Psychon Bull Rev
DOI 10.3758/s13423-014-0640-3

THEORETICAL REVIEW

Buy three but get only two: The smallest effect in a 2 × 2 ANOVA is always uninterpretable

Leonel Garcia-Marques · Teresa Garcia-Marques · Markus Brauer

Abstract Loftus (*Memory & Cognition* 6:312–319, 1978) distinguished between interpretable and uninterpretable interactions. Uninterpretable interactions are ambiguous, because they may be due to two additive main effects (no interaction) and a nonlinear relationship between the (latent) outcome variable and its indicator. Interpretable interactions can only be due to the presence of a true interactive effect in the outcome variable, regardless of the relationship that it establishes with its indicator. In the present article, we first show that same problem can arise when an unmeasured mediator has a nonlinear effect on the measured outcome variable. Then we integrate Loftus's arguments with a seemingly contradictory approach to interactions suggested by Rosnow and Rosenthal (*Psychological Bulletin* 105:143–146, 1989). We show that entire data patterns, not just interaction effects alone, produce interpretable or noninterpretable interactions. Next, we show that the same problem of interpretability can apply to main effects. Lastly, we give concrete advice on what researchers can do to generate data patterns that provide unambiguous evidence for hypothesized interactions.

Keywords Statistics · Statistical inference

L. Garcia-Marques (✉)
Faculdade de Psicologia, Universidade de Lisboa,
Alameda da Universidade, 1649-013 Lisboa, Portugal
e-mail: garcia_marques@sapo.pt

T. Garcia-Marques
ISPA-University Institute, Lisbon, Portugal

M. Brauer
University of Wisconsin–Madison, Madison, WI, USA

> Data do not speak for themselves—they need context, and they need skeptical evaluation.
> Allen Wilcox

Statistical interactions are frequent fellow travelers of the researcher in the psychological sciences. It is therefore striking that today, after more than 70 years of the introduction of factorial ANOVA models in psychology (Rubin-Rabson, 1937), researchers still struggle with the adequate way to make the most of this companionship. In this article, we join forces with those who have tried to clarify the interpretation of interactions. We will concentrate on one important distinction brought forth in psychology by Loftus (1978), the distinction between *interpretable* and *uninterpretable* interactions (for a recent restatement of Loftus's argument, see Wagenmakers, Krypotos, Criss, & Iverson, 2012). In a nutshell, this distinction refers to the fact that some interactions are inherently ambiguous to interpret because the observed difference in slopes may be due to a nonlinear relationship between the latent outcome variable and its indicator, rather than to a truly interactive effect of the independent variables on the dependent variable. These interactions are called *uninterpretable* (or *removable, ordinal*, *quantitative*, or *model dependent*). Other interactions, however, cannot be "explained away," because the data pattern can only be due to an interactive effect regardless of the type of monotonic relationship between the latent outcome variable and its indicator. This second type of interactions is referred to as *interpretable* (or *nonremovable*, *disordinal*, *qualitative*, or *model free*). We will discuss the difference between interpretable and uninterpretable interactions in more detail below.

In the present article, we extend Loftus's (1978) arguments in four ways. First, we show that the problem of uninterpretable interactions is by no means restricted to the case in which there is a nonlinear relationship between the latent outcome variable and its indicator. The same problem exists in the case

Published online: 20 May 2014

in which the effect of the independent variables on the dependent variable is mediated by an (unmeasured) variable and the relationship between the mediator and the outcome variable is nonlinear. Second, we integrate Loftus's arguments with a seemingly contradictory but very influential approach to interactions suggested by Rosnow and Rosenthal (1989, 1991, 1995; Rosenthal & Rosnow, 1985). According to this approach, statistical interactions cannot be directly interpreted by cell means because cell means are the result of both main and interaction effects. As we will demonstrate below, entire data patterns, not just interaction effects, produce interpretable or uninterpretable interactions. In fact, the size of the main effects relative to the interaction effect determines the extent to which the interpretation is ambiguous. Third, we show that the same problems of interpretability can arise for main effects. In fact, it turns out that in a 2 × 2 ANOVA, the smallest of the three effects is always uninterpretable. Fourth, we will give concrete advice on what researchers can do to generate data patterns that provide unambiguous evidence for a hypothesized interaction effect.

Some interactions are uninterpretable (Loftus, 1978)

Research in psychological sciences is typically interested in uncovering the systematic effect of the manipulated operationalized variables on a latent outcome variable or constructs that can only be approached indirectly—that is, by measuring one or more "indicators" or "manifest variables." When researchers can specify the exact relationship between the latent outcome variable and its indicator, they are able to make specific conditional claims about whether a statistically significant interaction will generalize from the indicator to the latent variable. However, most of the time, researchers cannot make such specifications, simply because they do not know enough about the latent variable. In these cases, the interpretation of certain interaction patterns is ambiguous, because the data are consistent with the hypothesized interaction effect only if the researchers assume a linear relationship between the latent variable and its indicator.

Loftus (1978) demonstrated that certain types of interactions are interpretable even when researchers do not know the nature of the relationship between the latent variable and its indicator. This is the case, for example, when the two lines representing the simple effects in a 2 × 2 ANOVA cross each other (a so-called "cross-over interaction"). In this case, the data pattern provides convincing evidence for the hypothesized interaction effect and cannot be "explained away" by a nonlinear monotonic relationship between the latent variable and its indicator.

The phenomenon of uninterpretable interactions is illustrated in Fig. 1. Let's imagine two manipulated independent variables, each with two levels, which have an additive effect, but no interactive effect, on some latent outcome variable (see the top left panel of Fig. 1).

Let's further imagine that the researchers measured their outcome variable with an indicator that happens to be related to its underlying latent construct in a curvilinear way (see the middle panel of Fig. 1 and the conversion chart in the Appendix).[1] Such a relationship may exist, for example, when participant are reluctant to express extreme feelings for social desirability reasons and report more moderate feelings than they actually have. In such a case, the researchers will observe the pattern of means depicted in the top right panel of Fig. 1. The interactive effect of the independent variables on the measured indicator may be statistically significant. And yet, the true interaction effect on the latent variable is zero.

When researchers obtain a pattern of means similar to the one shown in the top right panel of Fig. 1, the interpretation of the observed interaction is ambiguous. The results may be due to the hypothesized interaction hypothesis. Or they may be due to a nonlinear relationship between the latent variable and its indicator. Given that in most cases, researchers do not know how the latent variable is related to its indicator, the empirical evidence is weak at best. The interaction can be "explained away," or expressed differently, it is a "removable" interaction (Loftus, 1978). This phenomenon is graphically described in Fig. 2.

The researchers are in better position if the two lines representing the simple effects in a 2 × 2 ANOVA touch each other or, even better, cross each other when either one or both of the two independent variables is plotted on the abscissae. Consider the pattern of observed means displayed in the bottom right corner of Fig. 1. Regardless of the relationship between the latent outcome variable and its indicator—be it linear, curvilinear, cubic, or any other imaginable form (as long as it is monotonic)—there will always be an interactive effect of the independent variables on the latent outcome variable. In other words, this interaction cannot be explained away and is, thus, "nonremovable."

In Fig. 3, we have reproduced the original figure by Loftus (1978) showing a variety of interpretable and uninterpretable interactions. As can be seen in the figure, interactions in a 2 × 2 ANOVA are interpretable only if the two lines representing the simple effects touch or cross each other. Note that the results of a 2 × 2 ANOVA can be represented in two ways,

[1] To obtain the data patterns for this example, we just took four data points (40, 30, 30, and 20) as means in the latent variable and generated four unimodal, symmetrical samples with 70 % of the observations being included in interval between the mean and one standard deviation and N = 20. For instance, for the cell a, with M = 20, the sample was 37, 38, 38, 39, 39, 39, 40, 40, 40, 40, 40, 40, 40, 40, 41, 41, 41, 42, 42, 43. We then use the conversion chart to obtain the corresponding samples in the indicator variable.

