Maastricht University

The effects of practice schedule on learning a complex judgment task

Citation for published version (APA):

Helsdingen, A. S., van Gog, T., & van Merrienboer, J. J. G. (2011). The effects of practice schedule on learning a complex judgment task. *Learning and Instruction*, *21*(1), 126-136. https://doi.org/10.1016/j.learninstruc.2009.12.001

Document status and date:
Published: 01/02/2011

DOI:
10.1016/j.learninstruc.2009.12.001

Document Version:
Publisher's PDF, also known as Version of record

Document license:
Taverne

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy
If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Download date: 19 Apr. 2024

ELSEVIER

Learning and Instruction 21 (2011) 126–136

Learning and Instruction

www.elsevier.com/locate/learninstruc

The effects of practice schedule on learning a complex judgment task

Anne S. Helsdingen [a,*], Tamara van Gog [b], Jeroen J.G. van Merriënboer [c]

[a] Centre for Learning Sciences and Technologies, Open University of The Netherlands, PO Box 2960, 6401 DL Heerlen, The Netherlands
[b] Faculty of Social Sciences, Erasmus University, PO Box 1738, 3000 DR Rotterdam, The Netherlands
[c] School of Health Professions Education, University of Maastricht, PO Box 616, 6200 MD Maastricht, The Netherlands

Received 2 February 2009; revised 29 September 2009; accepted 1 December 2009

Abstract

The effects of practice schedule on learning a complex judgment task were investigated. In Experiment 1, participants' judgment accuracy on a retention test was higher after a random practice schedule than after a blocked schedule or operational schedule. Experiment 2 demonstrated that judgment on a transfer test was also better after a random practice schedule than after a blocked schedule. Both experiments failed to show any effects of practice schedule on performance during learning. These findings show that benefits of random practice for retention and transfer apply to learning a complex judgment task, and may be achieved without performance degradation during practice.
© 2009 Elsevier Ltd. All rights reserved.

Keywords: Practice schedule; Judgment; Transfer

1. Introduction

Most educational programs aim to achieve two major goals: Adequate *post-training* performance (retention) and transfer to related tasks and situations. However, very often, those goals are confused with enhancing performance and speed of skill acquisition *during* training. Research has shown that the opposite is often true: interventions that enhance performance during training may have detrimental effects on retention and transfer performance, and conversely, instructional manipulations that degrade performance during skill acquisition may support the long-term goals of training (for an overview, see Schmidt & Bjork, 1992).

An example of the latter is by providing a practice schedule where different variations of the learning tasks are sequenced randomly as opposed to sequenced in separate blocks (Shea & Morgan, 1979). This type of random sequencing is also often referred to as interleaving practice materials, or mixed practice (Hatala, Brooks, & Norman, 2003; Richland, Bjork, Finley, & Linn, 2005). Random practice may degrade performance during the learning phase but lead to better post-training performance and transfer (Lee & Simon, 2004). This technique has mainly been studied in motor tasks (Brady, 1998; Cross, Schmitt, & Grafton, 2007; Guadagnoli & Lee, 2004; Lee & Magill, 1983; Magill & Hall, 1990; Shea & Morgan, 1979; Simon, 2007). However similar findings have been obtained with, for example, procedural tasks (Carlson, 1989; Carlson & Schneider, 1989; Carlson, Sullivan, & Schneider, 1989; Carlson & Yaure, 1990), cognitive operational tasks, such as interacting with automatic teller machines (Jamieson & Rogers, 2000), language learning (Jacoby, 1978), foreign vocabulary learning (Schneider, Healy, & Bourne, 1998, 2002), learning logical rules (Schneider, Healy, Ericsson, & Bourne, 1995), learning problem-solving from worked examples (Paas & Van Merriënboer, 1994), and troubleshooting tasks (De Croock, Van Merriënboer, & Paas, 1998).

Only few studies have been conducted in the past to investigate the effects of practice schedule on learning complex judgment and decision-making, in which the goal is to learn the complex relationships between several phenomena and predict the value of a distal variable (e.g., clinical diagnosis, weather forecast, threat assessment; Brehmer, 1973, 1977, 1979). Moreover, these studies measured performance

* Corresponding author. Tel.: +31 619348915.
E-mail addresses: anne.helsdingen@ou.nl (A.S. Helsdingen), vangog@fsw.eur.nl (T. van Gog), j.vanmerrienboer@educ.unimaas.nl (J.J.G. Van Merriënboer).

0959-4752/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.learninstruc.2009.12.001

during training, not on retention or transfer tests. Although training principles that are effective with relatively simple tasks are not necessarily effective for complex tasks as well (Wulf & Shea, 2002), earlier studies on complex mathematical problem-solving have found benefits of mixing or interleaving practice materials on post-practice performance (Rohrer & Taylor, 2007; Simon, 2008). However, whether similar benefits can be expected in learning complex judgment tasks, and whether such effects also occur on transfer performance, remains an interesting question that is addressed in this study. It is important to establish the most optimal training sequence for such tasks, because of the far-reaching consequences that, for example, wrong clinical diagnoses or military judgments may have (Hogarth, 1980). Therefore, the present study explores the effects of practice schedule on learning and transfer of complex judgment tasks.

1.1. Complex judgment tasks

In the numerous choices or judgments people make every day, two distinct classes can be identified: value judgments, which express their preferences, and predictions, which reflect what they expect to happen (Hogarth, 1980). Value judgments encompass, for example, a choice for one house over another or one pair of shoes over another. Predictions concern future outcomes such as, for example, expectations regarding how someone might react to what you say or do, or who will win the next presidential elections. In this study we focus on predictive judgment tasks.

Consider, for example, a central executive officer (CEO) of an international company who has to make a decision on moving his production facility from The Netherlands to India. In making this judgment, several so-called 'points of reference' need to be considered that are expected to be related to the target of maximizing profit, such as labor costs, infrastructure, inflation rates, and so forth. These points of reference are part of the CEO's mental representation of the network of relationships between elements (objects, events) in the environment and the event to be predicted. Accuracy of the judgment depends on the extent to which the CEO's mental representation matches the real network of relationships (Hogarth, 1980). It is this match, or lack thereof, that is the object of study within the social judgment theory (SJT; Brehmer & Joyce, 1988; Hammond, McClelland, & Mumpower, 1980).

According to the SJT, which is modeled after Brunswik's (1943, 1955) theory of perception, a person does not have access to any direct information about the objects in the environment. Instead, perception is seen as an indirect process, mediated by a set of proximal cues (i.e., points of reference). The perceptual system uses these cues to make inferences about distal objects. In accordance with this view, SJT defines judgment as a process that involves the integration of information from a set of cues into a judgment about some distal state of affairs (Hogarth, 1980).

