Acta Psychologica 142 (2013) 1–5

Contents lists available at SciVerse ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/ locate/actpsy

The production effect benefits performance in between-subject designs: A meta-analysis ☆

Jonathan M. Fawcett *

Dalhousie University, Canada

ARTICLE INFO

Article history:
Received 4 April 2012
Received in revised form 17 September 2012
Accepted 7 October 2012
Available online 9 November 2012

PsycINFO classification:
2300 Human Experimental Psychology
2340 Cognitive Processes
2343 Learning & Memory

Keywords:
Production effect
Meta-analysis
Between-subjects
Distinctiveness
Human memory

ABSTRACT

Producing (e.g., saying, mouthing) some items and silently reading others has been shown to result in a reliable advantage favoring retention of the produced compared to non-produced items at test. However, evidence has been mixed as to whether the benefits of production are limited to within- as opposed to between-subject designs. It has even been suggested that the within-subjects nature of the production effect may be one of its defining characteristics. Meta-analytic techniques were applied to evaluate this claim. Findings indicated a moderate effect of production on recognition memory when varied between-subjects ($g = 0.37$). This outcome suggests that the production effect is not defined as an exclusively within-subject occurrence.

1. Introduction

The notion that producing (compared to silently reading) a word could benefit memory has been around for at least four decades (e.g., Hopkins & Edwards, 1972). Since then, a great deal of research has supported this assertion using tasks ranging from speaking the word aloud to silently mouthing it (see MacLeod, Gopie, Hourihan, Neary, & Ozubko, 2010). Recently branded the production effect, the performance benefit in recognition memory for produced compared to non-produced words has been attributed to the concept of distinctiveness. That is to say that producing a word results in a production trace that can be reconstituted at test to discriminate (produced) study items from (non-produced) distractor items (e.g., Dodson & Schacter, 2001). Therefore, the distinctiveness account attributes the production effect to an interaction between the distinctive processes (i.e., production) applied at study and the retrieval strategies employed at test.

Curiously the production effect has been demonstrated almost exclusively using within- as opposed to between-subject designs. This has resulted in the presumption that the benefits afforded by production are evident only when tested in relation to other non-produced items (e.g., Ozubko & Macleod, 2010). Hourihan and Macleod (2008) further argued that the absence of a reliable between-subjects production effect provides compelling evidence against a single process account, such as one based purely on the strength of the study item in memory (for further discussion, see Ozubko & Macleod, 2010). If producing an item merely strengthened the associated memory trace (see, e.g., Wickelgren, 1969) performance should favor produced relative to non-produced items regardless of the study design. The finding that production benefits memory only relative to other non-produced items from the same session is instead most congruent with a distinctiveness account such as the one summarized above (Hourihan & Macleod, 2008).

The issue of whether the production effect is limited to within-subject designs was most recently addressed by MacLeod et al. (2010) in an article delineating the production effect and its boundary conditions. They summarized three published articles manipulating production between-subjects (Dodson & Schacter, 2001; Gathercole & Conway, 1988; Hopkins & Edwards, 1972). Of those studies, only Gathercole and Conway (1988) reported a benefit of produced compared to silently read study items. MacLeod et al. (2010) then reported

☆ The author would like to thank Dr. John Christie, Dr. Vin LoLordo, Dr. Tracy Taylor, Dr. Glen Bodner, Chelsea Quinlan and Kate Thompson for their feedback, quantitative advice and encouragement. The author would also like to thank Dr. Colin MacLeod and Dr. Kathleen Hourihan for providing access to their raw data.

* Dalhousie University, Department of Psychology, Halifax, NS, Canada B3H 4J1. Tel.: +1 902 494 3001.
E-mail address: jmfawcet@dal.ca.

0001-6918/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.actpsy.2012.10.001

two experiments of their own in which they manipulated production between-subjects and then tested memory using either a yes–no (Experiment 2) or a two-alternative forced choice (Experiment 3B) recognition task. Neither experiment found a significant effect of production, resulting in the conclusion that the production effect is indeed a within-subjects phenomenon. They speculated that the absence of a between-subjects production effect is a defining characteristic of this paradigm (see also, Hourihan & Macleod, 2008; Ozubko & Macleod, 2010; Ozubko, Gopie, & MacLeod, 2011).

