

# On the interpretation of removable interactions: A survey of the field 33 years after Loftus

Eric-Jan Wagenmakers · Angelos-Miltiadis Kryptos ·  
Amy H. Criss · Geoff Iverson

© The Author(s) 2011. This article is published with open access at Springerlink.com

**Abstract** In a classic 1978 *Memory & Cognition* article, Geoff Loftus explained why noncrossover interactions are removable. These removable interactions are tied to the scale of measurement for the dependent variable and therefore do not allow unambiguous conclusions about latent psychological processes. In the present article, we present concrete examples of how this insight helps prevent experimental psychologists from drawing incorrect conclusions about the effects of forgetting and aging. In addition, we extend the Loftus classification scheme for interactions to include those on the cusp between removable and nonremovable. Finally, we use various methods (i.e., a study of citation histories, a questionnaire for psychology students and faculty members, an analysis of statistical textbooks, and a review of articles published in the 2008 issue of *Psychology and Aging*) to show that experimental psychologists have remained generally unaware of the concept of removable interactions. We conclude that there is more to interactions in a  $2 \times 2$  design than meets the eye.

**Keywords** Transformations · Measurement scale · Statistics in psychology · Literature review

---

E.-J. Wagenmakers (✉) · A.-M. Kryptos  
Department of Psychology, University of Amsterdam,  
Roetersstraat 15,  
1018 WB Amsterdam, The Netherlands  
e-mail: E.J.Wagenmakers@gmail.com

A. H. Criss  
Syracuse University,  
Syracuse, NY, USA

G. Iverson  
University of California,  
Irvine, CA, USA

Few statistical concepts appear to be as straightforward as an interaction in a  $2 \times 2$  design. Most statistical textbooks inform undergraduate psychology students that an interaction is indicated when “the lines are not parallel.” Thus, all psychologists are familiar with the concept of an interaction, and they often report and interpret interactions obtained in their own experiments. It is easy to conclude that experimental psychologists know what an interaction is, and how it should be interpreted. Unfortunately, there is more to an interaction than meets the eye.

More than three decades ago, Geoff Loftus published a *Memory & Cognition* article in which he summarized results from measurement theory (e.g., Krantz & Tversky, 1971; Luce & Tukey, 1964) and demonstrated that interactions are not created equal: Some interactions—the ones that cross over—are “nonremovable,” whereas the others are “removable” (Loftus, 1978; see also Anderson, 1961, 1963; Bogartz, 1976). A nonremovable interaction can never be undone by a monotonic transformation of the measurement scale, and it is therefore also known as qualitative, cross-over, disordinal, nontransformable, order-based, model-independent, or interpretable (Cox, 1984; De González & Cox, 2007; Neter, Wasserman, & Kutner, 1990). In contrast, a removable interaction can always be undone by a monotonic transformation of the measurement scale; such an interaction is also known as quantitative, ordinal, transformable, model-dependent, or uninterpretable.

Given the prominence of interactions in psychological research, it is important for experimental psychologists to be familiar with the Loftus (1978) article and realize that the only interactions that are nonremovable are the ones that cross over. Our personal experience, however, has led us to conjecture that experimental psychologists have forgotten about the difference between nonremovable and removable interactions. When told about the existence of

removable interactions and the role of scale transformations, colleagues commonly respond, “But why would I want to transform my measurement scale at all?” Therefore, the first goal of the present article is to answer this question and to reiterate the main message from Loftus (1978). The second goal is to introduce a classification scheme for interactions that refines the one proposed by Loftus (1978). The third goal is to demonstrate through various means (i.e., a study of citation histories for the Loftus article, a questionnaire for faculty and graduate students, an analysis of statistical textbooks, and a review of the literature) that, 33 years after the Loftus (1978) article, experimental psychologists are—to their peril—generally unaware of the fact that many interactions are removable.

### Why worry?

Psychologists are generally interested in latent processes, the workings of which they infer from changes in observed behavior (e.g., changes in a dependent variable). By definition, the latent process of interest is never directly observed; what is observed is the dependent variable. The dependent variable reflects merely the output of a latent psychological process. Thus, the scale of the unobserved psychological process is transformed to the scale of the observed dependent variable. Unfortunately, in most experimental paradigms the exact relationship between unobserved process and observed behavior is unknown; hence, the scale transformations may well be nonlinear.

To appreciate the counter-intuitive consequences that nonlinear transformations may have on the conclusions that we draw from data, consider the following real-life situation.<sup>1</sup> Suppose you want to travel by car to a destination that is 100 km away, in as little time as possible. Does it matter whether you increase your speed from 40 km/h to 50 km/h versus from 60 km/h to 70 km/h? Most people think of km/h as a fundamental measure of how long it takes to get from A to B, and they may therefore intuit that there is no interaction between initial speed (i.e., 40 versus 60 km/h) and speed increase (i.e., 10 km/h). However, this intuition is misleading. It makes more sense to compute that the trip time decreases by 30 min if you increase your speed from 40 to 50 km/h, but it decreases by only about 14 min if you increase your speed from 60 to 70 km/h. In other words, lack of interaction with respect to one dependent variable (km/h) implies an interaction in a monotonically but nonlinearly related dependent variable (h/km).

<sup>1</sup> We thank Geoff Loftus for proposing this example, which we have copied almost to the letter.

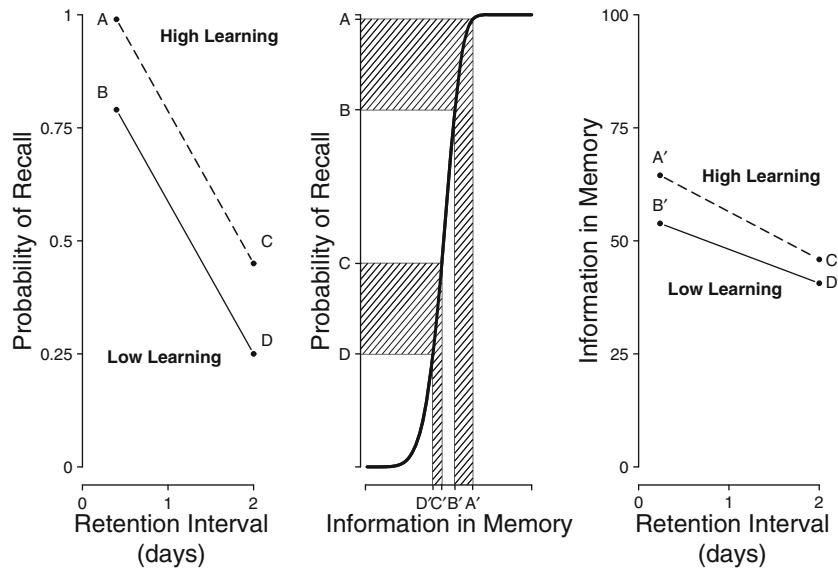
The following two examples demonstrate the relevance and practical ramifications of scale transformations for psychological research.

### Example 1: retention curves

Consider an experiment that seeks to establish whether the rate of forgetting depends on the initial level of learning (see Loftus, 1985a, Fig. 1; see also Anderson, 1963; for an in-depth discussion see Bogartz, 1990a, b; Laming, 1992; Lawrence, 1994; Loftus, 1985b; Loftus & Bamber, 1990; Slamecka, 1985; Slamecka & McElree, 1983; Wixted, 1990). Participants are presented with a list of study words, and their performance—measured in proportion of items correctly recalled—is assessed after two retention intervals, one short and one long. In the high learning condition, the study words are presented five times each, and in the low learning condition, the study words are presented only once. The left panel of Fig. 1 shows fictive but plausible results. Clearly, the reduction in recall probability does not depend on the initial level of study, and one might be tempted to conclude that the rate of forgetting is independent of the initial level of learning.

However, the middle panel shows a hypothetical function that translates the probability of successful recall to “information stored in memory.” This latter quantity could be measured in features, chunks, exemplars, or strength—The exact unit is not important here. Note that the translation is nonlinear, and the observed data occupy different positions on the function. Because of the one-to-one mapping between recall probability and information in memory, we can transform the observed data and express it on our new scale “information in memory.” The right panel of Fig. 1 shows the result. On the new scale, the data show an interaction: Information loss depends on the initial level of learning, and one might now be tempted to conclude that the rate of forgetting is steeper in the high learning condition than in the low learning condition.

This example shows that the conclusion one draws about an interaction may depend on the scale of measurement. One scale, “probability of recall,” shows that forgetting is independent of the level of initial learning, whereas another scale, “information in memory,” shows the opposite. The conflicting conclusions are caused by the nonlinear relationship between two scales. Nothing in psychological theories of cognition suggest an inherently linear relationship between observable behavior and latent cognitive processes (in fact, a linear relationship is a rare find). Whenever a valid model assumes a nonlinear relationship between process and behavior, the kind of interactions shown in Figs. 1 and 2 must be interpreted with caution, since the presence of the interaction hinges on the scale under consideration.

