A Meta-Analytic Review of the Distribution of Practice Effect: Now You See It, Now You Don’t

John J. Donovan and David J. Radosevich
University at Albany, State University of New York

The present review examined the relationship between conditions of massed practice and spaced practice with respect to task performance. A meta-analysis of 63 studies with 112 effect sizes yielded an overall mean weighted effect size of 0.46, indicating that individuals in spaced practice conditions performed significantly higher than those in massed practice conditions. Subsequent analyses, however, suggested that the nature of the task being practiced, the intertrial time interval, and the interaction between these two variables significantly moderated the relationship between practice conditions and performance. In addition, significantly higher effect sizes were found in studies with low methodological rigor as compared with those studies higher in rigor. Directions for future research and applications of the findings are discussed.

The notion that spaced practice conditions are superior to massed practice conditions in a variety of situations (i.e., the distribution of practice effect) has become one of the more strongly accepted ideas within the task performance and learning literatures (Baldwin & Ford, 1988). Massed practice conditions are those in which individuals practice a task continuously without rest, while spaced practice conditions are those in which individuals are given rest intervals within the practice session. Although research exploring the distribution of practice effect has been conducted for over one century, there remain several gaps in the research literature which limit our ability to make strong conclusions regarding this effect. For example, much of the literature has involved simple motor tasks where it is clear that spaced practice is superior to massed practice (Lee & Genovese, 1988). Research investigating the effects of spaced practice on the acquisition and retention of verbal skills and other nonmotor tasks, however, is not as abundant or conclusive (Goldstein, 1993). In addition, numerous studies have failed to produce the distribution of practice effect, raising the possibility of boundary conditions within which this effect operates. Despite these apparent gaps in our knowledge, the idea of spaced practice superiority continues to be widely accepted by researchers. The faith placed in spaced practice by these individuals may be due, in part, to their reliance on what they consider to be “common knowledge,” rather than on a careful consideration of the empirical findings. As noted by others (e.g., Eagly & Wood, 1994; Miller & Pollock, 1994), blind acceptance of common knowledge can be both misleading and counterproductive. In light of this situation, the purpose of the present article was to provide a much needed meta-analytic review of this area, with the hope of not only providing an estimate of the magnitude of the distribution of practice effect, but also identifying potential boundary conditions that influence this phenomenon.

It is important to note that although the research that has been conducted in this area has come almost completely from the educational and classroom settings, these findings have potentially important implications for both the design and the implementation of organizational training programs. Since the importance and frequency of training within organizational settings appears to be rapidly increasing (Goldstein, 1993), and the resources available for such programs are often scarce (Canon-Bowers, Rhodenizer, Salas, & Bowers, 1998), the potential gain from a review on the distribution of practice effect can be great. More specifically, a meta-analytic review that delineates the conditions under which the distribution of practice effect is strongest and/or weakest (e.g., task types, spacing between practice sessions) will provide training development specialists with specific information that can be used to guide the proper and efficient design of organizational training programs. For example, the simple knowledge that spaced practice enhances learning on a given task type as compared with
massed practice conditions would enable training developers to design their training program in a manner that allows trainees some degree of rest between practice sessions. In addition, if the training developer were aware of an "optimal" rest period for this task, they could then further structure their training program so as to achieve greater levels of trainee learning. Clearly, knowledge of such factors would allow for the development of more effective and efficient training programs.

Overview of Research on the Distribution of Practice

Research on the distribution of practice peaked around the middle of the century, with a considerable number of studies published from the early 1950s through the late 1970s. Since then, research interest has declined sharply as evidenced by the small number of articles on this topic appearing over the last fifteen years. However, theoretical and practical interest in the potential applications of this research has remained strong both within the educational literature (e.g., Adams, 1987; Dempster, 1988) and in the organizational training literature (e.g., Goldstein, 1993; Schneier, Russell, Beatty, & Baird, 1994). Although an exhaustive qualitative review of the literature is beyond the scope of the present article (for such a review, see Dempster, 1988), we feel it is important to highlight three main characteristics of the published research as they pertain to the present meta-analytic review.