Another possibility is that production does have an impact when manipulated between-subjects — it is merely very small. This hypothesis is supported by an examination of the directionality of the null findings from the between-subject studies described above. The majority of these comparisons – despite being non-significant – were still in the predicted direction. That is to say that even though the respective p-values were often above .05, performance tended to favor the produced relative to the non-produced study items. This leaves us with a simple count of the significant and nonsignificant outcomes which disagrees with the apparent reliability of the pattern observed within those comparisons. The goal of the current article was to resolve this tension by providing a brief meta-analytic evaluation of the evidence for (or against) the production effect in between-subject designs.

2. Method

2.1. Literature search

A search was conducted of the online resources Google, Google Scholar, PsycINFO, PsychARTICLES, and JSTOR using numerous combinations and variations of the keywords: produce, say, speak, aloud, mouth, read, pronounce, memory, recognition, recall, and between-subjects. Only articles containing between-subject comparisons fitting the definition of the production effect provided above were considered for inclusion. This search was conducted until July 2011 but succeeded in locating only two articles (Hopkins, Boylan, & Lincoln, 1972; Ozubko & Macleod, 2010) in addition to the four referenced above (Dodson & Schacter, 2001; Gathercole & Conway, 1988; Hopkins & Edwards, 1972; MacLeod et al., 2010). Two further unpublished studies were procured through direct communication with the authors (Major & MacLeod, 2008; Ozubko & Macleod, 2009). Therefore, the sample consisted of eight studies contributing twelve independent effect sizes which are summarized as a forest plot in Fig. 1. Articles contributing one or more effect sizes are indicated in the reference section by an asterisk (*). Data were coded for measures of yes–no recognition, two-alternative forced choice and list-discrimination as the proportion correct responses for the target items. Notably, none of the between-subject studies identified throughout the search employed recall as a dependent measure. Therefore, whereas the analyses which follow are applicable to recognition performance, caution must be used when generalizing these findings to recall performance.

2.2. Effect size calculation and analysis

Effect sizes were calculated as the standardized mean difference between the production and control groups using the escalc function from the metafor package (Viechtbauer, 2010) within R version 2.12.1 (R Development Core Team, 2010). This function employs the procedure recommended by Hedges (1982) with a correction for positive bias (see Hedges & Olkin, 1985). In most cases the group variances were either calculated from the raw data (when available) or estimated from the reported statistics. In one instance (Gathercole & Conway, 1988) only the means were available. In this case the group variances were approximated by pooling the variances from all other studies that used the same dependent measure (yes–no recognition). Importantly, sensitivity analyses demonstrate that the effects reported below are robust across a range of imputed values for the group variances within this study — and that the same pattern is evident even if this study were excluded.

Effect sizes have been calculated such that a positive value represents greater performance for produced as opposed to non-produced items. Therefore, higher (positive) effect sizes represent a larger production effect. A random-effects model (using a restricted maximum-likelihood estimator) was then fitted to the aggregate data to estimate the overall impact of production on memory performance. This model was generated using the rma function from the metafor package (see Viechtbauer, 2010).