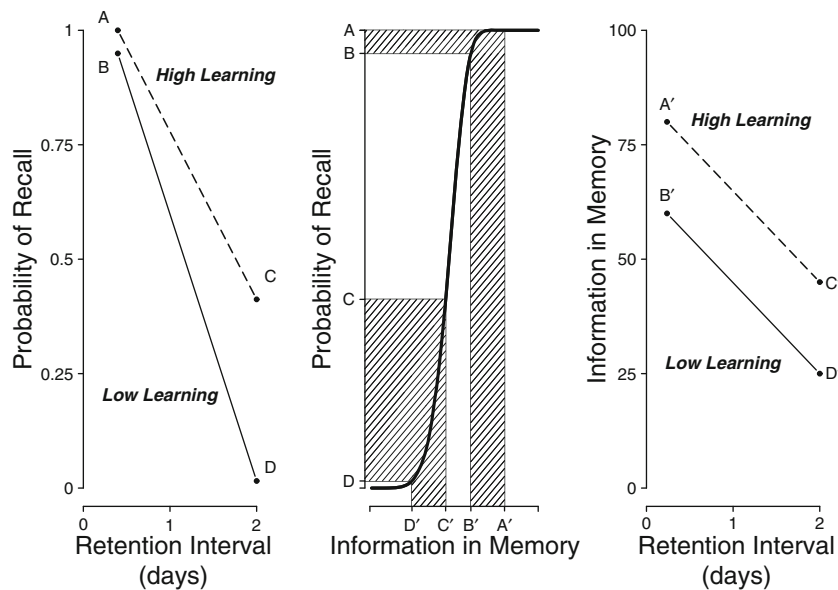


**Fig. 1** Additive effects on the probability of recall correspond to interaction effects on information in memory. The left panel shows data from a hypothetical  $2 \times 2$  experiment with high initial learning and low initial learning in which probability of recall was assessed after a few hours and after 2 days. The effect is additive, since high and low learning are associated with exactly the same delay-driven decrease in recall probability. The middle panel shows how probability

of recall could map on to information stored in memory (arbitrary units). We do not know this function, but for convenience we used the Weibull CDF, such that  $Pr(\text{recall}) = 1 - \exp(-(\text{information}/50)^6)$ ; that is, probability of recall increases with information stored in memory in a sigmoid fashion. The right panel shows that retention interval interacts with the initial level of learning

The reverse result is also possible. The left panel of Fig. 2 shows that in the low learning condition, participants go from performance near ceiling to performance near floor, whereas in the high learning condition, participants

show a much less pronounced decrease in recall probability. Thus, from these data, one may be tempted to conclude that the rate of forgetting is steeper in the low learning condition than in the high learning condition.



**Fig. 2** Interaction effects on the probability of recall correspond to additive effects on information in memory. The left panel shows data from a hypothetical  $2 \times 2$  experiment with high initial learning and low initial learning in which probability of recall was assessed after a few hours and after 2 days. The effect is interactive, since the high and low learning are associated with a very different delay-driven decrease

in recall probability. The middle panel shows how probability of recall could map on to information stored in memory (arbitrary units) according to  $Pr(\text{recall}) = 1 - \exp(-(\text{information}/50)^6)$ . The right panel shows that retention interval does not interact with the initial level of learning

The middle panel of Fig. 2 shows the function that transforms probability of recall to information in memory—note that this function is identical to the one shown in Fig. 1. The right panel plots the observed data on the new scale. On the new scale, the rate of forgetting is independent of the level of initial learning.

These examples show that additivity can turn in to an interaction, and an interaction can turn in to additivity, using a plausible nonlinear mapping from the observed variable (i.e., probability of recall) to the underlying psychological process that it intends to measure (i.e., information in memory). Thus, the intuitive interpretation of interactions in terms of psychological processes can easily lead to conclusions that are misleading. Of course, to argue that a conclusion is definitely misleading requires knowledge of the true scale, which is something we normally do not have access to. However, sometimes it is possible to argue that the correct scale transformation cannot be linear; for instance, recall performance approaches zero as the retention interval becomes very long, regardless of the initial level of performance. This means that large initial differences in performance have to become negligibly small when the retention interval is very large, so that an interaction is bound to occur. Observing such an interaction clearly does not warrant the conclusion that the rate of forgetting depends on the initial level of learning (Loftus, 1985b).

In the examples from Figs. 1 and 2, recall performance is sometimes at ceiling or floor, and researchers tend to be aware of scaling problems related to ceiling or floor effects. These particular ceiling and floor effects, however, are merely a manifestation of the deeper underlying problem. The next example on choice response times (RTs) shows that conflicting conclusions about interactions can also occur in the absence of clear floor and ceiling effects.

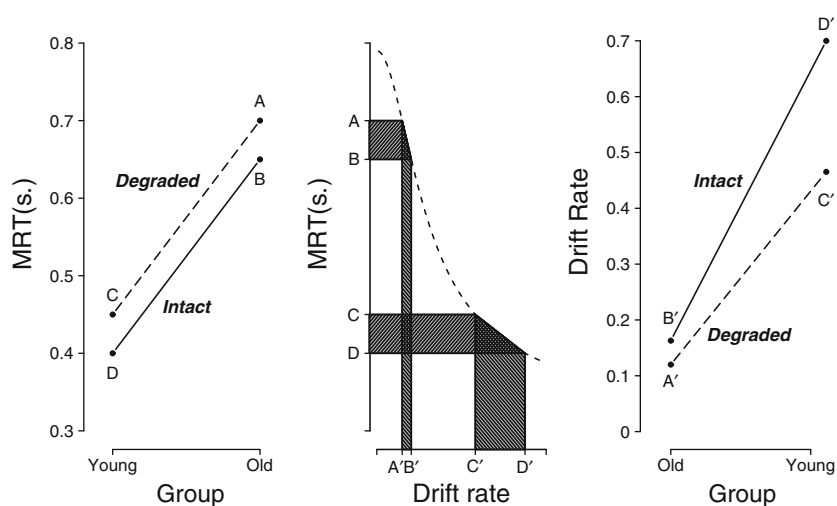
### Example 2: response time analysis

Consider a hypothetical experiment on the effects of aging (i.e., two groups: “old” and “young”) and stimulus quality (i.e., two stimulus types: “intact” and “degraded”) in the lexical decision task—a popular task that requires participants to quickly distinguish words such as *chair* from nonwords such as *drapa* (Rubenstein, Garfield, & Millikan, 1970). Here, we focus on hypothetical results for word stimuli only.

One possible outcome is shown in the left panel of Fig. 3: young people are faster than older people overall, a difference in mean response time (MRT) that is not affected by whether the stimuli are intact or degraded. Hence, the effect of stimulus quality on MRT is additive and does not interact with the effect of age on MRT. From this convincing statistical result, it is tempting to conclude that young and old participants are equally affected by reducing stimulus quality. As we already suggested in the previous example, this conclusion is likely to be misleading.

To illustrate why, consider the diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2008; Wagenmakers, 2009), arguably the most successful formal account of the way in which RTs are generated. The diffusion model has been successfully applied to many two-choice RT paradigms, including short-term and long-term recognition memory tasks, same/different letter-string matching, numerosity judgments, visual-scanning tasks, brightness discrimination, lexical decision, and letter discrimination (e.g., Criss, 2010; Ratcliff, 1978, 1981, 2002; Ratcliff & Rouder, 1998, 2000; Ratcliff, Van Zandt, & McKoon, 1999; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008); in addition, the model has been extensively applied to phenomena in the literature on aging (i.e., Ratcliff, Thapar, & McKoon, 2001, 2003, 2010; Ratcliff, Thapar, Gomez, & McKoon, 2004; Ratcliff, Thapar, & McKoon, 2004; Ratcliff, Thapar, Smith, &

**Fig. 3** Additive effects on MRT correspond to interaction effects on drift rate. The left panel shows an additive effect on MRT in a hypothetical  $2 \times 2$  experiment with young and old participants who are confronted with intact and degraded stimuli. The middle panel shows how mean RT maps onto drift rate, a diffusion model parameter that we assume is uniquely responsible for the observed differences in performance. The right panel shows the corresponding interaction effect on drift rate



McKoon, 2005; Ratcliff, Thapar, & McKoon, 2006a, b, 2007; Starns & Ratcliff, 2010; Thapar, Ratcliff, & McKoon, 2003).

A simplified, “EZ” version of the diffusion model is shown in Fig. 4 (Wagenmakers, van der Maas, & Grasman, 2007). In this model, the decision making process is conceptualized by the gradual accumulation of noisy information until a predetermined threshold is reached. The latent processes that are most often of interest are (1) drift rate  $v$ ; high drift rate signifies a high signal-to-noise ratio and leads to performance that is fast and accurate. Hence, drift rate quantifies the quality of evidence, that is, task easiness or subject ability. (2) Boundary separation  $a$ ; boundary separation modulates the speed–accuracy trade-off so that high boundary separation leads to performance that is relatively slow and accurate, since relatively much evidence needs to be accumulated before a decision is made (Bogacz, Wagenmakers, Forstmann, & Nieuwenhuis, 2010; Forstmann et al., 2008; Wagenmakers et al., 2008). Hence, boundary separation quantifies response caution. (3) Nondecision time  $T_{er}$ . Nondecision time is thought to capture processes not associated with the decision process itself, such as the time needed for stimulus encoding and motor execution (Luce, 1986).

