or reversed item. In the grouped-massed conditions (N = 186), the LOT items were also shown on the same page, but four regular items were followed by three reversed items (in the regular item first condition) or four reversed items were followed by three regular items (in the reversed item first condition). In both the grouped-alternated and grouped-massed conditions, the buffer items were shown on four separate pages with four items per page, counterbalancing whether the buffer items preceded or followed the LOT items. Counterbalancing had no effect on the results, so it is not considered as a design factor in the sequel. In the dispersed conditions (N = 205), each LOT item was shown on a separate page, and on each page the LOT item was followed by buffer items (i.e., on the first five pages, the LOT item was followed by three buffer items, and on the last page the LOT item was

followed by one buffer item, for a total of 16 buffer items). Depending on the 'First item reversed' (FIR) condition, the survey started with a regular or reversed item.

After the LOT and buffer items, we included eight items measuring frugality, all worded in the same direction (Lastovicka et al., 1999). These eight items were intended as a manipulation to induce the erroneous expectation among careless respondents that all items on this page measured the same underlying construct. However, the last item on this page served as an IMC and read, "For this statement, please do not check a response option, but simply click on the continue button below". Participants who clicked on one of the response options were coded as having failed the IMC (IMC=1), all others as having passed it (IMC=0; see Oppenheimer, Meyvis, & Davidenko, 2009). Only 4.2 percent of respondents failed the IMC.

Results

The distribution of the data was again sufficiently normal (all endogenous observed variables had skewness and kurtosis values between –1 and +1) and there were no missing values, so we used the ML estimator in MPlus 6.1. We started with a one-factor model in which the six observed variables loaded on a single substantive factor (LOT) and all estimated parameters were left unconstrained across the three item-arrangement conditions. This model fit quite poorly (χ^2 (27) = 320.50, $p < .0001$; RMSEA = .234; CFI = .862; TLI = .769); the sizable modification indices for the covariances of the unique factors again suggested the presence of reversed item bias.

We then estimated the full model shown in Figure 2. The loadings on the substantive factor were freely estimated across items (except for one unit loading for the marker item), but as in the previous study the loadings of corresponding items proved to be invariant across items arrangement conditions and were therefore constrained to be equal across groups. The loadings

on the consistency bias and confirmation bias factors were set to plus or minus unity, as shown in Figure 2. No covariances between any of the exogenous variables had to be introduced.

To test for the invariance of the γ parameters across the three item arrangement conditions, we conducted a series of model comparison tests. In support of the prediction that the effect of acquiescence on inconsistency bias would not depend on the positioning of the items in the questionnaire, γ_{NARS} was invariant across groups ($\Delta\chi^2$ (2) = .02, n.s.). Contrary to predictions, the effects of IMC (carelessness) and the keying direction of the first item (FIR) were also invariant ($\Delta\chi^2$ (2) = .51, n.s., for γ_{IMC} and $\Delta\chi^2$ (2) = 4.10, n.s., for $\gamma_{Confbias}$). The final model, in which all the γ parameters were specified to be invariant across groups, fit the data well, although the chi-square statistic was significant: χ^2 (94) = 150.08, p = .0002; RMSEA = .055 (90% C.I. .038 to .071); SRMR = .062; CFI = .974; TLI = .973. The estimation results are shown in Table 1.

NARS (γ_{NARS} = .33, p < .001) and IMC (γ_{IMC} = .31, p < .001) were highly significant determinants of inconsistency bias, supporting our predictions that both net acquiescence and careless responding would contribute to response inconsistency between regular and reversed items. The invariance of the effect of NARS on inconsistency bias was as expected, but contrary to prediction, the effect of IMC did not differ by item arrangement condition. The manipulation of whether or not the first target item was reversed (FIR) had no effect on responses to the LOT items, and there were no moderating effects of item arrangement. Thus, the hypotheses concerning confirmation bias were not supported in this study. The residual variance in the inconsistency bias factor was significant after accounting for NARS and IMC, which could mean that NARS and IMC do not fully capture acquiescence and careless responding, respectively, or that there are other influences on inconsistent responding besides acquiescence and carelessness.

The variance decomposition shown in Table 2 indicates that the substantive construct accounts for the largest portion of the variance in each observed item, although the proportions are smaller than in Study 1, particularly for the regular items. Even though the regular items contain somewhat less content variance than the reversed items, it should be noted that the labels 'reversed' and 'regular' are basically interchangeable as they depend on the way the construct is labeled (i.e., optimism vs. pessimism).

While the effects of NARS and IMC on inconsistency bias were highly significant, the variance proportions accounted for by acquiescence and careless responding were relatively minor (especially the latter). However, the residual of the inconsistency bias factor is a non-negligible contributor to observed item scores, similar in magnitude to Study 1 (with the exception of the grouped-massed condition). If we take the average variance proportion for NARS, IMC, and the residual in the inconsistency bias factor and set it in proportion to the relative substantive variance, the proportions are .13 in Study 1 and .14 in Study 2. In other words, the contribution of the substantive construct to the variability in observed scores is about 7 to 8 times greater than that of inconsistency bias. Interestingly, the unique factor component is much stronger in Study 2 than in Study 1. This may be due to the fact that the items used in Study 2 were less similar than those used in Study 1.