Within the SJT research paradigm, an experimental method was devised to study how people learn such difficult judgment tasks, namely the (Multiple) Cue Probability Learning (MCPL; Björkman, 1965; Brehmer, 1972; Brehmer & Brehmer, 1988; Brunswik & Herma, 1951; Hammond, Hursch, & Todd, 1964; Hursch, Hammond, & Hursch, 1964; Smedslund, 1955), also referred to as Multidimensional Functional Learning (MFL; Hoffman, Earle, & Slovic, 1981). During a typical MFL experiment a person makes judgments based on a number of probabilistic cues over a series of trials. Feedback may be given on each trial, or feedback may be given after subsets of trials. The aim is to correctly predict the quantitative or categorical criterion value on each trial.

The MFL studies have focused on how people learn to discover cues and judge the importance of these cues (Klayman, 1988a). In particular, it has been studied how learning and transfer performance are affected by parameters of the task (e.g., linearity of relationships between variables, predictability of cues, meaningful labels, time pressure; Edland, 1993; Koh, 1993), the nature and timing of given feedback (e.g., delayed feedback, cognitive feedback, outcome feedback, feedforward; Balzer, Sulsky, Hammer, & Sumner, 1992), and characteristics of the task performer (e.g., age, goal setting, prior knowledge or experience; Alm & Brehmer, 1982; Chasseigne, Mullet, & Stewart, 1997; Hoffman et al., 1981). From these studies, it became clear that performance is higher when (a) linear relationships exist between cues and the criterion (Alm, 1982b; Brehmer, 1979, 1987; Hammond & Summers, 1965), (b) these relationships are positive rather than negative (Björkman, 1965; Brehmer, 1977; Sheets & Miller, 1974), (c) cues have meaningful rather than abstract labels (Koele, 1980; Muchinsky & Dudycha, 1975; Ruble & Cosier, 1990), (d) there is no time pressure (Rothstein, 1986), (e) positive feedback is given rather than negative feedback (Klayman, 1988b), and (f) participants set goals for themselves (DeShon & Alexander, 1996).

Regrettably, in most MFL studies, participants' performance was only measured during the learning phase. This is problematic, because observed performance during the learning phase can be a notoriously poor guide to predicting learning outcomes, namely post-training performance (Bjork, 1994). Only a few MFL studies measured test performance after a retention interval (Alm, 1982a; Brehmer, 1973) or transfer of learning to new, unfamiliar tasks (Andersson & Brehmer, 1977; Brehmer, 1977, 1979; Brehmer & Almqvist, 1977; Lindberg & Brehmer, 1976).

In conclusion, MFL experiments use representative experimental tasks that adequately capture the major characteristics of complex judgment tasks. But although many task-related, feedback-related, and learner-related aspects were investigated for their effects on performance during learning and retention, few studies have focused on methods to increase transfer of MFL. One way to increase transfer performance might be to use random rather than blocked practice schedules.

1.2. Blocked vs. random practice schedules

Blocked task sequences, that is, sequences of learning tasks organised in blocks, with only one variation of a task being practiced in each block (e.g., AAA–BBB–CCC), have often been found to lead to better performance during training than random practice (e.g., A–B–C–B–C–A–A–C–B; see, e.g., Schneider, Healy, & Bourne, 2002). However, random practice often results in better retention and transfer of skills to related tasks and situations (Greeno, 1964; Healy et al., 2002). These beneficial effects of random practice have been observed in a variety of domains and tasks.

To explain the benefits of random practice for retention and transfer, several hypotheses have been formulated, such as the elaboration hypothesis (Shea & Morgan, 1979), the reconstruction hypothesis (Lee & Magill, 1983), and the retrieval hypothesis (Schmidt & Bjork, 1992). The major line of argument of these hypotheses is that in random practice, learners are challenged to compare the different procedures associated with the different tasks, whereas in a blocked schedule, only one task procedure has to be kept in mind during a block of tasks, forsaking the need for extra processing activities such as elaboration, reconstruction, or retrieval of procedures from long-term memory. These extra processing activities lead to abstraction, that is, to richer mental representations and more general knowledge about principles and procedures. In MFL, random practice schedules may encourage learners to abstract cue–criterion relations from the learning tasks, whereas in blocked practice schedules learners may rely on memory of specific cue–criterion observations from prior learning tasks, without attempting to abstract the underlying relationships between cues and criterion. Therefore, it is expected that although blocked practice may enhance performance during the learning phase of an MFL task, random practice will eventually lead to better retention and transfer.

1.3. Operational practice schedules

Studying the effects of different practice schedules on learning and transfer of complex judgment and decision-making skills may also provide insight into the effectiveness of the *train as you fight* paradigm that is being widely applied in decision-making training programs (e.g., for military command and control, crisis management, and general leadership and management). In this training approach, the real-world sequence and frequency of events serves as a basis for the scheduling of practice events (i.e., operational practice schedule), whereas in a test or exam, the less frequently presented cases may have a normal to high chance of occurrence because the event may have serious implications. Consider, for example, a medical student: if the student is trained in a real-world medical practice, that student will probably receive little practice diagnosing rare diseases. However, a subsequent medical exam may incorporate such serious cases because their correct diagnosis is very important. Within the research community, *train as you fight* is often considered an ineffective training methodology for several reasons (Farmer, Van Rooij, Riemersma, Jorna, & Moraal, 1999). One of the objections is that it may not provide the opportunity to practice rare or unusual tasks, which may yet be critical to effectively deal with emergency situations. However, if the real-world sequence of events increases variations in tasks when compared to a blocked schedule because it is random, it might lead to adequate post-training performance. Little research has addressed the effects of the *train as you fight* or operational practice schedules, especially in comparison with random or blocked schedules (Beaubien, Palev, Shadrick, Ennis, & Jacklin, 2006; Lussier, Shadrick, & Prevou, 2003).

1.4. The present study

The two experiments of the present study investigated the effects of different practice schedules on learning, retention (Experiment 1), and transfer (Experiment 2) of complex judgment tasks. It was hypothesized that a blocked practice schedule will yield better performance *during* the learning phase (Hypothesis 1), whereas random (Experiments 1 and 2) and operational (Experiment 1) practice schedules will yield better performance than a blocked one as regards learning *outcomes* measured in terms of retention- and transfer-test performance (Hypothesis 2).

2. Experiment 1

Experiment 1 tested Hypotheses 1 and 2. Specifically it investigated the effects of random, blocked, and operational practice schedules on performance during learning and on retention tasks.

2.1. Method

2.1.1. Participants – design

Participants were 54 students recruited from different faculties of the Universities of Utrecht and Amsterdam who volunteered to participate in the experiment (21 male, 33 female; mean age was 21 years, SD = 2.9); they had no prior knowledge or experience concerning the experimental tasks. Participants received 32 Euros for their participation in the experiment and could gain an additional bonus of between 0 and 12 Euros, based on their level of performance during the experiment. They were randomly assigned to the blocked practice schedule ($n = 18$), random practice schedule ($n = 18$), and operational practice schedule ($n = 18$).