First, the majority of research conducted within this area has focused on the learning and performance of simple motor tasks. This research has demonstrated that spaced practice conditions are superior to massed practice conditions, but we are still left wondering if this finding will generalize to other more complex, cognitive tasks. Although it is commonly proposed that spaced practice conditions will continue to demonstrate their superiority for complex, cognitive tasks, it has been suggested that the effects of learning for verbal tasks are much more complex than those for motor tasks. Therefore, the effects of distribution of practice may not be as substantial for such cognitive tasks (Goldstein, 1993). Others have suggested that spaced practice may actually hinder task acquisition and retention for nonmotor tasks. Specifically, shorter rest intervals may enhance performance when the learner is likely to forget critical responses or error tendencies are high (DeCecco, 1968). However, although hypotheses concerning the effects of distribution of practice on nonmotor tasks have been proposed, we have yet to see a convincing amount of research that either supports or refutes such claims.

A second characteristic of past research that is somewhat troubling concerns the apparent lack of concern for the identification of potential boundary conditions that limit or constrain the effectiveness of spaced practice conditions. Past research has generally focused on either confirming or disconfirming the superiority of spaced practice rather than attempting to understand what factors either (a) optimize or magnify this effect or (b) constrain, negate, or possibly reverse this effect. Clearly, identification of such factors could have important practical implications for organizational training, by not only enhancing the trainee learning that takes place, but also preventing the organization from spending resources on the development of inferior training programs.

A final area of concern relates to the conceptualizations of performance used by researchers. In most studies, task performance has been conceptualized and measured as performance immediately following the end of the practice sessions (i.e., acquisition performance). Very few studies have examined task performance after some period of time has elapsed since the completion of the practice conditions (i.e., retention performance). Further, those researchers who have examined retention performance have varied widely in their operationalizations of retention performance. This lack of structured and consistent empirical examination of retention performance has created considerable confusion. Although researchers generally agree that massed practice is inferior to spaced practice for motor skill acquisition, the effects of the distribution of practice on motor skill retention are much less clear (e.g., Magill, 1985; Sage, 1984). Given that the focus of most organizational training programs is on the application of the skills or knowledge learned at some later time in a potentially different environment (i.e., transfer of training), this omission is clearly an important oversight that must be remedied if the findings on distribution of practice are to be of practical use.

Previous Meta-Analytic Review

Lee and Genovese (1988) conducted an initial meta-analysis examining the effects of distribution of practice on motor skills. They reported that distributed practice not only enhanced acquisition of motor skills compared to massed practice ($d = 0.96$), but also resulted in greater retention than massed practice conditions ($d = 0.53$). However, Newell, Antoniou, and Carlton (1988) suggested that conclusions drawn from this meta-analysis regarding practice effects on retention are premature due to the ambiguous definitions of acquisition and retention that were used. Additionally, Lee and Genovese did not examine the effects of potentially important factors (e.g., task type) that may either constrain or magnify the effects of spaced practice. Clearly, an examination of such factors is necessary before conclusive statements can be made.

Consequently, the objective of the present meta-analysis was not only to provide an overall meta-analytic examination of the distribution of practice effect, but also to provide estimates of the magnitude of the distribution of practice...
effect under several different training scenarios (i.e., different task types, different intertrial intervals, different means of measuring performance). By providing an examination of these various factors that may either increase or reduce the magnitude of the distribution of practice effect, we hope to provide training development researchers and practitioners with specific information that would be of some practical use in designing more effective training programs.

Method

Literature Search

Four separate methods were used to locate appropriate studies for use in the present meta-analysis. First, a computer based literature search was conducted in PsycLit (1974–1996), ERIC (1966–1996), and Dissertation Abstracts (1981–1996) using the keywords massed practice, spaced practice, and distributed practice. Second, the reference list of the previous meta-analysis (Lee & Genovese, 1988) was reviewed for articles to include in the present meta-analysis. Third, a manual search was conducted in two journals: Journal of Experimental Psychology (1913–1996) and Research Quarterly (1969–1996). Finally, a citation search was conducted in which the reference sections from previously gathered articles were examined to identify any studies which may have been missed by earlier search methods.