DISCUSSION

The issue of whether or not reverse-keyed items should be included in measurement scales has been controversial. Some measurement experts recommend the routine use of items varying in keying direction, others advise against it. In this paper we contribute to the literature on reversed items in four ways. First, we distinguish three distinct mechanisms through which

reversed items can affect survey participants' responses (acquiescence, careless responding, and confirmation bias), and we discuss two sets of factors that may influence the extent to which these mechanisms operate (how the items measuring the focal construct are arranged in the questionnaire and whether the survey starts with a regular or reversed item). Second, we offer a review and classification of previous attempts to model method effects caused by the presence of reversed items. Third, we propose a multi-sample factor-analytic model containing two method factors (inconsistency bias, confirmation bias), which allows researchers to simultaneously investigate the operation of the three mechanisms in response to reversed items. Fourth, we report two empirical studies to demonstrate the usefulness of the proposed model and to provide evidence about the importance of the three sources of method bias under various conditions. The conditions studied were selected to reflect realistic variations in the way questionnaires are administered in practice rather than to maximize the biases of interest. Of course, the model can be applied to other situations in future research in order to further investigate the misresponse mechanisms studied in the current research.

We believe that the proposed model has several attractive features. In a general sense, our approach combines aspects of experimentation with response style measurement in order to gain additional insights into people's responses to questionnaires containing reverse-keyed items. Respondents are randomly assigned to surveys that vary how the focal items are distributed over the questionnaire and whether the first focal item is a regular or reverse-keyed item. We use a dedicated measure of net acquiescence response style based on many heterogeneous items and an explicit assessment of careless responding (in Study 2) to separate different mechanisms that might lead to inconsistency bias. The manipulation of whether the first measure of the target construct is a regular or reversed item makes it possible to examine the presence of confirmation

bias. The inclusion of antecedents of inconsistency bias and confirmation bias enhances the interpretability of the method factors, which has been a problem in some prior research. Finally, the multi-sample specification of the three types of item arrangement (grouped-alternated, grouped-massed, and dispersed) allows us to investigate the moderating influence of item positioning on the effect of acquiescence, careless responding, and the 'First item reversed' (FIR) manipulation on item scores. Although a six-group specification corresponding to the three item arrangements by two FIR conditions is feasible as well, especially when the sample size per condition is relatively large, we used the simpler three-group specification with a dummy variable representing the FIR manipulation because it sufficed to test the hypotheses implied by our conceptualization.

The findings of both studies clearly show that inconsistency bias is an important component of variation in observed measures of constructs, accounting for about 9 percent of the variance in Study 1 and 8 percent of the variance in Study 2 on average. If inconsistency bias were ignored, the fit of factor models would be rather poor, because the error introduced by the presence of reversed items is systematic. Although inconsistency bias has been demonstrated in previous research, we have attempted to move beyond prior work by identifying different sources of inconsistency bias. In particular, the two studies show that both acquiescence and careless responding contribute to inconsistency bias. Although the variance proportions accounted for by explicit measures of acquiescence (NARS) and careless responding (IMC) were small, this will not always be the case. For example, in the second study only 4 percent of respondents failed the IMC, whereas Oppenheimer et al. (2009) report failure rates of 14 to 46 percent in their studies. If a greater proportion of respondents had been careless, it is likely that careless responding would have been a more important contributor to inconsistency bias.

Acquiescence is sometimes treated as being synonymous with inconsistency bias (i.e., simultaneous agreement with regular and reversed items), but we have argued for a more limited interpretation of (net) acquiescence as a preference for the positive or negative side of the rating scale (i.e., the response options indicating agreement or disagreement on a Likert scale), because inconsistency bias may be due to other mechanisms besides acquiescence (e.g., careless responding). From this perspective, acquiescence is best measured as the extent of agreement with items that are heterogeneous in content (i.e., NARS). In our studies, NARS was a statistically significant but practically minor component of item scores. Possible reasons for this may be that respondents had well-established beliefs about the target constructs (self-esteem and optimism), that the questions were relatively straightforward, and that the survey did not impose undue demands on people's limited cognitive abilities. If a survey deals with issues about which respondents are less certain, the items are more ambiguous, and the questionnaire is completed under peripheral processing conditions, acquiescence is likely to be a more important determinant of observed responses. Survey researchers cannot control whether respondents have crystallized opinions about the focal concept, but they can avoid vaguely worded items and encourage respondents to engage in systematic processing. If these strategies are insufficient to prevent acquiescence, post hoc controls should be built into the questionnaire. For example, balanced scales consisting of an equal number of regular and reversed items should be used, or unrelated items should be included that enable the formation of an explicit measure of acquiescence (Baumgartner & Steenkamp, 2001).