2.1.2. Materials

2.1.2.1. Learning tasks. Three sets of six cases were developed. One set dealt with injury cases, one with damage cases, and one with traffic cases. Participants were presented with one set of cases, and had to prioritize each case on the urgency for the police to deal with it. The three sets were balanced over the three practice schedule conditions, that is, in each condition six participants worked on injury cases, six participants worked on damage cases, and six participants worked on traffic cases. Priorities depended on the values of two different cues. One cue had three possible values and the other cue had two different values, thus yielding six combinations of cue values. For injury cases these cues were (a) condition of victim (light injury,

Table 1
Cue values, priority scores, and presentation frequency in the operational practice schedule of the injury cases.

Cue value	Use of weapon	Priority score[a]	Presentation frequency (%)
Light injury	No Firearm	18	45
	Firearm	32	5
Heavy injury	No firearm	48	25
	Firearm	62	10
Dead	No firearm	78	3
	Firearm	92	12

[a] Min = 1, Max = 100.

Table 3
Cue values, priority scores, and presentation frequency in the operational practice schedule of the traffic cases.

Cue value	History	Priority score[a]	Presentation frequency (%)
Speeding	First time	4	30
	Recidivist	20	45
Uninsured	First time	24	7
	Recidivist	40	3
Drunk driving	First time	44	9
	Recidivist	60	6

[a] Min = 1, Max = 100.

heavy injury, dead), and (b) weapon (firearm, no firearm). For damage cases the cues were: (a) level of damage (light, heavy, irretrievable) and (b) nature of the crime (burglary, violence/holdup). For traffic cases the cues were: (a) nature of the offence (speeding, driving without insurance, driving drunk) and (b) history (first time, recidivist). Tables 1–3 present the priority scores for each combination of cue values.

The cases were presented one by one on a computer screen (see Fig. 1). Each participant received 96 learning tasks (i.e., each of the six cases was presented 16 times). The first six tasks were introduced by a short description of the crime, whereas the subsequent 90 tasks were presented as a set of cue values only. On the lower half of the screen, beneath the presentation of the case, a slide bar was presented covering the whole range of priority scores (1–100). Using the computer mouse, participants could manipulate an indicator on the slide bar to mark the priority of the case.

Feedback during the training session consisted of a second slide bar with the indicator at the position of the true priority score. This feedback slide bar was presented above the first one, immediately after participants had indicated their priority score. The bonus that participants had earned was calculated with the following formula and was continuously visible on the screen.

$$\text{Bonus} = \sum_{i=1}^{96}\left(1-\left(\frac{|\text{true priority}_i-\text{estimate}_i|}{\text{true priority}_i}\right)\right)\times .125$$

2.1.2.2. *Practice schedules.* In the random practice condition, the cases were presented in a fully randomized order (without replacement) and each combination of cue values was presented an equal number of times. Thus, during the 96 learning tasks, each of the six combinations of cue values (i.e., each case) was presented 16 times.

In the blocked practice condition, each combination of cues also appeared 16 times in total. The learning tasks in the first block of the learning phase (12 tasks) were sequenced in such a way that only one cue (the cue with 3 values) changed value from one task to the next, whereas the value of the second cue was kept constant. In the second block (12 tasks), the second cue had changed value, and again only the first cue changed value from one task to the next. In the third block (8 tasks) the first cue was kept constant and the second cue (with two values) changed value from one task to the next. In the fourth block (8 tasks), the value of the first block had changed, and again a sequence of tasks was presented where only the second cue changed value from one task to the next. In the fifth block (8 tasks), the value of the first cue changed again, and the second cue changed value from one task to the next. In the sixth block (24 tasks) and seventh block (24 tasks) two cues simultaneously changed their values from one task to the next.

The operational task sequence was realized by a random order of tasks in which some tasks were more likely to be presented to participants than others (i.e., the higher the occurrence in real-life situations, the more likely the case was to be presented). For the probability of selecting each case in the operational sequence, see Tables 1–3.

2.1.2.3. *Retention test.* The retention test consisted of a random selection of 24 tasks from the learning tasks, presented in

Table 2
Cue values, priority scores, and presentation frequency in the operational practice schedule of the damage cases.

Cue value	Level of damage	Priority score[a]	Presentation frequency (%)
Burglary	Light	19	18
	Heavy	25	31
	Irretrievable	31	21
Violence/holdup	Light	49	3
	Heavy	55	12
	Irretrievable	61	15

[a] Min = 1, Max = 100.

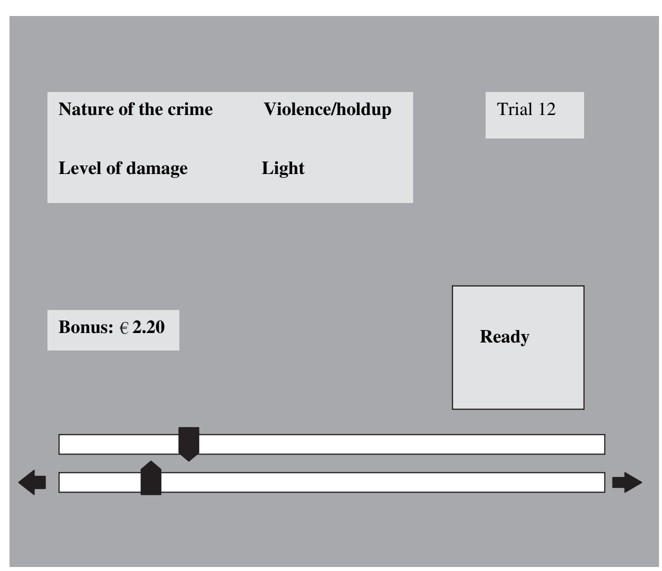

Fig. 1. Presentation of a case, including feedback on its true priority, on the participant's computer screen in Experiment 1 (translation, original in Dutch).

random order that was the same for each participant. Each combination of cue values was presented approximately four times. No feedback was given during the retention test.

2.1.3. Procedure

The experiment lasted approximately 1.5 h and was run in sessions with at most six participants. Before the experiment started, participants read a short instruction explaining how to rate the priority of the cases by moving the indicator on the slide bar. Then, the experimenter assigned each participant to a computer and started the learning phase. After participants had practiced the 96 learning tasks, they had a short break of approximately 10 min after which they completed the 24 tasks of the retention test. For both the learning phase and the retention test, participants could take as much time as they needed.

For each task in the learning phase and the retention test, participants' deviation score, defined as the absolute difference between the estimated priority and the true priority of the task, was automatically stored in a log file. For each participant, time-on-task for the total session (learning phase and retention test together) was logged.

2.2. Results – discussion

Table 4 presents the judgment performance (i.e., deviation) scores in the learning phase and the retention test per condition. Time-on-task (in seconds) did not significantly differ between random (M = 5166, SD = 1067), blocked (M = 5193, SD = 853), and operational (M = 5216, SD = 777) practice schedules, $F(2, 51) < 1$, ns. In the analyses reported here, a significance level of 0.05 was set, and partial eta-squared or Adjusted Hedges are reported as a measure of effect size.