Criteria for Inclusion

In order for a study to be included in the present quantitative review, several criteria had to be met. First, there had to be the presence of both a massed- and a spaced-practice condition within each study. Because the effect size of interest for the current review was the standardized difference (d) between spaced-practice conditions and massed-practice conditions, studies including only one of the conditions were necessarily excluded from consideration. Massed-practice conditions were defined as those in which subjects received continuous practice on the task of interest with no intertrial interval (i.e., no rest) between practice sessions except when necessary for reasons such as retesting equipment and recording performance scores. Spaced-practice conditions were defined as those in which subjects received practice sessions separated by a time interval beyond that necessary for the practical reasons mentioned above.

The second criterion for inclusion was that the study had to involve the acquisition of some skill or knowledge base on the part of the participants. Additionally, the dependent variable for the study had to be performance (either acquisition or retention) on a task utilizing the previously learned skill or knowledge. Studies using unrelated dependent variables (e.g., training satisfaction, willingness to participate in future training) were discarded. Finally, studies using nonhuman, child, or mentally disabled individuals as participants were excluded from the analysis in the interest of making the results of the present meta-analysis as generalizable as possible to a typical training setting in the workplace. A total of 63 studies with 112 effect sizes satisfied these inclusionary criteria, which represents a sizable increase in effect sizes relative to the 52 effect sizes utilized by Lee and Genovese (1988).

Coding of Potential Moderators and Study Characteristics

For the present meta-analysis, four study characteristics were coded as potential moderators of the distribution of practice effect. Each of the variables was independently coded by the authors. These ratings were compared to identify any disparities that had occurred concerning the coding for a particular study. A high degree of initial agreement was obtained (96%), with divergent ratings discussed until an appropriate solution was reached.

Acquisition versus retention performance. A dichotomous rating was created indicating whether the study had measured the subjects’ task performance in terms of acquisition or retention. Acquisition performance was defined as task performance that immediately followed the practice sessions. Retention performance was defined as task performance that was separated from the practice conditions by at least one day’s time (i.e., 24 hr).

Methodological rigor. The methodological rigor of the study was assessed using a nine-item checklist similar to that of Terpstra (1981). The nine items reflected study characteristics considered to be indicative of the rigor with which the study was carried out (Chalmers et al., 1987; Cook & Campbell, 1979; Kerlinger, 1986). Higher rigor scores indicated a more methodologically rigorous study. This nine item checklist is presented in the Appendix. Although the highest possible rigor score was nine points, the highest score actually obtained from the studies under examination was six (n = 2), giving the reader a realistic indication of the general level of rigor for this research area. Due to the small

1 It should be noted that this increase in effect sizes (112 vs. 52) was accomplished with the addition of a relatively small number of studies (16). For many articles included in the present review, enough data were reported to compute multiple effect sizes. These multiple effect sizes resulted from: (a) comparison of more than one spaced practice condition with the massed practice condition, (b) data on more than one performance-based dependent variable, and (c) two or more independent studies being reported within the framework of a single article. For the present review, the second and third sources were allowed to contribute additional effect sizes to the meta-analysis, while the first source was not. In light of the fact that the present review allowed studies to contribute multiple effect sizes for all analyses, while Lee and Genovese (1988) allowed each study to contribute only one effect size to their analyses, this increase in effect sizes is a function of the addition of new studies as well as the addition of new effect sizes from previously used studies.

2 While some may consider retention performance as performance that is separated from practice by a period of time that is substantially longer than 24 hr, the studies reviewed here necessarily prohibited such a definition of retention performance due to the nature of these studies. Of the 16 effect sizes for retention performance, only 3 contained a practice–performance separation interval of greater than 24 hr, whereas the remaining 13 studies used an interval of 24 hr. Any analysis of these three effect sizes would be generally uninterpretable due to problems associated with second order sampling error (Hedges & Olkin, 1985). Therefore, the somewhat limited definition of retention performance used in the present review was born out of necessity.
number of effect sizes for certain groups (e.g., rigor scores of 6 or 1), we collapsed the effect sizes for these six levels into three categories: Rigor Level 1 (RL1; checklist scores of 1 or 2, \( k = 14 \)), Rigor Level 2 (RL2; scores of 3 or 4, \( k = 69 \)), and Rigor Level 3 (RL3; scores of 5 or 6, \( k = 29 \)).