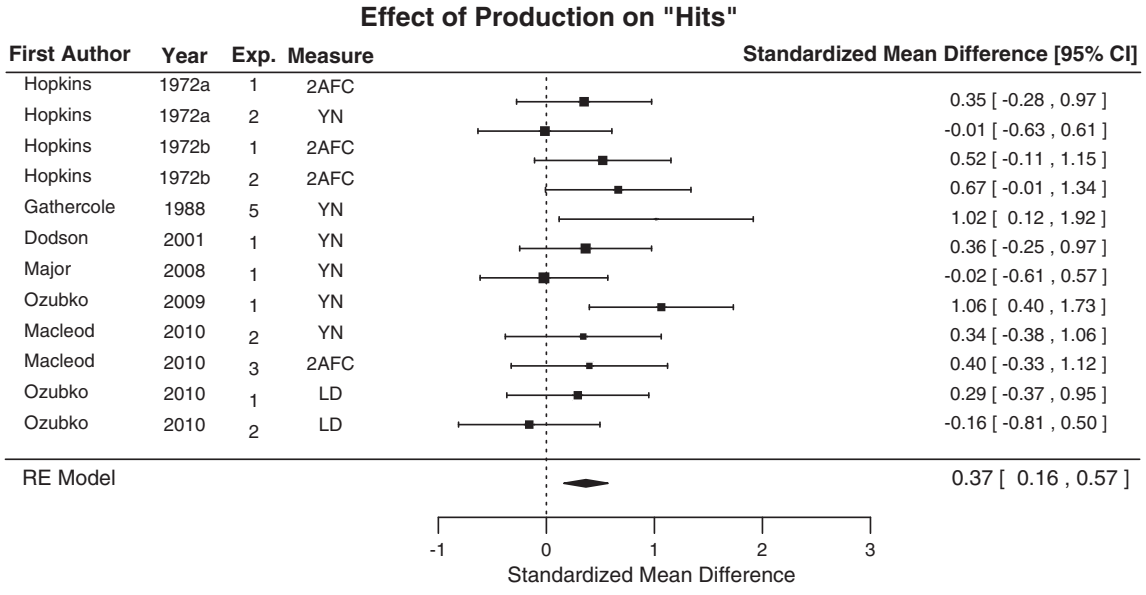

Fig. 1. Effect sizes and confidence intervals for the hits reported within each study. The polygon presented at the bottom represents the summary effect calculated using a random-effects model. Relative weight within the model is depicted by the size of the square representing the point estimate. 2AFC = two-alternative forced choice, YN = yes–no recognition, LD = list discrimination.

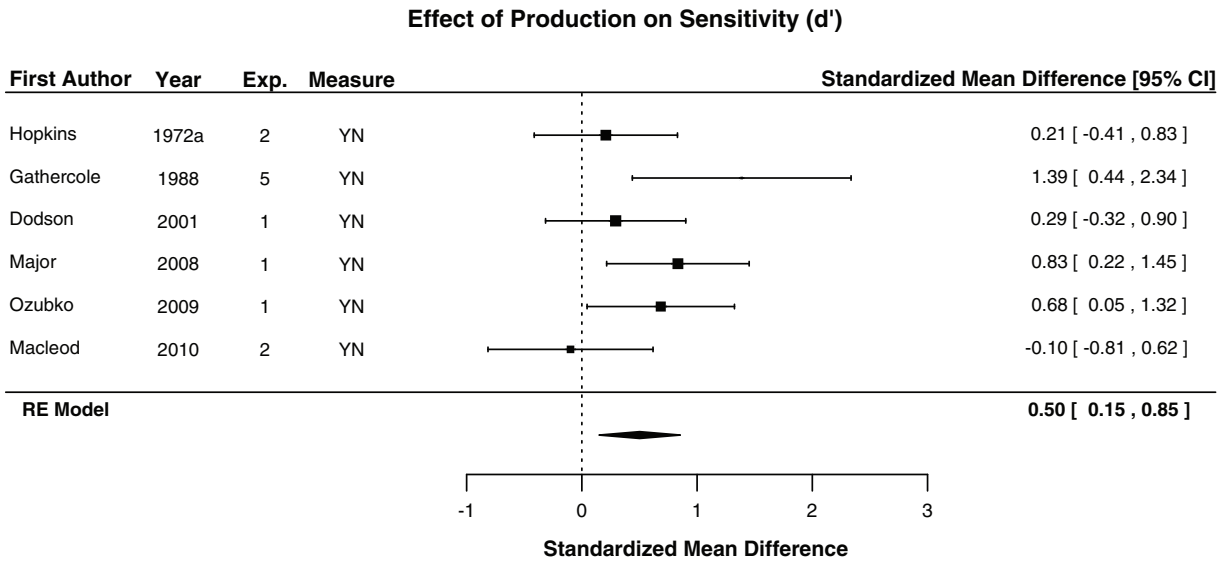

Fig. 2. Effect sizes and confidence intervals for estimated sensitivity (d′) within each study using yes–no recognition. The polygon presented at the bottom represents the summary effect calculated using a random-effects model. Relative weight within the model is depicted by the size of the square representing the point estimate.