A 3(practice schedule: random, blocked, operational) × 3(type of cases: injury, damage, traffic) ANOVA performed on the mean deviation scores during the learning phase showed no significant interaction between practice schedule and type of case, $F(4, 45) = 0.44$, ns, and no main effects of practice schedule, $F(2, 45) = 0.94$, ns, or case type, $F(2, 45) = 2.08$, ns, on performance during the learning phase. That is, contrary to Hypothesis 1, participants in the blocked condition did not perform better than participants in the random or operational conditions during the learning phase.

Regarding the mean deviation scores during the retention test, a 3(practice schedule: random, blocked, operational) × 3(type of cases: injury, damage, traffic) ANOVA did not show an interaction between case type and practice schedule, $F(4, 45) = 0.14$, ns, and no main effect of type of cases, $F(2, 45) = 1.43$, ns. The analysis did show a significant effect of practice schedule, $F(2, 45) = 4.46$, $p = 0.02$, partial $\eta^2 = 0.15$. Post hoc Tukey tests indicated that in line with our expectations, participants who followed random practice performed better (M = 3.12, SD = 1.84) on the retention test than participants who followed a blocked practice schedule (*M* = 5.48, SD = 2.81), that is, they had lower deviation scores. The other differences between conditions were not statistically significant. For the operational schedule condition, we also analyzed whether the frequency of presentation of cases (see Tables 1–3) had an effect on performance during the retention test. Application of Friedman's test showed that there were no significant differences in the scores of the retention test between cases that were presented with the highest frequency (31%, 45%, 60% for damage, traffic, and injury cases, respectively; M = 5.43, SD = 4.60), and cases that were presented with the lowest frequency (3% for all case types; M = 3.04, SD = 4.24), $\chi^2(1, N = 18) = 1.00$, $p = 0.32$, Adjusted Hedges $g = 0.59$.

In conclusion, Hypothesis 1 that a blocked practice schedule would generate better performance during learning than random or operational practice was not confirmed by our results. Hypothesis 2, however, that random and operational practice schedules would yield better performance than a blocked one in the retention test was partly confirmed. There are some other studies that also failed to find effects of practice schedule on performance during learning. Specifically, whereas some studies failed to find effects on retention-test performance as well (French, Rink, & Werner, 1990; Jones & French, 2007) others did show effects on retention-test performance, in line with our results (Immink & Wright, 1998; Ollis, Button, & Fairweather, 2005; Wrisberg & Liu, 1991). That is, we did find the expected benefits for participants in the random condition over the blocked condition on the retention test. However, the operational condition did not outperform the blocked practice group. When practice tasks were presented in an operational order, that is, random but with some cases being presented more frequently than others, the benefits of interleaving learning materials for retention performance disappeared. A possible explanation might be that because some cases appear more often than others in an operational schedule, not all cases that were presented in the retention test had been trained to the same extent. Nevertheless, the Friedman test showed that retention-test performance on cases that were practiced less frequently was not significantly worse than that on cases that were trained more frequently. Even though operational practice schedules are advocated by some practitioners as a means of providing more realistic training scenarios while increasing random practice, and as a consequence post-training performance, this experiment showed that a "normal" random schedule is to be preferred in terms of reaching good retention performance.

In Experiment 1, only effects on retention were measured. However, the aim of many education or training programs is to attain transfer of learning, that is, the adequate application of skills or knowledge acquired under specific conditions or with

Table 4
Means (and SD) of deviation scores for all conditions in Experiment 1.

Condition	Learning phase	Retention test
	M (SD)	M (SD)
Blocked practice schedule	7.25 (3.88)	5.48 (2.81)
Random practice schedule	6.15 (2.17)	3.12 (1.84)
Operational practice schedule	6.01 (2.61)	4.58 (2.41)

Deviation scores are the absolute difference between true priority (min. 1 and max. 100) and participant's estimate.

specific tasks in different tasks or conditions (Mayer & Wittrock, 1996; Roscoe & Williges, 1980). It has been argued that transfer is higher following random practice than following a blocked schedule, because in a random schedule participants may compare different tasks continuously, which may not only lead to a deeper understanding of the relationships between the different cues and the criterion, but also to abstract knowledge of how to approach this type of task (Wulf & Shea, 2002). Therefore, Experiment 2 investigated the effects of interleaving and blocked practice on transfer.

3. Experiment 2

Experiment 2 again tested Hypothesis 1 and 2. It investigated the effects of random and blocked practice schedules on performance during learning and on transfer in complex judgment tasks. Because an operational practice schedule did not generate any performance difference during learning and on the retention test in the first experiment, this condition was excluded from Experiment 2.

3.1. Method

3.1.1. Participants − design

Participants were 64 students (32 male, 32 female; mean age 22.0 years, SD = 4.4) recruited from different faculties of the Universities of Utrecht and Amsterdam who volunteered to participate in the experiment. They were randomly assigned to two groups that received either random practice ($n = 32$) or blocked practice ($n = 32$). Their prior level of schooling ranged from senior vocational education to university masters level. All participants received 32 Euros for their participation in the experiment and could gain an additional bonus of between 0 and 12 Euros, based on their level of performance during the learning phase.

3.1.2. Materials

3.1.2.1. Learning tasks. Learning tasks consisted of 32 descriptions of crimes. Each crime had to be prioritized on the urgency for the police to deal with it. Priorities could be determined on the basis of the dichotomous values on four different cues that occurred in each crime description: (a) the condition of the victim (injured, dead); (b) the use of a weapon (no firearm, firearm); (c) the nature of the crime (burglary, violence/holdup), and (d) available information concerning the perpetrator (description, known to the police). Table 5 presents the priority scores for each combination of cue values.

Two tasks were developed for each of the 16 combinations of cue values, resulting in 32 tasks.

The crime descriptions were presented to the participants one by one on a computer screen (see Fig. 2). The task was presented on the upper half of the screen. Below that, the four cues and possible cue values with tick boxes as well as a blank space for participants' estimate of the priority of the crime (to be given as a numerical value between 1 and 100) were presented. The correct cue values had to be ticked before entering priority scores.

Feedback in terms of the true priority of a crime was presented on the screen after completion of each task. The bonus that participants had earned was calculated with the following formula and was continuously visible on the screen.

$$\text{Bonus} = \sum_{i=1}^{32}\left(1 - \left(\frac{|\text{true priority}_i - \text{estimate}_i|}{\text{true priority}_i}\right)\right) \times .375$$

3.1.2.2. Practice schedules. In the random practice group, the task order was determined by random selection without replacement from the 32 available tasks. In the blocked practice group, the tasks in the first block of 8 tasks focused the participants' attention on the most influential cue, namely,

Table 5
The effect of cue values on the priority score in Experiment 2.

