Running head: META-ANALYSIS OF INTERLEAVED LEARNING

Similarity matters: A meta-analysis of interleaved learning and its moderators

Matthias Brunmair & Tobias Richter

University of Würzburg

Manuscript accepted for publication in *Psychological Bulletin* (2019)

Author Note

This research was funded by a grant from the state of Hessen in the State Offensive for Fostering Economic and Scientific excellence (LOEWE), research initiative "Desirable difficulties in learning". The data files and the R-script used in this meta-analysis are available in the repository of the Open Science Framework (https://osf.io/7u253/).

Address for correspondence:

Matthias Brunmair / Tobias Richter

University of Würzburg

Department of Psychology IV

Röntgenring 10

97070 Würzburg, Germany

E-mail: matthias.brunmair@uni-wuerzburg.de / tobias.richter@uni-wuerzburg.de

Abstract

An interleaved presentation of items (as opposed to a blocked presentation) has been proposed to foster inductive learning (interleaving effect). A meta-analysis of the interleaving effect (based on 59 studies with 238 effect sizes nested in 158 samples) was conducted to quantify the magnitude of the interleaving effect, to test its generalizability across different settings and learning materials, and to examine moderators that could augment the theoretical models of interleaved learning. A multilevel meta-analysis revealed a moderate overall interleaving effect (Hedges' $g = 0.42$). Interleaved practice was best for studies using paintings ($g = 0.67$) and other visual materials. Results for studies using mathematical tasks revealed a small interleaving effect ($g = 0.34$), whereas results for expository texts and tastes were ambiguous with nonsignificant overall effects. An advantage of blocking compared to interleaving was found for studies based on words ($g = -0.39$). A multiple meta-regression analysis revealed stronger interleaving effects for learning material more similar between categories, for learning material less similar within categories, and for more complex learning material. These results are consistent with the theoretical account of interleaved learning, most notably with the sequential theory of attention (attentional bias framework). We conclude that interleaving can effectively foster inductive learning but that the setting and the type of learning material must be considered. The interleaved learning, however, should be used with caution in certain conditions, especially for expository texts and words.

Key words: blocking, category learning, discrimination, inductive learning, interleaving effect

Public significance statement

In inductive learning, concepts are acquired by studying exemplars (e.g., mathematical procedures used for solving math problems, biological species by studying photographs of animals, or psychological disorders by reading case reports). Research suggests that interleaving exemplars from different categories (in contrast to presenting exemplars from the same category in a blocked fashion) might benefit inductive learning. This study presents a quantitative synthesis of extant studies of interleaving. We found a moderate positive effect of interleaving, but the benefits of interleaving depend on the type of learning material and setting. For educational applications, interleaving is promising but setting and learning material should be evaluated.

Inductive learning is a major way that humans acquire knowledge. Conceptual knowledge (e.g., rules, principles or facts associated with a concept) may be learned directly, for example, from a teacher or a text that conveys this information. In contrast, inductive learning refers to the acquisition of concepts based on observing exemplars. Inductive learning is a pervasive form of learning because it does not require formal instruction and can occur in a broad range of situations, from babies learning new words to doctors classifying x-rays (Kornell & Bjork, 2008). Accordingly, inductive learning is an important research topic in such diverse areas such as cognitive psychology and neuroscience (e.g., Ashby & Maddox, 2005, 2011; Holland, Holyoak, Nisbett, & Thagard, 1986), social psychology (e.g., Fiedler, 2000), educational psychology (e.g. Sana, Yan, & Kim, 2017), and developmental psychology (Poulin-Dubois & Pauen, 2017). Inductive learning is also the major type of learning in artificial intelligence approaches (machine learning, Michalski, 1986).

Given its broad range of applicability, the variety of subject matter in studies on inductive learning is not surprising. Experimental studies on inductive learning have included, for example, artificial categories or biological species that are learned by looking at pictures of exemplars (Birnbaum, Kornell, Bjork, & Bjork, 2013; Higgins & Ross, 2011; Lavis & Mitchell, 2006; Mitchell, Kadib, Nash, Lavis, & Hall, 2008; Wahlheim, Dunlosky, & Jacoby, 2011), the painting styles of artists that are learned by looking at paintings from these artists (Metcalfe & Xu, 2016; Kang & Pashler, 2012; Kornell, Castel, Eich, & Bjork, 2010; Verkoeijen & Bouwmeester, 2014; Zulkiply, 2015; Zulkiply & Burt, 2013a, 2013b), mathematical procedures that are acquired by solving math tasks (Higgins & Ross, 2011; Rau, Aleven, & Rummel, 2010, 2013; Rohrer & Taylor, 2007; Sana, Yan, & Kim 2017; Taylor & Rohrer, 2010), or psychological disorders that are learned by reading case studies (Zulkiply & Burt, 2013a). In some cases, categories are easier to distinguish than in others. The difficulty depends on, among other factors, the variation of features within and between categories

(Carvalho & Goldstone, 2014a, 2015a; Zulkiply & Burt, 2013b), which mainly vary as a function of the learning material and learner characteristics. These conditions are therefore difficult to change, but the positional relations of items are often easy to manipulate. One way to sequence the items during learning is to either present all exemplars of one category in a single block or interleave exemplars of different categories. Several studies have suggested that interleaving might be beneficial for inductive learning compared to a blocked presentation of items (Higgins & Ross, 2011; Kang & Pashler, 2012; Kornell & Bjork, 2008; Kornell et al. 2010; Rohrer & Taylor, 2007; Wahlheim et al., 2011). However, other studies have found no positive effect of interleaving (Carpenter & Mueller, 2013; Dobson, 2011; Higgins & Ross, 2011; Rau et al., 2010; Sorensen & Woltz, 2016), suggesting that the effect might depend on certain conditions.

The purpose of the present meta-analytic review was to systematically investigate the generalizability of the interleaving effect across different types of learning materials, characteristics of these learning materials, other aspects of the learning setting, and populations of learners. The theoretical perspective guiding this research was based on theoretical accounts of the interleaving effect (Carvalho & Goldstone, 2015b, 2017; Kang & Pashler 2012; Kornell and Bjork, 2008). We examined potential moderators that, according to these theories, might affect the magnitude of an interleaving effect or even cause the reverse effect, that is, an advantage of blocking over interleaving. The practical impetus of this research was to explore the conditions and extent that interleaving could also be used for designing inductive learning in school-like settings and with realistic educational materials.

Interleaving During Learning: An Illustration and a Definition

One way to manipulate positional relations during inductive learning is to space items temporally. Many studies have shown benefits of spacing between learning phases for memory and skill acquisition of various items in repetitive learning (spacing effect/distributed practice effect, Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Donovan & Radosevich, 1999;

Janiszewski, Noel, & Sawyer, 2003; Lee & Genovese, 1988). The spacing effect commonly refers to learning with materials in which items belong to the same categories and thus does not require the acquisition of category structures. Moreover, in many studies on the spacing effect, learning is repetitive such that exactly the same items (e.g., the same lists of words) are repeated in either a massed or a spaced fashion. In interleaved learning, different items representing two or more categories are presented, and the items from one category are spaced and interleaved with the presentation of items from other categories. Kornell and Bjork (2008) implemented a typical design of this interleaved presentation of items with impressionistic paintings of landscapes by 12 artists. In the learning phase, six paintings per artists were presented. The interleaved and massed condition were manipulated within participants. The paintings of six artists were presented blocked (M), that is, six paintings by the same artist were presented in one block. The paintings of six other artists were presented in an interleaved fashion, presented in blocks consisting of six paintings from different artists (S). Overall, the learning phase consisted of 12 blocks that alternated between a blocked and an interleaved presentation (MSSMMSSMMSSM). The different categories were defined by the artists, and the items (exemplars) were the artists' paintings. In the first experiment, each painting was presented for 3 sec and the artists' name was displayed below the painting. In a subsequent test phase, participants were required to assign new paintings to one of the 12 artists. Kornell and Bjork found that new paintings by artists for whom the paintings had been presented in an interleaved fashion during the learning phase were assigned more accurately than new paintings by artists for whom paintings had been presented in a blocked fashion ($\eta_p^2 = .39$). This *interleaving effect* has been replicated many times in experiments based on naturalistic paintings as learning materials with the same set of paintings as in the original study (Metcalfe & Xu, 2016; Kang & Pashler, 2012; Kornell et al., 2010; Verkoeijen & Bouwmeester, 2014; Zulkiply, 2015; Zulkiply, & Burt, 2013a; Zulkiply & Burt, 2013b). Other research demonstrated interleaving effects in other settings and with various types of

materials (e.g., Birnbaum et al., 2013; Higgins & Ross, 2011; Lavis & Mitchell, 2006; Rohrer & Taylor, 2007; Wahlheim et al., 2011), but other studies have reported no interleaving effect (e.g., Carpenter & Mueller, 2013; Dobson, 2011; Higgins & Ross, 2011; Rau et al., 2010; Sorensen & Woltz, 2016). Thus, a systematic quantitative review of the available research that also looks at potential moderating factors of the interleaving effect seems warranted.

For the purpose of this review, we define the *interleaving effect* as a positive effect of an interleaved compared to a blocked inductive learning condition on the performance in a subsequent category discrimination or classification task. In the interleaved condition, category items are presented in an interleaved fashion with items of at least one other category, whereas in the blocked condition all items of one category are presented consecutively before the items of another category are presented. Furthermore, performance in the criterion test is contingent on inductive learning. The counterpart of the interleaving effect (i.e., a negative effect of an interleaved compared to blocked inductive learning condition) will be called *blocking effect*.

This definition of the interleaving effect (and the blocking effect) excludes some studies, for example, Duggan and Payne (2001) who interleaved reading with acting as two modes of learning an interactive procedure from written instructions. Although Duggan and Payne investigated effects of interleaving on memory, the learning task required neither discrimination between categories nor inductive learning. For the same reason, studies on the spacing effect are not included, although items in spacing studies are often interleaved with other items.

For similar reasons, studies examining contextual interference were also excluded, unless they fulfilled the above definition of studies examining the interleaving effect. The contextual interference effect refers to the finding that motor skills are better learned (especially in the long term) when they are practiced in an interleaved instead of a blocked fashion (Brady, 2004; Magill & Hall, 1990). The interleaved practice is assumed to create

interference during learning, hence the name contextual interference effect. The contextual interference effect resembles the interleaving effect and both effects are often discussed in the same context (e.g., Bjork & Bjork, 2011). However, the focus of most studies on contextual inference is on motor learning for such skills used in basketball shots or rotary pursuit tasks. The practice of motor skills is sometimes regarded as a kind of inductive learning because learners use, among other processes, their proprioceptive feedback for practicing the skill, and the movements carried out in each practice trial overlap but are not identical (e.g., Kornell & Bjork, 2008). Despite these commonalities, noticeable differences can be observed between the inductive learning of categories and practicing motor skills in studies that have investigated interleaved learning. Motor learning is procedural and largely implicit learning, whereas the inductive learning of categories and concepts always involves a declarative component and is mainly explicit, that is, learners are aware that they are learning and what they have learned (for distinguishing features of explicit and implicit learning, see Cleeremans, 1996). Moreover, when learners acquire concepts inductively, one major criterion of successful learning is that learners are able to distinguish between exemplars from different categories and to categorize new stimuli correctly. Discriminating between exemplars from different categories might even be crucial for learning to occur (e.g., see the discriminative contrast hypothesis, Kang & Pashler, 2012, discussed below). None of these features seem to apply readily to contextual interference studies on motor learning. In these studies, the learning criterion is the correct execution of movements of varying complexity. Because of these obvious differences, we chose not to include these studies in the current meta-analysis. However, we included studies that referred to contextual interference in their theoretical argumentation if these studies matched our definition of interleaved learning. In other words, we included studies on contextual interference if these studies compared an interleaved and a blocked inductive learning condition and studied the effects of these conditions on the performance in a subsequent category discrimination or classification task

(e.g., de Croock & van Merriënboer, 2007; Rau, Aleven, & Rummel, 2013; Rau, Rummel, Aleven, Pacilio, & Tunc-Pekkan, 2012).

Key Theories on the Interleaving Effect

One theoretical approach to explain the effects of positional relations in inductive learning is the *discriminative contrast hypothesis* (Kang & Pashler, 2012). The discriminative contrast hypothesis is based on the observation that temporal spacing between items hinders inductive learning in interleaved conditions and has no advantage in blocked conditions. Accordingly, Kang and Pashler suggested that an interleaved presentation (with little or no time between the item presentations) promotes discriminative contrasts of subsequent items, which might benefit inductive learning.

The *attentional bias framework* (Carvalho & Goldstone, 2015b), also called *sequential attention theory* (Carvalho & Goldstone, 2017), refines the discriminative contrast hypothesis by adding assumptions about the conditions that render discriminative contrasts useful. This framework proposes that an interleaved presentation highlights differences between items, whereas a blocked presentation highlights similarities. Therefore, interleaving is assumed to be better for inductive learning when the differences between exemplars belonging to different categories are crucial for acquiring the category structure. This interleaving effect occurs in conditions of low-discriminability categories in which all exemplars presented during learning are highly similar to each other. Learning the category structure in this condition depends on finding the differences that discriminate between categories. In contrast, blocking is assumed to be better when the similarities of exemplars belonging to the same category are crucial. This blocking works in high-discriminability categories in which all exemplars are very different. Learning the category structure in this condition depends on finding the features that are similar or identical for all exemplars belonging to one category.

Another theoretical approach is the massing-aids-induction hypothesis. The term has been coined by Kornell and Bjork (2008; see also Kornell et al., 2010) who do not advocate

the hypothesis. According to the massing-aids-induction hypothesis, blocking enhances inductive learning by facilitating the recognition of similarities among nonidentical items within a category. The massing-aids-induction hypothesis cannot explain positive interleaving effects (such as those reviewed by Carvalho & Goldstone, 2015b) but addresses the advantage of blocked over interleaved presentation (i.e., a blocking effect). Thus, the hypothesis may be helpful to interpret the results from studies in which interleaving failed. For example, interleaving leads to poorer learning outcomes than blocking when items of different conceptual categories are presented (Sorensen & Woltz, 2016). In this condition, a specific form of inductive learning is required, which vastly depends on the identification of the shared concept between different words within a category.

Generalizability and Potential Moderators of Interleaving and Blocking Effects

Identifying moderators of the interleaving effect and the extent that the interleaving effect generalizes across different populations of learning materials, settings, and learners is important for determining the scope of the theoretical explanations that have been proposed for the effect. Moreover, investigating the generalizability and moderators of the effect is crucial for the application of interleaved learning in real-world educational settings. In the following section, we provide a brief overview of the potential moderators that potentially have theoretical meaning and were thus included in the meta-analysis (methodological study characteristics considered as potential moderators are described in the Method section).

Types of learning materials. Most often, *visual materials* are used to investigate the interleaving effect. These materials can be naturalistic pictures (paintings or photographs) or pictures of artificial objects. Research has shown that the interleaved presentation of naturalistic pictures benefits learning to discriminate between the styles of different painters (Metcalfe & Xu, 2016; Kang & Pashler, 2012; Kornell & Bjork; 2008; Kornell et al., 2010; Verkoeijen & Bouwmeester, 2014; Zulkiply, 2015; Zulkiply & Burt, 2013a, 2013b) and also naturalistic photos for distinguishing species of birds (Birnbaum et al., 2013; Wahlheim et al.,

2011). However, results are mixed for pictures of artificial objects, which are often used to test boundary conditions of the interleaving effect by manipulating category structures. Typical examples of this type of visual material are the multicolored 20 x 20 grids used by Lavis and Mitchell (2006; see also Mitchell et al., 2008). In this material, different categories are defined by six small, unique and connected elements of approximately five squares. No variations of these elements occur within categories and only slight variations between categories. Another typical example of pictures of artificial objects are the aliens used by Higgins and Ross (2011). In this material, the categories (different species of aliens) are defined by a prototype with six binary features (arms, tail, antennae, legs, eyes, and mouth).