3. Results

Results indicated a moderate effect of production, $g = 0.37$, $CI_{95\%} = [0.16, 0.57]$, as depicted by the polygon provided in Fig. 1. The current model found no evidence of heterogeneity across measures, $Q(11) = 12.85$, $p = .303$, suggesting that any variability observed within the current data is attributable to sampling error. Having established evidence of a between-subjects production effect, a conservative fail-safe N was calculated and found to suggest that twelve additional effects averaging to null ($g = 0$) are required for the current analysis to become non-significant (see Orwin, 1983). This number is large relative to the number of between-subject comparisons known to exist and when interpreting this value one should keep in mind that it is not the same as stating that twelve non-significant effects would be sufficient to eliminate the summary effect. Most of the comparisons included in this analysis failed to reach significance and yet the majority (nine of twelve) still support the presence of a production effect. Related to this point it is unlikely that the summary effect presented in Fig. 1 is attributable to publication bias favoring significant results because – after all – only two of the reported comparisons reached significance.[1]

Visual inspection of Fig. 1 reveals the effect of production to be highly consistent and robust across each of the measures included in our analyses. In fact, inclusion of the dependent measure as a moderator failed to account for any appreciable amount of heterogeneity within our effects ($p = .351$); comparing those studies using yes–no recognition to all other studies produced the same result ($p = .830$) and in fact even limiting the analysis to only the yes–no data (which is currently more common) still produces a significant effect, $g = 0.41$, $CI_{95\%} = [0.04, 0.79]$.

Recognizing that the analysis of "hits" alone does not fully encapsulate performance, a separate exploratory analysis was conducted for estimated values of sensitivity (d′). The raw data were available for very few of these experiments and insufficient information was available to estimate the variance of d′ from the aggregate measures of the remaining experiments. The studies for which raw data were available largely used yes–no recognition and therefore I have limited this analysis to studies having employed that dependent measure. Where the raw data were available, I calculated the relevant variance directly; where it was not available I estimated this value by pooling the variances from the available data. As depicted in Fig. 2, participants were significantly more sensitive to produced compared to non-produced items, $g = 0.50$, $CI_{95\%} = [0.15, 0.85]$. This analysis was robust across a range of imputed variances up to twice those observed in the experiments for which raw data were available.[2]

4. Discussion

The current study addressed whether the production effect is observed when manipulated between (as opposed to within) groups. This question was motivated in part by the recent suggestion that the absence of a between-subjects production effect is perhaps one of the defining characteristics of this phenomenon (see MacLeod et al., 2010). Counter to this proposal, a synthesis of the relevant between-subject comparisons revealed a significant effect on both hits ($g = 0.37$) and d′ ($g = 0.50$). Inspection of Fig. 1 reveals this effect to be surprisingly consistent. Despite their failure to reach significance, all but three comparisons favored memory for produced relative to non-produced items. The remaining comparisons appeared to show no effect of production (as opposed to an effect in the opposite direction). This outcome is surprising given the strong, long-standing sentiment that between-subject manipulations of production are ineffective.

The realization of a between-subjects production effect requires us to reconsider our evaluation of a single-process account. For example, production may increase the strength of the relevant memory trace and therefore improve performance for produced items at test. Such an account does not presuppose any particular retrieval

[1] When the fail-safe N is calculated using the methods proposed by Rosenthal (1979) or Rosenburg (2005) the total number of comparisons averaging to *null* required for the current analysis to become non-significant would be 55 or 32, respectively.

[2] A comparable analysis of the false alarms revealed a significant effect favoring fewer false alarms for produced relative to non-produced items, $g = -0.23$, $CI_{95\%} = [-0.44, -0.01]$. However, this finding was less robust than the analysis of hits described above and it failed to reach significance when the analysis was limited to only the yes–no data although the pattern was in the same direction, $g = -0.11$, $CI_{95\%} = [-0.47, 0.25]$. Likewise, an exploratory analysis of response bias (C) for the yes–no data revealed a non-significant trend favoring a more liberal response bias for produced compared to non-produced items, $g = -0.24$, $CI_{95\%} = [-0.68, 0.20]$.

heuristic (see Dodson & Schacter, 2001) and therefore may be considered to involve only a single process. The benefits of production may be more robust in the context of a within- as opposed to between-subjects design because produced items attract more attention and further encoding relative to the non-produced items when inter-mixed (for a summary of this view, see McDaniel, Dornburg, & Guynn, 2005). This would provide a compelling interpretation of the established within-subject production effect in relation to the between-subject production effect summarized above. Even so, such a single process account is still unable to address other findings within this literature. For instance, Ozubko and Macleod (2010) recently demonstrated that producing the distractor list either before or after the study phase (in which some portion of study items were also produced) eliminated the production effect within a list-discrimination task. Presentation of the distractor list with instructions to silently read each item still resulted in a production effect for the study items. They attributed this outcome to the removal of the production trace as a distinctive retrieval cue at test. Overall, their findings are difficult to reconcile with a single process account unless production of the distractor list were thought to somehow mitigate the degree to which produced study items attracted additional encoding. Even so, this perspective merits further empirical consideration.