Task type. An important study characteristic that was coded for was the type of task that was being learned during the practice sessions. In particular, we were interested in three aspects of the task being performed: overall complexity, mental requirements, and physical requirements of the task. Overall task complexity was defined by the degree to which the task requires a number of distinct behaviors, the number of choices involved in the performance of the task, and the degree of uncertainty involved in performance of the task. The mental requirements of the task were defined in terms of the degree to which the task requires the subject to use or demonstrate mental or cognitive skills and abilities in order to perform the task. The physical requirements of the task were defined as the degree to which the task requires the subject to use or demonstrate physical skills and abilities in order to perform or complete the task.

Because we were interested in classifying tasks using all three of these dimensions, we decided against using current taxonomies that focus on only one of these dimensions (e.g., Wood, 1986) to rate the various tasks. Instead we chose to use a cluster analysis in hopes of reducing the set of tasks into a usable set of subgroups, with each group defined by scores on the three dimensions mentioned above. To accomplish this, a list and short description of all 28 tasks used by the studies in this meta-analysis was generated and distributed to 95 graduate and undergraduate students from a large university in the northeast on a voluntary basis. Respondents were asked to rate each of the tasks along all three of the dimensions. The dimensions of overall complexity, physical requirements, and mental requirements were correlated as follows: overall complexity–mental requirements, \( r = .61 \); overall complexity–physical requirements, \( r = .16 \); mental requirements–physical requirements, \( r = -.63 \). The average correlation across judges for these ratings was .61, indicating adequate convergence among the 95 judges. The ratings were used as the basis for a cluster analysis in which the original list of tasks was reduced to four task clusters. To arrive at this final cluster solution, an agglomerative linking procedure was used to create a series of cluster solutions ranging from two task clusters to eight task clusters. These solutions were then examined to determine which cluster formation maximized the within-cluster homogeneity in terms of the three task ratings (overall complexity, physical requirements, and mental requirements) while at the same time maintaining a meaningful degree of heterogeneity in mean dimension ratings across the various clusters.

Of the seven solutions examined, the four cluster solution was the optimal solution according to these criteria. Cluster 1 contained tasks with high physical requirements, low mental requirements, and low overall complexity. Cluster 2 contained tasks with high mental requirements, low physical requirements, and average overall complexity. Cluster 3 contained tasks with low mental requirements, high physical requirements, and high overall complexity. Cluster 4 contained tasks with high mental requirements, high physical requirements, and high overall complexity. A list of the tasks contained within each of the clusters is presented in Table 1.

### Intenrial interval. A final study characteristic that was coded for was the intertrial interval that was given to participants under spaced practice conditions (i.e., how much time elapsed between spaced practice sessions). Unfortunately, 24 of the effect sizes previously identified for inclusion in this review failed to report the exact length of this intertrial interval, forcing us to drop these effect sizes. Consequently, this resulted in 88 effect sizes for this moderator analysis. Upon examination of the intertrial intervals, it became apparent that they did not represent a continuum of time, but rather represented a series of discrete groupings. More specifically, four discrete time categories emerged: (a) less than 1 min (Time Interval 1, \( k = 26 \)), (b) between 1 and 10 min (Time Interval 2, \( k = 31 \)), (c) between 10 min and 1 hr (Time Interval 3, \( k = 10 \)), and (d) greater than 1 day (Time Interval 4, \( k = 21 \)). Accordingly, separate effect sizes were generated for each of these intertrial interval categories.

### Computation of Effect Sizes

Following Hedges and Olkin (1985), we used \( g \) as the effect size index for this review, representing the standardized mean difference in performance between the spaced group and the massed group. These \( g \) values were then corrected for bias resulting from small sample size, resulting in a series of \( d \) values. A positive \( d \) value indicates that the individuals in the spaced practice condition