The diffusion model shown in Fig. 4 has a relatively straightforward relation between the model parameters and MRT (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006):

$$MRT = T_{er} + \frac{a}{2v} \tanh\left(\frac{av}{2s^2}\right), \quad (1)$$

where  $\tanh$  is the hyperbolic tangent,  $\tanh(x) = (e^{2x} - 1)/(e^{2x} + 1)$ , and  $s$  is a scaling factor that represents the within-trial variability in drift rate, a quantity that is often set to .1 for historical reasons.

Figure 3, middle panel, shows the nonlinear relation between MRT and drift rate (as per Eq. 1) for fixed values of  $T_{er} = .300$  and  $a = .14$ . Assuming, for the sake of the argument, that the effects of both aging and stimulus degradation reside exclusively in drift rate,<sup>2</sup> this panel also shows what happens when two equal differences on the MRT scale (i.e., A and B) are transformed to their corresponding differences on the drift rate scale (i.e., A' and B'): According to the diffusion model, a change from 300 to 350 ms is much more impressive in terms of drift rate than a change from 600 to 650 ms. Consequently, when we summarize performance not by MRT—as in the left panel—but by drift rate, the additivity is removed and we are instead confronted by an interaction, as is shown in the

right panel of Fig. 3. When comparing the left and right panels of Fig. 3, note that the group labels on the  $x$ -axis have been reversed because short MRTs correspond to high drift rates and long MRTs correspond to low drift rates.

Another possible outcome of the hypothetical lexical decision experiment is shown in the left panel of Fig. 5: Young people respond faster than older people overall, but now the difference in mean response time (MRT) is larger for degraded stimuli than it is for intact stimuli. Hence, the effect of stimulus quality on MRT interacts with the effect of age on MRT. From this convincing statistical result, it is tempting to conclude that a reduction in stimulus quality hurts old participants more than young participants.

However, the middle panel of Fig. 5 shows that the differences in performance for old and young participants are equal on the drift rate scale. The right panel of Fig. 5 confirms that, on the drift rate scale, the effect of stimulus quality is additive.

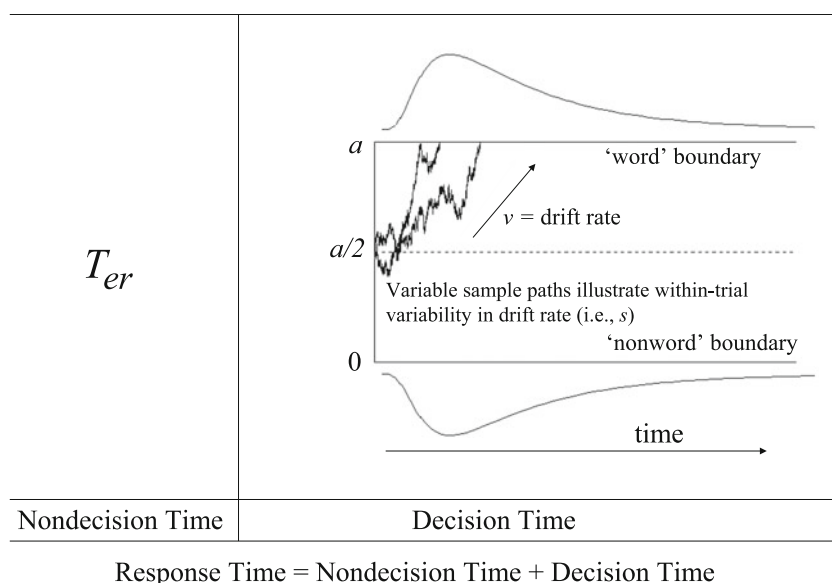
Both the example on the retention curves and the example on choice RT demonstrate that a monotonic nonlinear transformation from, say, MRT to drift rate may dramatically change the interpretation of an interaction (for additional examples see Loftus, 1978, 2002). The reason why experimental psychologists should worry about such transformations is that they are interested not in the specific dependent variable that is measured, per se (e.g., MRT, proportion correct,  $d'$ , priming scores, etc.), but rather in the latent psychological process that drives changes in that dependent variable. From the perspective of theory, experimental psychologists usually measure probability of recall not because they care about the number of items that people can reproduce after a delay, but because they care about the amount of information that is stored in memory. In the same vein, experimental psychologists usually measure RTs not because they care about how quickly people press buttons, but because they care about the efficiency of information processing. However, as was highlighted in the aforementioned examples, the relation between observed performance and latent process (e.g., information in memory, drift rate in a diffusion model) does not need to be linear.

Thus, the aforementioned concern is very general: Our particular examples were based on proportion correct and on a diffusion model analysis of MRT, but the same principle could have been illustrated with other models and other dependent measures such as confidence, galvanic skin response, event-related potentials, and so forth. In fact, it is difficult to find a quantitative model in which parameters that represent psychological constructs are related to dependent measures in a way that is strictly linear. Some experimental psychologists might be unwilling or unable to specify the underlying psychological construct of interest in a formal model, but the problem remains the same:

<sup>2</sup> In real data, effects of aging are usually evident from an increase in nondecision time  $T_{er}$  and an increase in boundary separation  $a$ ; only for perceptual discrimination tasks is there also a decrease in drift rate  $v$  (e.g., Ratcliff et al., 2006a, b; Starns & Ratcliff, 2010).



**Fig. 4** The EZ-diffusion model as applied to a lexical decision task. Noisy information is accumulated from a starting point equidistant between two response boundaries. A response is initiated as soon as the accumulation process reaches a boundary. Total response time is an additive combination of decision time and nondecision time,  $T_{er}$ . See text for details

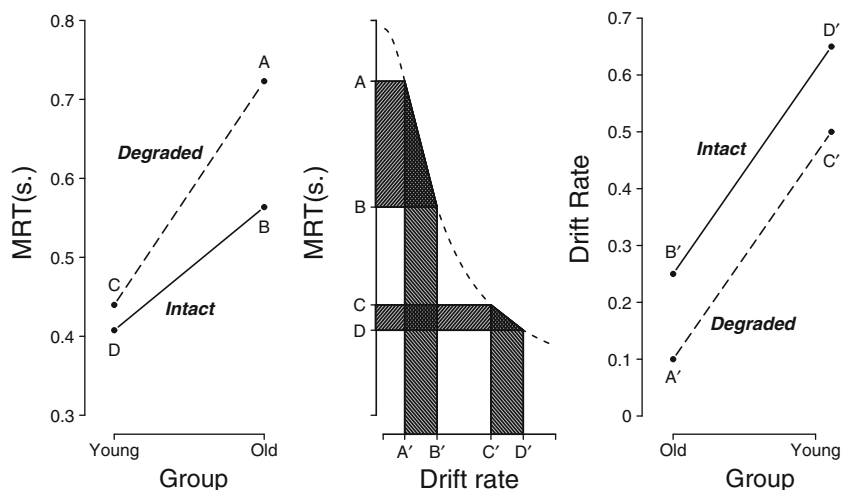


Experimental psychologists seek to draw conclusions about psychological constructs; when the relation between the psychological constructs and the dependent measure is nonlinear—as is likely to be the case—then conclusions about noncrossover interactions may be meaningless.

**Classification and extension**

As was shown in the previous examples, some interactions can be transformed away by a monotonic nonlinear change of the measurement scale; hence, interactions on the scale of the dependent variable may turn out to produce

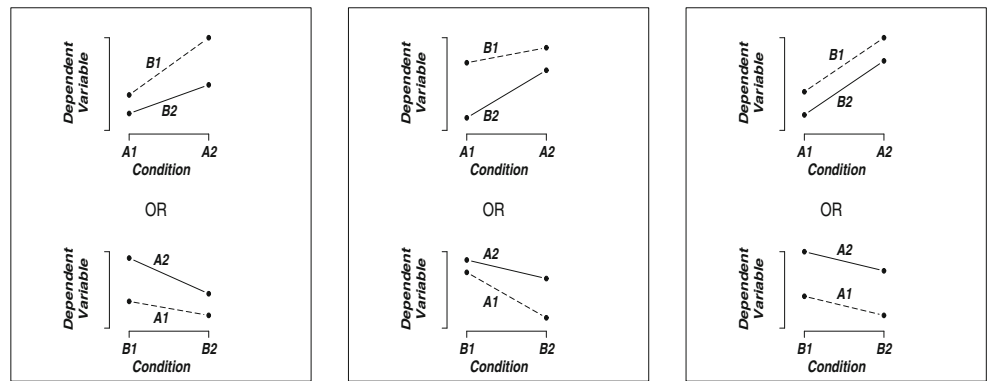
additivity on the scale of the underlying psychological process. Fortunately, however, not all interactions can be transformed away by changes in the measurement scale. The classification scheme of Loftus (1978, his Fig. 3) distinguishes two kinds of interactions. The first kind is sensitive to monotonic nonlinear changes in the measurement scale and is therefore deemed removable. As can be seen from Fig. 6, these are interactions that do not cross, regardless of what variable is plotted on the  $x$ -axis. Note that the interactions shown in the first and third column are of the same type as the ones used in the earlier diffusion model example (i.e., the left panels of Figs. 5 and 3, respectively). Note that the third column shows a null