Cue value	Use of weapon	Nature of crime	Information on perpetrator	Priority score[a]
Injury (10)	No Firearm (0)	Burglary (0)	Description (0)	10
			Known to police (8)	18
		Violence/holdup (12)	Description (0)	22
			Known to police (8)	30
	Firearm (4)	Burglary (0)	Description (0)	14
			Known to police (8)	22
		Violence/holdup (12)	Description (0)	26
			Known to police (8)	34
Death (43)	No Firearm (0)	Burglary (0)	Description (0)	43
			Known to police (19)	62
		Violence/holdup (29)	Description (0)	72
			Known to police (19)	91
	Firearm (9)	Burglary (0)	Description (0)	52
			Known to police (19)	71
		Violence/holdup (29)	Description (0)	81
			Known to police (19)	100

It can be inferred from the Table how the cues interact: The first cue, condition of the victim, determines the effect of the other cues on the priority. The use of a firearm has more effect on the priority when the victim is dead (injury: +4; dead +9), and so do the use of violence (injury: +12; death: +29) and the fact of the perpetrator being known to the police (injury: + 8; death: + 9).

[a] Min = 1, Max = 100.

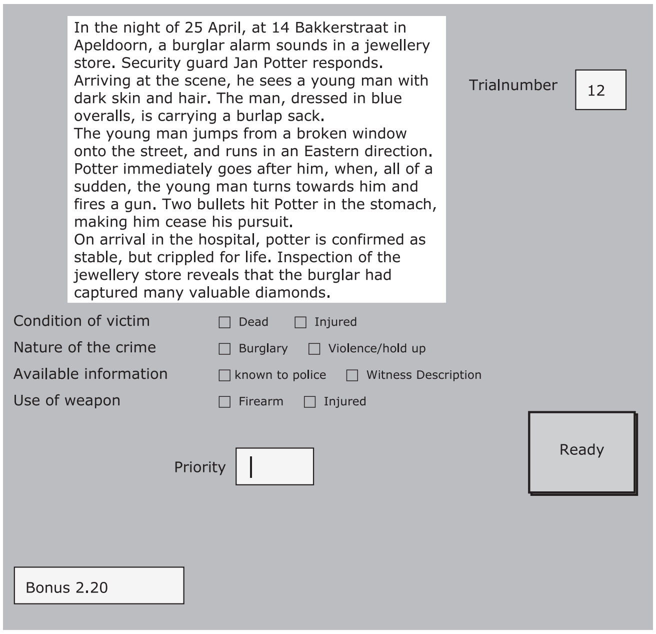

Fig. 2. Presentation of a case on the participant's computer screen in Experiment 2 (translation, original in Dutch).

the condition of the victim. The tasks in the second block (tasks 9–16) simultaneously varied the values of the condition of the victim and one other cue. The tasks in the third block (tasks 17–24) simultaneously varied the values of the condition of the victim and of the two other cues not yet used in the second block. The tasks in the fourth, and final, block (tasks 25–32) simultaneously varied the values of the condition of the victim and the three other cues.

3.1.2.3. Transfer test. The test consisted of eight transfer tasks, that is, tasks that were similar to the training tasks in structural features (same combination of cues) but different with regard to surface features (cover stories). The transfer tasks were presented in a random order that was the same for all participants. No feedback was given during the transfer test.

3.1.3. Procedure

The experiment lasted about one hour, with at most four participants per session. Before the experiment started, participants read an instruction explaining how to rate the priority of the tasks, and two exemplary task descriptions were discussed. Then, the experimenter assigned each participant to a computer and started the learning phase. The transfer test was conducted a few minutes after the last task in the learning phase was finished. For each task, participants' deviation scores, defined as the absolute difference between the estimated priority and the true priority of the task, were automatically stored in a log file. Time-on-task for the total session (learning phase and transfer test) was logged for each participant.

3.2. Results – discussion

Table 6 presents the mean deviation scores per condition in the learning phase and the transfer test. Time-on-task (in seconds) did not differ between random ($M = 3515$, $SD = 756$) or blocked ($M = 3564$, $SD = 680$) practice schedules, $t(62) = -0.26$, $p > 0.20$. In the analyses reported below, a significance level of 0.05 was set, and partial eta-squared is reported as a measure of effect size.

Table 6
Means (and SD) of deviation scores for the two conditions in Experiment 2.

Condition	Learning phase	Retention test
	M (SD)	M (SD)
Blocked practice schedule	15.30 (5.70)	15.09 (6.92)
Random practice schedule	6.89 (6.39)	10.78 (6.82)

Deviation scores are the absolute difference between true priority (min. 1 and max. 100) and participant's estimate.

As in Experiment 1, the mean deviation scores of the learning phase were analyzed with an ANOVA with practice schedule (random, blocked) as independent variable. The analysis showed that practice schedule did not have a significant effect on performance during the learning phase, $F(1, 62) = 1.106$, ns. Hence, Hypothesis 1 was not confirmed. In line with Hypothesis 2, participants in the random practice outperformed (i.e., lower deviation scores) participants in the blocked practice schedule on the transfer test, $F(1, 62) = 6.328$, $p = 0.014$, Cohen's $d = 0.63$.

In sum, Experiment 2 examined the effects of random and blocked practice schedules on learning and transfer of a multidimensional functional learning task. We expected that a blocked practice schedule would generate equal or better performance than a random practice schedule during the learning phase. However, as in Experiment 1, performance during the learning phase was equal, but not better in the blocked practice schedule compared to the random practice schedule. This unexpected finding is further discussed in the general discussion. In line with our expectations, this second experiment showed that a random practice schedule not only results in better retention as was shown in Experiment 1, but also in higher transfer test performance than a blocked schedule.

4. General discussion

The aim of the present study was to determine the effects of practice schedule on learning, retention and transfer in complex judgment tasks and to investigate the effectiveness of an operational practice schedule stemming from the train as you fight paradigm (i.e., a interleaved schedule but with some tasks being more likely to be presented than others, based on their frequency of occurrence in reality). For this, a multidimensional functional learning experiment (Experiment 1) was conducted in which participants had to learn how to judge the priority of crimes on the basis of a set of cues. It was expected (Hypothesis 1) that the blocked practice group would outperform both random and operational practice groups during learning and both random and operational practice groups would perform better during the retention test. However, it was found that the random practice group performed better than the blocked group on the retention test. Performance of the operational practice group did not differ from performance of the other groups and no differences were found on performance during the learning phase. A second experiment comparing random and blocked practice schedules on learning and transfer, also failed to find an effect during the learning phase, but did show beneficial effects of a random practice schedule for transfer test.