Careless responding means that respondents tend to assume that the keying direction of an item is the same as the keying direction of the items preceding it in a scale. Because of its contextual nature, careless responding was hypothesized to be especially problematic in the

grouped-massed condition, where reversed items are preceded by several regular items (or vice versa). We did not find conclusive evidence in support of this hypothesis, possibly because of the small number of careless responders in our data. If future research were to corroborate the hypothesized effect, it would imply that researchers should refrain from using unbalanced scales, in which a few reversed items are included among many regular items (possibly to counter the criticism that all items are keyed in the same direction). In particular, reversed items should not follow a long list of regular items as this may lead respondents to overlook variation in item polarity (Drolet & Morrison, 2001). Although the empirical evidence is currently somewhat ambiguous, we believe the best option is to disperse the items measuring the focal construct across the questionnaire and to mix them with unrelated buffer items. We recognize that this is more taxing for respondents (thus requiring greater involvement in the task) and may result in larger unique item variances and possibly lower reliability (as shown in Study 1). If this is undesirable and/or if the items measuring the same construct need to be grouped for other reasons (e.g., for the sake of the logical flow of the questionnaire), researchers should at least use balanced scales and alternate the keying direction of the items in the scale (i.e., switch between regular and reverse-keyed items).

Even after explicitly accounting for both acquiescence (using a direct NARS measure) and carelessness (as measured by an instructional manipulation check), most of the variance in the inconsistency bias factor remained unexplained in Study 2. There are several possible explanations for this. First, the empirical measures used probably did not fully capture acquiescence and carelessness. Second, important components of either acquiescence or carelessness may be scale-specific and can therefore not be measured in a general way across different scales. Third, other psychological mechanisms that lead to inconsistent responding to

regular versus reversed items may be at work, which were not considered in the current model. More research is needed to further explore these possibilities.

In addition to leading to inconsistent responses to items, variation in the keying direction of items may also affect the mean response to these items. Specifically, we hypothesized that confirmation bias would cause a mean shift in the direction of the keying of the first item encountered in the questionnaire (i.e., upward bias when a regular item comes first, downward bias when a reversed item comes first, after recoding), particularly in the two grouped conditions. Our empirical findings concerning confirmation bias were inconclusive, as the regression parameter had a wide confidence interval that came close to but did not include zero in Study 1 and a wide confidence interval centered about zero in Study 2. Kunda et al. (1993) showed that the direction of the question (e.g., being extraverted vs. being introverted) made a difference when respondents were able to retrieve information consistent with either position. When prior evidence consistently supported only one pole of the question, or when respondents rated themselves as low in variability across situations on the construct of interest, the wording of the question had no effect. In addition, in order for confirmation bias to be evident, respondents should not have an overall summary judgment readily available in memory (in which case the piecemeal retrieval of information is unnecessary), and they have to be sufficiently motivated to retrieve additional information in response to individual questions. If these conditions do not hold, confirmation bias is unlikely to emerge. Apparently, biased retrieval was not very strong in our studies, but future research should investigate when confirmation bias can be expected to have a more pronounced influence on the results.

Several findings were inconsistent between the two studies, but this is not entirely surprising because the two studies varied in many respects and should be thought of as

illustrative applications of the model to two different situations, not as simple replications. Study 1 used four ad hoc items that were rather similar in content and wording, so that respondents may have perceived the items to be somewhat redundant. This could be one of the reasons for the relatively low proportion of unique variance in the items. Study 2 used more items, and although the items came from a validated scale, they were less obviously similar, which probably increased the amount of unique variance. Study 1 used a seven-point rating scale, whereas Study 2 used a five-point rating scale. The former format has been found to result in greater misresponse (Weijters, Cabooter, & Schillewaert, 2010). The respondents and the assessment environment were also quite different. The first study used European participants who had agreed to be part of a nationally representative panel, whereas the second study was conducted with American respondents willing to answer some questions for a small payment. The first sample is quite experienced in filling out surveys and respondents may be more sophisticated in their use of rating scales, which may encourage confirmation biases and reduce acquiescence. We anticipated a relatively high rate of inattentive responding in the second sample, based on prior findings with similar populations (e.g., Oppenheimer, Meyvis, & Davidenko, 2009), but respondents were surprisingly attentive, which may account for the weak effect of careless responding and the near-absence of confirmation bias. Future research should investigate situations in which the three response biases studied in this paper are expected to have a stronger effect on the results. Our goal in the current studies was to simulate realistic variation in survey designs, but there will be conditions in which the three biases are expected to be stronger. We also want to emphasize that the main contribution of this research is not the empirical studies but the proposal of a general analysis strategy for investigating different sources of reversed item bias integrating response style measurement with survey design manipulations.