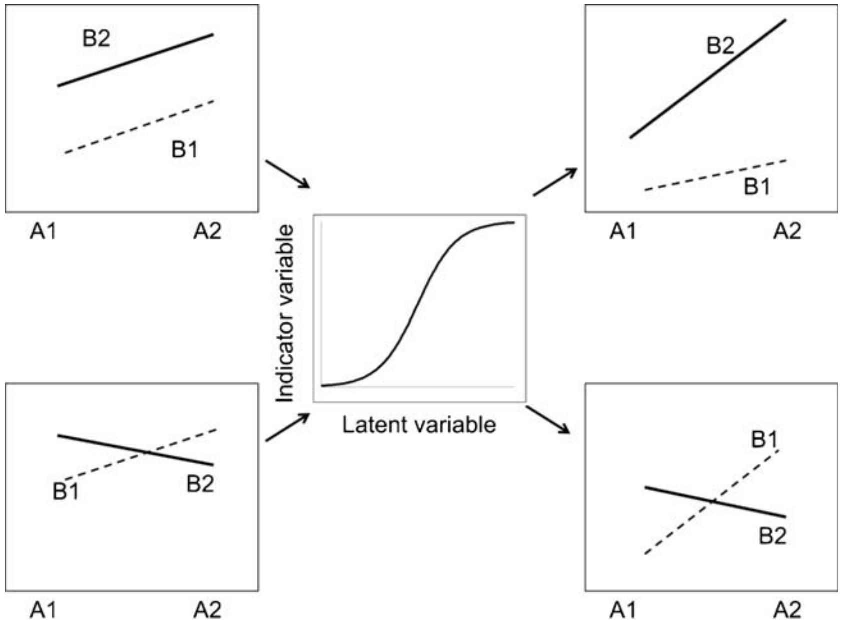


Fig. 1 The center panel shows the (nonlinear) relationship between the latent outcome variable and its indicator. The left panels show the effect of two factors A and B on the latent outcome variable in two hypothetical experiments with a 2 × 2 design. The right panels show these same effects on the observed indicator. The top panels show an uninterpretable interaction, the bottom panels an interpretable interaction

either factor A on the abscissa (top figure) and factor B in the graph or vice versa (bottom figure). An interaction is interpretable if the lines touch or cross in at least one of the two representations. This phenomenon is illustrated in the third column in the middle panel of Fig. 3. Although the two lines cross each other only in one of the two representations, the interaction is interpretable because it cannot be "explained away" by a nonlinear relationship between the latent outcome variable and its indicator.

Wagenmakers et al. (2012) distinguished between truly interpretable interactions, where the two lines cross, and borderline interpretable interactions, where the two lines merely touch each other (see the second column from the left in the middle panel of Fig. 3). The latter interactions are technically interpretable, but one must make the assumption that the absence of a simple effect at one level of A is meaningful (and not simply due to random error); that is, one must accept the point-null hypothesis. However, the absence of a simple effect is probably not meaningful at extreme levels of the independent variable, where it is difficult to statistically detect small differences. This phenomenon is illustrated in Fig. 4a, where there seems to be an interactive effect on the indicator but this effect is nonexistent with the latent outcome variable. Borderline interpretable interactions, which are characterized by two lines that touch but do not cross, thus have to be interpreted with caution, especially if it appears that the absence of a simple effect at one level of one of the factors is due to a ceiling effect, a floor effect, or insufficient statistical power.

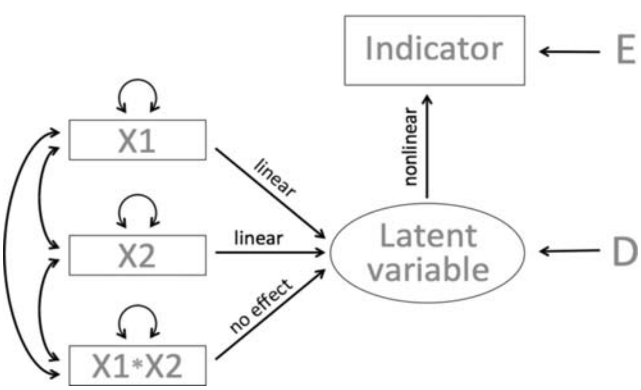


Fig. 2 Loftus's (1978) causal model with two additive effects on a latent variable and a nonlinear effect of the latent variable on its indicator

Finally, it should be noted that a nonlinear relationship between the latent outcome variable and its indicator can also mask an existing, meaningful interaction. Consider Fig. 4b, in which there is a significant interaction effect of the independent variables on the latent outcome variable. This effect is undetectable, however, when researchers perform statistical analyses with the indicator as the dependent variable. Researchers have always been advised to abstain from interpreting nonsignificant effects as evidence for an absence of an effect. This advice is even more relevant when dealing with nonsignificant interaction effects in a 2 × 2 ANOVA. The absence of a significant interaction may be entirely due to the fact that the latent outcome variable is nonlinearly related to its indicator. We will come back to the conditions that have to be satisfied in order to be able to interpret a nonsignificant interaction as a true absence of effect.

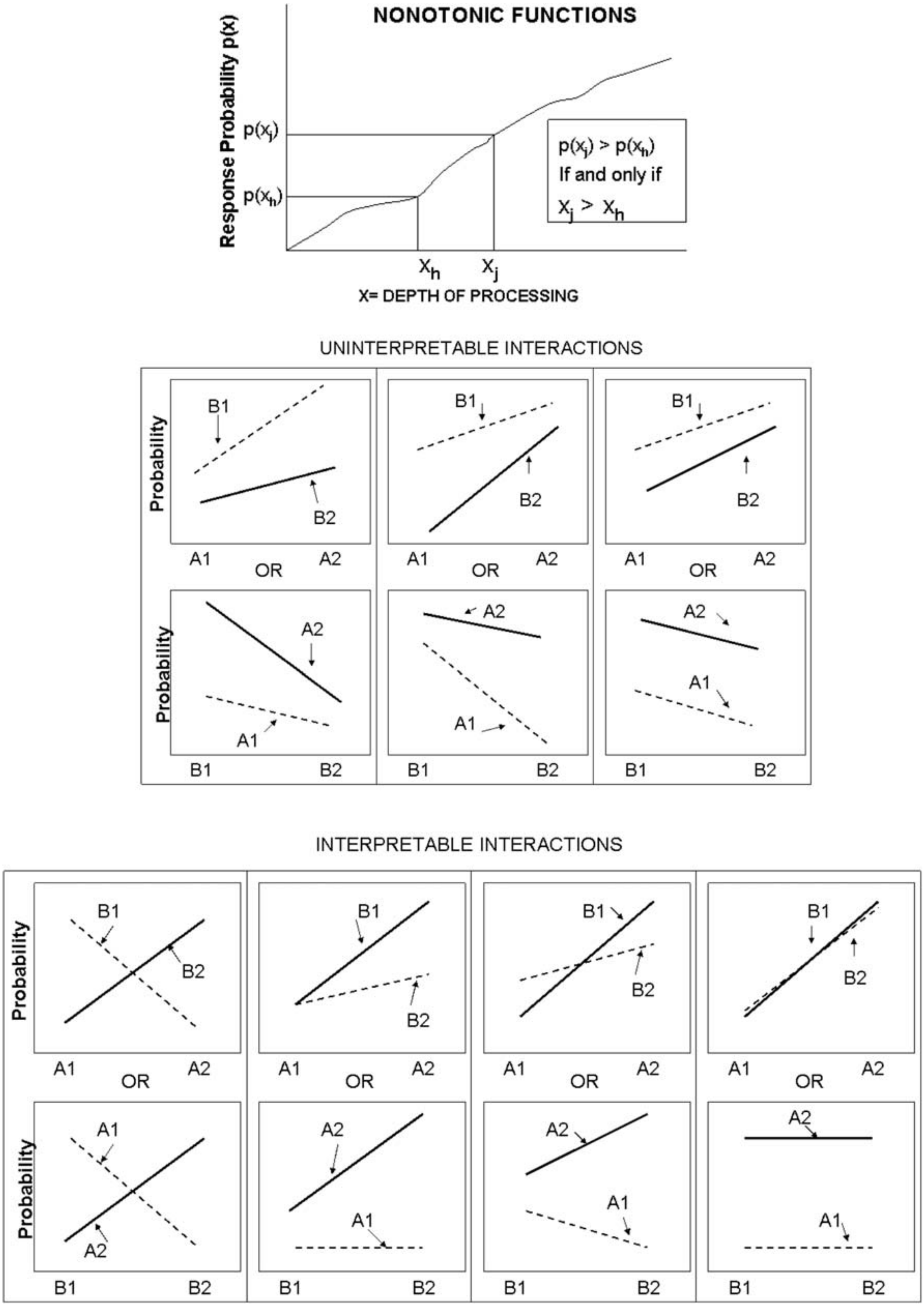


Fig. 3 Loftus's (1978) original figure distinguishing between interpretable and uninterpretable interactions. The results of each hypothetical experiment are shown twice, once with factor A on the abscissa (top figure) and factor B in the figure and once the other way around (bottom figure)

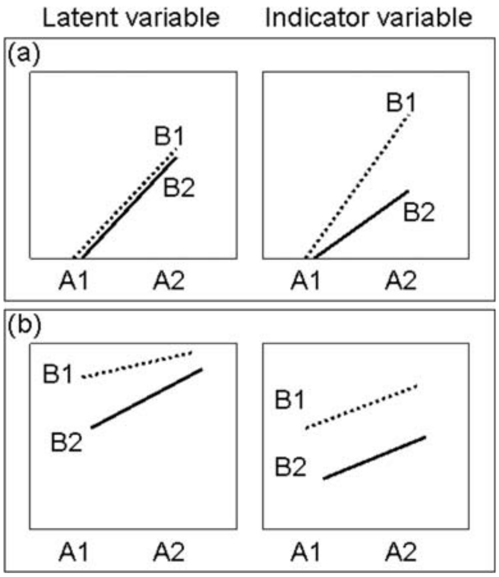


Fig. 4 Examples of a hypothetical experiment in which **a** the indicator suggests an interpretable interaction effect (because the two lines in the right panel touch, according to Loftus, 1978). It turns out, however, that they only touch because the indicator does not distinguish between minor differences on the latent variable. **b** A nonlinear relationship between the latent outcome variable and its indicator masks an existing interaction