Most studies with pictures of artificial objects demonstrated an interleaving effect (e.g., Higgins & Ross, 2011; Lavis & Mitchell, 2006; Mitchell et al., 2008), but studies have also reported inconsistent and contradictory findings. For example, the interleaving effect was fragile when categories were highly discriminable (Zulkiply & Burt, 2013b), when classification of categories required rule-based learning (Noh, Yan, Bjork, & Maddox, 2016), or when items were studied in passive learning conditions (Carvalho & Goldstone, 2015a).

Results seem to be mixed when *expository texts* were used. For example, Zulkiply and Burt (2013a) showed an interleaving effect when case studies were used for learning to distinguish between different psychopathological disorders. However, Dobson (2011) found no effect for interleaving in a study that interleaved a text about the human immune system with a text about female reproduction.

Results are also mixed for *mathematical tasks*. Positive effects of interleaving were found when categories were types of geometric solids that required the same formula for computing their volumes, and exemplars of geometric solids were the items (Rohrer & Taylor, 2007) or when categories were formulas to calculate faces, edges, corners and angles of prisms, and different prisms were the items (Taylor & Rohrer, 2010). Sana et al. (2017) found an interleaving effect when categories were statistical concepts (Wilcoxon signed-ranks

test, Chi-squared test, Kruskal-Wallis test), and verbal descriptions of different research designs were the items.

Studies have also reported no effect of interleaving on mathematical tasks or blocking effects, that is, findings suggesting that interleaving hinders learning mathematical concepts (Higgins & Ross, 2011; Rau et al., 2010). One explanation for the mixed results might be that mathematical tasks can be interleaved by different dimensions of categories (Rau et al., 2013). When dimensions are orthogonal, one dimension can be blocked while the other is interleaved. For example, Rau et al. (2013) examined interleaved vs. blocked practice on fractions by task type, such as identifying fractions from graphical representations or comparing fractions, and by graphical representation type such as number lines or circles. Performance of fifth and sixth graders were better when task type was interleaved than when representation type was interleaved.

Relatively few studies have used learning materials other than visual materials, expository texts, or mathematical tasks. One study produced mixed results for the discrimination of flavors (Dwyer, Hodder, & Honey, 2004). Blocking effects were found when categories were pronunciation rules and items were French words (Carpenter & Mueller, 2013). Moreover, blocking effects were found when categories were novel concepts defined by multiple features (e.g., is loud, can fly, smells, lives underground, lives in the desert, is green) and items were words like silence, eagle, rose, or cave (Sorensen & Woltz, 2016).

Material characteristics. According to the attentional bias framework, characteristics of the item materials such as within- and between-category similarities moderate the effects of interleaving. In particular, positive interleaving effects seem to be stronger when the different categories are more difficult to discriminate (Carvalho & Goldstone, 2014a, 2015a, 2017; Zulkiply & Burt, 2013b).

Retention interval. Learning success can be tested immediately after the learning phase or after a temporal delay. The retention interval refers to the temporal delay between learning phase and the final test. Few studies have varied retention interval when examining its potential role of moderating the interleaving effect. Studies with different types of materials have not provided evidence for an interaction of retention interval and interleaving (visual material; Carvalho & Goldstone, 2014a; Zulkiply & Burt, 2013a; case studies of psychological disorders; Zulkiply & Burt, 2013a; mathematical task; Rau et al., 2010). Contrary to these results, Rohrer et al. (2014b) found an interaction. The interleaving effect was higher in a 30-day retention interval (Cohen's $d = 0.79$) compared to a one-day retention interval (Cohen's $d = 0.42$).

Retention vs. transfer test. Inductive learning success can be tested with the same items that were also presented during the learning phase (retention test) or with new items that must be assigned to the categories acquired during the learning phase (transfer test). Many studies investigating interleaved learning have provided test results for both retention and transfer tests. For example, Kornell et al. (2010) directly compared the classification performance for previously studied paintings and new paintings of artists for whom a different set of paintings had been presented during the learning phase. They found no interaction between the interleaved vs. blocked condition and retention vs. transfer test.

Successive vs. simultaneous presentation. Items can be interleaved by presenting them successively, that is, one item of a category is followed by another item of a different category (e.g., Kornell & Bjork, 2008), or simultaneously, that is, two (or more) items of different categories are presented together at the same time (e.g., Carvalho & Goldstone, 2014b). Results for these different modes of presentation are unclear. Wahlheim et al. (2011) found that a simultaneous presentation benefitted interleaved learning more than blocked learning, but other studies have found that both learning conditions benefitted from a simultaneous presentation (Higgins & Ross, 2011; Mundy, Honey, & Dwyer, 2007). Kang

and Pashler (2012) found that simultaneous presentations enhanced neither interleaved nor blocked learning.

Temporal spacing. In interleaved learning, categorical items are temporally spaced. Some researchers have proposed that this temporal spacing may benefit inductive learning, implying that interleaving effects might partly be due to a spacing effect (Kornell & Bjork, 2008; Wahlheim et al., 2011). Interleaving inherently involves temporal spacing. Thus, these two potential sources of influence on inductive learning are to some extent impossible to disentangle. However, some studies purposely varied the intervals in which items were presented to examine the extent that additional spacing affects interleaved and blocked learning. These studies seem to suggest that interleaved learning is better when items of different categories are not additionally spaced, whereas blocked learning is not affected or even enhanced by temporal spacing (Birnbaum, et al., 2013; Kang & Pashler 2012; Mitchell, Nash, & Hall, 2008; Zulkiply & Burt, 2013b).

Age. Very few studies have examined age as a potential moderator of the interleaving effect. Kornell et al. (2010) found no associations between age and interleaving, whereas Lin et al. (2016) found an interaction between interleaving and age group, with younger adults benefitting more from interleaving. Furthermore, Lin et al. found interactions between age group and practice type in functional MRI data, suggesting interleaved learning is associated with a more efficient brain network for younger adults. Two studies examined 2- and 3-year-olds and reported no evidence for an interleaving effect in this age group (Sandhofer & Doumas, 2008; Vlach, Sandhofer, & Kornell, 2008).

Population. Most research on the interleaving effect examined college students, but online samples were also examined in some studies (e.g., Birnbaum et al., 2013; Noh et al., 2016, Hausman & Kornell, 2014) and samples of children or adolescents in classroom settings (e.g., Rau et al., 2013, Rohrer, Dedrick, & Burgess, 2014; Rohrer, Dedrick, & Stershic, 2014). Furthermore, one study was based on a sample of older adults (Kornell et al., 2010) and one

with pre-school samples (Sandhofer & Doumas, 2008). Note that these different populations are largely confounded with age.

Rationale of the Present Study

The present study undertook a comprehensive meta-analysis of studies on the effects of interleaved learning. Apart from providing an estimate of the magnitude of the interleaving effect, this meta-analysis followed two aims. The first aim was to gain information on the generalizability of the interleaving effect. Several studies have demonstrated the potential of interleaving as an educational technique to improve inductive learning (e.g., Kornell & Bjork, 2008; Kang & Pashler, 2012; Zulkiply, McLean, Burt, & Barth, 2012). However, the extent that this technique can be applied in different educational settings, in learners of different age groups (e.g., Kornell et al., 2010; Lin et al., 2016), and with different types of learning materials is unclear. Moreover, the usefulness of interleaving as an educational technique critically hinges on whether the benefits for learning persist over time and occur with different test formats.

The second aim was to clarify the role of two potential moderators – similarity between categories and similarity within categories – which are highlighted by theoretical accounts of the interleaving effect. The attentional bias framework/sequential theory of attention (Carvalho & Goldstone, 2015b, 2017) posits that similarity of stimuli or tasks affects the benefits of an interleaved vs. blocked presentation. More specifically, higher between-category similarities and higher within-category similarity should increase the interleaving effect. The discriminative-contrast hypothesis (Birnbaum et al., 2013) assumes that temporal spacing is critical for the interleaving effect. The effect is particularly assumed to be stronger when stimuli of different categories are juxtaposed closer in time and in space.

In addition to the two proposed moderators, we examined potential moderating effects of methodological and other study characteristics that are generally included in meta-analyses

and are important to gain a complete picture about boundary conditions of the interleaving effect.

Method

Review Criteria

We used different combinations of the search strings *interleav**, *block**, *random*, *schedul**, *pract** and *effect* to locate scientific publications in the PsycINFO and Google Scholar databases. We also used the same search strings to locate pertinent grey literature in the WorldCat and ProQuest databases. We also checked the reference lists of the review papers by Rohrer (2012) and Carvalho and Goldstone (2015b) and the supplemental material of their paper (Carvalho & Goldstone, 2015c). The research extended through June 30, 2018. We identified 176 potential eligible studies that were screened for inclusion in our meta-analysis.

Selection Criteria

Only studies meeting the following four criteria were included in the meta-analysis:

(a) The study design included an interleaved study condition and a blocked study condition, regardless of whether the study design was a within- or between-subjects design.

(b) Retention interval, outcome measurement, time per item, time interval between items and number of items per category were the same for the interleaved and the blocked condition.

(c) Studies reported appropriate statistical values to calculate the effect size for the comparison between the interleaved and the blocked condition. All relevant studies selected based on criteria (a) and (b) either met this inclusion criterion or the authors provided the information upon request.

Studies with deviating sample or item characteristics and manipulations and special study designs for both conditions were not excluded but were considered as moderators in the later statistical analysis. Likewise, dependency between effects (dependent samples) was

addressed by using appropriate statistical techniques (multilevel meta-analysis; see the section Meta-Analytic Strategies). Some data was duplicated because our research also included grey literature. In these instances, we used only the source published earlier as a reference. In total, 59 papers reporting 238 effect sizes based on 158 samples met our inclusion criteria.

Coded Variables

Study characteristics and characteristics of the intervention (interleaved vs. control condition) were coded individually for each effect size reported in the papers that met the inclusion criteria.

Study characteristics.

(a) *Published vs. unpublished studies*. Meta-analysis based only on published studies can overestimate the true population effect because of the publication bias. That is, studies with significant results (and studies with larger effects) are published more often than others. Unpublished studies are assumed to be free of a publication bias. Based on these assumptions, we dummy-coded journal articles and conference papers as published studies (coded 0) and theses and dissertations as unpublished studies (coded 1).

(b) *Student vs. nonstudent samples*. Most studies analyzed student samples (with k = 173 effect sizes). We distinguished between university students (coded 1) and other samples (coded 0) to assess generalizability across different samples. Other samples included online studies with no restrictions with regard to demographics, younger samples (usually school students), samples of elderly people and samples of other specific non-student communities or the general population.

(c) *Mean age*. We coded studies according to the mean age of participants. For most effect sizes (k = 141), participant's age was not reported. Therefore, we ran separate analyses to determine the impact of the mean age on the effects from studies that reported sufficient information.

Characteristics of the intervention. The second group of variables coded for the meta-analysis were characteristics of the intervention, including design, material and test-specific variables:

(a) *Design.* Measurement accuracy and effect sizes systematically differ between designs that vary conditions between participants and designs that vary conditions within participants. Even though statistical methods exist to transform within-subjects into between-subjects effects and vice versa, the estimates for the effect sizes may be biased by the type of design (Morris & DeShon, 2002). Therefore, we dummy-coded the study design, with between-participants designs coded as 0 and within-participants designs as 1.

(b) *Intentional vs. incidental Learning.* The attentional bias framework (Carvalho & Goldstone, 2015b) poses that interleaved practice directs attention to differences between items, whereas blocked practice directs it to similarities. Furthermore, when items can belong to more than one category in a study, they can be presented in an interleaved manner for one category and in a blocked manner in another. Given that learners can intentionally direct their attentional resources to certain aspects of the learning material, intentionality may be crucial for the interleaving effect to occur. Therefore, we dummy-coded intentional learning as 0 and incidental learning as 1. When no other information was provided, we assumed that intentional learning took place in the study setting.

(c) *Temporal spacing*. According to the discriminative contrast hypothesis (Birnbaum et al., 2013), the effectiveness of interleaved learning depends on how closely items are juxtaposed. To test this assumption, we coded the time that elapsed between the presentations of two successive items. Given the low variability between studies, we assigned them to one of three categories: (1) Temporal spacing with 10 or 30 sec time intervals between the presentation of successive

items. (2) Item presentation with distractors between the presentations of successive items. For example, Dwyer, Mundy, and Honey (2011) presented either a checkerboard, a male face, or no distractor between the blocked or interleaved presentation of female faces. (3) Immediate succession of items with less than 2 sec between item presentations. When no information about the spacing of items was given, we assumed immediate presentation. Only four studies were assigned to the first two categories. To minimize noise, we analyzed spacing of item presentation separately instead of including it in the main moderator analysis. Temporal spacing was dummy-coded. The reference-category were studies with immediate presentation.

(d) *Simultaneous vs. successive presentation*. Most studies presented items successively (e.g., Kornell and Bjork, 2008), but some studies presented two items simultaneously on the same screen (e.g., Carvalho & Goldstone, 2012). Simultaneously presented items belonged either to different categories in the interleaved condition or to the same categories in the blocked condition. Simultaneous vs. successive presentation of items may affect the processing of similarities and differences. Thus, we dummy-coded successive presentation as 0 and simultaneous presentation as 1.

(e) *Retention Interval*. We coded the total time between the end of the acquisition phase and the beginning of the final test. This variable is a measure of whether interleaving promotes durable learning. Most studies failed to report exact retention interval time, but we could identify in all studies whether the retention interval was short (< 20 min) or longer (dummy-coded with 0 vs. 1).

(f) *Retention vs. transfer tests*. Interleaving may differently affect performance in retention tests and transfer tests (Kornell et al., 2010). Therefore, we distinguished whether the effect size referred to transfer tests (coded as 1) or retention tests

(coded 0) as dependent variable. Most studies operationalized transfer by testing the classification of items that were not presented in the study phase and operationalized retention by testing the classification of items that were presented in the study phase.

Type of Learning Materials. We classified learning materials into the following seven categories:

(1) Paintings including mostly impressionistic paintings of different artists as first used by Kornell and Bjork (2008).

(2) Naturalistic photographs such as pictures of birds, butterflies (e.g., Birnbaum et al., 2013) and human faces (Dwyer et al., 2011).

(3) Artificial pictures, that is, pictures of artificial objects or creatures such as the ziggerins (Wong, Palmeri, & Gauthier, 2009) or fictitious aliens such as the deegers and koozles (Higgins & Ross, 2011).

(4) Mathematical tasks, such as calculating the volume of geometric solids or the use of significance tests (e.g., Rohrer & Taylor, 2007; Sana et al., 2017).

(5) Expository texts, which included plain expository texts and combinations of texts and other media in a multimedia/interactive media environment (e.g., about psychopathological disorders or the human immune system, Dobson, 2011; Zulkiply et al., 2012).

(6) Words, such as names that belonged to different conceptual categories, pronunciation rules, or translations in different languages (e.g., Carpenter & Mueller, 2013; Hausman & Kornell, 2014; Sorensen & Woltz, 2016).

(7) Tastes such as liquids with different tastes (Dwyer et al., 2004).

For some analyses, we combined categories of paintings, naturalistic photographs, and artificial pictures into the broader category of visual stimuli.