### Table 1

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Mental requirements</th>
<th>Overall complexity</th>
<th>Physical requirements</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
<td>Rotary pursuit, typing, ball toss, ladder climb, peg reversal, bilateral transfer, crank turning</td>
</tr>
<tr>
<td>2</td>
<td>Low</td>
<td>Average</td>
<td>High</td>
<td>Free recall task, video games, foreign language, slide bar task, voice recognition, classroom lecture, sound localization, word processing, stoop task, verbal discrimination, maze learning, connecting numbers, upside down alphabet printing</td>
</tr>
<tr>
<td>3</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>Gymnastic skills, balancing task</td>
</tr>
<tr>
<td>4</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>Air traffic controller simulation, milk pasteurization simulation, airplane control simulation, hand movement memorization, puzzle box task, music memorization and performance</td>
</tr>
</tbody>
</table>
performed at a higher level than those in the massed practice condition. Computations for the overall meta-analysis and all subsequent analyses were performed using the meta-analytic program DSTAT (Johnson, 1993).

Results

Overall Effect Size for the Superiority of Spaced Practice

The results of the overall meta-analysis and moderator analyses are presented in Table 2. The overall mean weighted effect size was 0.46, with a 95% confidence interval that extended from 0.42 to 0.50. With respect to this overall effect size, several points are worth noting. First, the 95% confidence interval for this effect size does not contain zero, indicating that spaced practice was significantly superior to massed practice in terms of task performance, a finding which is in agreement with the "common knowledge" that has been accepted by many researchers. Second, the overall magnitude of this effect was somewhat smaller than one would expect based on both Lee and Genovese’s (1988) meta-analysis and the general consensus that is held within this research area. According to Cohen’s (1988) standards, this is considered to be a medium effect size. In contrast, Lee and Genovese’s (1988) meta-analysis reported a mean weighted effect size of 0.96, indicating a strong effect for the superiority of spaced practice. Although this discrepancy is relatively large, we felt that this type of result could be expected given the more comprehensive nature of our review in terms of both the variety of tasks studied and the overall number of effect sizes. In fact, the 95% confidence interval for our obtained effect size did not include Lee and Genovese’s estimate of 0.96, indicating that the discrepancy between the two is in fact notable. Finally, significant heterogeneity existed among the 112 effect sizes, $Q(111) = 1025.25, p < .01$, indicating that the studies reviewed here could not be adequately described by a single effect size (Hedges & Olkin, 1985), and thus, a search for potential moderators was warranted.

Moderator Analyses

Task type. Mean weighted effect sizes were generated for each of the four task clusters: Task Cluster 1 (TC1), Task Cluster 2 (TC2), Task Cluster 3 (TC3), and Task Cluster 4 (TC4), with the results appearing in Table 2. The overall effect sizes for the four task categories were as follows: $d_{TC1} = 0.97$, $d_{TC2} = 0.42$, $d_{TC3} = 0.11$, $d_{TC4} = 0.07$. Simple contrasts among these effect sizes indicated that all of the task cluster effect sizes were significantly different from one another ($p < .05$), with the exception of TC3 and TC4. Therefore, task type appears to play an important role in studying the relationship between massed and spaced practice conditions. More specifically, the overall effect size generated for TC1 ($d = 0.97$) was higher than any of the other task cluster effect sizes (TC2, $d = 0.42$; TC3, $d = 0.11$; TC4, $d = 0.07$). TC1 contained tasks that were low in both complexity and mental requirements (i.e., tasks predominantly psychomotor in nature). The effect size for this task cluster is almost identical to the $d$ obtained by Lee and Genovese (1988) in their meta-analysis, which concentrated solely on studies using psychomotor tasks (0.97 compared with 0.96). Thus, there is a degree of consistency that

![Table 2](image)