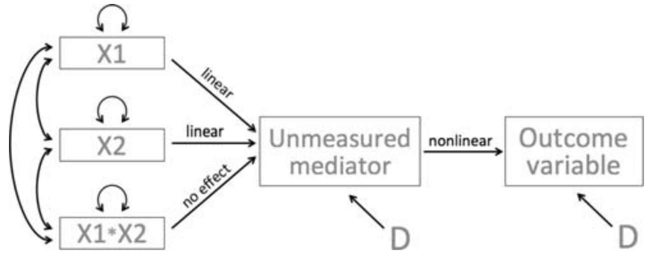


Fig. 5 A causal model with two additive effects on an unmeasured mediator and a nonlinear effect of the mediator on the outcome variable

Before we continue, we want to acknowledge that others have raised other problems in the interpretation of interactions An endemic form of this problem arises, for instance, when the construct variable has an interval nature but the measurement device is able to produce data only at the ordinal level of measurement (Stevens, 1946). In this case, spurious interactions may emerge (see, for instance, Anderson, 1961; Bogartz, 1976; Embretson, 1996; Kang & Waller, 2005). In particular, Davison and Sharma (1990) showed that in general, for factorial designs, the results of tests on a continuous, ordinal measure of a latent variable cannot be guaranteed to produce valid inferences about main effects and interactions on the latent variable even when the standard normality and equality of variance assumptions hold. Note, however, that the issue we are raising in this article is fundamentally different because, in our entire article, we refer to problems that may affect measurement devices even if they produce an interval level of measurement.

Extension to mediator models

Loftus's (1978) analysis can be extended to models in which all the variables at study maintain a linear relationship with their latent constructs. Take, for example, a model in which two independent variables have an additive effect on an unmeasured mediator and the mediator has a nonlinear effect on the observed (manifest) outcome variable. Such a meditational model is shown in Fig. 5.

Let's illustrate this point with a concrete example. Imagine a study in which occasional runners are contacted several months before a popular, city-wide 5 k race. All participants are offered a free registration for the race, but half of the participants receive the free registration 3 months before the race, whereas the other half receive it 1 week before the race (factor A). And participants are encouraged to run either alone or with a friend (factor B). The outcome variable of interest is how fast they run in the 5 k race. Let's assume that the true effects of the two independent variables on the dependent variable are purely additive; that is, there is no interactive effect on an unmeasured mediator, the number of training run per weeks. Knowing 3 months before that one will take part in a race causes participants to train more. And running with a friend causes them to train more. It is safe to assume that the number of training runs per week is nonlinearly related to the performance on the 5 k: The difference between running once a week versus running 3 times a week leads to a much greater performance increase than does the difference between running 5 times a week versus running 7 times a week. Table 1 shows the results of the hypothetical study. Although the interactive effect of the independent variables on the (unmeasured) mediator is zero, when analyzing his/her data, the researcher might find a statistically significant interaction in the analysis in which the running time in the 5 k race is the dependent variable. However, this apparent interaction

Table 1 Results of a hypothetical experiment in which the additive effects of two factors on the outcome variable are mediated by an unmeasured variable that is related in a logarithmically decreasing function with the measured outcome variable

	No. of Training Runs per Week (Unmeasured)	Time in the 5 k Race (Measured)
Registration 1 week before, run alone	2	30 min
Registration 1 week before, run with friend	3	25 min
Registration 3 months before, run alone	4	23 min
Registration 3 months before, run with friend	5	22 min

is due entirely to the fact that the (unmeasured) mediator has a nonlinear effect on the (observed) outcome variable. Note that similar results may be obtained in any study in which two independent variables have an additive effect on effort but unit increases in effort have diminishing returns on the outcome.

Let's consider another example. Imagine a study in which participants are requested to complete a list of word stems. Each word stems had a single possible completion and these completions corresponded to high- or low-frequency words (factor A) that were either old (corresponded to words previously presented in an earlier study list in the same experiment) or new (factor B). The dependent variable is the number of correct completions. Let's assume that the effects of the independent variables on the outcome variable are entirely mediated by level of activation (not directly measured in the study) and that the effects of the independent variables on the mediator are purely additive (no interaction). For the sake of our example, let's further assume that the relationship between the unmeasured variable, activation, and stem completion is nonlinear. As activation increases, word completion increases exponentially. As can be seen in Table 2, the researchers would falsely conclude that the independent variables exert an interactive effect on the dependent variable, but the pattern of means is entirely due to the fact that there is a nonlinear effect of level of activation (the unmeasured mediator) on stem completion (the measured outcome variable). Such a pattern of results is likely to occur in any study in which unit increases in a (mediating) psychological state lead to increasingly stronger [or weaker] behavioral responses.

As the two examples in the previous paragraphs demonstrate, the interpretation of a noncrossover interaction is ambiguous even if the relationship between the manifest outcome variable and its underlying latent construct is perfectly linear. An interaction with two noncrossing lines may be observed in a causal model in which the two independent variables have additive effects, in which these effects are mediated by an unmeasured variable, and in which the mediator has a nonlinear effect on the observed outcome variable.

The unmeasured mediator case may be even more problematic than the situation described by Loftus (1978) in which there is a nonlinear relationship between the latent outcome variable and its indicator. In the latter case, the researcher can perform the appropriate transformation before analyzing the data if enough is known about the density functions of the studied constructs. However, in the former case, the researchers may ignore the existence of the mediator altogether.

In sum, we contend that Loftus's (1978) distinction between interpretable and uninterpretable interactions is an important one, but we also think that it applies to a much wider range of psychological phenomena than initially assumed.

Table 2 Results of a hypothetical experiment in which the additive effects of two factors on the outcome variable are mediated by an unmeasured variable that is related with an exponential function to the measured outcome variable

	Activation (Unmeasured)	Word Completion (Measured)
Low frequency/new items	0	0
High frequency/new items	1	1
Low frequency/old items	3	9
High frequency/old items	4	16

Interactions and main effects (Rosnow & Rosenthal, 1989)

In a series of papers, Rosnow and Rosenthal (1989, 1991, 1995; Rosenthal & Rosnow, 1985) reminded us of the fact that cell means in a 2 × 2 ANOVA are determined by both main effects and the interaction effect. In fact, cell means can be reproduced using a linear combination of four components: the general mean, the first main effect, the second main effect, and the interaction effect:

$$Y = b_0 + b_1 A + b_2 B + b_3 (A * B).$$

Thus, in order to be able to inspect one of the effects, it is necessary to remove the other effects from the cell means. To illustrate this point, let's consider two hypothetical abstract examples that correspond to two contrasting interactions (Fig. 6):

Figure 6a illustrates each interaction based on cell means interpretation. The interaction depicted on the left is an interpretable interaction, whereas the one of the right is uninterpretable. They seem very different, but the figures represent

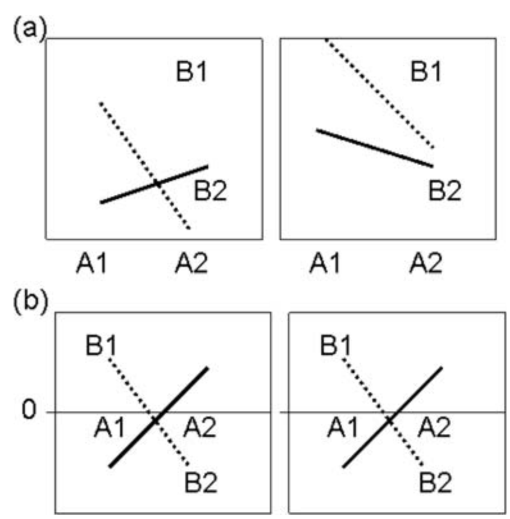


Fig. 6 **a** Examples of graphs representing cell means of a interpretable (left) and uninterpretable (right) interaction. **b** Graphs representing the interaction component after removing the two main effects, represented in the cell means (as suggested by Rosenthal & Rosnow, 1985)

cell means, which are a joint function of the general mean, the two main effects, and the interaction. If we want to understand the nature of the interaction by the inspection of the cell means, then, according to Rosenthal and Rosnow (1985), we should partial out the other effects first. Fig 6b displays the adjusted cell means; that is, the two main effects have been removed from the data. In the bottom panel, the two interactions look exactly the same. And indeed, after partialling out the other effects, all interactions appear interpretable according to Loftus's criterion.