Material Characteristics. Five raters provided a more detailed assessment of the nature of the learning materials on the following dimensions: (a) similarity within categories, (b) similarity between categories (c) complexity, (d) familiarity and (e) curiosity. For this purpose, we created a questionnaire to describe all materials that were used in the studies included in the meta-analysis. The categorization and ratings were based on verbal and pictorial descriptions obtained from the original papers. We also included supplemental online materials when available. For materials used in more than one study, we compiled information from several studies. The five raters rated on an analogous scale the similarity of items between categories, the similarity of items within one category, the complexity of material in the study phase, and the extent that the items appeared familiar and curious. For example, the 20 x 20 big checkerboards with six unique adjacent fields used by Mitchell, Nash and Hall (2008) received the highest rating of within-category similarity but also the highest between-category similarity. Complexity ratings, in contrast, were low to moderate and familiarity ratings were very low. In another example, the conceptual categories (e.g., is loud, can fly, smells, lives underground, lives in the desert, is green) with different words as exemplars (e.g., silence, eagle, rose, or cave) used by Sorensen and Woltz (2016) received very low within- and between-category similarity ratings, low complexity ratings, but very high familiarity ratings. The highest complexity ratings and the lowest familiarity ratings were given for the distiller simulation task used by de Crook and van Merriënboer (2007). In this task, categories were four different types of system failure. The similarity within these categories was rated slightly above average, whereas the similarity between categories was rated below average. For the impressionistic paintings used by Kornell and Bjork (2008) and other researchers, ratings of within-category similarity were slightly below average, and ratings of between-category similarity were slightly above average. Complexity and familiarity ratings were also slightly above average. Interrater reliability and intercorrelations of these ratings are displayed in Table 1. We used the mean of the standardized ratings to

assess the material characteristics. Curiosity was eliminated from the analysis because of the low interrater reliability scores.

Effect Size Calculation

We calculated effects for differences between a blocked and an interleaved condition. Furthermore, we converted all between-group effects in Cohen's *d* and within-group effects in Cohen's d_{av} that ignores the correlation between effect sizes and uses the average standard deviation (Lakens, 2013). We used means and standard deviations or other statistics (e.g., *t* values or *F* ratios) to calculate effect sizes. In the next step, we applied Hedges' (1981) correction term to convert *d* values into Hedges' *g*, because *d* tends to overestimate effect sizes in small samples. Hedges' *g* can be interpreted the same way as Cohen's *d*. Studies were excluded if the available statistical information was not sufficient to compute effect sizes. Dependent effects were coded separately. If a study reported several dependent effects and only one fit our research question, only this effect was included in the meta-analysis. For example, Kornell and Bjork (2008) reported results of multiple immediate test blocks. In each test block, feedback was provided to participants. The repeated testing with feedback possibly promotes learning, either independently from the interleaving effect or by moderating this effect. However, we used only the results from the first test block to avoid confounding, because effects of multiple testing and its interactions with interleaving were not the focus of our meta-analysis.

Meta-analytical Strategies

We used multilevel meta-analysis because of the hierarchical structure of our data. Several studies included in our meta-analysis provided more than one effect size (see Appendix A). Multiple effect sizes obtained within the same study depend on each other. The multilevel approach can address these dependencies and the nested structure without losing information through the exclusion of dependent effects or biases through weighting errors (Jackson, Riley, & White, 2011). Furthermore, we used a random-effects model to estimate

the overall effect across all studies and mixed-effects models for the subsequent analysis of potential moderators, which were included as fixed effects (Borenstein, Hedges, Higgins, & Rothstein, 2010).

We calculated I^2 and Q statistics to examine the variance of the residuals. Cochran's Q_B statistic (Cochran, 1954) is calculated by weighting the sum of the squared deviations of each study's effect size from the overall effect size. Q_B follows a χ^2 distribution with k-1 degrees of freedom (k represents the number of effect sizes) and can be used to test whether the variance of effect sizes is significantly different from 0. If Q_B is less than or equal to the degrees of freedom, complete homogeneity is assumed. Likewise, the Q_M statistic can be used to test the variance explained by the model through moderators. A significant Q_M means that the model explains a significant amount of variance between effects. The statistical power of the test, however, depends on the number of studies. For example, testing the heterogeneity of a large number of studies is more likely to result in a significant Q statistic even when heterogeneity remains the same. To overcome this limitation, we calculated τ^2 as a measure of total variance, I^2 as a measure of the proportion of variance that is due to heterogeneity between studies (Higgins, Thompson, Deeks, & Altman 2003) and R^2 to measure percentage of explained variance by meta-regression models (Raudenbush, 2009). For the multilevel meta-analysis, we used methods suggested by Cheung (2014) to calculate I^2 and R^2 between and within samples. For calculating I^2, we used the mean variance of all effect sizes to estimate the typical within-sample effect size variance as suggested by Xiong, Miller, and Morris (2010). In our meta-analysis, $I^2_{between}$ refers to the proportion of variation between samples, whereas I^2_{within} refers to the proportion of variation within samples. Likewise, $R^2_{between}$ refers to the amount of variance explained between samples, whereas R^2_{within} refers to the amount of variance explained within samples. I^2 values can range from 0 to 100%, and R^2 values range from 0 to 1. We used the package metafor in R (Viechtbauer, 2010) for all steps of the meta-analysis.

Results

We identified $k = 238$ effect sizes nested in 158 samples (59 papers) that met all inclusion criteria and were based on data from $N = 8466$ participants (Figures 1 to 7). Ninety-seven effect sizes nested in 62 samples reported the mean age of the participants with a mean age of 21.32 years. All studies were published between 2003 and 2018. We identified $n = 71$ different sets of learning materials that were used in the reported studies. The expert ratings of learning material characteristics are provided in Table 1.

In the following sections, we first report meta-analytic overall effect estimates of interleaving and its effect for different types of learning materials. We then report on the estimates of the more comprehensive moderator analyses, which include the ratings of additional material characteristics, study characteristics, and characteristics of the intervention.

Overall Effect of Interleaving and Generalizability across Learning Materials

We estimated mean Hedges' *g* across all studies as a measure of the overall effect of interleaving. In the next step, we estimated mean Hedges' *g* for different types of learning materials to gain information about the generalizability of the interleaving effect (Table 2). We used multi-level models to address the nested structure, that is, dependencies between effects obtained with the same samples. As a sensitivity analysis, we also used models that included only independent effects to further control for potential biases caused by including dependent effects. Finally, we estimated heterogeneity within and between samples for the multi-level models.

Effect of Interleaving. We found a moderate overall positive effect of interleaving ($k = 238$; $g = 0.42$; $p < .001$, 95% CI [0.34, 0.50]). The overall interleaving effect was nearly the same when estimated with a two-level random effects model that included only independent effects ($k = 138$; $g = 0.43$; $p < .001$, 95% CI [0.35, 0.51]). In addition, we estimated a two-level random effects model that included the dependent effect sizes. This approach ignores the

nested structure introduced by dependent effects, overweighing samples that provide more than one effect. In this model, the overall effect was biased downwards (k = 238; g = 0.36; p < .001, 95% CI [0.29, 0.42]). Accordingly, interleaving effects based on samples that provided more than one effect size were associated with a significant lower mean g value than samples that provided only one effect size (k = 238; Q_M = 9.7, p < .001). Overall, the multilevel approach seems to be an appropriate method to obtain an unbiased estimator without excluding valuable information.

Generalizability across learning materials. We found significant differences between types of learning materials (k = 238; Q_M = 52.64, p < .001). Interleaving had a positive effect in all types of visual stimuli, with the highest mean effect for paintings (g = 0.67, p < .001, 95% CI [0.57, 0.77]) and naturalistic photographs (g = 0.35, p < .001, 95% CI [0.16, 0.55]. In nonvisual stimuli, a positive interleaving effect was found only for mathematical tasks, g = 0.34 (p = .005, 95% CI [0.11, 0.57]) and no significant effect for expository texts. Furthermore, we found a negative effect (i.e., an advantage of blocking over interleaving) for studies based on words as learning materials, g = -0.39 (p > .001, 95% CI [-0.64, -0.14]).

Heterogeneity. We detected substantial heterogeneity between samples (Table 2) for the overall effect of interleaving (I^2 = 77.3%, τ^2 = .20) and for the effects within specific types of learning materials for paintings (I^2 = 43.4%, τ^2 = .07), naturalistic photographs (I^2 = 56.1%, τ^2 = .07), artificial pictures (I^2 = 73.7%, τ^2 = .15), expository texts (I^2 = 74.4%, τ^2 = .19), mathematical tasks (I^2 = 76.9 %, τ^2 = .23), and words (I^2 = 18.3 %, τ^2 = .03). The effects sizes ranged from -1.37 to 1.85 on the sample level. The significant variation of effect sizes within each type of learning material study suggests that there might be other moderating variables.

In sum, the results reported suggest that the generalizability of the interleaving effect is restricted to specific types of learning materials, particularly visual stimuli. Further

heterogeneity within learning materials suggests that sample, method, or intervention-specific characteristics moderate the effect of interleaving.

Moderator Analysis: Meta-Regression Models

We estimated a series of three nested meta-regression models to assess the impact of moderator variables on the magnitude of the interleaving effect (Table 3). In Model 1, we entered study and intervention-specific characteristics. In Model 2, we added quantitative material characteristics as potential moderators. These moderators included the effects of item similarity within and between categories, which are of primary theoretical interest for the validity of the attentional bias framework. In Model 3, we added the type of learning material as a set of dummy-coded predictors with paintings as the reference category.

In addition to analyses based on the total sample of effects, we analyzed subsets of effects to assess the robustness of the moderator effects (sensitivity analysis, Table 4). These subsets included effects from independent samples, from student samples, for visual stimuli and for artificial pictures. The number of effect sizes per analyses for the subsets ranged from 58 to 164. To avoid specification errors due to the smaller number of effect sizes, we excluded study and intervention-specific characteristics that were not significant moderators in previous Models 1-3.

Studies that analyzed the discrimination of tastes and studies that spaced time between item presentations were excluded from the moderator analysis. Because of their uniqueness, we believed that these studies would have added more noise than information to the models.

We used Model 3 (i.e., the full model with all predictors) to detect outliers. First, we ran the model with all effect sizes. Next, we used a stepwise analysis to exclude the most extreme outliers using influence diagnostics and rules of thumb for studentized deleted residuals, DFFITS values, Cook's Distance and hat-value (Viechtbauer & Cheung, 2010). We ran the model again, following this stepwise procedure until all effect sizes were in accordance with the rules of thumb. Twelve outliers from five studies were excluded from all

models and the additional analyses except for the sensitivity analysis (Table 4). However, the reader should note that the methods used for the outlier analysis were developed for independent effect sizes. Hence, the accuracy of this method for non-independent effect sizes is unclear.

In the remainder of this section, we first describe results for types of learning materials and the additionally coded material characteristics such as similarity (within and between categories) and complexity. We then describe results from the intervention characteristics analysis, followed by results for study characteristics.

Type of Learning Materials. We included type of learning materials in the full meta-regression model (Model 3) to test whether differences between types of learning materials are due to differences in study settings and material characteristics. The interleaving effect was still significantly smaller for all learning materials compared to the reference category of paintings. The negative weight was highest for expository texts (b = -0.56; SE = 0.18; p = .001), followed by words (b = -0.48; SE = 0.23; p = .014) and mathematical tasks (b = -0.43; SE = 0.18; p = .004). Overall, the different types of learning material explained 4.0% additional variance between samples compared to Model 2 that did not include learning materials as predictors.

Material Characteristics. As expected from the viewpoint of the attentional bias framework (Carvalho & Goldstone, 2015b), similarity between categories had a positive influence on the effect of interleaving, whereas similarity within categories had a negative influence, which decreased when learning materials were included in the model (Table 3). The sensitivity analyses (Table 4) show that the pattern of effects was very stable across different models and samples.

Familiarity had the highest influence of all material characteristics across all models (Table 3) and samples (Table 4) when the type of learning materials was also included in

Model 3. Finally, we found some indication that complexity might influence the interleaving effect, but the results differed across different models and samples.

Design. We found no effect of within-participants vs. between-participants designs on the interleaving effect.

Intentional vs. Incidental Learning. Incidental learning had a negative impact on the magnitude of the interleaving effect in the full model when outliers were included (Outlier Model, Table 3). However, it failed to reach significance in all other models, despite that the coefficient seems to indicate a rather large impact. This result might be due to the lack of power associated with this moderator. Only eight studies used interleaved practice combined with incidental learning.

Retention Interval. We found no evidence for a moderating effect of retention interval in the meta-regression analysis.

Simultaneous vs. successive presentation. We found no evidence for a moderating effect of simultaneous vs. successive presentation on the effect of interleaving in the meta-regression analysis (Table 3). As a follow-up test to confirm this null effect, we ran an additional analysis that included only studies that operationalized both presentation styles. The interleaving effect tended to be higher in simultaneously presented items ($g = 0.44$, $p < .001$, 95% CI [0.24, 0.64]) than in successively presented items ($g = 0.29$, $p = .005$, 95% CI [0.09, 0.50]), but the difference was not significant ($k = 26$, $Q_M = 1.0$ $p = 0.31$).

Retention vs. Transfer. We found no moderating effect of type of test items on the effect of interleaving in the meta-regression analysis (Table 3). We also conducted a follow-up test on this null effect by separately analyzing the effects from 31 studies that assessed both retention (classification of old items) and transfer (classification of new items) within the same sample. In a meta-regression of these studies, we found no overall effect of retention vs. transfer ($k = 64$, $Q_M = 0.56$, $p = .45$).

Student Samples. Interleaving effects were larger in pure university student samples compared to nonstudent samples. The moderating effect of student vs. nonstudent samples was obtained even when material characteristics and type of learning material were statistically controlled (Model 3, Table 3) and was also significant in the sensitivity analyses (Model 4).

Published vs. unpublished studies. We found no evidence for a moderating effect of publication in the meta-regression analysis.

Additional Analyses

We ran additional analyses for two moderators not included in the main meta-regression analysis because they were operationalized or reported in only a few studies. These analyses were based only on subsamples that provided relevant information for assessing the impact of the specific moderator.

Mean Age. Only 97 relevant effect sizes nested in 62 samples reported the mean age of the sample. For these studies, we analyzed the moderating role of mean age for the interleaving effect in a separate model. We excluded a study with a mean age of 77 years because of its distance to the other samples. The mean age of participants of remaining studies ranged from 9.5 to 37 years. Material characteristics and types of learning materials were statistically controlled. Because of the restricted sample size and lack of variability of other moderators, we excluded other moderators but analyzed separate samples of effect sizes. We first analyzed all studies together, then analyzed visual learning materials in a second step to control for bias due to the type of learning materials and then analyzed university student samples in a third step to avoid bias due to demographic variables. These analyses suggest that interleaving works better in younger samples, both in the overall sample in which age explained 13.4 % additional variance (ΔR^2) between samples (k = 85, b = -0.04; SE = 0.01; p < .001, $R^2_{between}$ = 58.4 %) and in the sample of studies that included only visual material in which age explained 30.9 % additional variance between samples (k = 42, b = -0.03; SE =

0.01; $p < .001$, $R^2_{between} = 75.8$ %). In the sample with only university students the effect was not significant ($k = 49$, $b = -0.07$; $SE = 0.05$; $p = .19$, $R^2_{between} = 78.3$ %).).

Temporal spacing. We analyzed the moderating role of temporal spacing only for studies in which the results of its effects, compared either in within-subjects designs or between different samples, were reported in the same article.