There are two primary ways in which a regular item can be reversed. On the one hand, the meaning of an item can be switched by using a negation (e.g., by inserting the particle ‘not’). On the other hand, a reversal can be achieved by using an antonymic expression (see Bentler, Jackson, & Messick, 1971; Schriesheim et al., 1991). Research has shown that negations can be confusing to respondents (see Swain et al., 2008, for some recent evidence) and should probably be used sparingly. In our studies, we did not use reversed items based on particle negations (‘not’), but future research should investigate to what extent our findings can be generalized to negated reversals (e.g., items that are reversed simply by adding ‘not’), and how the type of reversal influences the misresponse mechanisms studied in this paper (see Weijters and Baumgartner, forthcoming, for an extended conceptual discussion of some of these issues).

Another topic that may merit further research is the question of the longitudinal stability of the different effects found in the present cross-sectional investigation. The literature suggests that acquiescence has a stable component (Alessandri et al., 2012; Weijters, Geuens, & Schillewaert, 2010), but less is known about the longitudinal stability of careless responding to reversed items and confirmation bias. Some recent work has extended the traditional method-factor and correlated-uniqueness models to a longitudinal context (Geiser & Lockhart, 2012), and it would be interesting to consider similar extensions of the proposed model incorporating multiple sources of reversed item bias.

Some researchers have recommended that problems caused by reversed items can be avoided altogether if only items keyed in one direction are used to measure a target construct. We want to emphasize that this is not a valid argument supporting the elimination of reversed items from measurement scales. When all items are worded in the same direction, acquiescence, careless responding, and confirmation bias may still be present, but the method effects generated

by these mechanisms are completely confounded with content variance and may become undetectable, unless direct measures of the method effects of interest are available (Podsakoff et al., 2003). Although it is best not to have method effects at all, it is better to be aware of them and to be able to take corrective action rather than to ignore them completely.

REFERENCES

Alessandri, G., Vecchione, M., Fagnani, C., Bentler, P.M., Barbaranelli, C., Medda, E., Nisticò, L., Stazi, M.A., & Caprara, G.V. (2010). Much more than model fitting? Evidence for the heritability of method effect associated with positively worded items of the Life Orientation Test Revised. *Structural Equation Modeling,* 17, 642-653.

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 38, 143–156.

Bentler, P. M., Jackson, D. N., & Messick, S. (1971). Identification of content and style: A two-dimensional interpretation of acquiescence. *Psychological Bulletin*, 76, 186-204.

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling,* 7, 608-628.

Cole, D. A., Ciesla, J. A., & Steiger, J. H. (2007). The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis. *Psychological Methods*, 12, 381-398.

Davies, M. F. (2003). Confirmatory bias in the evaluation of personality descriptions: Positive test strategies and output interference. *Journal of Personality and Social Psychology*, 85 (4), 736–744.

DiStefano, C., & Motl, R. W. (2009). Personality correlates of method effects due to negatively worded items on the Rosenberg self-esteem scale. *Personality and Individual Differences,* 46, 309-313.

Drolet, A., & Morrison, D.G. (2001). Do we really need multiple-item measures in service research? *Journal of Service Research* 3, 196–204.

Geiser, C., & Lockhart, G. (2012, February 6). A comparison of four approaches to account for method effects in latent state–trait analyses. *Psychological Methods*. Advance online publication. doi: 10.1037/a0026977.

Geiser, C., Eid, M., & Nussbeck, F.W. (2008). On the meaning of the latent variables in the CT-C(M-1) model: A comment on Maydeu-Olivares and Coffman (2006). *Psychological Methods* 13 (1), 49-57.

Goodman, J. K., Cryder, C.E., & Cheema, A. (forthcoming). Data collection in a flat world: Strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, forthcoming.

Greenleaf, E. A. (1992). Measuring Extreme Response Style. *Public Opinion Quarterly* 56, 328-50.

Harris, M. M., & Bladen, A. (1994). Wording effects in the measurement of role conflict and role ambiguity: A multitrait-multimethod analysis. *Journal of Management*, 20 (4), 887-901.

Horan, P. M., DiStefano, C., & Motl, R.W. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? *Structural Equation Modeling*, 10 (3), 435-455.

Hui, C. H., & Triandis, H. C. (1985). The instability of response sets. *Public Opinion Quarterly*, 49, 253-260.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, 213-236.

Kunda, Z., Fong, G. T., Santioso, R., & Reber, E. (1993). Directional questions direct self-conceptions. *Journal of Experimental Social Psychology*, 29, 63-86.

Lastovicka, J.L., Bettencourt, L.A., Hughner, R.S., & Kuntze, R.J. Lifestyle of the tight and frugal: Theory and measurement. *Journal of Consumer Research*, 26 (1), 85-98.

Levin, J., & Montag, I. (1989). The bipolarity of the Comrey Personality Scales: A confirmatory factor analysis. *Personality and Individual Differences*, 10 (11), 1989, 1115-1120.

Marsh, H. W. (1986). Negative item bias in ratings scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology*, 22 (1), 37-49.