This insight suggests that it is not a characteristic of the interaction effect (b_3) itself that makes a statistically significant interaction interpretable, but a characteristic of the general pattern of cell means (as shown by Loftus, 1978). Given that cell means are determined by both main effects (b_1 and b_2) and the interaction effect (b_3), both the main effects and the interaction effect must jointly contribute to the interpretability of an interaction. We will illustrate this point in the following paragraphs.

Rosenthal and Rosnow (1985) have offered us a pedagogical way to understand the data pattern of factorial designs, building data by adding each treatment component to each cell. In Fig. 7, we have generated several data patterns using this approach. In Fig. 7a, all three effects are of identical magnitude and of size 1. In Fig. 7b, one of the main effects is stronger in magnitude than the other two effects. In Fig. 1c, the interaction effect is greater in magnitude than both main effects. In Fig. 1d, the interaction effect is smaller in magnitude than both main effects.

As can be seen in Fig. 7a, the emerging interaction is considered interpretable according to Loftus's (1978) classification (or at least borderline interpretable according to the classification of Wagenmakers and colleagues, 2012), because the lines touch in both of its possible graphic depictions. The same is true for the interaction presented in Fig. 7b, because the lines touch in one of its possible graphic depictions. The data from Fig. 7c also correspond to an interpretable interaction (even in the more stringent Wagenmakers classification), because the lines cross in both of the two possible graphic representations. The example presented in Fig. 7d corresponds to an uninterpretable interaction, because the two lines representing the simple effects do not cross or touch in either of the two representations.

As Fig. 7 shows, it is the relative position of cell means, and not any particular aspect of the interaction effect (b_3) alone, that determines whether the interpretation of an observed

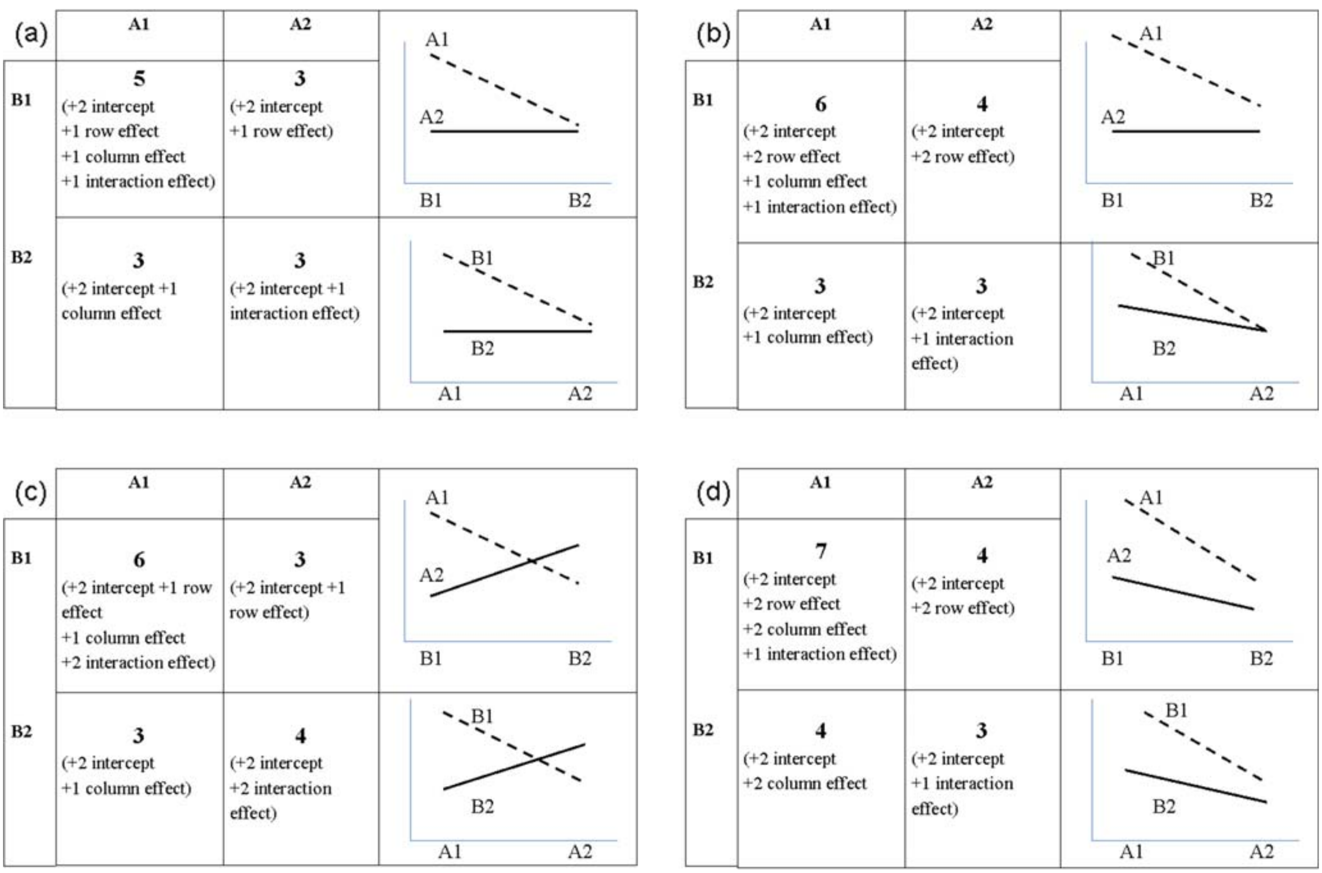


Fig. 7 Pattern of means of a hypothetical experiment in which **a** the first main effect, the second main effect, and the interaction effect all have identical magnitude; **b** the second main effect and the interaction have the identical magnitude but are smaller than the first main effect; **c** the interaction effect is greater than the two main effects; and **d** the interaction effect is smaller than the two main effects

interaction is ambiguous or not. Although the interaction effect modeled in Fig. 7a, b is exactly the same as the one modeled in Fig. 7d—1 was added to the cells in the diagonal—the resulting interaction is interpretable in the former examples but not in the latter. By modeling main effects and interaction effects of different sizes using the Rosnow and Rosenthal approach, one can come to the following conclusions (see the addendum for a proof):

- When the interaction effect (b_3) is greater than the smaller of the two main effects, then the interaction is interpretable.
- When the interaction effect (b_3) is smaller than the smaller of the two main effects, then the interaction is uninterpretable.
- When the interaction effect (b_3) is of equal magnitude as the smaller of the two main effects, then the interaction is borderline interpretable.

What matters is the relative size of the interaction effect, as compared with the smaller of the two main effects. Of course, if both main effects are of equal size, then the relative size of the interaction effect, as compared with either main effect, determines the interpretability of the interaction. The insight that relative size plays a key role leads to another important conclusion that we will discuss in more detail in the next section of the article: If researchers want to demonstrate an interaction, they should try to design an experiment so that at least one of the two main effects is smaller than the interaction effect.

It is important to point out that when talking about the "smallest of the three effects," we are referring to the absolute size of the unstandardized regression coefficients b_1, b_2, and b_3, or, expressed differently, the differences between groups of cell means. In the present context, relative size is *not* determined by the effect size or the significance level of each of the three effects.

The Rosnow and Rosenthal approach also makes clear that main effects influence the interpretability of "the absence of an interaction"—that is, a purely additive effect. Using Rosnow and Rosenthal's approach, one can show that the lack of an interaction is interpretable only when at least one of the two main effects is zero. If both main effects are nonzero, an additive pattern of means cannot be interpreted as a lack of interaction (in addition to the recommendation that we should never interpret the absence of an effect as evidence for the nonexistence of this effect). Loftus (1978) mentioned this point as well: "The lack of an interaction is uninterpretable when both independent variables show a main effect; however, when one or both of the independent variables fail to show a main effect, this lack of interaction is interpretable" (p. 315).

Uninterpretable main effects

Rosenthal and Rosnow (1985) approach allows us to take the analysis one step further and address "uninterpretable main effects." If one cannot exclude the possibility that there is a nonlinear monotonic relationship between a latent outcome variable and its indicator (or between an unmeasured mediator and a measured outcome variable), then there is always one effect that is uninterpretable, and this effect is not necessarily the interaction. As it turns out, the smallest of the three effects in a 2 × 2 ANOVA is always uninterpretable. If the interaction effect is the smallest of the three effects, the interaction effect is uninterpretable. If one of the two main effects is the smallest of the three effects, this main effect can be "explained away" by a nonlinear but monotonic relationship between a latent outcome variable and its indicator. Loftus's (1978) argument is thus not specific to the interaction effect but applies to the smallest of the three effects. Let's consider a data pattern in which the scores of the measured indicator variable are obtained by log-transforming the scores of an unmeasured latent construct.