We found a significant effect of temporal spacing on the effect of interleaving ($k = 17$, $Q_M = 9.64$, $p = .008$). In line with the discriminative contrast hypothesis, we found a significant interleaving effect only for immediate succession of items (g = 0.73, 95% CI [0.51, 0.95]) but not for temporally spaced items ($g = 0.22$, 95% CI [-0.13, 0.45]). We also found a significant interleaving effect for item presentation with distractors ($g = 0.51$, 95% CI [0.07, 0.96]).

Publication Bias

Methods to detect publication bias in multi-level meta-analysis have not been published. All available methods were developed under the assumption of independent effect sizes. As a workaround, we analyzed only one effect size per study when several outcome variables were reported in a study. We always included the effect size of the outcomes or conditions that were more common. Therefore, we included successive presentation, transfer tests, low retention intervals, and immediate succession of items. We used funnel plots (see Sterne & Egger, 2001) and the trim-and-fill method recommended by Duval and Tweedie (2000). The standard errors of the effect sizes differed between types of learning materials ($F(1, 231) = 5.81$; $p < .001$). Furthermore, types of learning materials differed strongly in their effect sizes. In combination, this may lead to an asymmetrical funnel plot. Therefore, we additionally ran separate analyses for the different learning materials.

The trim-and-fill analysis for the total sample of independent effect sizes suggested a publication bias ($g_{Trim\ \&\ Fill} = 0.29$, $p < .001$, 95% CI [0.20, 0.38]) with 23 studies missing on the left side. However, the separate trim-and-fill analyses for different types of learning

materials suggested publication biases only for artificial pictures ($g_{\text{Trim \& Fill}} = 0.17$, $p = .10$, 95% CI [0.02, 0.31]) with nine studies missing on the left side (Figure 8). For all other types of learning materials, the results revealed no evidence of a publication bias.

To assess evidential value and detect potential *p*-hacking, we used *p*-curve analysis (Simonsohn, Nelson, Simmons, 2014; Simonsohn, Simmons, & Nelson, 2015). We found no evidence for *p*-hacking (Figure 9). The test of right skewness indicated evidential value for both the full p-curve ($Z = -16.89$; $p < .001$) and for the half *p*-curve ($Z = -16.44$; $p < .001$). The test for flatness was nonsignificant for the full *p*-curve ($Z = 11.55$; $p > .99$) and for the half *p*-curve ($Z = 16.31$; $p > .99$). Thus, we found no indication for inadequate or absent evidential value. The power of tests included in the *p*-curve was satisfying with 98% (90% CI [97%, 99%]).

Discussion

The aims of this meta-analysis were to determine the overall effect that interleaving has on inductive learning, to clarify the generalizability of the interleaving effect to different learning materials and settings, and to examine how the available evidence relates to major theories in the research area. Overall, we found a moderate and positive effect of interleaving vs. blocking ($g = 0.42$). The analysis also revealed that the benefits of interleaving seem to be restricted to specific learning materials and settings. Consistent with the attentional bias framework (Carvalho & Goldstone, 2015b), interleaving is most effective in inductive learning with complex visual stimuli, such as paintings ($g = 0.67$), and when similarity between categories is high and similarity within categories is low. This finding is also corroborated by the more detailed analyses that included ratings of similarity between and within categories as moderators. In line with the discriminative contrast hypothesis (Kang & Pashler, 2012) interleaving is effective only when items are presented in immediate succession without spacing. Furthermore, intentional learning seems to strengthen the

interleaving effect. In the following sections, we discuss the meta-analytic findings in more detail.

Generalizability

Learning materials. Our analysis revealed moderate to large effects for interleaved presentation of visual materials, which included naturalistic photographs and paintings but also pictures of artificial objects. For visual materials, interleaved presentation of exemplars seems to be a promising measure to foster inductive learning of category structures. However, the analysis revealed no evidence for the beneficial effects of interleaving when learning materials are expository texts, or tastes. When the learning materials are words, interleaving even seems to have negative effects compared to blocking. In addition to systematic differences between types of learning materials, the meta-analysis revealed significant heterogeneity of the interleaving effect within all types of learning materials. In the context of the random effects model, this result suggests that a number of studies deviate systematically from the estimated mean effect, particularly the large effects in some of the primary studies (e.g., Rohrer, Dedrick, & Burgess, 2014; Rohrer & Taylor, 2007), which were unlikely due to random variation. Therefore, the lack of an overall effect for expository texts and the small effect found for mathematical tasks does not imply that interleaving is ineffective when used with these learning materials. Instead, interleaving can benefit but also harm inductive learning, depending on the implementation, the measure of learning outcomes, and the specific characteristics of the learning material. For example, in the study by Rau et al. (2013), mathematical tasks were blocked on one dimension and interleaved on another dimension. Interleaving on one dimension had positive effects on learning but negative effects on learning on another dimension.

The reviewed literature for words and tastes included too few empirical studies that varied in setting and task type to be able to obtain generalizable results for all kind of tasks using words or tastes. Therefore, the results should be interpreted with caution. The negative

effect for words, for example, might not be caused by the use of words but rather by the task types or the specific categories or other (unknown) variables that characterized the few studies using words as learning materials.

Population and age. Our meta-analysis revealed that interleaving works best for student populations. However, it remains an open question whether this finding can be generalized or is merely caused by the composition of student samples. Such samples are often homogeneous, which decreases noise and increases the effect size. Overall, relatively few studies have employed nonstudent samples. Thus, the greater effect sizes for student samples compared to samples from other populations should be interpreted with caution. Regarding age, our analysis revealed a strong association between the mean age of the samples and the size of the interleaving effect, which was larger for younger participants. This result departs from the inconclusive findings from primary research (Kornell et al., 2010; Lin et al., 2016; Sandhofer & Doumas, 2008; Vlach et al., 2008). Considering that the association of age and the interleaving effect was also found when only student samples or visual materials were analyzed, confounds with other sociodemographic characteristics (e.g., educational background) or learning materials are unlikely.

One possible explanation for the moderating effect of age is age-related changes in cognitive functions. Lin et al. (2016) found psychophysiological interactions for functional MRI data between age group and interleaved versus blocked practice of a motor skill. In the interleaved condition, younger participants showed a network that exhibited efficient small world topology, stronger functional segregation, and a significant association between higher network centrality and better learning after interleaved practice. Older adults did not exhibit those favorable network properties. Likewise, theories of category learning, such as the COmpetition between Verbal and Implicit Systems theory (COVIS; Ashby, Alfonso-Reese, Turken, & Waldron, 1998) assume age-related changes through neuronal development. Another possible explanation for the moderating effect of age are generational effects. For

example, such generational differences may be rooted in different experiences and different amount of exposure to various media and technology. Among other influences, this exposure may affect participants' familiarity with the research environment and task or affect their processing of the stimuli presented in an experiment.

Interestingly, the effect of age on interleaved learning differs from that found for other types of learning subsumed under the label of desirable difficulties. For example, theories of the spacing and testing effects suggest age invariance, which has been supported by empirical studies and systematic reviews (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Rowland, 2014; Toppino, Fearnow-Kenney, Kiepert, & Teremula 2009; Toppino & Gerbier 2014; Toppino, Kasserman, & Mracek, 1991). Furthermore, systematic reviews of the generation effect revealed an opposite effect of age, with older subjects benefitting more from generation (Bertsch, Pesta, Wiscott, & McDaniel, 2007). The age-related differences found for interleaving might hint at a unique mechanism of interleaved learning, a question that should be pursued in future investigations.

Retention vs. transfer tests. Our meta-analysis revealed no evidence for differential effects of interleaving on retention tests and transfer tests, which is consistent with previous research (cf. Kornell et al., 2010). Retention tests are based on previously studied items, whereas transfer tests are based on new items (e.g., the classification of new exemplars to the learned categories) that come from the same item pool. The finding that interleaving improves transfer performance strongly supports the assumption that interleaving fosters inductive learning and not just associative learning (e.g., the association of exemplars with category names).

Theoretical and Practical Implications

Theoretically, the results of this meta-analysis support the discriminative contrast hypothesis and the attentional bias framework (sequential theory of attention). Consistent with the discriminative contrast hypothesis (Kang & Pashler, 2012), our analysis revealed that interleaving was more effective when no spacing or distraction between the item presentations

was included in the method. Note that these findings do not rule out the possibility that interleaving effects are (at least partly) based on the spacing of materials (i.e., exemplars from the same category), which is inherent to interleaving. However, the assumption is plausible that intervals between item presentations must not become too large because this spacing might hamper the identification of features that discriminate between categories. One factor that might play a role here is the retrievability of previously presented items and their features (analogous to distributed learning, Benjamin & Tullis, 2010).

The attentional bias hypothesis (Carvalho & Goldstone, 2015b), also known as sequential theory of attention (Carvalho & Goldstone, 2017) refines the discriminative contrast hypothesis by distinguishing learning situations in which detection of similarities versus the detection of dissimilarities are crucial for inductive learning. In particular, the theory predicts that an advantage of interleaved learning should occur for categories with low discriminability, whereas an advantage of blocked learning should occur for categories with high discriminability but low similarity of exemplars belonging to the same category. In line with these predictions, our meta-analysis revealed that the interleaving effect increased with similarity between categories and with the complexity of the material, which could have lowered the discriminability of categories. Furthermore, the interleaved effect decreased with the similarity of items within a given category.

For practical applications, our results underscore that interleaving can be fruitfully implemented as a way to assist inductive learning. However, this recommendation comes with several qualifications depending on the type of learning materials. Interleaved presentation seems to work well with naturalistic visual stimuli (naturalistic paintings or photographs) are used as learning materials, when the stimuli are complex, when categories are difficult to discriminate, and when the similarity of exemplars within categories is low. The learning materials used in the studies included naturalistic paintings, photographs of birds and butterflies, and pictures of human faces. Given that our estimates of effect sizes are based on a

random effects model, makes us confident that the estimates can be generalized to the populations of visual stimuli from which samples were drawn in the experiments included in the meta-analysis. Thus, interleaving may be a very useful didactic tool in disciplines such as history of arts, biology, medicine, geology and other disciplines that require the differentiation and classification of complex naturalistic visual stimuli. The complexity and similarity of objects should also be taken into consideration for visual stimuli depicting artificial objects. Interleaving may foster inductive learning with artificial pictures such as logical or schematic diagrams illustrating abstract processes, theoretical models, geometric objects, or chemical elements. However, to the extent that such diagrams are designed with the aim to reduce complexity and highlight differences between concepts, interleaving may be less beneficial.

We are reluctant to recommend interleaving for learning materials such as mathematical tasks, expository texts, grammar rules, foreign languages or words (i.e., category names). When interleaving is used with these materials, it could even impede learning compared to a blocked presentation of items.

Limitations and Directions for Future Research

One purpose of this meta-analysis was to examine the generalizability of the interleaving effect across different materials, settings, and populations of learners. The ideal database for tackling the question of generalizability would be many primary studies for each of the different types of learning materials and settings. However, this ideal database was not available. For example, most of the extant studies have relied on visual materials, whereas educationally interesting materials, such as mathematical tasks or expository texts, are clearly underrepresented. Moreover, most studies have employed samples of university students, which limits the generalizability of the present results for younger learners in primary and secondary education but also for adults beyond the college age.

A second limitation is that different types of learning materials are confounded with methodological characteristics, especially when scrutinizing the differences between studies

based on visual stimuli and studies based on nonvisual materials in their study design. For example, the retention intervals and the presentation times of each item are often longer for studies based on nonvisual learning materials compared to studies that have used visual learning materials. Furthermore, most of the available studies on interleaved learning have focused on learning outcomes. In contrast, relatively few studies have attempted to examine the cognitive processes underlying interleaving effects. One notable exception are the experiments conducted in the context of the attentional bias framework/sequential theory of attention (e.g., Carvalho & Goldstone, 2015b, 2017) that used pictures of artificial objects. To derive sound practical recommendations and to determine the scope of theoretical models of interleaving, more experiments are needed that directly focus on psychological mechanics (e.g., by collecting process measures and relating these to learning outcomes) across a wider range of materials.

Our study revealed a robust impact of mean age on the interleaving effect. However, associations found in aggregated data on the study level do not necessarily correspond to associations at the individual level. For example, Kornell et al. (2010) found no moderating effect of age on the interleaving effect, but they only compared two age-groups. Thus, the role of age for interleaved learning should be clarified by further studies. This research should consider that age-related changes over the lifespan are more often nonlinear than linear.

Another potential issue for the interpretation of the present results is the reliability and validity of the rating scales used to assess specific characteristics of the learning materials. In many meta-analyses, measures with suboptimum rating scales are avoided in favor of categorical variables that are defined in a way that perfect interrater agreement such that Cohen's κ approximates 1. In this meta-analysis, we deliberately chose a different approach for several reasons. Using categorical judgments to capture a continuous variable leads to a loss of information. Even when interrater reliability is high for a categorical variable, the relationship to the continuous variable still depends on how closely the subjective rating

matches the underlying (continuously scaled) construct. Therefore, the artificial grouping of metric variables may maximize reliability, but the procedure may also decrease the validity. To avoid this limitation, we created a questionnaire that included all learning materials and five independent raters rated the material characteristics. These ratings need to be interpreted with caution, because they are prone to limitations like any other subjective rating scale. Nevertheless, the ratings were highly reliable and the associations between the interleaving effect and the material characteristics were very robust and were consistent with previous research and the theoretical assumptions of the attentional bias framework. We hope that these findings stimulate primary studies that use continuous manipulations of similarity and complexity within and between categories. These studies may provide new insights into the role of category structure in interleaved learning.

On a general level, our meta-analysis revealed that the extant research on interleaved learning has concentrated on a relatively small range of learning materials, learning situations, and age groups. Given the pervasiveness of induction, especially in everyday informal learning but also in many formal learning contexts, the limited scope of the extant research is surprising. For example, the theoretical approaches discussed here suggest that the sequential order of information might play a large role in concept formation and language learning in small children. However, to our knowledge, no systematic research exists on interleaving vs. blocking in children's concept formation. Likewise, interleaving has received little attention, if any, in social-cognitive research on social categorization and impression formation. In these and other fields, the effects and the potential of interleaved learning remain uncovered.

Conclusion

The findings from this meta-analysis fit well to the previous research and current theoretical explanations of the interleaving effect in inductive learning. In particular, the results support the attentional bias framework and the discriminative contrast hypothesis, whereas temporal spacing impedes interleaved learning. For practical applications in

educational settings, the generalizability of the results is most relevant. Interleaving is clearly effective for inductive learning of complex visual material with high similarity between categories but not within categories. In contrast, results concerning the interleaving effect for learning with expository texts are mixed. Thus, interleaving should be used with caution when using this type of materials. For mathematical tasks, a small overall effect was found, and the effects were overall heterogeneous, indicating the need for further research to clarify the conditions that are crucial for interleaving to be effective with these types of learning materials. In sum, interleaved learning shows great potential, but the type of learning materials and the category structure that characterizes the learning materials need to be taken into consideration for practical implementations. Moreover, the effects of interleaving vs. blocking are still largely unexplored in other research areas in which inductive learning plays a major role such as children's concept formation and social categorization.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56*, 149-178. doi: 10.1146/annurev.psych.56.091103.070217

Ashby, F. G., & Maddox, W. T. (2011). Human category learning 2.0. *Annals of the New York Academy of Sciences, 1224*, 147-161. doi: 10.1111/j.1749-6632.2010.05874.x

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*, 442–481. doi: 10.1037//0033-295X.105.3.442

*Archambault, K. B. (2014). *How stimulus similarity impacts spacing and interleaving effects in long-term memory* (Doctoral Dissertation, University of Minnesota). Retrieved from http://conservancy.umn.edu/handle/11299/164690

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology, 61*, 228-247. doi: 10.1016/j.cogpsych.2010.05.004

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35*, 201-210. doi: 10.3758/BF03193441

*Birnbaum, M. S. (2013). *Understanding and optimizing the inductive learning of categories and concepts* (Doctoral Dissertation, University of California). Retrieved from http://escholarship.org/uc/item/8396m4g0#page-1

*Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition*, *41*, 392–402. doi: 10.3758/s13421-012-0272-7

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, &

J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). New York: Worth.