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, 70 (4), 810–819.

Martin , J. (1964). Acquiescence – measurement and theory. *British Journal of Social and Clinical Psychology*, 3, 216-225.

Maydeu-Olivares, A., & Coffman, D.L. (2006). Random intercept item factor analysis. *Psychological Methods*, 11 (4), 344–362.

McClendon, M. J. (1991). Acquiescence: Tests of the cognitive limitations and question ambiguity hypotheses. *Journal of Official Statistics*, 7, 153–166.

McPherson, J., & Mohr, P. (2005). The role of item extremity in the emergence of keying-related factors: An exploration with the Life Orientation Test. *Psychological Methods*, 10 (1), 120–131.

Meade, A.W., & Craig B. (2012). Identifying careless responses in survey data. *Psychological Methods*, online advance publication, doi: 10.1037/a0028085.

Motl, R. W., & DiStefano, C. (2002). Longitudinal invariance of self-esteem and method effects associated with negatively worded items. *Structural Equation Modeling*, 9 (4), 562-578.

Mulaik, S. A. (2009). Linear causal modeling with structural equations. Chapman Hall/CRC.

Nunnally, J. C. (1978). *Psychometric theory, 2nd ed.* New York: McGraw-Hill.

Oppenheimer, D.A., Meyvis, T., & Davidenko, N. (2009), Instructional manipulation checks: Detecting satisficing to increase statistical power, *Journal of Experimental Social Psychology*, 45, 867–872.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17-59). San Diego, CA: Academic Press.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88 (5), 879-903.

Rosenberg, M. (1965). *Society and adolescent self-image*. Princeton, NJ: Princeton University Press.

Scheier, M.F., Carver, C.S., & Bridges, M.W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the Life Orientation Test. *Journal of Personality and Social Psychology*, 67 (6), 1063-1078.

Scheier, M. F, & Carver, C. S. (1985). Optimism, coping, and health: Assessment and implications of generalized outcome expectancies. *Health Psychology*, 4, 219-247.

Schmitt, N., & Stults, D.M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, 9, 367-373.

Schriesheim, C. A., Eisenbach, R. J., & Hill, K. D. (1991). The effect of negation and polar opposite item reversals on questionnaire reliability and validity: An experimental investigation. *Educational and Psychological Measurement*, 51, 67-78.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording and context.* San Diego, CA: Academic Press.

Swain, S.D., Weathers, D., & Niedrich, R.W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research*, 45, 116–131.

Tomás, J. M., & Oliver, A. (1999). Rosenberg's Self-Esteem Scale: Two factors or method effects. *Structural Equation Modeling* 6(1), 84-98.

Tourangeau, R. Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge, UK: Cambridge University Press.

Watson, D. (1992). Correcting for acquiescent response bias in the absence of a balanced scale: An application to class consciousness. *Sociological Methods and Research*, 21 (1), 52-88.

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing* 27, 236-247.

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. *Psychological Methods*, 15 (1), 96-110.

Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research*, accepted for publication: 1–42. doi :10.1509/jmr.11.0368.

Welkenhuysen-Gybels, J., Billiet, J., & Cambré, B. (2003). Adjustment for acquiescence in the assessment of the construct equivalence of Likert-type score items. *Journal of Cross-cultural Psychology*, 34 (6), 702-722.

Woods, C.M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, 28 (3), 186-191.

Table 1
Estimation Results (Studies 1 and 2)

Parameter	Study 1			Study 2		
	Estimate	95% C.I.		Estimate	95% C.I.	
		Lower	Upper		Lower	Upper
γ_{NARS}	0.168	0.045	0.291	0.329	0.245	0.413
γ_{IMC}	n.a.	n.a.	n.a.	0.306	0.163	0.449
σ^2_Z	0.166	0.093	0.239	0.060	0.036	0.084
	0.367	0.244	0.490	0.083	0.056	0.110
	0.097	0.019	0.175	0.076	0.049	0.103
$\gamma_{Confbias}$	-0.285	-0.567	-0.003	0.000	-0.131	0.131
$\sigma^2(\xi)$	1.736	1.185	2.287	0.393	0.283	0.503
	1.585	1.064	2.106	0.569	0.410	0.728
	1.505	1.007	2.003	0.419	0.303	0.535
σ^2(NARS)	0.280	0.206	0.354	0.112	0.090	0.134
	0.262	0.188	0.336	0.122	0.097	0.147
	0.289	0.211	0.367	0.114	0.092	0.136
σ^2(IMC)	n.a.	n.a.	n.a.	0.060	0.048	0.072
	n.a.	n.a.	n.a.	0.026	0.020	0.032
	n.a.	n.a.	n.a.	0.033	0.027	0.039
σ^2(FIR)	0.247	0.180	0.314	0.250	0.201	0.299
	0.250	0.179	0.321	0.250	0.199	0.301
	0.245	0.178	0.312	0.250	0.201	0.299
$\sigma(\xi$, NARS)	-0.241	-0.386	-0.096	n.a.	n.a.	n.a.
	-0.233	-0.374	-0.092	n.a.	n.a.	n.a.
	-0.222	-0.363	-0.081	n.a.	n.a.	n.a.