Figure 8 shows the case of an uninterpretable interaction. The interaction effect is the smallest of the three effects, and, as a consequence, the two lines do not cross in either of the two representations. A (monotonic) log_2 transformation of data makes the interaction disappear, as we can see in the right-hand side of the figure. But what if one of the main effects (the row effect), and not the interaction, were the weakest effect?

Such a case is represented in Fig. 9. Figure 9a shows the scores on the latent outcome variable, and Fig. 9b the scores on the measured indicator. As one can see in Fig. 9b, the row main effect was removed by the log transformation, whereas the other effects (the column and the interaction) were preserved. The reason why this happens is the following. All of the effects of 2 × 2 design can be defined by two contrasts between differences—namely,

Row effect: (A1B1– A1B2) – (A2B2–A2B1) **or** (A1B1–A2B2) – (A1B2– A2B1)
Column effect: (A1B1– A2B1) – (A2B2– A1B2) **or** (A1B1– A2B2) – (A2B1– A1B2)
Interaction: (A1B1– A2B1) – (A1B2– A2B2) **or** (A1B1–A1B2) – (A2B1– A2B2).

Note that every effect, except the weakest, can be defined by a contrast between differences in which the smaller is included in the larger difference. For instance, the column effect can be defined as a difference between (A1B1 – A2B1) [(16 – 1)] and (A2B2 – A2B1) [(4 – 4)] (=15) and the interaction effect by the difference between (A1B1 – A2B1) [(16 – 1)] and (A1B2 – A2B2) [(4 – 4)] (=15). Only

(a)

	A1	A2
B1	16	4
B2	4	1

(b)

	A1	A2
B1	4	2
B2	2	0

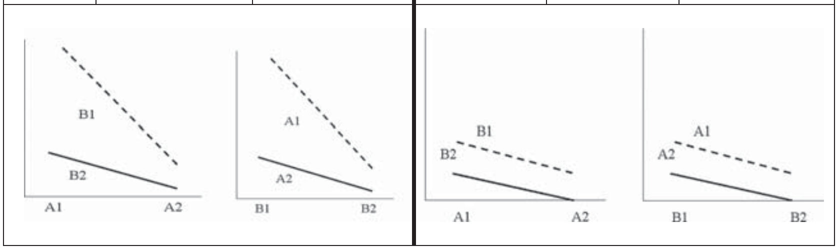


Fig. 8 **a** Data and graphs representing an interaction that is removable by a log transformation. **b** Data and graph following the log transformation of the data

the row effect (=9), in this case, cannot be defined by a difference between the largest difference in this case (16 – 1) and the other difference (4 – 4).

Generally speaking we can say that when an effect of a 2 × 2 design can be defined by a contrast between the largest difference between the two of the four cells and the difference between the remaining cells, the effect will be preserved by any monotonic transformation, because the latter is included in the former. An easy way to understand this is to note that by definition, a monotonic transformation preserves the rank ordering of the observations. Thus, the only operations that we know will be preserved with a monotonic transformation are the operations that produce results that are necessarily preserved after a rank transformation of the observations. In our case, the only difference between differences that is preserved after we transform cell means in ranks is the difference between the biggest difference (rank 1 – rank 4) and (rank 2 – rank 3). In the same vein, we could provide a different definition to the uninterpretable interactions refer to by Loftus (1978): Uninterpretable interactions are interactions that are removed by a rank transformation.

Perhaps Loftus (1978) was not aware of the possibility of removable main effects because he only considered cases in which the largest difference was always between cells A1B1 and A2B2 and, therefore, only main effects could be defined by the contrast between this difference and the difference between the two remaining cells. In these cases, one can observe only uninterpretable interactions.

Two examples from real research

As was the case in Loftus (1978), in the present article, we resort to abstract cases and hypothetical examples. The reason is simple: Hypothetical examples can be made straightforward and cleaner. More important, we "know" both the conceptual variables and their empirical outlets in hypothetical examples, whereas we have access only to the latter with real research examples. However, it is easy to find published 2 × 2 designs in which all three effects are statistically significant and to inquire about the reliability of the smallest of these effects (a main effect in one case and an interaction in the other). We

(a)

	A1	A2
B1	16	1
B2	4	4

(b)

	A1	A2
B1	4	0
B2	2	2

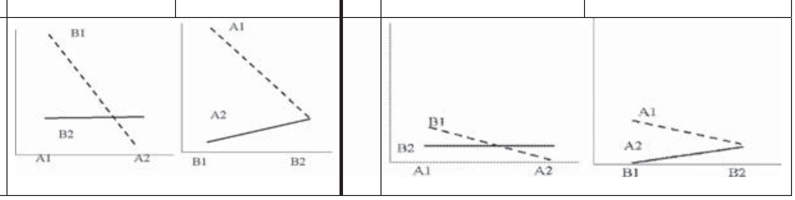


Fig. 9 **a** Data and graphs representing a main effect as the weakest effect. **b** The log_2 transformation of the same data

chose two examples, in which although the researchers were not of course dealing with the problems raised in this article, their subsequent research nevertheless provided interesting suggestions for our inquiry and how to deal with these problems.

In Beaman, Hanczakowski, Hodgetts, Marsh, and Jones (2013), the authors used a distraction paradigm (Neely & LaCompte, 1999) to investigate the extent to which arbitrary perceptual features, which vary between lures and targets, can help discriminate between semantically related lures and targets. The participants had to ignore visually presented words written in italic and remember words written in normal typeset font. The design of Experiment 1 included two semantic relatedness conditions (targets belonged to the same category or to different categories) crossed with perceptual similarity (targets and lures appear in the same color or in different colors). Both main effects and the respective interaction were significant in the analyses with false recall as the dependent variable (lures recalled as targets). However, the main effect of perceptual similarity was the smallest of the three effects. Is the perceptual similarity effect uninterpretable? The answer to this question turns out to be important, because this effect was new and did not appear in true free recall (targets recalled as targets). Experiment 2 conceptually replicated Experiment 1, but the study modality was changed from visual to auditory. Thus, participants had to ignore words spoken in a loud voice and remember words spoken in a quiet voice. And perceptual similarity was manipulated in terms of the gender of the voice (same gender voice reading both targets and lures vs. different genders reading targets and lures). The same main effects and interaction emerged from the false free recall data. Moreover, the perceptual similarity effect was no longer the effect with the smallest magnitude. The most parsimonious account of these data is that the perceptual similarity effect obtained is interpretable because not only did the effect prove to be robust with two completely different manipulations of perceptual similarity, but also its smallest magnitude status did not replicate across experiments. In the next section, we will use this example to draw implications for research.

In Acheson, Postle, and McDonald (2010), the authors wanted to explore the effects of a lexical-semantic variable, word concreteness, in the moderation of the phonological similarity effect (PSE). The PSE corresponds to the finding that immediate memory for lists containing phonological overlap is worse than for those that do not (Conrad & Hull, 1964). In Experiment 1, the authors crossed phonological overlap (lists composed of words that were phonologically similar or not) with word concreteness (two levels of concreteness according to pretest ratings). Both main effects and the interaction emerged as significant. However, the interaction had a smaller magnitude than the two main effects. Is this interaction interpretable? In Experiment 2, the authors made participants perform a concurrent articulatory suppression task (voicing the word "the") during stimuli presentation, but the manipulation of the factors remained the same. As was expected, the articulatory suppression task made phonological similarity effect disappear, but the interaction remained significant. These results strongly suggest that interaction is, in fact, interpretable, because the interaction emerged in both experiments but its smallest status varied across the two. In the next sections, we discuss the methodological strategies that can be used when all effects of a 2 × 2 design are significant and the authors want to interpret all three effects.

Implications for research

One of the goals of this article is to raise awareness among researchers about the problems we face when interpreting effects in a 2 × 2 ANOVA. As previous publications have made clear, there is a need to carefully attend to the pattern of data associated with interactions. The most important take-home message of this article is that any pattern of four cell means generates an interpretative problem that researchers should be aware of. When the interaction effect is smaller than the smaller of the two main effects, an observed interaction is uninterpretable because it may be due to a nonlinear relationship between the latent outcome variable and its indicator or to a nonlinear relationship between an unmeasured mediator and the observed outcome variable. When one of the two main effects is the smallest of the three effects, this main effect is uninterpretable. Thus, Loftus's (1978) warning does not apply exclusively to interaction effects, but to overall data patterns. It is the pattern of four cell means that determines which of the three effects is the smallest (and thus uninterpretable), not some characteristic of one of the three effects. In this article, we focus mainly on interaction effects because we assume that most researchers are interested in the interaction effect when conducting a study with a 2 × 2 design.

Statistical training should include the learning and recognition of overall data patterns and not only of isolated effects. As Rosenthal and Rosnow (1985) have shown, what is often interpreted as an interaction is a composite of main and interaction effects. Being able to recognize problematic data patterns should be an important addition to the researcher's data analytical toolbox. To that end, the approach advocated by Rosenthal and Rosnow (1985)—consisting of constructing or reconstructing data patterns by adding main effects and interactions—seems particularly promising. Our examples were restricted to probably the most popular design in psychology—namely, the 2 × 2 factorial design—but the same approach can easily be used in more complex designs.