**Fig. 5** Interaction effects on MRT may correspond to additive effects on drift rate. The left panel shows an interaction effect on MRT in a hypothetical  $2 \times 2$  experiment with young and old participants who are confronted with intact and degraded stimuli. The middle panel shows how mean RT maps onto drift rate, a diffusion model parameter

that we assume is uniquely responsible for the observed differences in performance. The right panel shows the corresponding additive effect on drift rate. When comparing the left and right panels, note that short MRTs correspond to high drift rates and vice versa

**Fig. 6** Removable interactions. These interactions can be transformed to additivity (or vice versa) by a monotonic change of the measurement scale. Note: A1 and A2 refer to two levels of factor A; B1 and B2 refer to two levels of factor B. Within each column, the top graph plots factor A on the x-axis, and the corresponding bottom graph plots factor B on the x-axis



interaction, a pattern of additivity that is commonly thought to reflect the fact that the impact of one factor does not depend on the level of the other (e.g., Sternberg, 1969). As was illustrated in Fig. 3, such a pattern of additivity can be removed by a monotonic transformation of the measurement scale.

The second kind of interactions is insensitive to monotonic nonlinear changes in the measurement scale and is therefore deemed nonremovable. Figure 7 shows that these interactions do cross. For these interactions, one factor moderates the effect of the other factor in terms of direction (i.e., up or down). Note that the interaction in the bottom right panel does not cross when the B factor is plotted on the x-axis, but it does cross when the A factor is plotted on the x-axis (top right panel).

This leaves a set of interactions shown in Fig. 8; strictly speaking, these interactions are nonremovable: They do not cross, but they do touch, meaning that one of the factors has no effect whatsoever. Loftus (1978) classified these interactions as nonremovable. However, in practice, the strict equality of two conditions needs to be established by statistical means. In the traditional framework of null-hypothesis testing, this is problematic, since nonsignificant *p* values do not necessarily support the null hypothesis that a particular simple main effect is absent (e.g., Wilkinson & the Task Force on Statistical Inference, 1999). In other words, our confidence in the interactions shown in Fig. 8 hinges on the statistical evidence in favor of the null hypothesis, something that cannot be quantified with a *p* value (Busemeyer, 1980, footnote 1).

In contrast with *p* values, statistical evidence in favor of a null hypothesis can be quantified within a Bayesian framework (Gallistel, 2009; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009). Nevertheless, even Bayesians are sometimes reluctant to attach probabilities to exact equalities (i.e., point nulls). Also note that the interactions from Fig. 8 are on the cusp between interactions that are removable and nonremovable. Therefore, we decided to label these in-between interactions as “borderline nonremovable.”

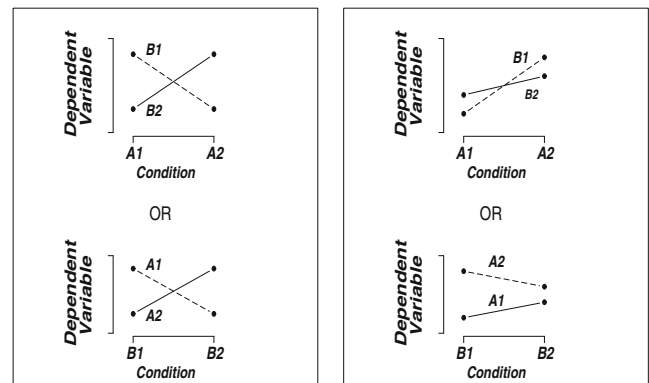
We believe this latter category is useful because it prevents researchers from claiming a nonremovable inter-

action on the basis of the inclusion of a condition that is too hard or too easy. When performance is at chance or at ceiling, it is very difficult to statistically detect small differences. In other words, extreme performance leads to statistical tests that are severely underpowered, and such tests should not be used to support the null hypothesis or to support the existence of a nonremovable interaction.

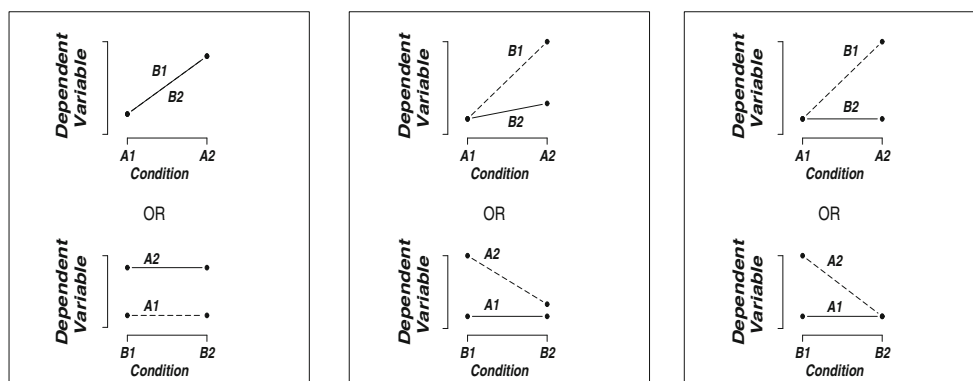
### Acknowledgement and scientific practice

We have shown why it is important to consider scale transformations of the dependent variables, and how the effect of such transformations divides interactions into three categories: nonremovable, removable, and borderline nonremovable. The aforementioned examples highlighted the fact that the interpretation of removable interactions can easily lead to conclusions that are incorrect and misleading.

Given the ease with which researchers may draw incorrect and misleading conclusions from interactions, one would hope that experimental psychologists are aware of the 1978 Loftus article and its theoretical and practical



**Fig. 7** Nonremovable interactions. These interactions cannot be transformed to additivity by a monotonic change of the measurement scale. Note: A1 and A2 refer to two levels of factor A; B1 and B2 refer to two levels of factor B. Within each column, the top graph plots factor A on the x-axis, and the corresponding bottom graph plots factor B on the x-axis



**Fig. 8** Borderline nonremovable interactions. These interactions are on the cusp between removable and nonremovable. Theoretically, these interactions are nonremovable, but in practice their classification hinges on the statistical evidence in favor of a point-null hypothesis.

consequences. To investigate this issue, we took four steps: (a) We recorded the citation history for the 1978 Loftus article; (b) we browsed introductory books on statistics for any mention of the classification of interactions; (c) we conducted a survey amongst psychology students and faculty members to test their knowledge of the different categories of interactions; (d) we considered all articles published in the 2008 volume of *Psychology and Aging* and analyzed the extent to which researchers report and interpret the different categories of interactions.

Citation history for the Loftus 1978 article

To report the citation history of Loftus (1978), we used Web of Knowledge and Google Scholar (work done September 2010). After excluding duplicates and self-citations, both search engines returned 144 hits. Figure 9 shows the number of articles that have cited Loftus (1978) over the years. The number of citations has increased somewhat over time. Overall, the citation rate is rather modest. For instance, *Psychology and Aging* has featured only six citations of Loftus (1978) in 32 years, and the entire field of aging and child development has featured only 19 citations.<sup>3</sup>

Given the importance of the topic and the prevalence of removable and borderline nonremovable interactions in empirical practice, an average score of 4.5 citations per year suggests that most experimental psychologists are probably unaware of the fact that interactions can be removable.

<sup>3</sup> Journals from the field of aging and child development that feature citations to Loftus (1978) are *Psychology and Aging* (6), *Aging, Neuropsychology, and Cognition* (3), *Child Development* (2), *The Journals of Gerontology Series B* (2), *Age and Aging* (1), *Developmental Neuropsychology* (1), *Developmental Psychology* (1), *Experimental Aging Research* (1), *Journal of Experimental Child Psychology* (1), and *The Journal of Gerontology* (1).

Note that the top-left panel features two lines that overlap. Note: A1 and A2 refer to two levels of factor A; B1 and B2 refer to two levels of factor B. Within each column, the top graph plots factor A on the x-axis, and the corresponding bottom graph plots factor B on the x-axis

Nevertheless, Loftus (1978) has not been the only author to point out the existence of removable interactions, and we therefore took additional steps to corroborate this first suggestion.