(Continued on the next page)

Table 1 *continued*

Parameter	Study 1			Study 2		
	Estimate	95% C.I.		Estimate	95% C.I.	
		Lower	Upper		Lower	Upper
λ_{p1}	1.000	n.a.	n.a.	1.000	n.a.	n.a.
λ_{p2}	1.059	0.971	1.147	1.120	0.998	1.242
λ_{p3}	n.a.	n.a.	n.a.	1.170	1.048	1.292
λ_{n1}	-0.981	-1.106	-0.856	-1.169	-1.314	-1.024
λ_{n2}	-1.070	-1.199	-0.941	-1.348	-1.501	-1.195
λ_{n3}	n.a.	n.a.	n.a.	-1.366	-1.523	-1.209
σ^2 (ε_{p1})	0.393	0.211	0.575	0.501	0.385	0.617
	0.376	0.166	0.586	0.427	0.319	0.535
	0.688	0.423	0.953	0.459	0.351	0.567
σ^2 (ε_{p2})	0.286	0.110	0.462	0.468	0.356	0.580
	0.213	0.013	0.413	0.331	0.239	0.423
	0.464	0.227	0.701	0.372	0.276	0.468
σ^2 (ε_{p3})	n.a.	n.a.	n.a.	0.296	0.210	0.382
	n.a.	n.a.	n.a.	0.326	0.232	0.420
	n.a.	n.a.	n.a.	0.308	0.222	0.394
σ^2 (ε_{n1})	0.413	0.217	0.609	0.409	0.313	0.505
	0.267	0.091	0.443	0.333	0.249	0.417
	0.873	0.552	1.194	0.402	0.306	0.498
σ^2 (ε_{n2})	0.425	0.213	0.637	0.240	0.166	0.314
	0.207	0.023	0.391	0.213	0.142	0.284
	0.657	0.359	0.955	0.253	0.175	0.331
σ^2 (ε_{n3})	n.a.	n.a.	n.a.	0.311	0.227	0.395
	n.a.	n.a.	n.a.	0.260	0.182	0.338
	n.a.	n.a.	n.a.	0.354	0.260	0.448

Note: 95% C.I. = 95% confidence interval; Lower = lower bound; Upper = upper bound. NARS = net acquiescence response style; IMC = instruction manipulation check; FIR = First Item Reversed dummy variable; see Figure 2 for an explanation of the other parameter names. In case of multiple cell entries for a given parameter, the first line refers to the grouped-alternated condition, the second to the grouped-massed condition, and the third to the dispersed condition.

Table 2
Variance decomposition for regular and reversed items (Studies 1 and 2)

Variance component	Study 1						Study 2					
	Average of regular items			Average of reversed items			Average of regular items			Average of reversed items		
	GA	GM	DIS	GA	GM	DIS	GA	GM	DIS	GA	GM	DIS
Substantive construct (ξ)	.804 [.729, .878]	.734 [.647, .822]	.719 [.625, .813]	.724 [.651, .798]	.701 [.623, .779]	.621 [.529, .714]	.487 [.421, .553]	.599 [.536, .663]	.518 [.453 .583]	.624 [.564 .685	.723 [.671 .775]	.622 [.562, .682]
NARS	.003 [-.002, .009]	.003 [-.002, .008]	.004 [-.002, .009]	.003 [-.002, .009]	.003 [-.002, .008]	.003 [-.002, .008]	.012 [.005, .019]	.012 [.005, .018]	.013 [.006, .020]	.011 [.005, .018]	.010 [.004, .016]	.011 [.005, .017]
IMC	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	.006 [.000, .011]	.002 [.000, .004]	.003 [.000, .006]	.005 [.000, .010]	.002 [.000, .004]	.003 [.000, .005]
Residual of inconsistency bias factor (σ^2_Z)	.073 [.038, .107]	.160 [.103, .218]	.044 [.008, .080]	.066 [.035, .097]	.154 [.098, .210]	.038 [.007, .067]	.062 [.037, .087]	.072 [.047, .098]	.078 [.050. .106]	.057 [.033, .080]	.063 [.040, .085]	.067 [.043, .091]
FIR	.009 [-.009, .026]	.009 [-.009, .027]	.009 [-.009, .027]	.008 [-.008, .024]	.009 [-.009, .026]	.008 [-.008, .023]	.000 [.000, .000]	.000 [.000, .000]	.000 [.000, .000]	.000 [.000, .000]	.000 [.000, .000]	.000 [.000, .000]
Unique variances ($\sigma^2(\varepsilon_i)$)	.148 [.098, .198]	.129 [.085, .173]	.259 [.178, .341]	.166 [.111, .221]	.099 [.064, .135]	.300 [.210, .389]	.433 [.369, .496]	.315 [.259, .371]	.389 [.329, .449]	.302 [.249, .355]	.203 [.161, .244]	.297 [.245, .349]
Covariance of NARS and ξ	-.036 [-.070, -.002]	-.035 [-.068, -.002]	-.035 [-.068, -.001]	.033 [.004, .062]	.034 [.004, .063]	.030 [.003, .057]	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.