As we have shown, only the weakest of the three effects in a 2 × 2 ANOVA is uninterpretable. In all of Loftus's (1978) examples, the interaction was the weakest of the effects, but the problem also arises for a given main effect if the other main effect and the interaction effect are larger. The smallest of the three effects will, on average, be nonsignificant more often than the other two effects, but even if it is significant, there is an interpretational ambiguity.

In addition, there is an important set of circumstances that should make the researcher particularly wary of the interpretative status of the smallest effect. As a reviewer from a previous version of this manuscript suggested, the mapping of the latent into the empirical variable will do more than leading to the emergence of uninterpretable interactions or main effects; it will cause heterocedasticity and/or skewness in the overall data pattern. Thus, we should be particularly distrustful of the smallest effect of a 2 × 2 ANOVA (with all effects being significant) when the homoscedasticity or asymmetry assumptions are violated. As we all know, testing ANOVA assumptions should be part of our standard analytic practice, but we now add the possibility of testing violations revealing nonlinear mapping consequences of measurement. Unfortunately, there is no reason to assume either that the latent variable will always follow a normal probability density function (e.g., just think of the law of practice or the forgetting curve) or that every transformation will produce heterocedasticity or skew. After all, most of the transformations we routinely perform on our data are used to *reduce* hetereocedasticity and skewness. Thus, it is quite possible that mapping functions will sometimes approximate the empirical variable to the shape of a normal distribution, and not the opposite. In these more treacherous cases, problems with the interpretability of the smallest effect can emerge with no violation of ANOVA assumptions.

What can researchers do when, in a 2 × 2 design, both main effects and their respective interaction are significant? How can they set up their experiment so that they end up with interpretable interactions and main effects? We suggest six different approaches.

The first suggestion is to increase efforts in multiple operationalizations of the same latent construct, exactly as Beamon et al. (2013) did in our first real research example. If we use different operationalizations of the same construct, it is unlikely that the same problems of nonlinearity emerge in each case or, at least, that different manipulations of the same factors always produce a pattern in which the weakest factor remains the same (be it an interaction or a main effect).

Second, it may useful to try several plausible monotonic transformations of the same data and to see whether the weakest of the three effects is robust or disappears. However, this approach does not guarantee that the interpretational ambiguity can be removed. The fact that the effect remains robust with several different transformations provides us with valuable insight, but not with any definitive conclusion. It informs us either about the nature of the function that maps the construct to an indicator or about the shape of the distributional function of the latent construct. Different transformations will provide us with this information.

Third, in the case of interactions, we should try to make the interaction effect as big as possible. By using strong manipulations in order to create extreme levels of the independent variables, the size of the interaction effect can sometimes be increased. Also, it is worth it to think very hard about the levels of factor A at which the simple effect of factor B will be minimum and maximum in size, and likewise with factor B. The goal will then be to manipulate factors A and B in a way that maximizes the interaction effect.

The fourth suggestion follows up on the third: Researchers can try to make at least one of the main effects as small as possible. For instance, Acheson et al. (2010), in our second real-world example, first found the interaction to be smallest of the three significant effects in their 2 × 2 ANOVA and then replicated their experiment adding a manipulation known to suppress one of the two main effects of their design. Because the interaction remained significant, whereas the added manipulation made one of the main effects disappear, the interaction can be said to be interpretable.

Consider another hypothetical example. A group of researchers runs an experiment in which depressed and nondepressed participants are asked to find words in a letter grid. For half of the participants, the letter grid contains sad words; for the other half, it contains happy words. The researchers thus employ a 2 (factor A: “Type of participant: depressed vs. nondepressed”) × 2 (factor B: “Type of word: happy vs. neutral”) between-subjects design. The researchers’ prediction is that depressed individuals will find fewer words if these words have a happy meaning. The researchers’ hypothesis is shown in Table 3 (see the column labeled “Hyp”): They predict an interaction between type of participant and type of word. Let’s further assume that the researcher’s hypothesis is correct.

The researchers might run the experiment without devoting a lot of thought to the main effects. After all, they think, any main effect will be statistically removed prior to testing the interaction effect. What they seem to forget, however, is that large main effects can transform an interpretable interaction into an uninterpretable one. Consider the results displayed in the column “Exp. 1” in Table 3. They confirm the researchers’ prediction, but in addition, we have “added” two main effects: Sad words are easier to find than happy words (maybe sad words were shorter), and nondepressed

Table 3 Hypothesis and possible outcomes of a hypothetical experiment with an interaction hypothesis

	Hyp.	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Depressed participants, happy words (a):	6	6	6	6	6
Depressed participants, sad words (b):	10	12	10	10	8
Nondepressed participants, happy words (c):	10	11	10	10	10
Nondepressed participants, sad words (d):	10	13	10	11	8

Note. The reported values are cell means.

participants find more words than depressed women (maybe nondepressed participants had better language skills). The result is an uninterpretable interaction.

Researchers are better off devoting effort to pretesting their material to make sure that sad words are as easy or as difficult to find as happy words and to find participants who do not differ in their language abilities across groups. If they do that, they may end up with the pattern of means displayed in the "Exp. 2" column of Table 3—that is, exactly their prediction. The interaction is borderline interpretable. They would now have to show that the two lines representing the simple effects truly touch each other. In other words, they would have to demonstrate (1) that at least one of the two simple effects is statistically nonsignificant (either the type of word effect for nondepressed participants, the type of participant effect for sad words, or both; demonstrating that one of them is nonsignificant is sufficient because, as soon as the two lines touch each other in one of the two possible graphic representations, the interaction is borderline interpretable). The researchers would also have to demonstrate (2) that they had enough statistical power to detect the simple effect if it were to exist and (3) that the nonsignificance of this simple effect replicates across two or more experiments. In short, they have to demonstrate that it makes sense to accept the point-null hypothesis (see Cook & Campbell, 1979, for a discussion of the conditions that have to be satisfied before we can reasonably accept a null hypothesis).

The previous strategy has two drawbacks, however. First, it may be difficult to come up with convincing evidence in favor of the point-null hypothesis. Second, minor random variation will turn the borderline interpretable interaction into an uninterpretable one (as shown in Table 3, column labeled "Exp. 3"). The researchers' case would be strengthened considerably if they could generate a data pattern that corresponds to an interpretable interaction. There is one way to do precisely that: The researchers could "create" a main effect that is likely to result in a cross-over interaction. For example, they could use sad words that are slightly more difficult to find than the happy words for control participants. In doing so, they might obtain a data pattern comparable to the one shown in the last column of Table 3 (labeled "Exp. 4"). Nondepressed participants find more happy words than sad words, and vice versa for depressed participants. The resulting cross-over interaction is interpretable. Alternatively, they may pretest participants' language skills and include participants so that the depressed participants in the experiment have, on average, slightly better language skills than the nondepressed participants. The result would again be an interpretable cross-over interaction.

Our fifth suggestion is to use more than two levels of at least one of the two factors. For instance, in the stem completion example described earlier, researchers instead of using a single representation, may manipulate the number of previous presentations of words that could be used to complete the word stems (from one to four repetitions). The resulting pattern of means would allow the researchers to eliminate any ambiguity with regard to the interpretation of the interaction.

This point is illustrated in Fig. 10. In the top panels, we have reproduced the data from the top panels of Fig. 1, but we have added two levels of factor A so that this factor now has four levels instead of two. If the effects of the two independent variables on the latent outcome variable are purely additive and the observed interaction in the top right panel of Fig. 1 is due to a nonlinear relationship between the latent outcome variable and its indicator, then an experiment with four levels of factor A would yield two lines that are curved differently. In the bottom right panel of Fig. 10, we show a data pattern in which there is a linear effect of factor A on the observed indicator but this linear effect is moderated by factor B. It is unlikely that such a data pattern would be produced by a nonlinear relationship between the latent outcome variable and its indicator. As can be seen in the bottom left panel of Fig. 10, one would have to assume that factor A has a curvilinear effect of one kind for one level of B (unit increases of A lead to increasingly smaller increases in the latent outcome variable) and a curvilinear effect of another kind for the other level of B (unit increases of A lead to increasingly larger increases in the latent outcome variable). Such relationships are highly unlikely. With a 4 × 2 design, researchers can thus argue convincingly that an interaction between the linear