Reference to removable interactions in statistical textbooks

As a second step, we investigated whether statistical textbooks for undergraduate psychology students mention that certain types of interactions can be transformed away. To this end, we selected 14 popular course books (Agresti & Finlay, 2009; Aron, Aron, & Coups, 2006; Bluman, 2007; Dunn, 2000; Everitt, 2001; Field, 2005; Goodwin, 1997; Greene & D'Oliveira, 1999; Howell, 1992; Howitt & Cramer, 2008; Moore & McCabe, 2006; Nolan & Heinzen, 2007; Pagano, 1998; Shaughnessy & Zechmeister, 1994) and carefully considered the discussion of interactions in each book.<sup>4</sup> Not a single textbook mentioned that certain interactions can be transformed away and should therefore be interpreted with caution. In most books, the authors just point out that interpretation of interactions should be based on a graphical display of the results combined with post hoc tests. We have already seen that a graphical display of the data, however informative, also invites the interpretation of removable interactions.

This disappointing conclusion is consistent with our personal experience: We know of no introductory textbook that discusses the impact of transformations on the interpretation of interactions. The only exception we came across was Winner, Brown, and Michels (1991, pp. 355–356), in which the issue receives cursory discussion. This state of affairs is all the more disappointing because statistical textbooks for undergraduate students do mention

<sup>4</sup> We thank Dr. Lourens Waldorp, who teaches first-year statistics at the University of Amsterdam, for suggesting these materials.





**Fig. 9** Number of peer-reviewed articles that cite Loftus (1978) in 5-year intervals

the difference between various measurement scales (e.g., nominal, ordinal, and interval scales), a distinction that is central to the interpretive problems discussed in this article.

To investigate this issue more thoroughly, we then turned to more advanced textbooks (Abelson, 1995; Carlson & Thorne, 1997; Dobson, 1990; J. O. Dunn & Clark, 1987; Edwards, 1979; Hair, Black, Babin, & Anderson, 2010; Howell, 1992; Krzanowski, 1990; Neter et al., 1990; Sachs, 1982; Sprent, 1998; Stevens, 2002). We found that some of these books discuss the difference between ordinal and disordinal interactions in terms of their shape when plotted; few books, however, explicitly mention the fact that one type of interaction allows stronger, more general conclusions than the other. The exceptions are Neter, Wasserman, and Kutner (1990, pp. 688–691), Sprent (1998, pp. 193–203), and Abelson (1995, pp. 117–118), books that briefly address the issue under consideration.

#### Questionnaire for graduate students and faculty

The lack of coverage in introductory textbooks suggests once more that students and faculty members may not know about the existence of removable interactions. To quantify this more directly, we created a questionnaire in which three hypothetical experiments were described—one for each different type of interaction.<sup>5</sup>

Each questionnaire featured three  $2 \times 2$  factorial designs, one with a nonremovable interaction, one with a borderline nonremovable interaction, and one with a removable interaction. After reading a cover story about the data, each participant was confronted with a line plot of the data; in the plot, data points were surrounded by small confidence

<sup>5</sup> The entire questionnaire, including the cover stories, is available at <http://www.ejwagenmakers.com/misc/Loftus/Loftus.html>. The questionnaire can also be experienced first-hand at <http://www.surveymonkey.com/s/WBVJJVG>, <http://www.surveymonkey.com/s/WBVHGBZ>, or <http://www.surveymonkey.com/s/WBKBBSS>.

intervals so as to prevent any uncertainty about the statistical evidence for the crucial comparisons. Next, a statement about the data was presented, always framed in terms of an underlying psychological process, never in terms of the dependent variable that was plotted. The statement postulated the existence of an interaction, and participants had to indicate their level of agreement on a 5-point Likert scale (1 = *totally disagree*, 5 = *totally agree*). Finally, participants had to briefly explain their answers.

Figure 10 shows an example item for a removable interaction.<sup>6</sup> The associated cover story and research statement were as follows:

Dr. Doyle conducted an experiment on age differences in long-term memory. His experiment featured a group of young adults and a group of old adults. Participants had to read a list of words and recall it later. Long-term memory was estimated by the proportion of words recalled. Every individual participated in two conditions: the short study-test interval condition and the long study-test interval condition. The results are summarized in Figure 1. The interaction is statistically significant ( $p < 0.001$ ). Based on the results, Dr. Doyle concluded:

“An increase in study-test interval affects long-term memory of young adults more than it affects that of older adults.”

Do you agree with Dr. Doyle?

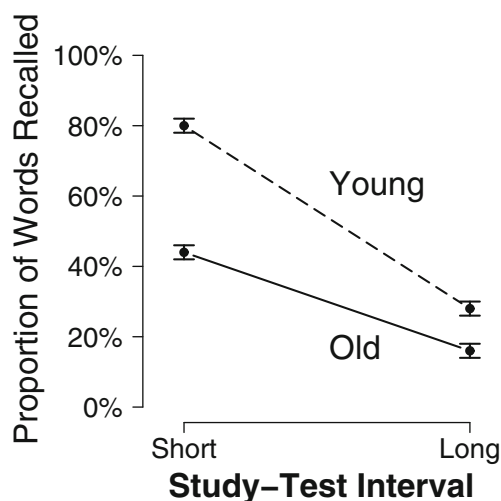
Note that we did not explicitly ask about the distinction between nonremovable and removable interactions. Instead, the goal of the questionnaire was to mimic the process by which researchers draw conclusions from interactions. In the case of the Dr. Doyle cover story, for instance, some students and faculty members may intuit that the conclusion is too general, even though they do not know or remember the distinction between nonremovable and removable interactions.

A total of 100 participants completed the questionnaire. Among the participants were 37 Master’s students, 36 PhD students, and 19 professors. Participants were asked to fill out the questionnaire at the psychology department of the University of Amsterdam and at a seminar on formal models in psychology.<sup>7</sup>

The results of the questionnaire are twofold. First, students and faculty members tended to agree with the

<sup>6</sup> In order to counterbalance the different situations across the types of interactions, the Dr. Doyle cover story was sometimes presented in the context of a nonremovable interaction, sometimes in the context of a borderline nonremovable interaction, and sometimes in the context of a removable interaction.

<sup>7</sup> Descriptive statistics on participants’ educational level and specialization are shown in an online appendix available at <http://www.ejwagenmakers.com/misc/Loftus/Loftus.html>.



**Fig. 10** Example item from a questionnaire that tests knowledge of removable interactions. After reading a cover story, participants were confronted with this figure and had to indicate their level of agreement with the statement “An increase in study–test interval affects long-term memory of young adults more than it affects that of older adults”

statement that an interaction is present, and this tendency holds across all types of interactions. A comparison of the middle and right panel of Fig. 11 shows that students and faculty members agreed that a removable interaction was present about as often as they agreed that a borderline nonremovable interaction was present. The left panel of Fig. 11 shows that students and faculty members agreed most often with the statement that a nonremovable interaction was present.

Second, a detailed analysis of their answers confirmed that students and faculty members based their decisions primarily on the graphical representation of the data, ignoring the fact that this representation may depend critically on the measurement scale: In their open-ended responses, only four out of 100 participants correctly identified the removable interaction as such.<sup>8</sup>

In sum, the results of the questionnaire confirm that psychology students and faculty members are not generally aware of the difference between nonremovable and removable interactions. Instead, students and faculty members appear to interpret interactions by eye, thereby ignoring the possibility that the graphical representation is qualitatively dependent on the measurement scale.

#### Literature review

To demonstrate that removable interactions go undetected in scientific practice, we reviewed all 88 articles published in the 2008 volume of *Psychology and Aging*. We selected

*Psychology and Aging* because we expected the field of aging research to contain a relatively large amount of interactions; it is unlikely that young and older adults will react to a particular experimental manipulation in exactly the same way.

In the 88 articles published in the 2008 issue of *Psychology and Aging*, we found a total of 66 significant  $2 \times 2$  interactions. Each of these interactions was plotted and categorized in one of six categories: (a) nonremovable; (b) removable; (c) borderline nonremovable; (d) nonremovable vague; (e) removable vague; and (f) borderline nonremovable vague. Categories 4–6 are “vague” in the sense that, strictly speaking, the interactions could not be classified because not all post hoc tests were reported. This was true for 31 out of 66 interactions. In these cases, we based our assessment on visual inspection.<sup>9</sup>

The left panel of Fig. 12 shows the frequencies of interactions that could be unambiguously classified; most interactions fall in the category “borderline nonremovable.” The right panel of Fig. 12 shows the frequencies of interactions for which classification had to be based on visual inspection. Most of these interactions are either “borderline nonremovable” or “removable.” It is interesting that in almost half of the cases, too little statistical information was provided in order to decide whether an interaction was nonremovable, borderline nonremovable, or removable. This again suggests that researchers do not recognize the importance of the distinction.