As mentioned before, other studies have noted lack of performance differences during the learning phase with blocked and random practice schedules as we did in both experiments, but sometimes these studies also failed to find effects on a retention test (French et al., 1990; Jones & French, 2007). However, some other studies also failed to find an effect of a random practice schedule on performance during learning, but did find positive effects on transfer, as in the present study. Such findings were reported in a study conducted by Wrisberg and Liu (1991) on the effects of practice schedules on learning badminton skills. They demonstrated better retention and transfer under random practice schedule but no difference was found between groups during acquisition. And using knot-tying skills in professional fire-fighters training, Ollis et al. (2005) found that the detriment to acquisition performance as a result of a random practice schedule was not as great as previous laboratory findings would suggest. Similarly, Immink and Wright (1998) failed to find any performance detriment during acquisition of a movement task as a result of a random practice sequence. They hypothesized that, in accordance with the reconstruction hypothesis by Lee and Magill (1985), a blocked schedule benefits performance during acquisition because it obviates the need to re-plan movements between tasks, whereas participants in a random practice schedule need to engage in these time consuming re-planning activities. They subsequently showed that when participants in a random practice schedule were given sufficient time in between learning tasks to plan the upcoming response the acquisition benefit, often apparent with blocked practice schedule, disappeared while keeping the benefits for random practice schedule for transfer performance. As participants could work self-paced, a similar mechanism might underlie the lack of performance difference that was found between random and blocked practice schedules during learning phase. However, whether this assumption is correct remains an open question, but an interesting one for future research.

As mentioned above, the performance on the retention test in Experiment 1 was improved as result of a random practice schedule compared to a blocked one. This is probably due to the contextual interference that arises due to variations in tasks in a random practice schedule, but spacing may also be a relevant factor (see also Kornell & Bjork, 2008). The spacing effect refers to the consistent finding that in a given amount of study time, memory for repeated stimuli is mediated by the interval between the first and second occurrence of the stimulus, with spaced stimulus presentations leading to better memory than massed presentations (for a review, see Dempster, 1988; Donovan & Radosevich, 1999). In a random practice schedule, spacing is also automatically introduced. Richland, Finley, and Bjork (2004) conducted a study in which they investigated the relationship between contextual interference and spacing in a foreign vocabulary learning task. They came to the conclusion that contextual interference and spacing are distinct and additive effects, which both tend to lead to detrimental performance during practice but better learning outcomes.

The present study found no improvement on retention test performance after having followed an operational practice schedule. This operational practice schedule, represented by the train as you fight paradigm, failed to show higher performance compared to a blocked one. This may have been due to performance impairment on the less frequently practiced cases. In line with the elaboration hypothesis (Shea & Morgan, 1979) participants in the operational practice schedule did not have the opportunity to elaborate upon the similarities and differences between procedures for all different task variations. Therefore, as intuitively attractive as operational practice schedules may seem, a random practice schedule seems preferable over an operational one.

In the two experiments of the present study, the time interval between the learning and test (retention or transfer) was quite short, that is, in the first experiment 10 min and in the second experiment only a few minutes. Although a short or no interval between learning and test is common in many multidimensional functional learning experiments (Brehmer, 1977, 1979; Carroll, 1963; Chasseigne et al., 1997), it is known from educational and training research that effects on test performance may start to differ between conditions (or vice versa, to disappear) after a longer interval (e.g., a week; see, for example, Roediger & Karpicke, 2006), although effects may also remain stable across conditions after a longer interval (Nückles, Hübner, & Renkl, 2009). Moreover, Brehmer and Lindberg (1973) have shown that duration of retention interval hardly influenced performance on retention tests of multidimensional functional learning.

Another consideration is the complexity of the judgment task. These tasks are relevant for training real-world complex judgment tasks, because even though they may be a simplified version of a real-world task, they allow novice learners to practice with *whole* tasks, that can be increased to real-world complexity levels in later stages of training (cf. Van Merriënboer & Sweller, 2005). The exact relationships between cues and priorities in our experimental tasks may not fully correspond to real-world judgment tasks, but the type of relationships that needed to be learned resemble real-world judgment and decision-making problems (cf. distinction between physical and psychological fidelity; Van Merriënboer, 1997). Moreover, even though these tasks were not complex in the sense that they take many hours of practice, they are complex in terms of the number of interacting information elements that need to be simultaneously considered (cf. Ayres, 2006; Sweller, 1988; Sweller, Van Merriënboer, & Paas, 1998), that is, the different cues and their relationship to the criterion value. It should be noted though, that the tasks in the two experiments differed with respect to task complexity. In

Experiment 1 participants were only required to consider the additive effects of two cues that were given to them, whereas in Experiment 2, participants had to identify the relevant cues (which also interacted) from among irrelevant information, which is closer to real-world judgment tasks.

In considering the practical implications of the results of the present study, the operationalization of a blocked practice schedule, which is somewhat different from blocked schedules as they are traditionally designed, needs to be taken into account. In a blocked practice schedule, the sequence of learning tasks is such that only one task variation per block is practiced (e.g., block 1: AAAAAA, block 2: BBBBB, where A and B are variations of the learning task). As we stated in the Introduction, in complex judgment, the task for the judge is to predict a future outcome on the basis of a few cues. In learning to predict such future outcomes, the judge has to learn which cues are relevant (cue selection), how each cue relates to the criterion to be predicted, and whether cues are intercorrelated (i.e., cue–criterion functions). Thus, the creation of blocks for these complex judgment tasks is less straightforward than it would be for less complex tasks. The blocked schedules in our experiments were such that in the first blocks one cue-changed value from one task to the next. In terms of predicting the future outcomes, this would not seem to be a blocked sequence as the resulting sequential tasks are different task variations. However, in terms of the learning task, that is, to learn cue–criterion functions, participants are provided the opportunity to learn the effects of one cue–criterion function before moving on to learning the next. And in the last blocks, two (or more) cues changed value from one task to the next, providing the opportunity to learn the interaction between the cues. It may, however, be that the, albeit limited, trial-by-trial variability within a block resembles interleaved practice more than "regular" blocked practice schedules, which might perhaps explain why no performance difference was found between random and blocked practice groups during learning.

Nevertheless, in the light of learning cue–criterion relationships, this is a blocked practice schedule, and moreover, one that resembles real-world training approaches for complex judgment and decision-making. For example, sonar image operators, in a typical training program for identifying and judging sonar contacts, first learn how a sonar image depends on ocean bottom patterns, then how water temperatures influence sonar image, and only after that, how ocean bottom pattern and water temperatures interact and how that influences the sonar image (see, e.g., www.mosaichydro.com). The data presented here, however, show that such training programs better present the trainee with a random sequence of tasks, leaving it to the trainee to identify and categorize relevant cues and cue–criterion relationships.

To conclude, the present study showed that random practice schedule benefits both retention and transfer in learning a complex judgment task, and that these effects may occur without detrimental effects on performance during learning. This study focused on retention and near transfer test performance, that is, the transfer tasks were different from the learning tasks on a superficial level, but not on a deep structural level. Far transfer in complex judgment tasks, that is, when the transfer tasks also differ from the learning tasks on a deep, structural level, was not investigated, but remains an interesting subject for future research.