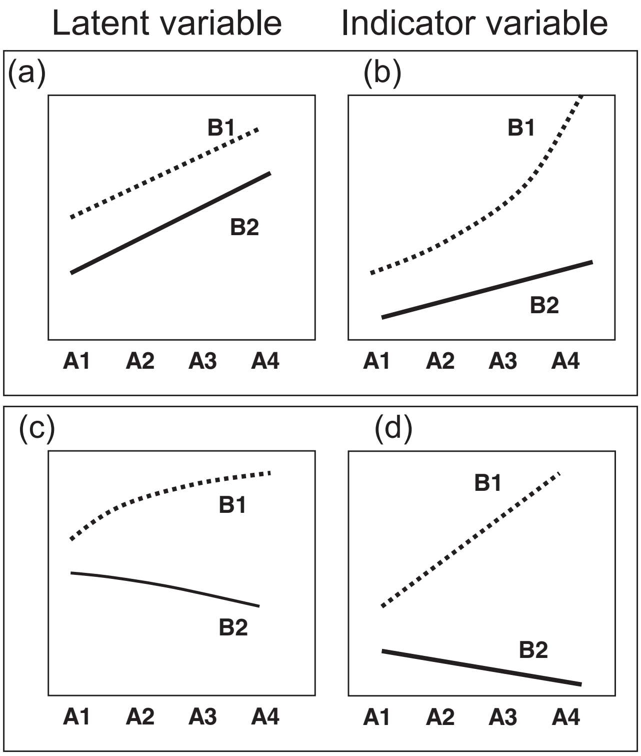


Fig. 10 Examples of a hypothetical experiment with a 4 × 2 design in which the effects of two factors A and B on the latent outcome are shown (**a** and **c**) and these same effects on the observed indicator are shown (assuming the nonlinear relationship between the latent outcome variable and the indicator shown in Fig. 1) (**b** and **d**). Panels a and b show a purely additive effect on the latent outcome variable, and panels c and d show a data pattern that is impossible to attribute to a nonlinear relationship between the latent outcome variable and the indicator

trends of factor A and factor B is truly an interactive effect, rather than a by-product of a nonlinear relationship between the latent outcome variable and its indicator (or between an unmeasured mediator and the observed outcome variable).

Our sixth and final suggestion is to invoke theory, including measurement theory, to argue against the alternative interpretation of an observed interaction. If researchers have a good measurement model, they can argue against a nonlinear relationship between the latent outcome variable and its indicator. Maybe the same latent outcome variable has been measured with multiple indicators in earlier studies, and it has been shown that the indicator used in the present experiment is related linearly to the latent variable as measured with the other indicators. Or maybe the researchers have already identified a mediating variable that accounts for a large proportion of the total effect of the independent variables on the outcome variable and can now show that this mediator is linearly related to the outcome variable. When attempting to publish an interaction that is "uninterpretable," researchers should explicitly mention in the article that they are aware of the interpretational ambiguity and ideally generate some arguments for why they think it is unreasonable to assume that there is a nonlinear relationship between the latent outcome variable and its indicator or between an unmeasured mediator and the observed outcome variable.

Conclusion

In the present article, we have shown that certain interactions (so-called cross-over interactions) are inherently more interpretable than other interactions. We have shown that the interpretational ambiguity of non-cross-over interactions is not limited to the case in which the latent outcome variable is related to its indicator in a nonlinear way but exists also in the case where there is a nonlinear relationship between an unmeasured mediator and the outcome variable. We have further shown that there is nothing specific about the interaction effect (b_3) that is problematic but that entire data patterns (all four cell means of a 2 × 2 ANOVA) conspire to make it interpretable or not. Using an approach introduced by Rosnow and Rosenthal (1989), we showed that the smallest of the three effects is always uninterpretable even if it is statistically significant. If the interaction effect is larger than at least one of the two main effects, it is interpretable. This insight leads to a number of recommendations. If researchers predict an interaction and want to able to interpret this interaction, they should devote considerable attention to the main effects of their experiment and make sure that they are as small as possible. Researchers may also attempt to construct their stimulus material with the goal of obtaining a cross-over interaction. Other possible solutions are to include more than two levels of at least one of the two factors or to use theoretical considerations to argue against the idea that an observed interaction can be "explained away" by the existence of a nonlinear relationship in the causal model.

Sometimes, researchers tend to think of theoretical knowledge as conjectural and continuously evolving, but they take scientific methods as definitive. We join those who reject this difference and see methods and statistics as endlessly evolving (Laudan, 1977). Statistics and methods are excellent aids to critical thinking, not easier alternatives to critical thinking (Abelson, 1995). We hope that the present article furthers the understanding of the interpretation of interaction effects.

Acknowledgments The writing of this article was supported, in part, by Grant PTDC/PSI-PSO/111992/2009 from the FCT (Portuguese National Science Foundation) awarded to the first author. The authors want to thanks the valuable contribution of Geoffrey Iverson, a much appreciated reminder that sometimes the review system really works.

Appendix

latent	indicator	latent	indicator	latent	indicator	latent	indicator	latent	indicator
0	.7	21	5.2	41	28.9	61	75.0	81	95.7
1	.7	22	5.7	42	31.0	62	76.9	82	96.1
2	.8	23	6.3	43	33.2	63	78.6	83	96.4
3	.9	24	6.9	44	35.4	64	80.2	84	96.8
4	1.0	25	7.6	45	37.8	65	81.8	85	97.1
5	1.1	26	8.3	46	40.1	66	83.2	86	97.3
6	1.2	27	9.1	47	42.6	67	84.6	87	97.6
7	1.3	28	10.0	48	45.0	68	85.8	88	97.8
8	1.5	29	10.9	49	47.5	69	87.0	89	98.0
9	1.6	30	11.9	50	50.0	70	88.1	90	98.2
10	1.8	31	13.0	51	52.5	71	89.1	91	98.4
11	2.0	32	14.2	52	55.0	72	90.0	92	98.5
12	2.2	33	15.4	53	57.4	73	90.9	93	98.7
13	2.4	34	16.8	54	59.9	74	91.7	94	98.8
14	2.7	35	18.2	55	62.2	75	92.4	95	98.9
15	2.9	36	19.8	56	64.6	76	93.1	96	99.0
16	3.2	37	21.4	57	66.8	77	93.7	97	99.1
17	3.6	38	23.1	58	69.0	78	94.3	98	99.2
18	3.9	39	25.0	59	71.1	79	94.8	99	99.3
19	4.3	40	26.9	60	73.1	80	95.3	100	99.3
20	4.7								

Note. We assumed that the relationship between the indicator and its latent variable was a modified inverse logit function. The exact conversion formula is *indicator* = (e $^{(latent\text{-}50)/10}$/ (e $^{(latent\text{-}50)/10}$+ 1)) * 100

Addendum by Geoffrey Iverson

The purpose of the following remarks is to examine the meaning of the term "interaction" as it occurs in connection with tables of experimental means. The general setup is familiar. A pair of two-level factors $A=(A_1,A_2)$, $B=(B_1,B_2)$ is studied in a factorial (within-subjects) design, and the experimental means (assuming for simplicity equal sample sizes in each condition) are plotted in a 2×2 table of the sort

Treatment	B_1	B_2	Marginals
A_1	$\overline{X}_{11}=x$	$\overline{X}_{12}=z$	$\overline{X}_{1.}=(x+z)/2$
A_2	$\overline{X}_{21}=y$	$\overline{X}_{22}=w$	$\overline{X}_{2.}=(y+w)/2$
Marginals	$\overline{X}_{.1}=(x+y)/2$	$\overline{X}_{.2}=(x+y)/2$	$\overline{X}_{..}=(x+y+z+w)/4$

The usual ANOVA model is assumed—namely,
$X_{ijk}\sim\mu_{ij}+\sigma\varepsilon_{ijk}$, with ε_{ijk} i.i.d $N(0,1)$.

Here, the subscript $i\in\{1,2\}$ indexes the levels of factor A, the subscript $j\in\{1,2\}$ indexes the levels of factor B, and the subscript k distinguishes participants.

The ANOVA is a saturated model and decomposes treatment means as a sum of main effects and a 2×2 interaction matrix:

$$\mu_{ij} = \mu_{..} + \alpha_i + \beta_j + \gamma_{ij}.$$

The nice feature of this linear statistical model is that this same decomposition carries over to treatment means (which are estimators for the estimands μ_{ij}):

$$\overline{X}_{ij} = \overline{X}_{..} + \widehat{\alpha}_i + \widehat{\beta}_j + \widehat{\gamma}_{ij}.$$

The structure of the interaction matrix is very simple. Because the rows and columns of γ_{ij} sum to zero, one has

$$\left(\gamma_{ij}\right) = \varepsilon\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

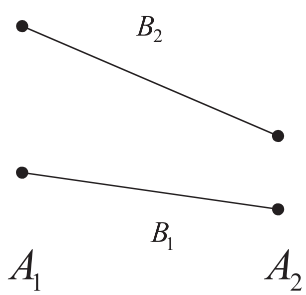
One has a corresponding estimator $\left(\widehat{\gamma}_{ij}\right) = \widehat{\varepsilon}\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}$. Note that even if $\varepsilon=0$ (so the model is additive and lacks an interaction term), the data will almost surely show an interaction—that is, $\widehat{\varepsilon}\neq 0$.