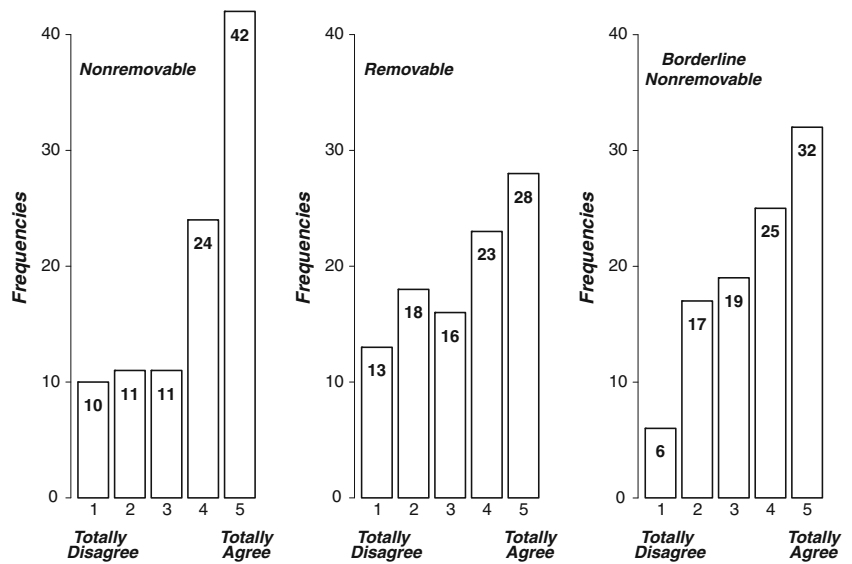
In general, Fig. 12 suggests that the majority of interactions reported in the 2008 issue of *Psychology and Aging* should be interpreted with caution: Out of 66 significant interactions, only four (i.e., 6%) were classified as nonremovable. But how cautious are the authors when it comes to the interpretation of their interactions? To address this question, we examined the results and conclusion sections of the relevant articles. Unfortunately, it proved difficult to arrive at a fair classification of authors’ interpretations; often, a particular interaction is interpreted not in isolation, but in relation to a substantial body of other findings. However, none of the articles referred to Loftus (1978). The difficulties that we encountered in classifying the interpretations are illustrated in the next example.

*Interpretation of a removable interaction* In an experiment on recall and aging, Henkel (2008) discussed data that we replotted in Fig. 13. The author summarized these data as follows (p. 256): “As in Experiment 1, both age groups showed significant increases in recall from Test 1 to Test 3, and the increases were larger for young adults than for older adults.”

<sup>8</sup> A detailed classification of participants’ responses is available at <http://www.ejwagenmakers.com/misc/Loftus/Loftus.html>.

<sup>9</sup> Plots of all 66 interactions are available at <http://www.ejwagenmakers.com/misc/Loftus/Loftus.html>.

**Fig. 11** Students and faculty members in psychology generally agree with the statement that synthetic data show an interaction, even when this statement is formulated in terms of a latent psychological process. Note: Out of 100 participants, two did not answer the question about the nonremovable interaction, and one did not answer the question about the borderline nonremovable interaction



From this description and the figure, we can conclude that the interaction is removable, meaning that it is tied to the particular measurement scale (i.e., proportion of items recalled). Nevertheless, it appears that the author interpreted the removable interaction in terms of the latent psychological construct of recall ability. In the author’s defense, it could be argued that the term “recall” was used only as a shorthand for “proportion of items recalled,” not as a generalization that refers to the latent psychological construct. This could well be true, but it serves only to highlight the ease with which one can shift from a conclusion that is true but specific (i.e., a conclusion in terms of proportion of items recalled) to a conclusion that is probably false but general (i.e., a conclusion in terms of recall ability). Because experimental psychologists seek conclusions that are general and therefore independent of

the measurement scale, it becomes even more tempting to adopt the interpretation that is not strictly supported by the data.

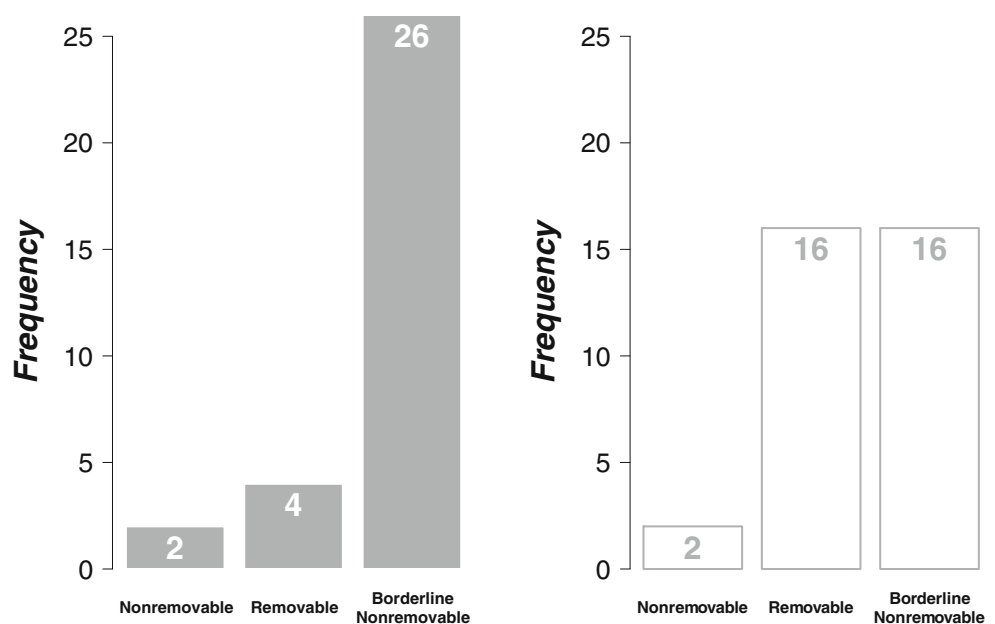
In sum, our questionnaire and our literature reviews suggest that researchers in experimental psychology currently do not acknowledge the existence of removable interactions.

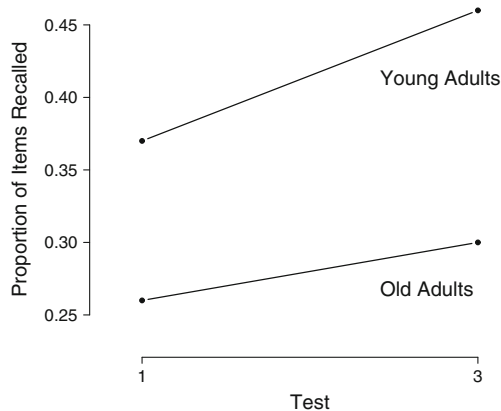
**Objections and alternative approaches**

In the course of developing and presenting the present research, we have encountered a range of objections and opinions, some of which we briefly discuss here:

1. “If we have to worry about monotonic non-linear transformations of the dependent variable, this could

**Fig. 12** Frequencies of nonremovable, removable, and borderline nonremovable interactions reported in the 2008 issue of *Psychology and Aging*. The left panel contains the interactions for which post hoc tests allowed a definite classification. The right panel contains the interactions for which post hoc tests were not reported and classification had to be based on visual inspection instead





**Fig. 13** Example of a removable interaction that is interpreted in terms of increases in recall (Henkel, 2008)

change many things in an analysis, not just whether an interaction is nonremovable.”

This statement is true, but it is not a valid reason to ignore the problem. From observed patterns of interactions or additivity, experimental psychologists often draw general, substantive conclusions about unobserved cognitive processes. It is therefore good to realize that a monotonic transformation may radically change the substantive conclusions, as illustrated in Figs. 3 and 5.

2. “Transformations should not be carried out randomly: We should choose transformations that stabilize the variance and make the data suitable for ANOVA modeling.”

This is a statistical solution, and it ignores the particular problems that arise when researchers want to draw conclusions that are independent of the measurement scale. Yes, certain transformations may make the data adhere to the assumptions of ANOVA modeling, but this does not make these transformations psychologically meaningful. In particular, there is no guarantee that the chosen scale is privileged such that it—and no other scale—allows general conclusions about unobserved psychological processes. For example, one may apply a logit transformation to proportion correct and find an additive pattern of results; this does not mean that patterns based on other plausible scales (e.g., the interactive pattern for the associated drift rates in a diffusion model) are somehow less valuable. In general, statistical models such as ANOVA may not be particularly appropriate to describe the psychological processes that drive behavior.

3. “Transformations should not be carried out randomly: We should choose transformations that are based on process models to provide a meaningful and accurate characterization of the dependent variable.”

This is in fact the approach that we have chosen here to illustrate the problem in the first place (Figs. 3 and

5). However, the large majority of researchers do not base their conclusions on process models, but rather on the output of ANOVA. It should also be noted that even if one uses a process model that fits the data well, removable interactions can still be transformed away by considering a different level of description (see point 4 below).

4. “Transformations should not be carried out randomly: researchers should carefully select and calibrate the dependent variable to have the proper scale.”

This sounds admirable, but what exactly is the proper scale? Perhaps, in the diffusion model example (Figs. 3 and 5), one can argue that drift rate is the proper scale, or at least more proper than mean RT. But, then again, drift rates may be generated by the accumulation of neural spike trains, which themselves may be influenced by neural changes in electrical or chemical activity, perhaps brought about by changes in neurotransmitters. All these scales are proper in the sense that they correspond to a well-defined computational or physiological process that is ultimately responsible for the observed data. Thus, the problem is not to find a single proper scale—it may not exist—but to realize that conclusions that critically depend on the scale of measurement cannot be generalized to other scales.