References

Alm, H. (1982a). Effects of pretraining on the construction of complex rules in probabilistic inference tasks. *Umea Psychological Reports, 162*, 1–13.

Alm, H. (1982b). Learning of probabilistic inference tasks: effects of uncertainty and function form. *Umea Psychological Reports, 9*, 1–36.

Alm, H., & Brehmer, B. (1982). Hypotheses about cue–criterion relations in linear and random inference tasks. *Umea Psychological Reports, 164*, 1–20.

Andersson, H., & Brehmer, B. (1977). Transfer effects in the SJT interpersonal learning paradigm. *Umea Psychological Reports, 131*, 1–42.

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction, 16*, 389–400.

Balzer, W. K., Sulsky, L. M., Hammer, L. B., & Sumner, K. E. (1992). Task information, cognitive information, or functional validity information: which components of cognitive feedback affect performance? *Organizational Behavior and Human Decision Processes, 53*, 35–54.

Beaubien, J. M., Shadrick, S. D., Paley, M. J., Badugu, S., Ennis, C. W., & Jacklin, S. (2006). Using deliberate practice to train military-civilian interagency coordination. In *Proceedings of the interservice/industry training, simulation and education conference (I/ITSEC)* (pp. 1–11). Orlando, FL: ITSA.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe, & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Björkman, M. (1965). Studies in predictive behavior: explorations into predictive judgments based on functional learning and defined by estimation, categorization, and choice. *Scandinavian Journal of Psychology, 6*, 129–156.

Brady, F. (1998). A theoretical and empirical review of the contextual interference effect and the learning of motor skills. *Quest, 50*, 266–293.

Brehmer, B. (1972). Cue utilization and cue consistency in multiple-cue probability learning. *Organizational Behavior and Human Decision Processes, 8*, 286–296.

Brehmer, B. (1973). Note on the relation between single-cue probability learning and multiple-cue probability learning. *Organizational Behavior and Human Decision Processes, 9*, 246–252.

Brehmer, B. (1977). General and specific effects of practice in using functional rules in inference tasks. *Umea Psychological Reports, 114*, 1–13.

Brehmer, B. (1979). Effect of practice on utilization of nonlinear rules in inference tasks. *Scandinavian Journal of Psychology, 20*, 141–149.

Brehmer, B. (1987). Note on subjects' hypotheses in multiple-cue probability learning. *Organizational Behavior and Human Decision Processes, 40*, 323–329.

Brehmer, B., & Almqvist, P. (1977). Learning two multiple-cue probability learning tasks in succession. *Umea Psychological Reports, 130*, 4–32.

Brehmer, A., & Brehmer, B. (1988). What have we learned about human judgment from thirty years of policy capturing? In B. Brehmer, & C. R. B. Joyce (Eds.), *Human judgment: The SJT view* (pp. 75–114) Amsterdam: North-Holland Elsevier.

Brehmer, B., & Joyce, C. R. B. (Eds.). (1988). *Human judgment: The SJT view*. Amsterdam: North-Holland Elsevier.

Brehmer, B., & Lindberg, L. A. (1973). Retention of single-cue probability learning tasks as a function of cue validity, retention interval, and degree of learning. *Journal of Experimental Psychology, 101*, 404–406.

Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review, 50*, 255–272.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62*, 193–217.

Brunswik, E., & Herma, H. (1951). Probability learning of perceptual cues in the establishment of a weight illusion. *Journal of Experimental Psychology, 41*, 281–290.

Carlson, R. A. (1989). Processing nonlinguistic negation. *American Journal of Psychology, 102*, 211–224.

Carlson, R. A., & Schneider, W. (1989). Acquisition context and the use of causal rules. *Memory and Cognition, 17*, 240–248.

Carlson, R. A., Sullivan, M. A., & Schneider, W. (1989). Practice and working memory effects in building procedural skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 517–526.

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.

Carroll, J. D. (1963)*Functional learning: The learning of continuous functional mappings relating stimulus and response continua, Vol. RB-63-26*. Princeton, NJ: Educational Testing Service.

Chasseigne, G., Mullet, E., & Stewart, T. R. (1997). Aging and multiple cue probability learning: the case of inverse relationships. *Acta Psychologica, 97*, 235–252.

Cross, E. S., Schmitt, P. J., & Grafton, S. T. (2007). Neural substrates of contextual interference during motor learning support a model of active preparation. *Journal of Cognitive Neuroscience, 19*, 1854–1871.

De Croock, M. B. M., Van Merriënboer, J. J. G., & Paas, F. (1998). High vs low contextual interference in simulation based training of troubleshooting skills: effects on transfer performance and invested mental effort. *Computers in Human Behavior, 14*, 249–267.

Dempster, F. N. (1988). The spacing effect. *American Psychologist, 43*, 627–634.

DeShon, R. P., & Alexander, R. A. (1996). Goal setting effects on implicit and explicit learning of complex tasks. *Organizational Behavior and Human Decision Processes, 65*, 18–36.

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.

Edland, A. (1993). The effects of time pressure on choices and judgments of candidates to a university program. In N. J. Castellan, Jr. (Ed.), *Individual and group decision making: Current issues* (pp. 43–64). Hillsdale, NJ: Erlbaum.

Farmer, E. W., Van Rooij, J. C. G. M., Riemersma, J. B. J., Jorna, P. G. A. M., & Moraal, J. (1999). *Handbook of simulator-based training*. London: Ashgate.

French, K. E., Rink, J. E., & Werner, P. H. (1990). Effects of contextual interference on retention of three volleyball skills. *Perceptual and Motor Skills, 71*, 179–186.

Greeno, J. G. (1964). Paired-associate learning with massed and distributed repetitions of items. *Journal of Experimental Psychology, 67*, 286–295.

Guadagnoli, M. A., & Lee, T. D. (2004). Challenge point: a framework for conceptualizing the effects of various practice conditions in motor learning. *Journal of Motor Behavior, 36*, 212–224.

Hammond, K. R., Hursch, C. J., & Todd, F. J. (1964). Analyzing the components of clinical inference. *Psychological Review, 71*, 438–456.

Hammond, K. R., McClelland, G. H., & Mumpower, J. (1980). *Human judgment and decision making: Theories, decisions, and procedures*. New York: Praeger.

Hammond, K. R., & Summers, D. A. (1965). Cognitive dependence on linear and non-linear cues. *Psychological Review, 72*, 215–224.

Hatala, R. M., Brooks, L. M., & 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.

Healy, A. F., Buck-Gengler, C. J., Barshi, I., Parker, J. T., Schneider, V. I., Raymond, W. D., et al. (2002). Optimizing the durability and generalizability of knowledge and skills. In S. P. Shohov (Ed.), *Trends in cognitive psychology* (pp. 123–192). Hauppauge, NY: Nova Science.