<table>
<thead>
<tr>
<th></th>
<th>$k$</th>
<th>$N$</th>
<th>$g$</th>
<th>$d$</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>112</td>
<td>8980</td>
<td>0.46</td>
<td>0.46</td>
<td>0.42/0.50</td>
</tr>
<tr>
<td>Rigor level 1</td>
<td>14</td>
<td>714</td>
<td>1.23</td>
<td>1.22</td>
<td>1.05/1.38</td>
</tr>
<tr>
<td>Rigor level 2</td>
<td>69</td>
<td>6084</td>
<td>0.41</td>
<td>0.40</td>
<td>0.35/0.46</td>
</tr>
<tr>
<td>Rigor level 3</td>
<td>29</td>
<td>2182</td>
<td>0.41</td>
<td>0.40</td>
<td>0.32/0.49</td>
</tr>
<tr>
<td>Task cluster 1</td>
<td>30</td>
<td>2209</td>
<td>0.98</td>
<td>0.97</td>
<td>0.88/1.06</td>
</tr>
<tr>
<td>Task cluster 2</td>
<td>58</td>
<td>4715</td>
<td>0.43</td>
<td>0.42</td>
<td>0.36/0.48</td>
</tr>
<tr>
<td>Task cluster 3</td>
<td>10</td>
<td>781</td>
<td>0.11</td>
<td>0.11</td>
<td>-0.03/0.26</td>
</tr>
<tr>
<td>Task cluster 4</td>
<td>14</td>
<td>1275</td>
<td>0.07</td>
<td>0.07</td>
<td>-0.05/0.18</td>
</tr>
<tr>
<td>Acquisition performance</td>
<td>96</td>
<td>7999</td>
<td>0.46</td>
<td>0.45</td>
<td>0.41/0.50</td>
</tr>
<tr>
<td>Retention performance</td>
<td>16</td>
<td>981</td>
<td>0.52</td>
<td>0.51</td>
<td>0.39/0.64</td>
</tr>
<tr>
<td>Time interval 1</td>
<td>26</td>
<td>2383</td>
<td>0.71</td>
<td>0.71</td>
<td>0.62/0.79</td>
</tr>
<tr>
<td>Time interval 2</td>
<td>31</td>
<td>2686</td>
<td>0.48</td>
<td>0.48</td>
<td>0.48/0.55</td>
</tr>
<tr>
<td>Time interval 3</td>
<td>10</td>
<td>578</td>
<td>0.26</td>
<td>0.26</td>
<td>0.09/0.43</td>
</tr>
<tr>
<td>Time interval 4</td>
<td>21</td>
<td>2386</td>
<td>0.17</td>
<td>0.16</td>
<td>0.08/0.25</td>
</tr>
</tbody>
</table>

Note. $k$ = the number of effect sizes contributing to each effect size estimate; $N$ = the total sample size used to derive the effect size; $g$ = the uncorrected mean weighted effect size estimate for each category; $d$ = the mean weighted effect size estimate for each category corrected for bias due to sample size; 95% CI = the 95% confidence interval for each mean weighted effect size.
emerges when the two meta-analyses are compared, indicating that for these simple motor tasks, spaced practice is clearly superior to massed practice. However, it should be noted that this consistency may be due in part to the considerable overlap of studies used in the previous meta-analysis and the studies used within TC1 in the present meta-analysis. According to Cohen (1988), the effect size for TC1 represents a strong effect, while the remaining effect sizes would be classified as moderate for TC2, and weak for TC3 and TC4.

In order to more clearly determine which task dimensions influenced the effectiveness of distribution of practice, we computed semipartial correlations (holding the other dimensions constant) between overall ratings for our three task dimensions with the observed study level effect sizes. The results of this analysis indicated that overall complexity was negatively and significantly correlated with the effect sizes generated \( r = -0.25, p < 0.05 \). Mental and physical requirements of the task were not significantly correlated with the effect sizes \( p > 0.35 \). Thus, the overall complexity of the task appears to be a key factor in determining the overall superiority of spaced practice over massed practice. Tasks high in overall complexity were associated with smaller mean differences between spaced and massed practice conditions.

**Acquisition and retention performance.** Separate effect sizes were computed for those studies using retention measures of performance and those using acquisition measures of performance, with the results presented in Table 2. The effect sizes were: \( d_{\text{acquisition}} = 0.45 \), \( d_{\text{retention}} = 0.51 \). These effect sizes were not significantly different from one another, \( Q_g(1) = 0.86, p > .35 \). Thus, it appears that the distribution of practice effect does not differ for acquisition and retention performance, exhibiting a moderate effect size in both cases. However, it should be noted that this equivalence of effect sizes may be due in part to the definition of retention performance utilized in the present review. More specifically, 24 hours may not be a sufficient interval between practice and performance to truly assess retention performance.