Thus, we acknowledge that there are several ways to transform data; for our point, however, this is irrelevant. Almost never do we know the transformation between a specific dependent variable and a latent process of interest, although process models such as the diffusion model allow us to specify at least one such transformation. When the data are not transformed, current practice is based on the implicit assumption that the relation between unobserved process and observed behavior is linear, which is almost certainly false. In the absence of a specific process model, the safest conclusion is that interactions that do not cross over cannot be interpreted in terms of underlying psychological constructs. Even when the data have been analyzed with the help of a process model, the safe conclusion is to acknowledge that the interpretation of the data is specific to the process model used.

In sum, we believe that Loftus’ summary is as relevant as it was 33 years ago. The key realization is that sweeping conclusions, such as those about unobserved psychological processes, are warranted only for a privileged subset of interactions.

## General discussion

Mathematical psychologists have long known that there is more to an interaction than meets the eye (Krantz & Tversky, 1971; Loftus, 1978): Interactions that do not cross

over are removable and should be interpreted with caution. In the present article, we have explained why the concept of removable interactions is important to experimental psychologists: In contrast with statisticians, experimental psychologists are often interested not in the dependent variable itself, but in what that variable reveals about the underlying psychological process. When we allow for the possibility that the dependent variable is monotonically but nonlinearly related to the psychological process of interest, all interactions that do not cross over can be transformed away. Concretely, this means that an interaction in terms of MRT, for example, can turn out to be additive in terms of a construct such as drift rate in a diffusion model (Ratcliff, 1978).

Our analysis showed that few experimental psychologists may be aware that interactions can be removable, and this means that experimental effects may easily lead to conclusions about psychological processes that are incorrect or misleading.

### Remedies

In order to improve the current situation, we suggest several courses of action. First, students in experimental psychology can learn to distinguish not only additive effects from those that interact, but also nonremovable from removable interactions. In other words, the difference between nonremovable and removable interactions should be included in the academic curriculum. Second, more research effort and attention can be devoted to the various techniques that allow general conclusions under changes of the measurement scale (e.g., Bamber, 1979; J. C. Dunn & James, 2003; Hoijtink, Klugkist, & Boelen, 2008; Iverson & Falmagne, 1985; Luce & Tukey, 1964; Pettitt, 1982; Wagenmakers, Molenaar, Grasman, Hartelman, & van der Maas, 2005). Third, experimental psychology as a field can encourage the use of quantitative models; these models bridge the gap between observed data and latent psychological processes and thereby promote a better understanding of how the two relate. At the very least, quantitative models promote the understanding that the relationship between unobserved process and observed behavior is often nonlinear. Fourth, experimental psychologists should be more careful when interpreting interactions (including null interactions—i.e., additivity) and explicitly indicate when these interactions can be removed by monotonic transformations of the measurement scale.

The aforementioned courses of action may be laudable but perhaps they are not of immediate benefit to the researcher who seeks to draw strong conclusions from a set of removable interactions. We now mention several experimental procedures that one can use to probe the true nature of an interaction.<sup>10</sup> One procedure is known as matching (Anderson, 1963; Bitterman, 1975), and it seeks to

vary an experimental factor so that different groups of participants have approximately the same score on the dependent variable of interest. For instance, to study whether older people forget words more quickly than younger people, one might first determine the amount of study time that each participant needs in order to recall, say, 50% of the study list after 1 hr. In the second stage of the experiment, one uses the individual study times and varies the retention interval. This procedure ensures that the forgetting curves of older people and young people overlap initially, and this greatly increases the opportunity to observe (borderline) nonremovable interactions. Another, less controlled method is to use a range of different study times and create matching samples after data collection has concluded.

Another strategy to study the robustness of the conclusions by using several plausible transformations; for RT, these transformations include  $\log(\text{RT})$  and  $1/\text{RT}$ ; for proportion correct, these include  $\logit(p)$  and  $d'$ ; for RT and proportion correct jointly, one could use the drift rate parameter from the diffusion model (e.g., Wagenmakers et al., 2007). Finding the same pattern across a wide range of transformations increases one's confidence in the robustness of the conclusions. In the same vein, one can examine an interaction for subsets of participants or items that vary along the scale—for instance, one could confirm that a particular interaction pattern holds for groups of high, medium, and low performance.

### Assumptions, models, and interpretations

Our analysis rests on the assumption that the measurement scale is ordinal—that is, that there exists some monotonic transformation between the observed dependent variable and a unidimensional unobserved psychological process of interest (Anderson, 1961). We believe that this assumption is plausible, or at least considerably more plausible than the assumption that the transformation is linear. It should be clear that the assumption of linearity is the statistical model that researchers tacitly adopt whenever they interpret removable interactions as if they had been nonremovable. Thus, it is a mistake to blame psychological process models for creating the interpretive problem and to argue that as long as we refrain from using models and focus on the data, all is well. Data analysis and interpretation simply cannot proceed in the complete absence of a model.

For example, in the context of the analysis of retention functions, Slamecka (1985, p. 813) argued that "...no psychological theory or cognitive model is a precondition for investigating forgetting.... In our own work, we did not speculate about underlying processes." However, Bogartz (1990a, p. 142) believed that the term "forgetting" refers to a psychological process and a theory, "however informal and unarticulated." Also, Paul (1994, p. 993) pointed out

<sup>10</sup> We thank the reviewers for pointing out these options.



that "... any comparison of forgetting between experimental conditions usually requires some rudimentary form of modeling" (see also Anderson, 1963; Laming, 1992, p. 1343; Loftus, 1985a, b; Wixted, 1990). In the case of Slamecka and McElree (1983), the rudimentary model posits a linear relation between the psychological process of forgetting and the delay-induced reduction in the number of items recalled.

The fact of the matter is that removable interactions are model dependent, and this remains true when one commits, implicitly or explicitly, to a model that postulates a linear mapping between the dependent variable and the psychological process. The extent to which the interaction is meaningful or interesting naturally depends on the veracity of the model: If the model is clearly false, the interpretation of the data is also suspect. However, the problem is persistent: Even when one uses a model that is (approximately) correct, removable interactions allow an interpretation only in terms of the parameters of the model. As mentioned above, the fact that drift rates in the diffusion model are additive could be used to argue that a particular experimental manipulation affects two groups of participants equally (e.g., Fig. 5); however, this claim is valid only on the level of drift rates. When a more detailed model is constructed in which drift rate is a nonlinear function of neural firing rates, it is likely that the claim needs to be revised, and one needs to conclude that—on the level of neural firing rates—the manipulation affects two groups of participants differentially.

A reviewer remarked that, regardless of our classification of interactions, it is useful to know, for instance, that older people forget more of what they learned over a week than younger people. We agree, and we have no problem whatsoever with the interpretation of interactions on the level of the observed data. The problem occurs when the interaction for the observed data is thought to hold also for the underlying psychological process.

## Conclusion

In sum, the concept of interactions in a  $2 \times 2$  design is considerably more complicated than is suggested by the textbook definition that "the lines are not parallel." And, even though it is tempting to use the eye to interpret interactions that do not cross, it is better to resist temptation and acknowledge that general claims about psychological processes might require designs more intricate than those afforded by approaches that are currently standard.

**Author Note** The present research was supported by a Vidi grant from the Dutch Organization for Scientific Research (to E.J.W.) and National Science Foundation Grant 0951612 (to A.H.C.). A-M. K. is a scholar of the Alexander S. Onassis Public Benefit Foundation. We

thank Rene Zeelenberg, Doug Hintzman, Geoff Loftus, and Roddy Roediger for helpful comments on an earlier draft of this article. Correspondence concerning this article may be addressed to E.-J. Wagenmakers, University of Amsterdam, Department of Psychology, Roetersstraat 15, 1018 WB Amsterdam, the Netherlands.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