The current outcome also requires us to reconsider the role of distinctiveness in this paradigm. Some theorists have referenced the absence of a between-subjects production effect as compelling evidence that the production effect is attributable to distinctiveness (e.g., Hourihan & Macleod, 2008; Ozubko & Macleod, 2010). Therefore one might wonder if the realization of a between-subjects production effect should work against the distinctiveness account instead. This is not necessarily the case. The distinctiveness account summarized above would expect the magnitude of the production effect to vary according to (a) the degree to which access to a production trace is capable of discriminating "old" study items from "new" distractor items, and, (b) the degree to which access to a production trace is used to discriminate "old" study items from "new" distractor items. In other words, both the utility and the application of the production trace must be considered. As mentioned earlier, having participants produce the distractor list (in addition to some portion of the study list) has been shown to eliminate the production effect (Ozubko & Macleod, 2010). This is because access to a production trace is no longer useful — even if it were applied to guide recognition. Even when the presence of a production trace could be used to discriminate study items from distractor items the potential of this information is only actualized to the extent that the necessary heuristic is applied. Production is more likely to be perceived as a useful heuristic when manipulated within-subjects because this design juxtaposes the produced and non-produced items against each other. It is possible that whereas production could benefit performance when manipulated between-subjects these benefits are less robust because participants are less likely to use production to guide recognition. It follows that should participants be encouraged to utilize production as a retrieval cue the between-subjects production effect would become more robust.

It is also important to consider how the current findings impact the status of the production effect relative to other effects attributed to distinctiveness (for discussion, see Hunt & Worthen, 2006). Schmidt (1991) provides a taxonomic framework organizing these effects into one of four theoretical categories. Of relevance to the current discussion are those of primary and secondary distinctiveness. The former category is reserved for effects limited to within-subject manipulations. The latter category includes effects attainable using between- as well as within-subject manipulations. McDaniel and Geraci (2006) have argued that effects categorized as arising from primary distinctiveness are attributable to retrieval processes whereas those categorized as arising from secondary distinctiveness are attributable to both encoding and retrieval processes. The realization that the production effect is attainable between-subjects shifts its categorization (according to the flowchart provided by Schmidt, 1991) from primary to secondary distinctiveness — and therefore from a purely retrieval-based account to an account including a mixture of encoding and retrieval processes. McDaniel and Geraci (2006) explicitly assume that effects arising from primary distinctiveness involve the application of similar or even identical encoding processes to the distinct and non-distinct items. This is certainly not the case in the production paradigm because the encoding processes applied to distinct (produced) and non-distinct (non-produced) items are precisely the intended manipulation. The theoretical framework associated with secondary distinctiveness is more consistent with the processes currently thought to drive the benefits of production. The production effect is currently thought to be an emergent property of the interaction between processes applied at encoding (production) and strategies applied at retrieval (distinctiveness heuristic). While it is possible to imagine production strengthening the relevant memory trace to some degree, the distinctive information provided by the application of these encoding processes provides a retrieval cue capable of magnifying performance for produced compared to non-produced items.