**Methodological rigor.** Separate effect sizes were calculated for the three rigor categories, RL1, RL2, and RL3. The results are presented in Table 2. The effect sizes generated for these three categories are as follows: \( d_{\text{RL1}} = 1.22 \), \( d_{\text{RL2}} = 0.40 \), \( d_{\text{RL3}} = 0.40 \). Simple contrasts among these effect sizes revealed that the RL1 effect size was significantly different from the effect sizes for both RL2 and RL3, \( Q_g(1) = 86.04 \) and 74.77, respectively, \( p < .01 \). The effect sizes generated by RL2 and RL3 were not significantly different from one another, \( Q_g(1) = 0.00, p > .70 \). Thus, higher levels of methodological rigor are associated with smaller effect sizes. In fact, the correlation between the original methodological rigor scores and the effect sizes generated by the studies under review was significant \( r = -0.17, p < .05 \), lending further support to the results of this moderator analysis.

**Intertrial interval.** Mean weighted effect sizes were calculated for each of the four intertrial interval categories: Time Interval 1 (TI1), Time Interval 2 (TI2), Time Interval 3 (TI3), and Time Interval 4 (TI4). The results are presented in Table 2. The overall effect sizes are as follows: \( d_{\text{TI1}} = 0.71 \), \( d_{\text{TI2}} = 0.48 \), \( d_{\text{TI3}} = 0.26 \), and \( d_{\text{TI4}} = 0.16 \). Simple contrasts among these effect sizes revealed that the TI1 effect size was significantly higher than the effect sizes generated for TI2, TI3, or TI4, \( Q_g(1) = 14.83, 21.47 \), and 77.51, respectively, \( p < .05 \), and that TI2 was significantly higher than TI4, \( Q_g(1) = 28.53, p < .05 \). It appears that as the intertrial interval between spaced practice conditions becomes shorter (moving closer to the massed practice conditions), the resulting standardized mean differences between these groups actually tend to increase (from 0.16 to 0.71). This finding is somewhat contrary to what one might expect given the traditional view on the distribution of practice. If, in fact, the time period between the practice sessions is what makes spaced practice conditions superior to massed practice conditions, then one would expect that shortening this time interval would be associated with a decrease in effect sizes, indicating the superiority of spaced practice. One potential reason for this observed relationship is that for the tasks examined in this review (many of which were relatively simple motor tasks), a relatively short rest period may be all that is needed by participants in the spaced practice conditions. Any additional time between practice sessions may actually have been detrimental to their subsequent performance. In other words, it may not be the case that “more is better” with respect to intertrial time intervals, but rather there are different “optimal” intertrial time periods for different types of tasks. If this were in fact the case, then an examination of only the overall means for the different time intervals could be potentially misleading because task specific influences on the relationship between the intertrial interval and the effect size are obscured. A more appropriate analysis would be to look at the effects of these different intertrial intervals within the context of a single task type.

**Exploratory Moderator Interaction Analysis**

In order to provide a more appropriate examination of the findings concerning intertrial time interval, an exploratory moderator interaction analysis was conducted in which each level of the time interval variable (TI1–TI4) was crossed with each of the task cluster variables (TC1–TC4), resulting in a \( 4 \times 4 \) matrix of effect sizes presented in Table 3. An examination of these effect sizes provides some support for the proposition that task type moderates the relationship between the intertrial interval and the effect size. Specifically, TC1 and TC2 (the only task categories for which
The purpose of the present review was to meta-analytically examine the distribution of practice effect with respect to task performance. An overall effect size of 0.46 was obtained indicating that individuals in spaced practice conditions outperformed those in massed practice conditions by almost one half of a standard deviation. Perhaps more interesting were the findings that the type of task being practiced, the length of the intertrial time interval, and the interaction of these two factors played an important role in determining the magnitude of the distribution of practice effect. Correlational analyses revealed that the overall complexity of the task being practiced was negatively related to the magnitude of the study level effect sizes (i.e., tasks high in overall complexity were associated with smaller effect sizes). Although the influence of task type has been proposed by researchers to be a potentially critical boundary condition of the superiority of spaced practice, the results of the present review provide actual empirical evidence to support this notion. The strong distribution of practice effects previously reported appears to be limited to relatively simple tasks. The results also indicated that the optimal intertrial interval appears to partially be a function of the type of task being learned. Paradoxically, stronger effects were found for simple tasks when using very brief rest periods. For more complex tasks, including those that are representative of organizational training settings, longer rest periods appeared to be more beneficial for task learning. These findings indicate that one should be aware that the type of task being practiced and the amount of time given between practice sessions can have a profound effect on the overall superiority of spaced practice over massed practice. In addition, it was found that studies low in methodological rigor tended to report larger effect sizes than those studies that were moderate or high in methodological rigor.