Borenstein, M., Hedges, L. V., Higgins, J. P.T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*, 97–111. doi: 10.1002/jrsm.12

Brady, F. (2004). Contextual interference: A meta-analytic study. *Perceptual and Motor Skills, 99*, 116-126. doi: 10.2466/pms.99.1.116-126

*Carlson, R. A., & Shin, J. C. (1996). Practice schedules and subgoal instantiation in cascaded problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 157–168. doi: 10.1037/0278-7393.22.1.157

*Carlson, R. A., & Yaure, R. G. (1990). Practice schedules and the use of component skills in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 484–496. doi: 10.1037/0278-7393.16.3.484

*Carpenter, S. K., & Mueller, F. E. (2013). The effects of interleaving versus blocking on foreign language pronunciation learning. *Memory & Cognition*, *41*, 671–682. doi: 10.3758/s13421-012-0291-4

*Carvalho, P. F., & Albuquerque, P. B. (2012). Memory encoding of stimulus features in human perceptual learning. *Journal of Cognitive Psychology*, *24*, 654–664. doi: 10.1080/20445911.2012.675322

*Carvalho, P. F., & Goldstone, R. L. (2012). Category structure modulates interleaving and blocking advantage in inductive category acquisition. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 186–191). Austin, TX: Cognitive Science Society. Retrieved from http://escholarship.org/uc/item/0nw0d367

*Carvalho, P. F., & Goldstone, R. L. (2014a). Effects of interleaved and blocked study on delayed test of category learning generalization. *Frontiers in Psychology*, *5*, 936. doi: 10.3389/fpsyg.2014.00936

*Carvalho, P. F., & Goldstone, R. L. (2014b). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, *42*, 481–495. doi: 10.3758/s13421-013-0371-0

Carvalho, P. F., & Goldstone, R. L. (2015a). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review*, *22*, 281–288. doi: 10.3758/s13423-014-0676-4

Carvalho, P. F., & Goldstone, R. L. (2015b). What you learn is more than what you see: What can sequencing effects tell us about inductive category learning? *Frontiers in Psychology, 6*, 505. doi: 10.3389/fpsyg.2014.00936

*Carvalho, P. F., & Goldstone, R. L. (2015c). What you learn is more than what you see: What can sequencing effects tell us about inductive category learning? [Supplemental Material]. *Frontiers in Psychology, 6*, 505. doi: 10.3389/fpsyg.2014.00936

*Carvalho, P. F., & Goldstone, R. L. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*, 1699–1719. doi: 10.1037/xlm0000406

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132,* 354–380. doi: 10.1037/0033-2909.132.3.354

Cheung, M. W. L. (2014). Modeling dependent effect sizes with three-level meta-analyses: a structural equation modeling approach. *Psychological Methods, 19*, 211-229. doi: 10.1037/a0032968

Cleeremans, A. (1996). Principles of implicit learning. In D. Berry (Ed.), *How implicit is implicit learning?* (pp. 196–234). Oxford: Oxford University Press.

Cochran, W.G. (1954). The combination of estimates from different experiments. *Biometrics, 10*, 101–129. doi: 10.2307/3001666

*de Croock, M. B., & van Merriënboer, J. J. (2007). Paradoxical effects of information presentation formats and contextual interference on transfer of a complex cognitive skill. *Computers in Human Behavior, 23*, 1740–1761. doi: 10.1016/j.chb.2005.10.003

*Dobson, J. L. (2011). Effect of selected "desirable difficulty" learning strategies on the retention of physiology information. *Advances in Physiology Education*, *35*, 378–383. doi: 10.1152/advan.00039.2011

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology, 84*, 795–805. doi: 10.1037/0021-9010.84.5.795

Duggan, G. B., & Payne, S. J. (2001). Interleaving reading and acting while following procedural instructions. *Journal of Experimental Psychology: Applied, 7*, 297. doi: 10.1037/1076-898X.7.4.297

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14,* 4-58. doi: 10.1177/1529100612453266

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455–463. doi: 10.1111/j.0006-341X.2000.00455.x

*Dwyer, D. M., Hodder, K. I., & Honey, R. C. (2004). Perceptual learning in humans: Roles of preexposure schedule, feedback, and discrimination assay. *The Quarterly Journal of Experimental Psychology Section B, Comparative and Physiological Psychology*, *57*, 245–259. doi: 10.1080/02724990344000114

*Dwyer, D. M., Mundy, M. E., & Honey, R. C. (2011). The role of stimulus comparison in human perceptual learning: Effects of distractor placement. *Journal of Experimental Psychology. Animal Behavior Processes*, *37*, 300–307. doi: 10.1037/a0023078

*Eglington, L. G., & Kang, S. H. (2017). Interleaved presentation benefits science category learning. *Journal of Applied Research in Memory and Cognition, 6,* 475–485. doi: 10.1016/j.jarmac.2017.07.005

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review, 107*, 659-676.

*Friedman, M. C. (2013). *The importance of selectivity in memory: The influence of value on monitoring, learning, and cognitive aging*. (Doctoral Dissertation, University of California). Retrieved from http://escholarship.org/uc/item/4b98r6mw

*Gane, B. D. (2006). *Can modular examples and contextual interference improve transfer?* (Master Thesis, Georgia Institute of Technology). Retrieved from http://hdl.handle.net/1853/11451

*Guzman-Munoz, F. J. (2017). The advantage of mixing examples in inductive learning: A comparison of three hypotheses. *Educational Psychology, 37,* 421–-437. doi: 10.1080/01443410.2015.1127331

*Hatala, R. M., Brooks, L. R., & Norman, G. R. (2003). Practice makes perfect: The critical role of mixed practice in the acquisition of ECG interpretation skills. *Advances in Health Sciences Education, 8,* 17–26. doi: 10.1023/A:102268740

*Hausman, H., & Kornell, N. (2014). Mixing topics while studying does not enhance learning. *Journal of Applied Research in Memory and Cognition*, *3*, 153–160. doi: 10.1016/j.jarmac.2014.03.003

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107–128. doi: 10.1037/edu0000119

*Higgins, E. J., & Ross, B. H. (2011). Comparisons in category learning: How best to compare for what. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 1388-1393). Austin, TX: Cognitive Science Society. Retrieved from http://escholarship.org/uc/item/7zz764kx

Higgins, J., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327,* 557–560. doi: 10.1136/bmj.327.7414.557

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1989). *Induction: Processes of inference, learning, and discovery*. Boston, MA: MIT press.

Jackson, D., Riley, R., & White, I. R. (2011). Multivariate meta-analysis: Potential and promise. *Statistics in Medicine, 30*, 2481–2498. doi: 10.1002/sim.4172

Janiszewski, C., Noel, H., & Sawyer, A. G. (2003). A meta-analysis of the spacing effect in verbal learning: Implications for research on advertising repetition and consumer memory. *Journal of Consumer Research, 30*, 138–149. doi: 10.1086/374692

*Kang, S. H. K., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, *26*, 97–103. doi: 10.1002/acp.1801

*Kenney, A. (2009). *The spacing effect in inductive learning*. Retrieved from http://akenney.fastmail.fm/works/9.61Paper.pdf

*Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the „enemy of induction“? *Psychological Science*, *19*, 585–592. doi: 10.1111/j.1467-9280.2008.02127.x

*Kornell, N., Castel, A. D., Eich, T. S., & Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and Aging*, *25*, 498–503. doi: 10.1037/a0017807

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology, 4,* 863. doi: 10.1037//1082-989X.7.1.105

*Lavis, Y., & Mitchell, C. (2006). Effects of preexposure on stimulus discrimination: An investigation of the mechanisms responsible for human perceptual learning. *The Quarterly Journal of Experimental Psychology*, *59*, 2083–2101. doi: 10.1080/17470210600705198

Lee, T. D., & Genovese, E. D. (1988). Distribution of practice in motor skill acquisition: Learning and performance effects reconsidered. *Research Quarterly for Exercise and Sport, 59*, 277–287. doi: 10.1080/02701367.1988.10609373

Lin, C. H. J., Knowlton, B. J., Wu, A. D., Iacoboni, M., Yang, H. C., Ye, Y. L., ... & Chiang, M. C. (2016). Benefit of interleaved practice of motor skills is associated with changes in functional brain network topology that differ between younger and older adults. *Neurobiology of Aging*, *42*, 189–198. doi: 10.1016/j.neurobiolaging.2016.03.010

Magill, R. A., & Hall, K. G. (1990). A review of the contextual interference effect in motor skill acquisition. *Human Movement Science, 9*, 241-289. doi: 10.1016/0167-9457(90)90005-X

*Metcalfe, J., & Xu, J. (2016). People mind wander more during massed than spaced inductive learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 978–984. doi: 10.1037/xlm0000216

Michalski, R. S. (1984). A theory and methodology of machine learning. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (pp. 83-134). Berlin: Springer.

*Mitchell, C., Kadib, R., Nash, S., Lavis, Y., & Hall, G. (2008). Analysis of the role of associative inhibition in perceptual learning by means of the same-different task. *Journal of Experimental Psychology. Animal Behavior Processes*, *34*, 475–485. doi: 10.1037/0097-7403.34.4.475

*Mitchell, C., Nash, S., & Hall, G. (2008). The intermixed-blocked effect in human perceptual learning is not the consequence of trial spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 237–242. doi: 10.1037/0278-7393.34.1.237

*Monteiro, S., Melvin, L., Manolakos, J., Patel, A., & Norman, G. (2017). Evaluating the effect of instruction and practice schedule on the acquisition of ECG interpretation skills. *Perspectives on Medical Education, 6,* 237–245. doi: 10.1007/s40037-017-0365-x

*Moors, A. C. (2013). *The spacing effect in category learning: When is spaced practice advantageous?* (Master Thesis, Villanova University). Retrieved from https://www.worldcat.org/title/spacing-effect-in-category-learning-when-is-spaced-practice-advantageous/oclc/869463863&referer=brief_results

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods, 7*, 105–125. doi: 10.1037//1082-989X.7.1.105

Mundy, M. E., Honey, R. C., & Dwyer, D. M. (2007). Simultaneous presentation of similar stimuli produces perceptual learning in human picture processing. *Journal of Experimental Psychology: Animal Behavior Processes, 33*, 124–138. doi: 10.1037/0097-7403.33.2.124

*Noh, S. M., Yan, V. X., Bjork, R. A., & Maddox, W. T. (2016). Optimal sequencing during category learning: Testing a dual-learning systems perspective. *Cognition*, *155*, 23–29. doi: 10.1016/j.cognition.2016.06.007

*Patel, R., Liu, R., & Koedinger, K. (2016). When to block versus interleave practice? Evidence against teaching fraction addition before fraction multiplication. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2069–2074). Philadelphia, PA: Cognitive Science Society. Retrieved from https://pdfs.semanticscholar.org/4cf2/940b550c85a4df34bf1c68bf965be7edb415.pdf

*Phillips, F. (2016). Are you making learning too easy? Effects of grouping accounting problems on students' learning. *Issues in Accounting Education, 32*, 81–93. doi: 10.2308/iace-51589

Poulin-Dubois, D., & Pauen, S. (2017). The development of object categories: What, when and how. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (2nd. ed., pp. 653-671). Amsterdam: Elsevier.

*Rau, M. A., Aleven, V., & Rummel, N. (2010). Blocked versus interleaved practice with multiple representations in an intelligent tutoring system for fractions. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Intelligent tutoring systems* (pp. 413–422). Berlin, Germany: Springer. Retrieved from http://link.springer.com/10.1007/978-3-642-13388-6_45

Rau, M. A., Aleven, V., & Rummel, N. (2013). Interleaved practice in multi-dimensional learning tasks: Which dimension should we interleave? *Learning and Instruction, 23*, 98–114. doi: 10.1016/j.learninstruc.2012.07.003

*Rau, M. A., Rummel, N., Aleven, V., Pacilio, L., & Tunc-Pekkan, Z. (2012). How to schedule multiple graphical representations? A classroom experiment with an intelligent tutoring system for fractions. In J. V. Aalst, K. Thompson, M. J. Jacobson, & P. Reimann (Eds.), *The future of learning: Proceedings of the 10th International Conference of the Learning Sciences* (pp. 64–71). Sydney: International Society of Learning Sciences. Retrieved from https://pdfs.semanticscholar.org/9657/42372ddae1593df9e9af5ed4b194f328ccc5.pdf

Raudenbush, S. W. (2009). Analyzing effect sizes: Random effects models. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 295–315). New York: Russell Sage Foundation.

*Richland, L. E., Bjork, R. A., & Finley, J. R. (2005). Linking cognitive science to education: Generation and interleaving effects. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Twenty-seventh Annual Conference of the Cognitive Science Society* (pp. 1850–1855). Mahwah, NJ: Erlbaum. Retrieved from http://learninglab.uchicago.edu/Publications_files/5-CogsciIddeas2005.pdf

Rohrer, D. (2012). Interleaving helps students distinguish among similar concepts. *Educational Psychology Review, 24*, 355–367. doi: 10.1007/s10648-012-9201-3

*Rohrer, D., Dedrick, R. F., & Burgess, K. (2014). The benefit of interleaved mathematics practice is not limited to superficially similar kinds of problems. *Psychonomic Bulletin & Review*, *21*, 1323–1330. doi: 10.3758/s13423-014-0588-3

Rohrer, D., Dedrick, R. F., & Stershic, S. (2014). Interleaved practice improves mathematics learning. *Journal of Educational Psychology, 107*, 900–908. doi: 10.1037/edu0000001

*Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*, 481–498. doi: 10.1007/s11251-007-9015-8

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140,* 1432 -1436. doi: 10.1037/a0037559

*Sana, F., Yan, V. X., & Kim, J. A. (2017). Study sequence matters for the inductive learning of cognitive concepts. *Journal of Educational Psychology*, *109*, 84–98. doi: 10.1037/edu0000119

Sandhofer, C. M., & Doumas, L. A. (2008). Order of presentation effects in learning color categories. *Journal of Cognition and Development, 9*, 194–221. doi: 10.1080/15248370802022639

*Shah, R., Sibbald, M., Jaffer, N., Probyn, L., & Cavalcanti, R. B. (2016). Online self-study of chest X-rays shows no difference between blocked and mixed practice. *Medical Education, 50,* 540–549. doi: 10.1111/medu.12991

Shea, J. B., & Morgan, R. L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 179-187. doi: 10.1037/0278-7393.5.2.179.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*, 534-547. doi: 10.1037/a0033242

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General, 144,* 1146–1152. doi: 10.1037/xge0000104

*Sorensen, L. J., & Woltz, D. J. (2016). Blocking as a friend of induction in verbal category learning. *Memory & Cognition*, *44*, 1000–1013. doi: 10.3758/s13421-016-0615-x

Sterne, J. A., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology, 54*, 1046–1055. doi: 10.1016/S0895-4356(01)00377-8

*Taylor, K. M. (2008). *The benefits of interleaving different kinds of mathematics practice problems*. (Doctoral Dissertation, University of South Florida). Retrieved from http://scholarcommons.usf.edu/etd/529/

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology*, *24*, 837–848. doi: 10.1002/acp.1598