Recognizing that both encoding and retrieval processes contribute to the production effect also provides insight as to why this effect might be relatively less robust in between-subjects designs: How the production instruction is presented may independently influence the manner in which the study items are encoded and retrieved. For retrieval processes, it is possible that the presentation of both produced and non-produced items at study provides a context in which the production trace serves as a more potent retrieval cue, in part because fewer items are associated with a recent production trace. This is similar to the earlier point that exposure to both produced and non-produced items could encourage the adoption of a production-based retrieval strategy. For encoding processes, participants might allocate additional or relatively more efficient rehearsal to the produced items in a within-subjects design because they "stand-out" in relation to the non-produced items. Of course, another concern might be that instead of encoding the produced items more efficiently, participants might also encode the non-produced items less efficiently than if those items had been presented in a pure list: This has been referred to as the "lazy reading" hypothesis (see MacLeod et al., 2010; cf. Begg & Snider, 1987). MacLeod et al. (2010) tested whether lazy reading of the non-produced items could account for the production effect by instructing participants to make a semantic judgment for each item prior to receiving the within-subjects production instruction. A production effect was still observed, resulting in the conclusion that lazy reading of the non-produced items is not a sufficient explanation. Nonetheless, the semantic orienting task resulted in a production effect that was of lesser magnitude ($M_{Produced} - M_{Non\text{-}produced} = .06$ for Experiment 8) compared to the other within-subjects experiments reported in the same publication ($M_{Produced} - M_{Non\text{-}produced} = .16$, .13, .17, and .19 for Experiments 1, 3A, 5, and 6, respectively). In fact, the semantic orienting task resulted in a production effect of numerically comparable magnitude to the between-subject experiments reported by those authors ($M_{Produced} - M_{Non\text{-}produced} = .06$ and .03 for Experiments 2 and 3B, respectively).

To summarize the above, one might speculate that the within-subject production effect arises from (a) differences in the degree to which participants encode or rehearse the produced and non-produced items at study, either through greater emphasis on the produced items (what might be thought of as the "lively reading" hypothesis) or lesser emphasis on the non-produced items (the "lazy reading" hypothesis) relative to a pure produced or non-produced list; and, (b) a production-based retrieval strategy such that access to a production trace is used to discriminate between targets and distractors at test. In a typical within-subjects experiment, these processes aggregate to produce a robust difference in recognition memory favoring produced over non-produced items; however, in a between-

subjects experiment only the latter process contributes to the effect. The fact that MacLeod et al. (2010) observed a numerically smaller production effect when they attempted to equalize study-phase processing across conditions is consistent with mitigating the former but not the later process, resulting in an effect of similar magnitude to their between-subjects experiments (e.g., Experiment 2).

On a final methodological note the goal of the current article was to resolve the tension between the number of significant as opposed to non-significant between-subject comparisons of the production effect and the apparent reliability of the pattern within the same comparisons. The act of counting significant and non-significant results (as opposed to synthesizing the effects) is known as vote counting (Hedges & Olkin, 1980; cf., Light & Smith, 1971).[3] Whichever side receives the plurality of votes dictates the acceptance or rejection of the relevant finding. As noted by Hedges and Olkin (1980), the statistical power of such a procedure tends to be lower than that of the individual studies it counts. While it might seem reasonable to suspect that low statistical power is an issue only when dealing with relatively few studies (as in the current case), the statistical power of vote counting may even decrease as the number of studies increases (Hedges & Olkin, 1980). Therefore we are left with Borenstein's (2000) advice that we should interpret the absence of a significant effect as "more information is required" as opposed to "no effect exists" — at least until sufficient literature has emerged to support a more rigorous analysis.

In conclusion, the current article evaluated recent claims that the production effect is limited to within-subject designs (e.g., Hourihan & Macleod, 2008; MacLeod et al., 2010; Ozubko & Macleod, 2010; Ozubko et al., 2011). This was not the case. Instead, a reliable effect of production was evident throughout all but three comparisons (see Fig. 1). Even so, although the production effect in recognition memory does not appear to be characterized by the absence of a between-subjects effect it may be characterized as demonstrating a relatively more robust within-subjects effect. Therefore, the weight of the current analysis should be viewed as qualifying as opposed to rejecting this notion. Further research is required exploring the mechanisms underlying the between- and within-subjects production effect. In particular the current findings raise questions as to the relative contributions of encoding and retrieval processes. It is my hope that this work will spark further investigations evaluating this possibility.

References

Begg, I., & Snider, A. (1987). The generation effect: Evidence for generalized inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(4), 553–563. http://dx.doi.org/10.1037//0278-7393.13.4.553.

Borenstein, M. (2000). The shift from significance testing to effect size estimation. In A. S. Bellack, & M. Hersen (Eds.), *Comprehensive Clinical Psychology*, *Volume 3*. (pp. 313–349)Oxford, UK: Pergamon.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). *Introduction to meta-analysis.* Hoboken, NJ: John Wiley & Sons.