Hoffman, P. J., Earle, T. C., & Slovic, P. (1981). Multidimensional functional learning (MFL) and some new conceptions of feedback. *Organizational Behavior and Human Performance, 27*, 75–102.

Hogarth, R. M. (1980). *Judgment and choice: The psychology of decision*. New York: Wiley.

Hursch, C. J., Hammond, K. R., & Hursch, J. L. (1964). Some methodological considerations in multiple-cue probability studies. *Psychological Review, 71*, 42–60.

Immink, M. A., & Wright, D. L. (1998). Contextual interference: a response planning account. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 51A*, 735–754.

Jacoby, L. L. (1978). On interpreting the effects of repetition: solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior, 17*, 649–667.

Jamieson, B. A., & Rogers, W. A. (2000). Age-related effects of blocked and random practice schedules on learning a new technology. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 55*, 343–353.

Jones, L. I., & French, K. E. (2007). Effects of contextual interference on acquisition and retention of three volleyball skills. *Perceptual and Motor Skills, 105*, 883–890.

Klayman, J. (1988a). Cue discovery in probabilistic environments: uncertainty and experimentations. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14*, 317–330.

Klayman, J. (1988b). On the how and why (not) of learning from outcomes. In B. Brehmer, & C. R. B. Joyce (Eds.), *Human judgment: The social judgment theory view* (pp. 115–162). Amsterdam: North-Holland Elsevier.

Koele, P. (1980). The influence of labeled stimuli on nonlinear multiple-cue probability learning. *Organizational Behavior and Human Decision Processes, 26*, 22–31.

Koh, K. (1993). Induction of combination rules in two-dimensional function learning. *Memory and Cognition, 21*, 573–590.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: is spacing the "enemy of induction"? *Psychological Science, 19*, 585–592.

Lee, T. D., & Magill, R. A. (1983). The locus of contextual interference in motor-skill acquisition. *Journal of Experimental Psychology: Learning, Memory and Cognition, 9*, 730–746.

Lee, T. D., & Magill, R. A. (1985). Can forgetting facilitate skill acquisition? In D. Goodman, R. B. Wilberg, & I. M. Franks (Eds.), *Differing perspectives in motor learning, memory, and control* (pp. 3–22) Amsterdam: Elsevier.

Lee, T. D., & Simon, D. A. (2004). Contextual interference. In A. M. Williams, & N. J. Hodges (Eds.), *Skill acquisition in sport: Research, theory and practice* (pp. 29–44). London: Routledge.

Lindberg, L. A., & Brehmer, B. (1976). Transfer in single-cue probability learning. *Organizational Behavior and Human Decision Processes, 16*, 177–192.

Lussier, J., Shadrick, S., & Prevou, M. (2003). *Think like a commander prototype: Instructor's guide to adaptive thinking. Research product 2003–02*. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Magill, R. A., & Hall, K. G. (1990). A review of contextual interference effect in motor skill acquisition. *Human Movement Science, 9*, 241–289.

Mayer, R. E., & Wittrock, M. C. (1996). Problem-solving transfer. In D. C. Berliner, & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 47–62). New York: Simon & Schuster Macmillan.

Muchinsky, P. M., & Dudycha, A. (1975). Human inference behavior in abstract and meaningful environments. *Organizational Behavior and Human Performance, 13*, 377–391.

Nückles, M., Hübner, S., & Renkl, A. (2009). Enhancing self-regulated learning by writing learning protocols. *Learning and Instruction, 19*, 259–271.

Ollis, S., Button, C., & Fairweather, M. (2005). The influence of professional expertise and task complexity upon the potency of the contextual interference effect. *Acta Psychologica, 118*, 229–244.

Paas, F., & Van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: a cognitive-load approach. *Journal of Educational Psychology, 86*, 122–133.

Richland, L. E., Bjork, R. A., Finley, J. R., & Linn, M. C. (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). Mahwah, NJ: Erlbaum.

Richland, L. E., Finley, J. R., & Bjork, R. A. (2004). Differentiating the contextual interference effect from the spacing effect. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the twenty-sixth annual conference of the cognitive science society* (pp. 1624). Mahwah, NJ: Erlbaum.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning. Taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255.

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics practice problems boosts learning. *Instructional Science, 35*, 481–498.

Roscoe, S. N., & Williges, B. H. (1980). Measurement of transfer of training. In S. N. Roscoe (Ed.), *Aviation psychology* (pp. 182–193). Ames, IA: Iowa State University Press.

Rothstein, H. G. (1986). The effects of time pressure on judgment in multiple cue probability learning. *Organizational Behavior and Human Decision Processes, 37*, 83–92.

Ruble, T. L., & Cosier, R. A. (1990). Effects of cognitive styles and decision setting on performance. *Organizational Behavior and Human Decision Processes, 46*, 283–295.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*, 207–217.

Schneider, V. I., Healy, A. F., & Bourne, L. E., Jr. (1998). Contextual interference effects in foreign language vocabulary acquisition and retention. In A. F. Healy, & L. E. Bourne, Jr. (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 78–90). Mahwah, NJ: Erlbaum.

Schneider, V. I., Healy, A. F., & Bourne, L. E., Jr. (2002). What is learned under difficult conditions is hard to forget: contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language, 46*, 419–440.

Schneider, V. I., Healy, A. F., Ericsson, K. A., & Bourne, L. E., Jr. (1995). The effects of contextual interference on the acquisition and retention of logical rules. In A. F. Healy, & L. E. Bourne, Jr. (Eds.), *Learning and memory of knowledge and skills: Durability and specificity* (pp. 95–131). Thousand Oaks, CA: Sage.

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

Sheets, C. A., & Miller, M. J. (1974). The effect of cue–criterion function form on multiple-cue probability learning. *American Journal of Psychology, 87*, 629–641.

Simon, D. A. (2007). Contextual interference effects with two tasks. *Perceptual and Motor Skills, 105*, 177–183.

Simon, D. A. (2008). Scheduling and learning. *Advances in Psychology, 139*, 61–72.

Smedslund, J. (1955). *Multiple-probability learning: An inquiry into the origins of perception*. Oxford, UK: Akademisk.

Sweller, J. (1988). Cognitive load during problem-solving: effects on learning. *Cognitive Science, 12*, 257–285.

Sweller, J., Van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–295.

Van Merriënboer, J. J. G. (1997). *Training complex cognitive skills: A four component instructional design model for technical training*. Englewood Cliffs, NJ: Educational Technology Publications.

Van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: recent developments and future directions. *Educational Psychology Review, 17*, 147–177.

Wrisberg, G. A., & Liu, Z. (1991). The effect of contextual variety on the practice, retention, and transfer of an applied motor skill. *Research Quarterly for Exercise and Sport, 62*, 406–412.

Wulf, G., & Shea, C. H. (2002). Principles derived from the study of simple skills do not generalize to complex skill learning. *Psychological Bulletin & Review, 9*, 185–211.