Toppino, T. C., Fearnow-Kenney, M. D., Kiepert, M. H., & Teremula, A. C. (2009). The spacing effect in intentional and incidental free recall by children and adults: Limits on the automaticity hypothesis. *Memory & Cognition, 3*, 316-325. doi: 10.3758/MC.37.3.316

Toppino, T. C., & Gerbier, E. (2014). About practice: repetition, spacing, and abstraction. *The Psychology of Learning and Motivation, 60*, 113–189. doi: 10.1016/B978-0-12-800090-8.00004-4

Toppino, T. C., Kasserman, J. E., & Mracek, W. A. (1991). The effect of spacing repetitions on the recognition memory of young children and adults. *Journal of Experimental Child Psychology, 51*, 123-138. doi: 10.1016/0022-0965(91)90079-8

*Verkoeijen, P. P. J. L., & Bouwmeester, S. (2014). Is spacing really the „friend of induction“? *Frontiers in Psychology*, *5*, 259. doi: 10.3389/fpsyg.2014.00259

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*, 1-48. doi: 10.18637/jss.v036.i03

Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*, 112–125. doi: 10.1002/jrsm.11

Vlach, H. A., Sandhofer, C.M., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition, 109,* 163–167. doi: 10.1016/j.cognition.2008.07.013

*Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & Cognition*, *39*, 750–763. doi: 10.3758/s13421-010-0063-y

*Weitnauer, E., Carvalho, P. F., Goldstone, R. L., & Ritter, H. (2013, January). Grouping by similarity helps concept learning. *Proceedings of the 35th Annual Conference of the Cognitive Science Society.* (pp. 3747-3752). Austin, TX: Cognitive Science Society. Retrieved from https://escholarship.org/uc/item/4574x59n

Wong, A. C.-N., Palmeri, T. J., & Gauthier, I. (2009). Conditions for face-like expertise with objects: Becoming a Ziggerin expert – but which type? *Psychological Science, 20*, 1108–1117. doi: 10.1111/j.1467-9280.2009.02430.x

*Wright, E. G. (2017). *Combining blocked and interleaved presentation during passive study and its effect on inductive learning* (Master Thesis, University of Dayton). Retrieved from https://etd.ohiolink.edu/pg_10?0::NO:10:P10_ACCESSION_NUM:dayton1492602935015415

Xiong, C., Miller, J. P., & Morris, J. C. (2010). Measuring study-specific heterogeneity in metaanalysis: application to an antecedent biomarker study of Alzheimer's disease. *Statistics in Biopharmaceutical Research, 2*, 300–309. doi:10.1198/sbr.2009.0067

*Yan, V. (2014). *Learning concepts and categories from examples how learners' beliefs match and mismatch the empirical evidence*. (Doctoral Dissertation, University of California). Retrieved from http://escholarship.org/uc/item/91q7z7z4

*Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General*, 145, 918–933. doi: 10.1037/xge0000177

*Yan, V. X., Soderstrom, N. C., Seneviratna, G. S., Bjork, E. L., & Bjork, R. A. (2017). How should exemplars be sequenced in inductive learning? Empirical evidence versus learners' opinions. *Journal of Experimental Psychology: Applied, 23*, 403–416. doi: 10.1037/xap0000139

*Ziegler, E., & Stern, E. (2014). Delayed benefits of learning elementary algebraic transformations through contrasted comparisons. *Learning and Instruction, 33*, 131–146. doi: 10.1016/j.learninstruc.2014.04.006

*Ziegler, E., & Stern, E. (2016). Consistent advantages of contrasted comparisons: Algebra learning under direct instruction. *Learning and Instruction, 41*, 41–51. doi: 10.1016/j.learninstruc.2015.09.006

*Zulkiply, N. (2013). Effect of interleaving exemplars presented as auditory text on long-term retention in inductive learning. *Procedia - Social and Behavioral Sciences*, *97*, 238–245. doi: 10.1016/j.sbspro.2013.10.228

*Zulkiply, N. (2015). The role of bottom-up vs. top-down learning on the interleaving effect in category induction. *Pertanika Journal of Social Sciences & Humanities*, *23*, 933–944. Retrieved from http://www.pertanika.upm.edu.my/

*Zulkiply, N., & Burt, J. S. (2013a). Inductive learning: Does interleaving exemplars affect long-term retention? *Malaysian Journal of Learning and Instruction*, *10*, 133–155. Retrieved from http://mjli.uum.edu.my/

*Zulkiply, N., & Burt, J. S. (2013b). The exemplar interleaving effect in inductive learning: Moderation by the difficulty of category discriminations. *Memory & Cognition*, *41*, 16–27. doi: 10.3758/s13421-012-0238-9

*Zulkiply, N., McLean, J., Burt, J. S., & Bath, D. (2012). Spacing and induction: Application to exemplars presented as auditory and visual text. *Learning and Instruction*, *22*, 215–221. doi: 10.1016/j.learninstruc.2011.11

Table 1

Interrater Reliability and Intercorrelations for Standardized Mean Ratings of Characteristics of the Learning Materials

	Cronbach's α	1	2	3	4
1. Similarity within categories	.91				
2. Similarity between categories	.92	.57***			
3. Complexity	.92	.12	.17		
4. Familiarity	.94	-.14	-.39**	-.17	
5. Curiosity	.72	.11*	.21	.15	-.69***

Note. Intercorrelations of the ratings of 71 different learning materials. Cronbach's α estimated for four raters who rated all 71 materials.

* $p < .05$; ** $p < .01$; *** $p < .001$

Table 2

Interleaved Learning: Overall Effect and Mean Effects by Different Types of Learning Materials

Learning materials					95 % CI		Heterogeneity I^2	
	k	*g*	*p*	Q_B	*LL*	*UL*	Between	Within
Overall	238	0.42	<.001	1201.1***	0.34	0.50	77.3 %	0.0 %
Paintings	70	0.67	<.001	278.0***	0.57	0.77	43.4 %	27.1 %
Naturalistic photographs	19	0.35	.001	40.1**	0.16	0.55	56.1%	0.0 %
Artificial pictures	63	0.31	<.001	262.4***	0.18	0.44	73.7 %	0.0 %
Expository texts	25	0.21	.119	78.4***	-0.06	0.47	74.4 %	0.0 %
Mathematical tasks	45	0.34	.005	202.3***	0.11	0.57	76.9 %	0.0 %
Words	13	-0.39	.005	28.8**	-0.64	-0.14	18.1%	39.7 %
Tastes	3	0.57	.24	4.3	-0.11	1.26	36.9 %	

Note. *k* = number of effect sizes, *g* = mean Hedges' *g*, CI = confidence interval for Hedges' *g*, *LL* = Lower Limit, *UL* = Upper Limit. All effect sizes were estimated using a multilevel random-effects model.
** $p < .01$; *** $p < .001$

Table 3

Multilevel Mixed-Effect Meta-Regression Estimating the Effects of Study Characteristics, Characteristics of the Intervention, Material Characteristics, and Type of Learning Materials

	Model 1	Model 2	Model 3	Outlier Model
Intercept	0.28*	0.33**	0.54***	0.52***
Grey literature	0.12	-0.05	-0.07	-0.01
Design	0.11	-0.02	-0.02	0.07
Student samples	0.13	0.14	0.17*	0.19*
Retention interval	-0.13	-0.10	-0.08	-0.05
Incidental vs. intentional learning	-0.46	-0.52	-0.37	-0.55**
Retention vs. transfer tests	-0.02	-0.06	-0.05	-0.04
Simultaneous vs. successive presentation	0.01	-0.05	0.09	0.10
Category similarity				
within		-0.33***	-0.17*	-0.07
between		0.41***	0.22*	0.08
Complexity		0.06	0.12*	0.15**
Familiarity		0.19***	0.20***	0.20**
Type of learning material				
Naturalistic photographs			-0.22	-0.26
Artificial pictures			-0.20	-0.21
Expository texts			-0.56**	-0.73***
Mathematical tasks			-0.43*	-0.45*
Words			-0.48*	-0.63**
df	208	204	199	211

Q_B	1014.1	722.9	617.2	691.3
Q_M	11.8	88.0	105.8	98.9
$\boldsymbol{R}^2$				
Between studies	.05	.46	.50	.44

Note. Dichotomous predictors were dummy-coded with 0 or 1. Reference categories coded with 0 were published studies, between-subjects designs, nonstudent samples, short retention intervals, intentional learning, retention tests, and successive presentation. Quantitative predictors were centered around the mean. Type of learning material was dummy-coded with naturalistic paintings as the reference group. $\boldsymbol{R}^2$ within studies was 1 for all models.
The outlier model was the same as Model 3 except for the inclusion of outliers.
* $p < .05$; ** $p < .01$; *** $p < .001$

Table 4

Sensitivity Analysis Assessing Robustness of Moderators of Interleaving Effect Across different Samples.

	All Studies	Independent Studies	Students	Visual Material	Artificial Pictures
Intercept	0.44***	0.46***	0.69***	0.22*	0.11
Student samples	0.21**	0.21**		0.37***	0.14
Category similarity					
within	-0.16*	-0.12	-0.18*	-0.29**	-0.28*
between	0.20*	0.18*	0.28**	0.34***	0.38**
Complexity	0.12*	0.11*	0.11*	0.18***	0.22***
Familiarity	0.18**	0.20**	0.26***	.26**	0.14
Type of learning material					
Naturalistic photographs	-0.22	-0.35*	-0.35*	-0.12	
Artificial pictures	-0.21	-0.26*	-0.24	-0.12	
Expository texts and interactive tasks	-0.60***	-0.58***	-0.62***		
Mathematical tasks	-0.43*	-0.46**	-0.69***		
Words	-0.52*	-0.26	-0.65*		
df	205	127	154	128	52
Q_B	625.0	383.1	379.9	318.6	156.8
Q_M	101.7	88.1	121.5	81.6	22.3
$\boldsymbol{R}^2$					
Between studies	0.51	0.50	0.64	0.55	0.38

Note. Dichotomous predictors were dummy-coded with 0 or 1. Reference category was coded with 0 for nonstudent samples. Quantitative predictors were centered around the mean. Type of learning material was dummy coded-with naturalistic paintings as reference group. R^2 within studies was 1 for all models.
* $p < .05$; ** $p < .01$; *** $p < .001$

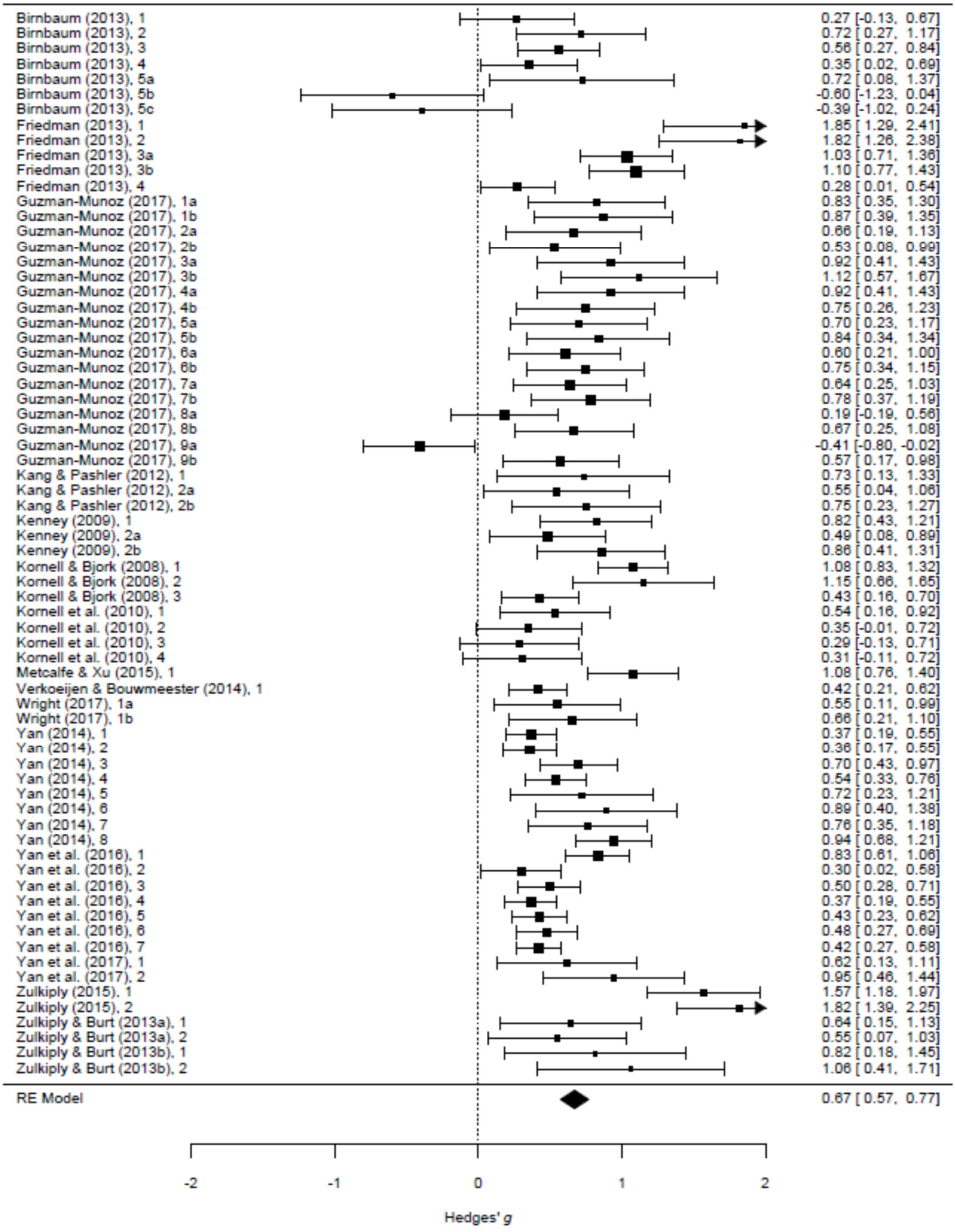

Figure 1. Forest plot of effect sizes of the interleaving effect (with 95% confidence interval) for naturalistic paintings as learning materials. The diamond shape represents the overall effect for naturalistic paintings (estimated with a random effects model).