Note. The numbers between brackets are the lower and upper bounds of the 95% confidence intervals. GA = grouped-alternated condition; GM = grouped-massed condition; DIS = dispersed condition. NARS = net acquiescence response style; IMC = instruction manipulation check; FIR = First Item Reversed dummy variable; see Figure 2 for an explanation of the other terms in the first column.

FIGURE CAPTIONS

Captions for Figure 1:

Figure 1. Method factor models described in the literature. X and Y are substantive latent factors; M1 and M2 are method factors. The unique terms are indicated by arrows leading to the items, but they are not labeled explicitly and their variances are not shown for simplicity. Items p1 to p6 are regular (non-reversed items); items n1 to n6 are reversed items. If the loadings are set to unity (freely estimated but constrained to be equal across items), the corresponding factor variance is freely estimated (set to unity). All method factor loadings are assumed to be positive.

Captions for Figure 2:

Figure 2. Graphical depiction of the proposed integrated model of reversed item bias for a specific item configuration. ξ = substantive construct; η_1 = inconsistency bias; NARS = net acquiescence response style; IMC = instruction manipulation check; Z = other influences on inconsistency bias; η_2 = confirmation bias; FIR = First Item Reversed dummy variable.

SUPPLEMENTAL MATERIALS

MODEL SPECIFICATION FOR STUDY 1

The model of Study 1 can be specified as follows:

$$p_1 = (+) \quad \xi + \eta_1 + \eta_2 + \varepsilon_{p1}$$

$$p_2 = (+)\, \lambda_{p2}\, \xi + \eta_1 + \eta_2 + \varepsilon_{p2}$$

$$n_1 = (-)\, \lambda_{n1}\, \xi + \eta_1 - \eta_2 + \varepsilon_{n1}$$

$$n_2 = (-)\, \lambda_{n2}\, \xi + \eta_1 - \eta_2 + \varepsilon_{n2}$$

$$\eta_1 = \gamma_{NARS}\, \mathrm{NARS} + Z$$

$$\eta_2 = \gamma_{Confbias}\, \mathrm{FIR}$$

See Figure 2 in the paper for variable labels.

Although not shown explicitly, the model is a three-group specification for the grouped-alternated, grouped-massed and dispersed conditions. In the final model of Table 1, group-invariant parameters are estimated for λ_{p2}, λ_{n1}, λ_{n2}, γ_{NARS} and $\gamma_{Confbias}$ and group-specific parameters are estimated for $\sigma^2(\xi)$, $\sigma^2(\mathrm{NARS})$, $\sigma^2(\mathrm{FIR})$, $\sigma^2(\varepsilon_{p1})$, $\sigma^2(\varepsilon_{p2})$, $\sigma^2(\varepsilon_{n1})$, $\sigma^2(\varepsilon_{n2})$, $\sigma(\xi, \mathrm{NARS})$, and $\sigma^2(Z)$. The last term is the residual variance in η_1, which is interpreted as the variation in careless responding in Study 1. The total number of parameters estimated is thus 32, and since there are 63 distinct variances and covariances across the three conditions, the model has 31 degrees of freedom.

MODEL SPECIFICATION FOR STUDY 2

The model of Study 2 can be specified as follows:

$$p_1 = (+) \quad \xi + \eta_1 + \eta_2 + \varepsilon_{p1}$$

$$p_2 = (+)\,\lambda_{p2}\,\xi + \eta_1 + \eta_2 + \varepsilon_{p2}$$

$$p_3 = (+)\,\lambda_{p3}\,\xi + \eta_1 + \eta_2 + \varepsilon_{p3}$$

$$n_1 = (-)\,\lambda_{n1}\,\xi + \eta_1 - \eta_2 + \varepsilon_{n1}$$

$$n_2 = (-)\,\lambda_{n2}\,\xi + \eta_1 - \eta_2 + \varepsilon_{n2}$$

$$n_3 = (-)\,\lambda_{n3}\,\xi + \eta_1 - \eta_2 + \varepsilon_{n3}$$

$$\eta_1 = \gamma_{NARS}\,\mathrm{NARS} + \gamma_{IMC}\,\mathrm{IMC} + Z$$

$$\eta_2 = \gamma_{Confbias}\,\mathrm{FIR}$$

See Figure 2 in the paper for variable labels.