**Implications of Findings**

Our findings have implications for the development and implementation of organizational training programs and suggest that the optimal design of training programs will depend on factors such as the type of task being learned. The magnitude of the distribution of practice effect for certain task types (e.g., Task Cluster 1 and Task Cluster 2) suggests that the benefits that could be realized through the use of spaced practice condition could be quite substantial. For example, Goldstein (1993) noted that the lecture method is one of the more commonly used training methods in organizations when attempting to teach employees factual knowledge. According to the classifications utilized in the present review, the learning of factual knowledge via a lecture would fall into Task Cluster 2, with a resulting effect size of 0.42. Thus, according to these results, we could expect an increase in learning of approximately one half of a standard deviation with a lecture method through the use of spaced material presentation as opposed to massed material presentation. Similarly, distance learning accomplished via the Internet could also be considered to be a part of Task Cluster 2 and thus would demonstrate a similar effect size. Thus, the simple transition from massed material presentation to spaced material presentation sessions with these two training methods could have a relatively large impact on the amount of learning that occurs as a result of the training sessions. Given the popularity of lectures as a training method, as well as the increasing popularity of...
Web-based distance learning, it becomes clear that the results of the present review have important implications for the design of such programs.

A second implication to be drawn from the present review is that future research should be directed towards identifying additional boundary conditions under which the distribution of practice effect operates (e.g., initial level of trainee ability). Clearly, there is no one "universal effect" resulting from the distribution of practice. In addition, future research should be directed at areas that have been relatively unexplored, such as the effects of distribution of practice on retention measures of performance. Considering the increased attention being given to the transfer of training in the workplace (e.g., Baldwin & Ford, 1988), this sort of information could have potentially large implications for the design of organizational training environments.

**Limitations of the Current Review**

One important limitation of the current review is that we were unable to code for other potentially important moderator variables of the distribution of practice effect. Factors such as the motivation of the participants to perform the task and what the participants were asked to do during the intertrial interval could impact the distribution of practice effect. Unfortunately, this type of information was rarely included within the study reports used for the present review. A second limitation is that we were unable to examine other facets of task performance that are directly relevant to the organizational training environment, such as long-term retention performance and the actual degree to which the knowledge or skills learned through either massed or spaced practice are transferred back to the job. Clearly such information would have provided additional insight into the distribution of practice phenomenon.

In closing, we emphasize that although the current meta-analysis confirms the superiority of spaced practice over massed practice under certain conditions, this relationship is not as strong or pervasive as many researchers in the past have been inclined to accept. In fact, under certain conditions, this effect may even disappear or reverse itself. The results of the present review should be used by future researchers to further delineate the boundaries of the distribution of practice effect, as well as develop theoretical models to guide research in this area.

**References**

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


RESEARCH REPORTS


Checklist Used to Assess Methodological Rigor

Did the study use:
(1) the use of a pretest-posttest design;
(2) the use of a significance level of .05 or less when testing proposed relationships;
(3) a sample size of thirty or greater (to satisfy the appropriate statistical assumptions of the hypothesis testing procedures);
(4) random assignment to experimental conditions within the study;
(5) comparison of massed and spaced practice groups on potentially important covariates (e.g., intelligence, familiarity with the task being performed);
(6) control for Type I error rate when performing multiple comparisons;
(7) use of a randomized procedure for selecting participants;
(8) controls for experiment expectancies (i.e., the use of “blind” testing procedures when collecting performance data);
(9) a high level of control over the experimental manipulation and environment?

Received June 17, 1997
Revision received December 10, 1998
Accepted December 10, 1998