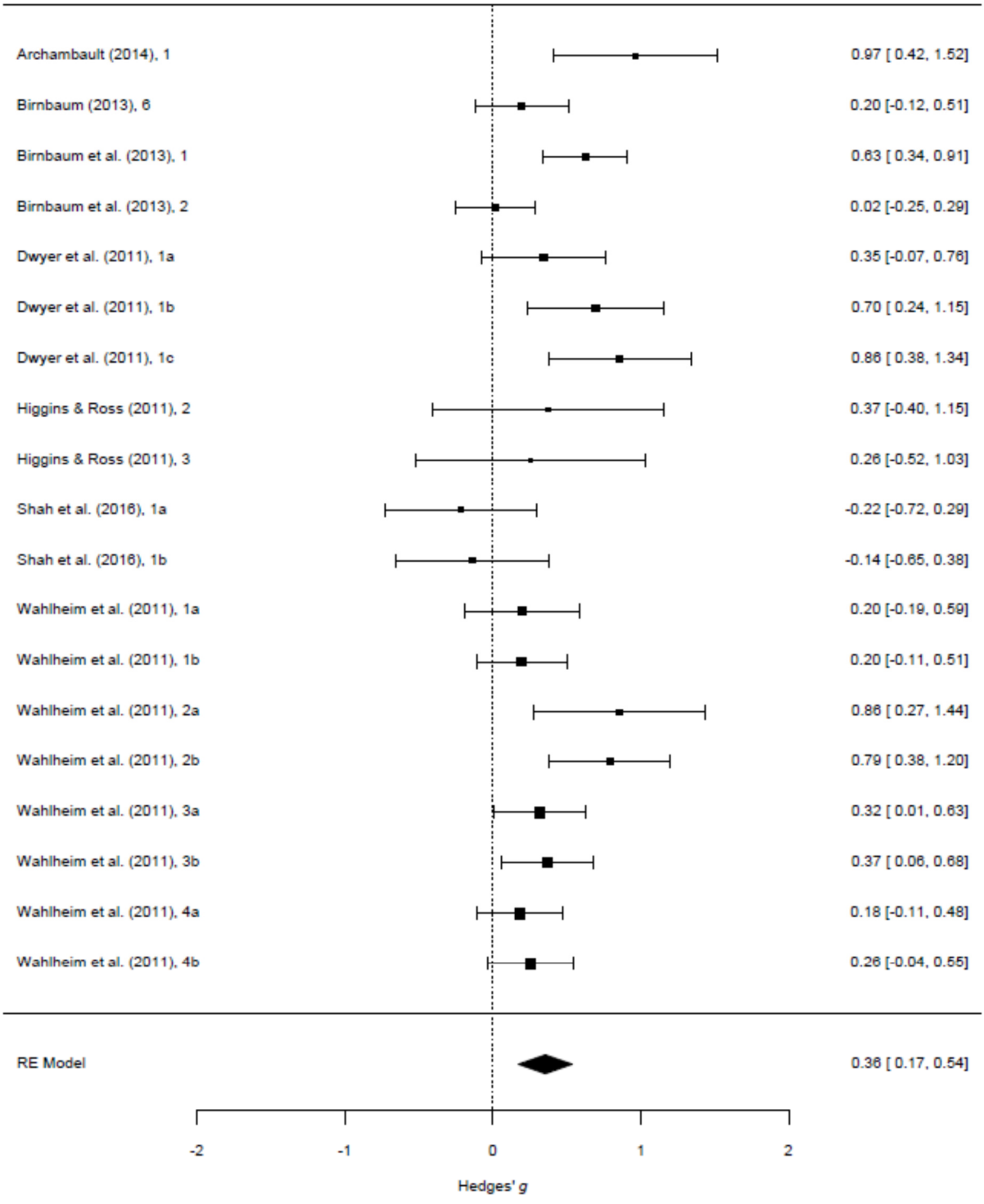

Figure 2. Forest plot of effect sizes of the interleaving effect (with 95% confidence interval) for naturalistic photographs as learning materials. The diamond shape represents the overall effect for naturalistic photographs (estimated with a random effects model).

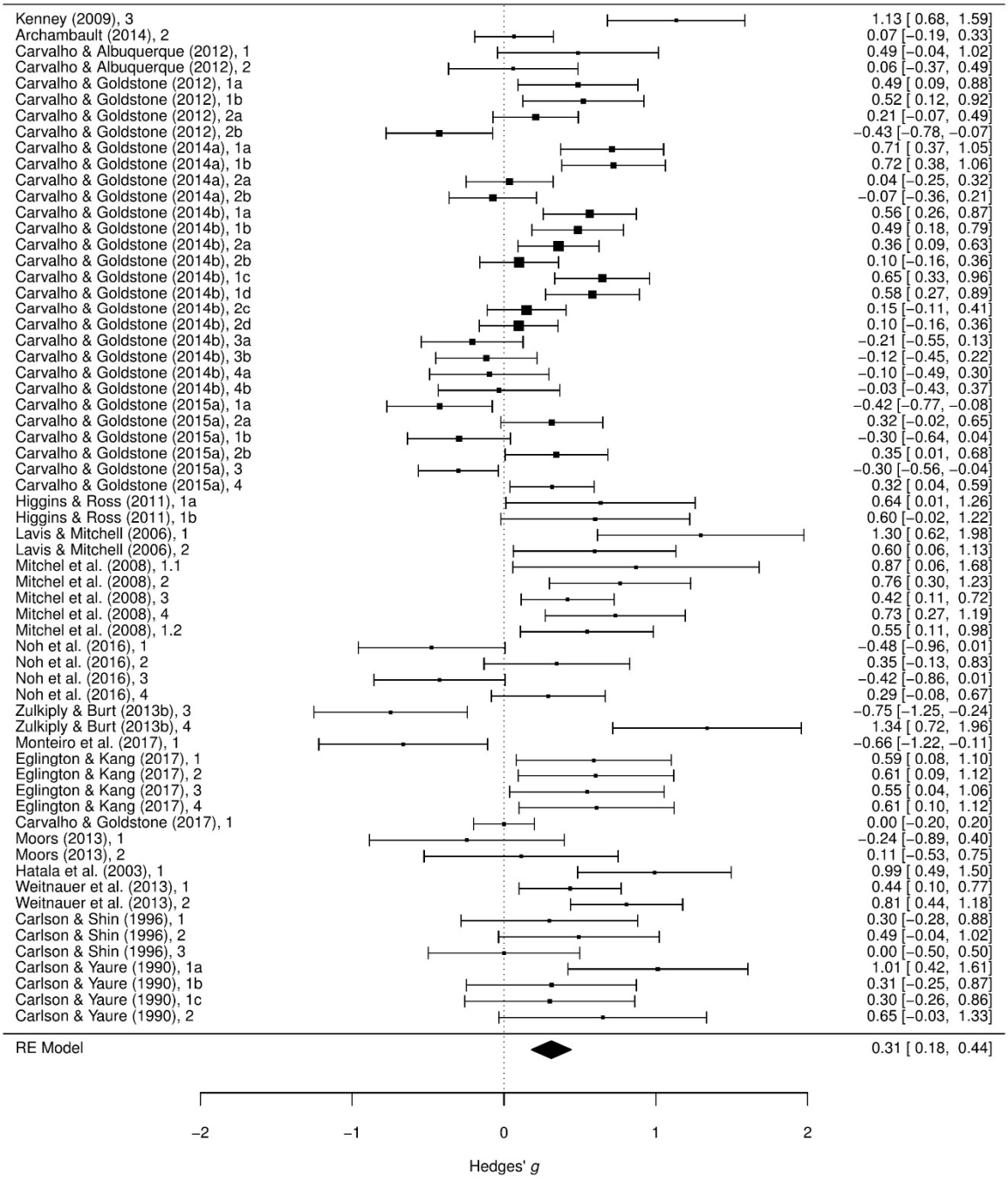

Figure 3. Forest plot of effect sizes of the interleaving effect (with 95% confidence interval) for artificial pictures as learning materials. The diamond shape represents the overall effect for artificial pictures (estimated with a random effects model).

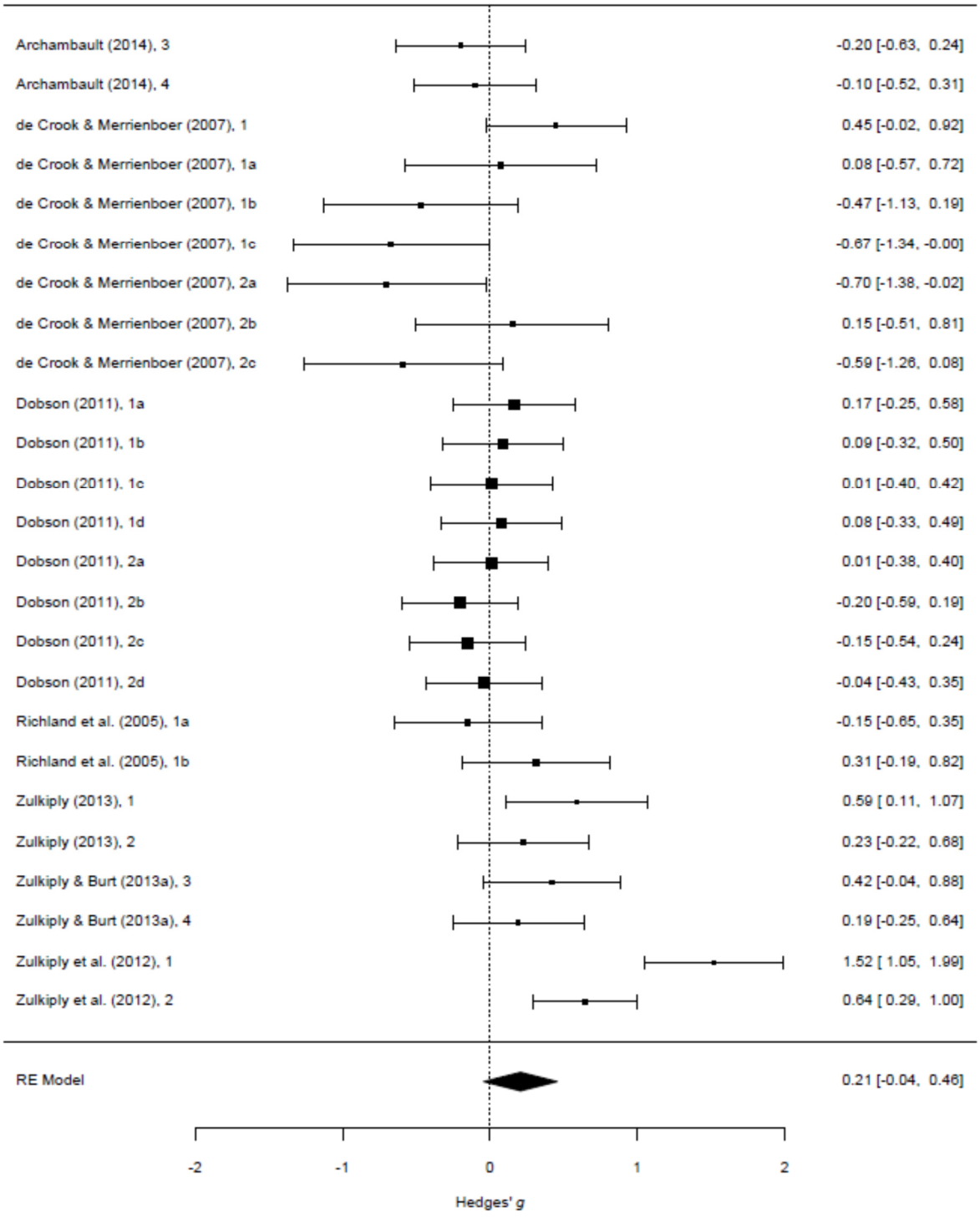

Figure 4. Forest plot of effect sizes of the interleaving effect (with 95% confidence interval) for expository texts as learning materials. The diamond shape represents the overall effect for expository texts (estimated with a random effects model).

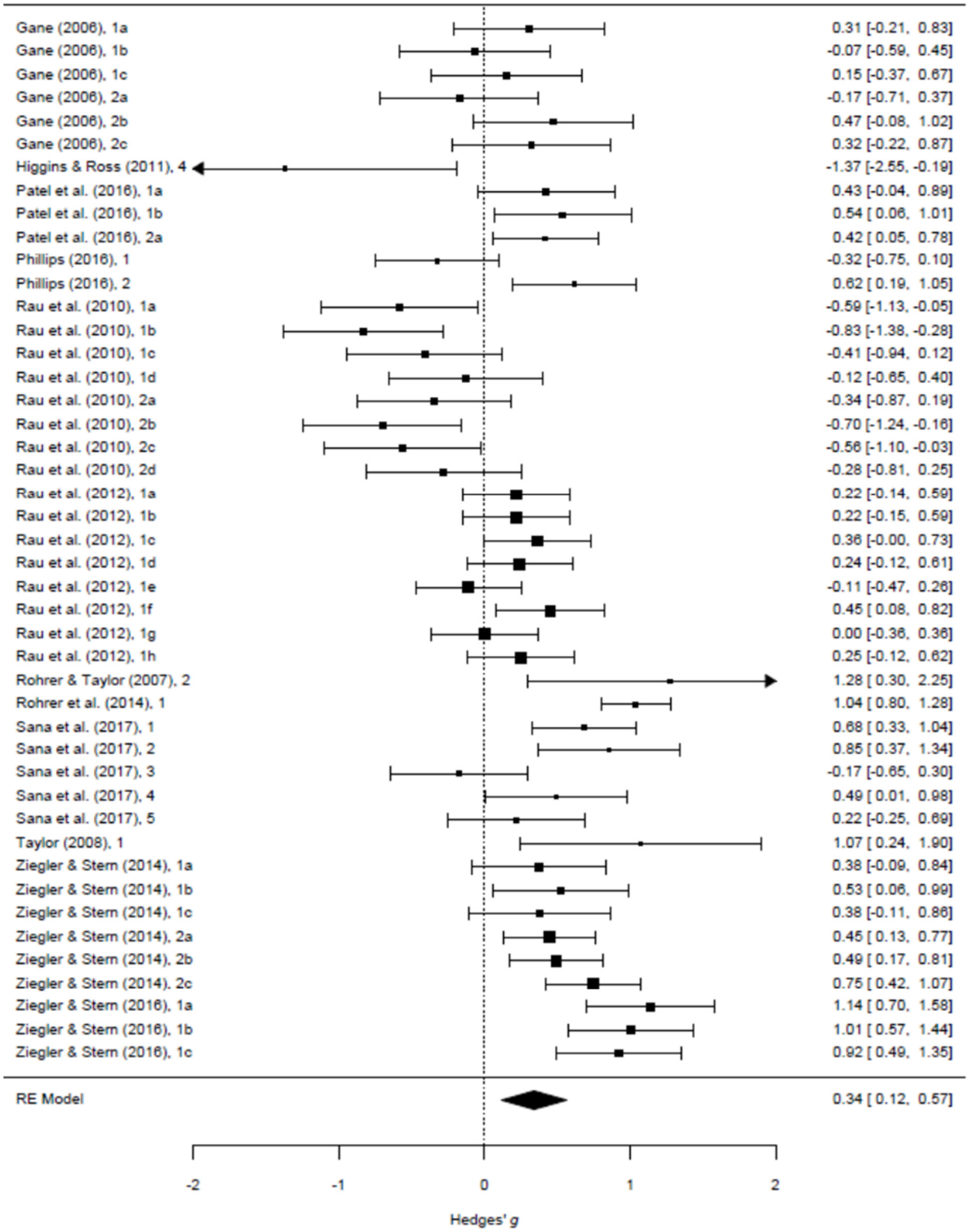

Figure 5. Forest plot of effect sizes of the interleaving effect (with 95% confidence interval) for mathematical tasks as learning materials. The diamond shape represents the overall effect for mathematical tasks (estimated with a random effects model).