Although not shown explicitly, the model is a three-group specification for the grouped-alternated, grouped-massed and dispersed conditions. In the final model of Table 1, group-invariant parameters are estimated for λ_{p2}, λ_{p3}, λ_{n1}, λ_{n2}, λ_{n3}, γ_{NARS}, γ_{IMC}, and $\gamma_{Confbias}$, and condition-specific parameters are estimated for $\sigma^2(\xi)$, $\sigma^2(\mathrm{NARS})$, $\sigma^2(\mathrm{IMC})$, $\sigma^2(\mathrm{FIR})$, $\sigma^2(\varepsilon_{p1})$, $\sigma^2(\varepsilon_{p2})$, $\sigma^2(\varepsilon_{p3})$, $\sigma^2(\varepsilon_{n1})$, $\sigma^2(\varepsilon_{n2})$, $\sigma^2(\varepsilon_{n3})$, and $\sigma^2(Z)$. $\sigma^2(Z)$ is the residual variance in η_1, after controlling for NARS and IMC. The total number of parameters estimated is thus 41, and since there are 135 distinct variances and covariances across the three conditions, the model has 94 degrees of freedom.

ITEM DESCRIPTIVE STATISTICS BY CONDITION (STUDY 1)

			variance-covariance matrix					
		M	p1	n1	p2	n2	FIR	NARS
GA	p1	4.811	2.288					
	n1	3.085	-1.774	2.536				
	p2	4.755	1.925	-1.693	2.244			
	n2	3.274	-1.624	2.167	-1.704	2.620		
	FIR	0.443	-0.001	0.019	-0.043	0.125	0.249	
	NARS	-0.553	-0.191	0.272	-0.245	0.324	0.038	0.282
GM	p1	4.562	2.207					
	n1	3.042	-1.129	2.314				
	p2	4.740	1.990	-1.357	2.384			
	n2	3.010	-1.269	2.189	-1.492	2.558		
	FIR	0.510	-0.09	0.073	-0.139	0.026	0.253	
	NARS	-0.498	-0.135	0.290	-0.183	0.328	-0.014	0.265
DIS	p1	4.712	2.324					
	n1	2.933	-1.214	2.199				
	p2	4.750	1.714	-1.299	2.267			
	n2	3.038	-1.698	1.692	-1.796	2.814		
	FIR	0.433	-0.117	0.185	-0.066	0.061	0.248	
	NARS	-0.426	-0.173	0.225	-0.197	0.302	-0.006	0.292

Note. GA = Grouped-alternated condition (N = 106); GM = Grouped-massed condition (N = 96); DIS = dispersed condition (N = 104).

ITEM DESCRIPTIVE STATISTICS BY CONDITION (STUDY 2)

			Variance-covariance matrix								
		Mean	p1	n1	p2	n2	p1	n3	FIR	NARS	IMC
GA	p1	3.137	1.045								
	n1	3.078	-0.349	1.009							
	p2	3.373	0.643	-0.429	1.062						
	n2	3.255	-0.462	0.724	-0.506	1.087					
	p3	3.623	0.530	-0.498	0.550	-0.540	0.857				
	n3	3.314	-0.494	0.645	-0.518	0.826	-0.533	1.093			
	FIR	0.505	-0.030	0.025	0.003	0.070	-0.050	0.016	0.251		
	NARS	3.458	0.071	0.024	0.064	0.013	0.060	0.007	-0.011	0.112	
	IMC	0.064	0.016	0.035	0.011	0.036	0.000	0.030	-0.003	0.011	0.060
GM	p1	3.022	1.135								
	n1	3.301	-0.621	1.228							
	p2	3.242	0.854	-0.667	1.233						
	n2	3.333	-0.620	0.991	-0.730	1.283					
	p3	3.586	0.766	-0.799	0.890	-0.787	1.239				
	n3	3.403	-0.629	0.964	-0.745	1.086	-0.914	1.355			
	FIR	0.505	0.076	-0.004	0.110	-0.014	0.043	-0.001	0.251		
	NARS	3.452	0.082	0.004	0.109	-0.016	0.079	-0.021	0.015	0.123	
	IMC	0.027	0.005	0.014	0.010	0.009	0.006	0.016	0.003	0.004	0.026
DIS	p1	3.137	1.001								
	n1	3.298	-0.346	1.014							
	p2	3.332	0.675	-0.401	1.036						
	n2	3.337	-0.429	0.777	-0.471	1.146					
	p3	3.517	0.556	-0.448	0.627	-0.600	0.957				
	n3	3.317	-0.486	0.655	-0.615	0.858	-0.703	1.188			
	FIR	0.512	-0.021	0.001	-0.014	-0.003	-0.011	0.001	0.251		
	NARS	3.392	0.053	0.029	0.083	0.015	0.049	-0.021	-0.008	0.114	
	IMC	0.034	0.005	0.010	0.008	0.031	0.007	0.011	-0.003	-0.001	0.033

Note. GA = Grouped-alternated condition (N = 204); GM = Grouped-massed condition (N = 205); DIS = Dispersed condition (N = 205).