## References

- Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale: Erlbaum.
- Agresti, A., & Finlay, B. (2009). *Statistical methods for the social sciences*. New York: Pearson.
- Anderson, N. H. (1961). Scales and statistics: Parametric and nonparametric. *Psychological Bulletin*, *58*, 305–316.
- Anderson, N. H. (1963). Comparison of different populations: Resistance to extinction and transfer. *Psychological Review*, *70*, 162–179.
- Aron, A., Aron, E., & Coups, E. (2006). *Statistics for psychology* (4th ed.). New York: Pearson.
- Bamber, D. (1979). State-trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology*, *19*, 137–181.
- Bitterman, M. E. (1975). The comparative analysis of learning. *Science*, *188*, 699–709.
- Bluman, G. B. (2007). *Elementary statistics. A step by step approach* (6th ed.). New York: McGraw-Hill.
- Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced choice tasks. *Psychological Review*, *113*, 700–765.
- Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U., & Nieuwenhuis, S. (2010). The neural basis of the speed–accuracy tradeoff. *Trends in Neurosciences*, *33*, 10–16.
- Bogartz, R. S. (1976). On the meaning of statistical interactions. *Journal of Experimental Child Psychology*, *22*, 178–183.
- Bogartz, R. S. (1990a). Evaluating forgetting curves psychologically. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 138–148.
- Bogartz, R. S. (1990b). Learning–forgetting rate independence defined by forgetting function parameters or forgetting function form: Reply to Loftus and Bamber and to Wixted. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 936–945.
- Busemeyer, J. R. (1980). Importance of measurement theory, error theory, and experimental design for testing the significance of interactions. *Psychological Bulletin*, *88*, 237–244.
- Carlson, W. L., & Thome, B. (1997). *Applied statistical methods for business, economics and social sciences*. New Jersey: Prentice Hall.
- Cox, D. R. (1984). Interaction. *International Statistical Review*, *52*, 1–31.
- Criss, H. A. (2010). Differentiation and response bias in episodic memory: Evidence from reaction time distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 484–499.
- De González, A. M., & Cox, D. R. (2007). Interpretation of interaction: A review. *The Annals of Applied Statistics*, *1*, 371–385.
- Dobson, A. J. (1990). *An introduction to generalized linear models*. Great Britain: Chapman and Hall.
- Dunn, D. S. (2000). *Statistics and data analysis for the behavioral sciences*. New York: McGraw-Hill.
- Dunn, J. C., & James, R. N. (2003). Signed difference analysis: Theory and application. *Journal of Mathematical Psychology*, *47*, 389–416.
- Dunn, J. O., & Clark, V. A. (1987). *Applied statistics: Analysis of variance and regression* (2nd ed.). New York: Wiley.

- Edwards, A. L. (1979). *Multiple regression and the analysis of variance and covariance*. San Francisco: W. H. Freeman and Company.
- Everitt, D. S. (2001). *Statistics for psychologists. An intermediate course*. Mahwah: Erlbaum.
- Field. (2005). *Discovering statistics using SPSS* (2nd ed.). London: Sage.
- Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., von Cramon, D. Y., Ridderinkhof, K. R., et al. (2008). Striatum and pre-SMA facilitate decision-making under time pressure. *Proceedings of the National Academy of Sciences*, *105*, 17538–17542.
- Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*, 439–453.
- Goodwin, C. J. (1997). *Research in psychology: Methods and design*. New York: Wiley.
- Greene, J., & D'Oliveira, M. (1999). *Learning to use statistical tests in psychology*. Buckingham-Philadelphia: Open University Press.
- Hair, J. F. J., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis*. Upper Saddle River: Pearson.
- Henkel, L. A. (2008). Maximizing the benefits and minimizing the costs of repeated memory tests for older adults. *Psychology and Aging*, *23*, 250–262.
- Hojtink, H., Klugkist, I., & Boelen, P. (2008). *Bayesian evaluation of informative hypotheses that are of practical value for social scientists*. New York: Springer.
- Howell, D. C. (1992). *Statistical methods for psychology* (3rd ed.). Belmont: Duxbury.
- Howitt, D., & Cramer, D. (2008). *Introduction to statistics in psychology* (4th ed.). Harlow: Prentice Hall.
- Iverson, G., & Falmagne, J.-C. (1985). Statistical issues in measurement. *Mathematical Social Sciences*, *10*, 131–153.
- Krantz, D. H., & Tversky, A. (1971). Conjoint-measurement analysis of composition rules in psychology. *Psychological Review*, *78*, 151–169.
- Krzanowski, W. J. (1990). *Principles of multivariate analysis: A user's perspective*. Oxford: Oxford Science Publications.
- Laming, D. (1992). Analysis of short-term retention: Models for Brown-Peterson experiments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 1342–1365.
- Lawrence, M. P. (1994). Making interpretable forgetting comparisons: Explicit versus hidden assumptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 992–999.
- Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition*, *6*, 312–319.
- Loftus, G. R. (1985a). Consistency and confoundings: Reply to Slamecka. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 817–820.
- Loftus, G. R. (1985b). Evaluating forgetting curves. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 397–406.
- Loftus, G. R. (2002). Analysis, interpretation, and visual presentation of experimental data. In H. Pashler & J. Wixted (Eds.), *Stevens' handbook of experimental psychology: Vol. 4. Methodology in experimental psychology* (3rd ed., pp. 339–390). New York: Wiley.
- Loftus, G. R., & Bamber, D. (1990). Learning–forgetting independence, unidimensional memory models, and feature models: Comment on Bogartz (1990). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 916–926.
- Luce, R. D. (1986). *Response times*. New York: Oxford University Press.
- Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*, 1–27.
- Moore, D., & McCabe, G. (2006). *Introduction to the practice of statistics*. New York: Freeman.
- Neter, J., Wasserman, W., & Kutner, M. H. (1990). *Applied linear statistical models* (3rd ed.). Homewood: Richard D. Irwin.
- Nolan, S. A., & Heinzen, T. E. (2007). *Statistics for the behavioral sciences*. New York: Worth Publishers.
- Pagano, R. R. (1998). *Understanding statistics in the behavioral sciences*. Belmont: Brooks/Cole Publishing Company.
- Paul, L. M. (1994). Making interpretable forgetting comparisons: Explicit versus hidden assumptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 992–999.
- Pettitt, A. N. (1982). Inference for the linear model using a likelihood based on ranks. *Journal of the Royal Statistical Society Series B*, *44*, 234–243.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.
- Ratcliff, R. (1981). A theory of order relations in perceptual matching. *Psychological Review*, *88*, 552–572.
- Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review*, *9*, 278–291.
- Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922.
- Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*, 347–356.
- Ratcliff, R., & Rouder, J. N. (2000). A diffusion model account of masking in two-choice letter identification. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 127–140.
- Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004a). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and Aging*, *19*, 278–289.
- Ratcliff, R., Thapar, A., & McKoon, G. (2001). The effects of aging on reaction time in a signal detection task. *Psychology and Aging*, *16*, 323–341.
- Ratcliff, R., Thapar, A., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on brightness discrimination. *Perception & Psychophysics*, *65*, 523–535.
- Ratcliff, R., Thapar, A., & McKoon, G. (2004b). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language*, *50*, 408–424.
- Ratcliff, R., Thapar, A., & McKoon, G. (2006a). Aging and individual differences in rapid two-choice decisions. *Psychonomic Bulletin & Review*, *13*, 626–635.
- Ratcliff, R., Thapar, A., & McKoon, G. (2006b). Aging, practice, and perceptual tasks: A diffusion model analysis. *Psychology and Aging*, *21*, 353–371.
- Ratcliff, R., Thapar, A., & McKoon, G. (2007). Application of the diffusion model to two-choice tasks for adults 75–90 years old. *Psychology and Aging*, *22*, 56–66.
- Ratcliff, R., Thapar, A., & McKoon, G. (2010). Individual differences, aging, and iq in two-choice tasks. *Cognitive Psychology*, *60*, 127–157.
- Ratcliff, R., Thapar, A., Smith, P. L., & McKoon, G. (2005). Aging and response times: A comparison of sequential sampling models. In J. Duncan, P. McLeod, & L. Phillips (Eds.), *Measuring the mind: Speed, control, and age* (pp. 3–32). Oxford: Oxford University Press.
- Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *102*, 261–300.
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.
- Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, *9*, 487–494.

- Sachs, L. (1982). *Applied statistics. A handbook of techniques*. New York: Springer.
- Shaughnessy, J. J., & Zechmeister, E. B. (1994). *Research methods in psychology*. Singapore: McGraw-Hill.
- Slamecka, N. J. (1985). On comparing rates of forgetting: Comment on Loftus (1985). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 812–816.
- Slamecka, N. J., & McElree, B. (1983). Normal forgetting of verbal lists as a function of their degree of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*, 384–397.
- Sprent, P. (1998). *Data driven statistical methods*. Great Britain: Thomson Science.
- Starns, J. J., & Ratcliff, R. (2010). The effects of aging on the speed–accuracy compromise: Boundary optimality in the diffusion model. *Psychology and Aging*, *25*, 377–390.
- Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, *30*, 276–315.
- Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences*. Mahwah: Erlbaum.
- Thapar, A., Ratcliff, R., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on letter discrimination. *Psychology and Aging*, *18*, 415–429.
- Wagenmakers, E.-J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, *21*, 641–671.
- Wagenmakers, E.-J., Molenaar, P. C. M., Grasman, R. P. P. P., Hartelman, P. A. I., & van der Maas, H. L. J. (2005). Transformation invariant stochastic catastrophe theory. *Physica D*, *211*, 263–276.
- Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58*, 140–159.
- Wagenmakers, E.-J., van der Maas, H. J. L., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*, 3–22.
- Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t-test. *Psychonomic Bulletin & Review*, *16*, 752–760.
- Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.
- Winner, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.
- Wixted, J. T. (1990). Analyzing the empirical course of forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 927–935.