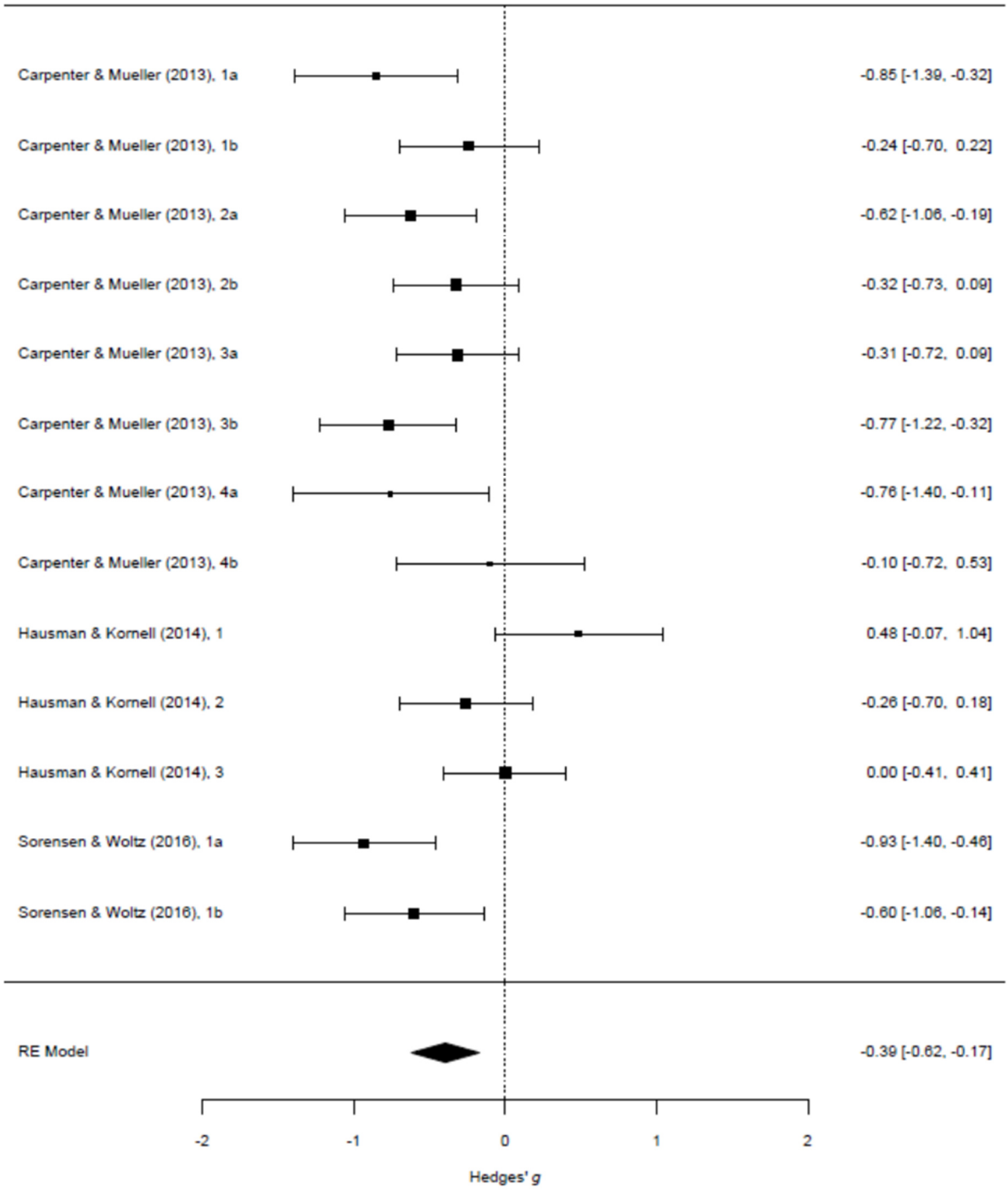

Figure 6. Forest plot of effect sizes of the interleaving effect (with 95% confidence interval) for words as learning materials. The diamond shape represents the overall effect for words (estimated with a random effects model).

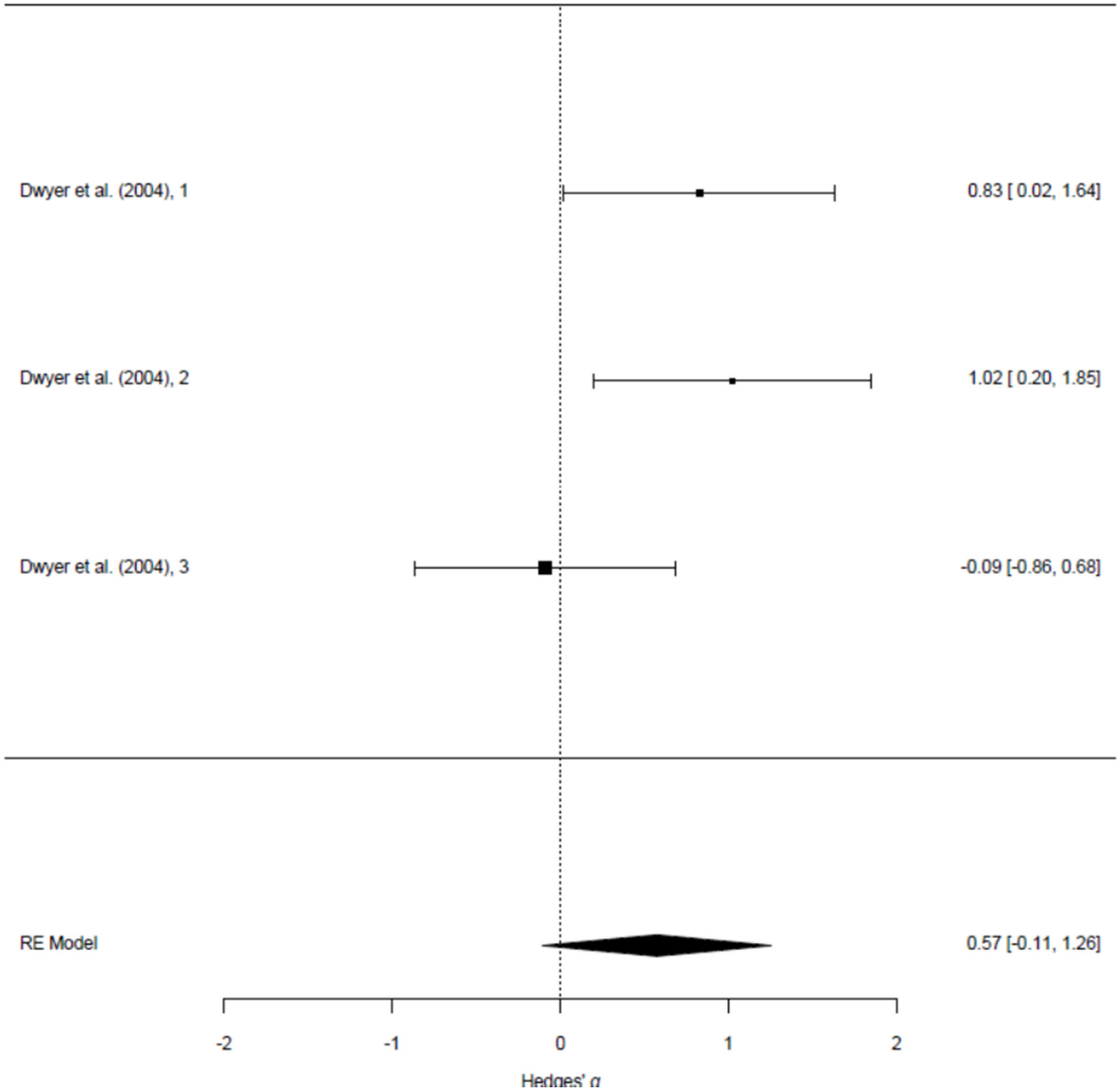

Figure 7. Forest plot of effect sizes of the interleaving effect (with 95% confidence interval) for tastes as learning materials. The diamond shape represents the overall effect for tastes (estimated with a random effects model).

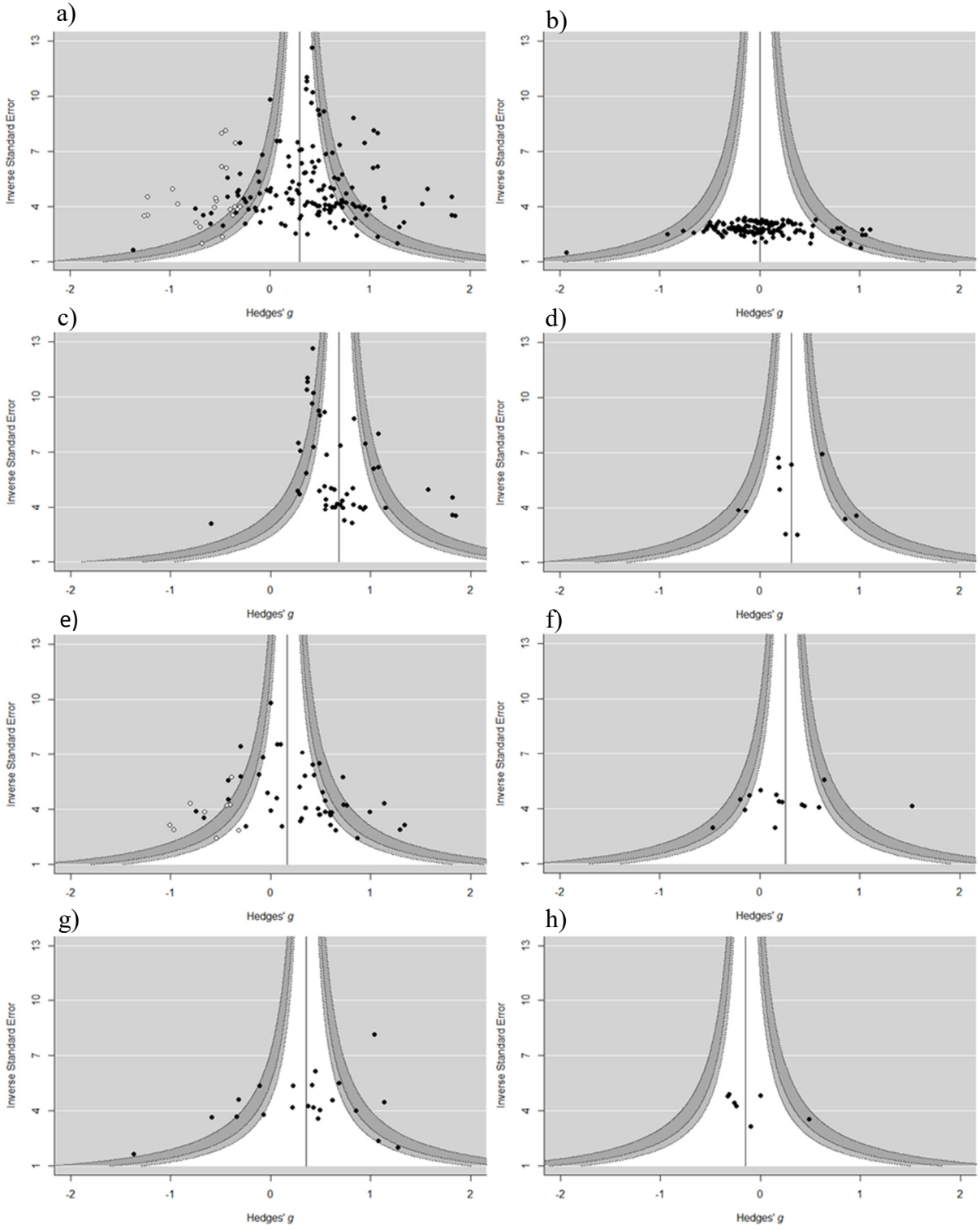

Figure 8. Counter-enhanced funnel plots for independent effect sizes. (a) Overall effect sizes estimated without moderators, (b) overall effects sizes with the moderators in the main model controlled for, and effect sizes for (c) naturalistic paintings, (d) naturalistic photographs, (e) artificial pictures, (f) expository texts, (g) mathematical tasks, and (h) words as learning materials.

Black dots represent observed effect sizes and white dots represent effect sizes imputed by the trim-and-fill procedure. The probability that effect sizes fall by chance in the particular area

are $p < .01$ in lightest gray area; $.01 < p < .05$ in the light-grey area; $.05 < p < .10$ in the dark-grey area and $.10 < p < 1.00$ in the white area.)

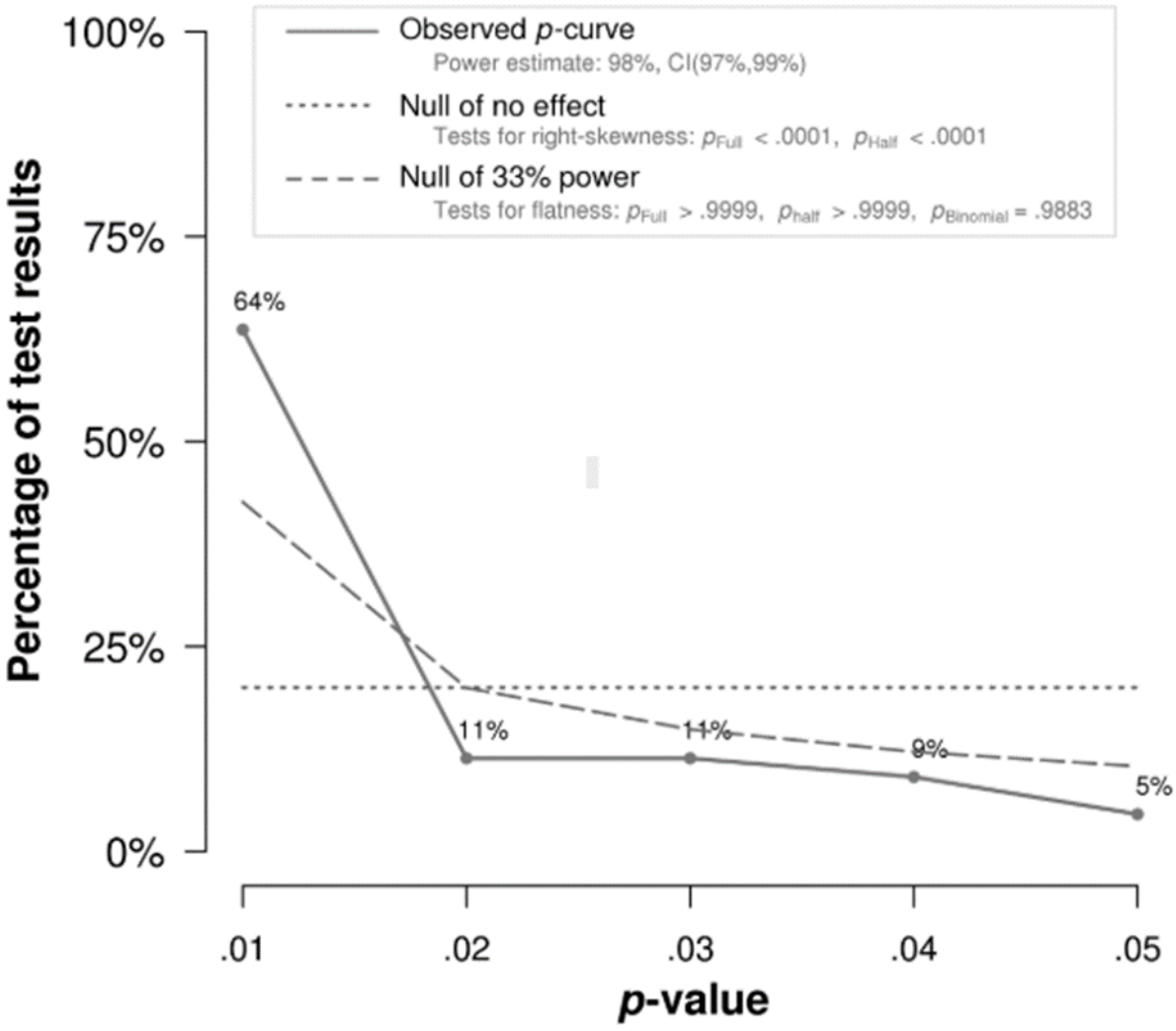

Figure 9. *P*-curve of significant (p <.05) studies included in the meta-analysis (k = 44). Nonsignificant results were not included. The solid line represents the observed significant p values. The dotted line shows the expected p-values given a Null effect. The dashed line represents the expected p-values for a set of studies with a power of .33 (for 37 studies, p < .025).

Supplements

The full data file of the meta-analysis is provided as electronic Supplement 1. The test statistics for the P-curve analysis are provided as electronic Supplement 2.