Bushman, B. J., & Wang, M. C. (2009). Vote-counting procedures in meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 207–220). (Second Edition). New York: Russell Sage Foundation.

*Dodson, C. S., & Schacter, D. L. (2001). "If I had said it I would have remembered it": Reducing false memories with a distinctiveness heuristic. *Psychonomic Bulletin & Review*, *8*(1), 155–161.

Fawcett, J. M., Russell, E. J., Peace, K. A., & Christie, J. (2011). Of guns and geese: a meta-analytic review of the "weapon focus" literature. *Psychology, Crime & Law*. http://dx.doi.org/10.1080/1068316X.2011.599325.

*Gathercole, S. E., & Conway, M. a (1988). Exploring long-term modality effects: Vocalization leads to best retention. *Memory & Cognition*, *16*(2), 110–119 Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3352516

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, *92*(2), 490–499. http://dx.doi.org/10.1037/0033-2909.92.2.490.

Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin*, *88(2)*. (pp. 359–369): American Psychological Association. http://dx.doi.org/10.1037//0033-2909.88.2.359.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* San Diego, CA: Academic Press.

*Hopkins, R., Boylan, R., & Lincoln, G. L. (1972). Pronunciation and apparent frequency. *Journal of Verbal Learning and Verbal Behavior*, *11*, 105–113.

*Hopkins, R., & Edwards, R. (1972). Pronunciation effects in recognition memory. *Journal of Verbal Learning and Verbal Behavior*, *11*(4), 534–537. http://dx.doi.org/10.1016/S0022-5371(72)80036-7.

Hourihan, K. L., & Macleod, C. M. (2008). Directed forgetting meets the production effect: distinctive processing is resistant to intentional forgetting. *Canadian Journal of Experimental Psychology*, *62*(4), 242–246. http://dx.doi.org/10.1037/1196-1961.62.4.242.

Hunt, R. R., & Worthen, J. B. (2006). *Distinctiveness and memory.* New York: Oxford University Press.

Light, R. J., & Smith, P. V. (1971). Accumulating evidence: Procedures for resolving contradictions among different research studies. *Harvard Educational Review*, *41*(4), 429–471.

*MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(3), 671–685. http://dx.doi.org/10.1037/a0018785.

*Major, J. C., & MacLeod, C. M. (2008). *Recognition test performance in production*. Unpublished raw data.

McDaniel, M. A., Dornburg, C. C., & Guynn, M. J. (2005). Disentangling encoding versus retrieval explanations of the bizarreness effect: Implications for distinctiveness. *Memory & Cognition*, *33*(2), 270–279.

McDaniel, M. A., & Geraci, L. (2006). Encoding and retrieval processes in distinctiveness: Toward an integrative framework. In R. R. Hunt, & J. B. Worthen (Eds.), *Distinctiveness and memory* (pp. 65–88). New York: Oxford University Press.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, *8*(2), 157–159.

Ozubko, J. D., Gopie, N., & MacLeod, C. M. (2011). Production benefits both recollection and familiarity. *Memory & Cognition*. http://dx.doi.org/10.3758/s13421-011-0165-1.

*Ozubko, J. D., & Macleod, C. M. (2009). *Recognition test performance in production*. Unpublished raw data.

*Ozubko, J. D., & Macleod, C. M. (2010). The production effect in memory: Evidence that distinctiveness underlies the benefit. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1543–1547. http://dx.doi.org/10.1037/a0020604.

R Development Core Team (2010). *R: A language and environment for statistical computing. Vienna, Austria.* 3-900051-07-0.

Rosenberg, M. S. (2005). The file-drawer problem revisited: A general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution*, *59*, 464–468.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Schmidt, S. R. (1991). Can we have a distinctive theory of memory? *Memory & Cognition*, *19*(6), 523–542 Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1758300

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Wickelgren, W. A. (1969). Associative strength theory of recognition memory for pitch. *Journal of Mathematical Psychology*, *6*, 13–61.

[3] For further discussion of the dangers of statistical vote counting see Borenstein, Hedges, Higgins, and Rothstein (2010, Chapter 28). For a similar discussion of vote counting in relation to a recent meta-analysis of the *weapon focus* literature, see Fawcett, Russell, Peace, and Christie (2011). Finally, for discussion of more sophisticated vote counting procedures and their applications see Bushman and Wang (2009).