The three canonical forms of data

A collection of data means fall into one or other of three disjoint classes; each is illustrated below by way of a typical example.

Class 1 Order-independence (removable interaction)

The factors A and B are each ordered, and these orders are *independent*. For example, the diagram below illustrates that $\overline{X}_{11} < \overline{X}_{12}$ and $\overline{X}_{21} < \overline{X}_{22}$, and also that $\overline{X}_{11} > \overline{X}_{21}$ and $\overline{X}_{12} > \overline{X}_{22}$ (we might summarize these inequalities in the form $A_1 \succ A_2$ and $B_2 \succ B_1$).



This pattern of data means corresponds to the order $\overline{X}_{12} > \overline{X}_{22} > \overline{X}_{11} > \overline{X}_{21}$.

The importance of this order-independent class (which consists of eight orders all of the above sort) is that (1) the magnitude of the interaction is smaller than the magnitude of either main effect—that is, $|\widehat{\varepsilon}| < \min\left\{|\widehat{\alpha}_1|, \left|\widehat{\beta}_1\right|\right\}$ —and (2) it allows the observed interaction to be "transformed away." (This last fact does not obtain for the other two classes [irremovable interactions] that we consider below.) Let us be concrete about this claim: Using the abbreviations x,y,z,w for treatment means as per the 2 × 2 table above, we have

$$\widehat{\alpha}_1 = (x-y+z-w)/4 = -\widehat{\alpha}_2$$
$$\widehat{\beta}_1 = (x+y-z-w)/4 = -\widehat{\beta}_2$$
$$\widehat{\gamma}_{11} = (x-z-y+w)/4 = -\widehat{\gamma}_{12} = -\widehat{\gamma}_{21} = \widehat{\gamma}_{22} = \widehat{\varepsilon}.$$

Because the mean data are ordered such that $z>w>x>y$, we see that $-\widehat{\beta}_1 > \widehat{\alpha}_1 > 0$. . The above claim (1) is established if $\widehat{\alpha}_1 > |\widehat{\varepsilon}|$. Suppose $\varepsilon>0$. Then the claim (1) reads

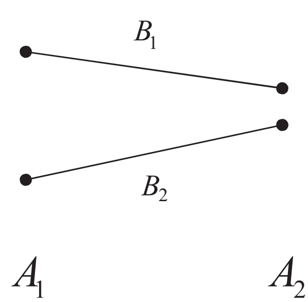
$$x-y+z-w > x-z-y+w,$$

which is true since $z>w$. Suppose $\varepsilon<0$. Then the claim (1) reads $x-y+z-w>-x+z+y-w$, which holds because $x>y$.

Suppose now that the mean data are transformed by a 1–1 mapping φ: $\overline{Y}_{ij} = \varphi\left(\overline{X}_{ij}\right)$. The transformed means $\overline{Y}_{ij}$ are ordered in exactly the same way as the original means if φ is strictly increasing and take on the reverse order if φ is strictly decreasing. In either case, the magnitude of the interaction observed in the transformed data is $|\varphi(x)-\varphi(z)-\varphi(y)+\varphi(w)|$. Suppose now that φ is strictly increasing, so that $\varphi(z)>\varphi(w)>\varphi(x)>\varphi(y)$. Pick values of φ such that $\varphi(z)>\varphi(w)>\varphi(y)$ and define $\varphi(y)=\varphi(x)+\varphi(w)-\varphi(z)<\varphi(x)$. Then the interaction among the transformed means is zero, establishing claim (2). The case where φ is strictly decreasing is similar.

Class 2 Partial Interaction (irremovable interaction, removable main effect).

For all members of this class, one factor is consistently ordered at each level of the other factor, but the same is not true for the other factor. For instance, in the example below, we have $\overline{X}_{11} > \overline{X}_{12}$ and $\overline{X}_{21} > \overline{X}_{22}$ (which we might write $B_1 \succ B_2$); however, we also have $\overline{X}_{11} > \overline{X}_{21}$ and $\overline{X}_{12} < \overline{X}_{22}$ (so that neither $A_1 \succ A_2$ nor $A_2 \succ A_1$). In all, there are eight possible mean orders that fall into this class. It is enough to describe in detail any one of these eight orders, and we do so in the following example. Note that in the diagram, the lines do not cross. However, interchanging the roles of factors A and B and replotting the same mean data produces a diagram in which the lines do cross.



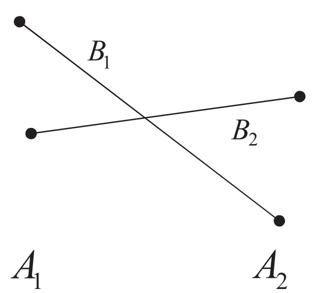
These mean data correspond to the simple order $x>y>w>z$.

In this example, the smallest effect magnitude is that of $|\alpha_1|$. This claim is easily confirmed. For it is readily established that $\widehat{\beta}_1 > \widehat{\varepsilon} > 0$. Thus, one needs to verify only that $\widehat{\varepsilon} > |\widehat{\alpha}_1|$. Suppose $\widehat{\alpha}_1 > 0$; then $\widehat{\varepsilon} > \alpha_1 > 0$ since $w > z$. If, however, $\widehat{\alpha}_1 < 0$, the claim is verified in virtue of the fact that $x > y$.

Under a strictly increasing transformation φ the data means are transformed so that $\varphi(x)>\varphi(y)>\varphi(-w)>\varphi(z)$, and these values are otherwise arbitrary. Choose values of $\varphi(x),\varphi(y),\varphi(w)$ at will, subject to $\varphi(x)>\varphi(y)>\varphi(w)$, and define $\varphi(z)=\varphi(y)+\varphi(w)-\varphi(x)$, noting that this definition entails $\varphi(z)<\varphi(w)$. The main effect of factor A is thus zero following the transformation.

Class 3 Full (crossover) Interaction (removable main effect)

This class is exemplified in the diagram below. To be concrete about things, consider data means satisfying $x>w>z>y$. (There are eight such orders in this class). Neither factor A nor factor B is consistently ordered in such data. For this class of data means, the interaction is the largest effect, and the smallest effect corresponds to a main effect that can be transformed to zero by a 1–1 function φ. These claims are easily established along lines detailed above for classes 1 and 2.



References

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Erlbaum.

Acheson, D. J., Postle, B. R., & MacDonald, M. C. (2010). The interaction of concreteness and phonological similarity in verbal working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(1), 17.

Anderson, N. H. (1961). Scales and statistics: Parametric and nonparametric. *Psychological Bulletin, 58,* 305–316.

Beaman, C. P., Hanczakowski, M., Hodgetts, H. M., Marsh, J. E., & Jones, D. M. (2013). Memory as discrimination: What distraction reveals. Memory & cognition, 41(8), 1238–1251.

Bogartz, R. S. (1976). On the meaning of statistical interactions. *Journal of Experimental Child Psychology, 22,* 178–183.

Conrad, R., & Hull, A. J. (1964). Information, acoustic confusion and memory span. *British journal of psychology, 55*(4), 429–432.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field settings*. Chicago, IL: Rand McNally.

Davison, M. L., & Sharma, A. R. (1988). Parametric statistics and levels of measurement. *Psychology Bulletin, 104,* 137–144.

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement, 20,* 201–212.

Kang, S.-M., & Waller, N. G. (2005). Moderated multiple regression, spurious interaction effects, and IRT. *Applied Psychological Measurement, 29,* 87–105.

Laudan, L. (1977). *Progress and its problems: Towards a theory of scientific growth*. Berkeley, CA: University of California Press.

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition, 6,* 312–319.

Neely, C. B., & LeCompte, D. C. (1999). The importance of semantic similarity to the irrelevant speech effect. *Memory & Cognition, 27*(1), 37–44.

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. Cambridge: Cambridge University Press.

Rosnow, R. L., & Rosenthal, R. (1989). Definition and interpretation of interaction effects. *Psychological Bulletin, 105,* 143–146.

Rosnow, R. L., & Rosenthal, R. (1991). If you're looking at the cell means, you're not looking at only the interaction (unless all main effects are zero). *Psychological Bulletin, 110,* 574–576.

Rosnow, R. L., & Rosenthal, R. (1995). "Some things you learn aren't so": Cohen's paradox, Asch's paradigm, and the interpretation of interaction. *Psychological Science, 6,* 3–9.

Rubin-Rabson, G. (1937). The influence of analytical prestudy in memorizing piano music. *Archives of Psychology, 31*(No. 220).

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677–680.

Wagenmakers, E.-J., Krypotos, A.-M., Criss, A. M., & Iverson, G. (2012). On the interpretation of removable interactions: A survey of the field 33 years after Loftus. *Memory & Cognition, 40,* 145–160.