

The Proportion Heuristic: Problem Set Size as a Basis for Performance Judgments

DAVID H. SILVERA,^{1*} ROBERT A. JOSEPHS² and R. BRIAN GIESLER³

¹*The University of Tromsø, Norway*

²*The University of Texas at Austin, USA*

³*Indiana University Schools of Medicine and Nursing, Indianapolis, USA*

ABSTRACT

How do people evaluate their degree of mastery over a task? A series of four studies demonstrated that a potentially irrelevant cue can have a strong influence on such evaluations. In these studies, the total amount of work given to participants (the problem set size) influenced both (a) the amount of work participants completed before feeling that they had performed well and were adequately prepared for a related future task, and (b) participants' assessments of their performance and their feelings of preparedness for a related future task. These effects occurred even when a randomization procedure was used to emphasize the arbitrary nature of the problem set size. The effects vanished, however, when participants were given extra time to evaluate their progress after completing each problem. Copyright © 2001 John Wiley & Sons, Ltd.

KEY WORDS calibration; performance judgments; problem size

An auto worker finishes 23 carburetors on her first day on the job. How does she decide whether she did good work or not? A student solves 23 practice problems one evening in preparation for an upcoming calculus examination. How does he know how prepared he is for the examination? After a few days on the job, the auto worker will be able to evaluate her work on the basis of her performance on previous days, and perhaps on the basis of the performance of coworkers who she comes to know (i.e. social comparison information; see Festinger, 1954). The student will eventually get feedback about his degree of preparation after he takes the calculus examination and gets his grade. However, it seems reasonable to expect that these individuals will want to evaluate their performances before valid environmental feedback becomes available. How do they make these evaluations?

Ideally, these individuals would have an internal sense of their degree of mastery over their respective tasks. The auto worker could then know that she is 'good' at putting together carburetors, and thereby assume that she has assembled a 'reasonable' number of carburetors in a day's work. Similarly, the student might feel that he understands the material and is prepared to do well on his calculus examination. Unfortunately for both of these individuals, self-assessments of skill acquisition represent a very difficult judgment problem

*Correspondence to: David H. Silvera, Institute of Psychology, University of Tromsø, 9037 Tromsø, Norway. E-mail: davids@psyk.uit.no

(e.g. Campione, 1987; Glaser and Pellegrino, 1987; Greeno, 1983), perhaps because this type of judgment depends largely on the ability to identify complex or ill-defined goal states (e.g. Sternberg and Frensch, 1991). Consequently, it is not surprising that research in educational psychology suggests that predictions of performance and judgments of learning are often only weakly correlated with actual performance (e.g. Maki and Berry, 1984; Maki and Serra, 1992; Weaver, 1990). Similarly, research in organizational psychology indicates that workers' assessments of their performance are often poorly calibrated when compared with objective performance measures and evaluations made by supervisors (Farh and Dobbins, 1989; Gioia and Sims, 1985; Schrader and Steiner, 1996). Poor calibration between judgment and performance has clear practical implications: inadequate preparation can lead to poor performance on the task being practiced, whereas excessive preparation is inefficient and can lead to poor performance on other tasks that compete for resources within the limited time frames that characterize many academic and work settings (e.g. Josephs and Hahn, 1995; Mazzoni and Cornoldi, 1993).

Several explanations have been offered for the poor calibration of self-assessments of performance and skill acquisition in different contexts, including the use of general domain familiarity in place of text-specific knowledge (e.g. Glenberg *et al.*, 1987), underutilization of normative item difficulty (e.g. Nelson *et al.*, 1986), and lack of prior knowledge of the task domain (e.g. Maki and Serra, 1992; Josephs and Hahn, 1995). In addition, a substantial body of research suggests that poor calibration might be caused by the use of invalid evaluation criteria – people often use simple judgment rules that rely on readily apparent context information (cf. Kahneman *et al.*, 1982) in preference to normative learning strategies that depend on information that is computationally complex (e.g. Pelham *et al.*, 1993) or difficult to access (Bryson *et al.*, 1991; Funke, 1991; Schooler *et al.*, 1993). Thus, for example, novice writers tend to be guided by text production rather than by idea production (e.g. Collins and Gentner, 1980), emphasizing simple, quantitative features of their writing (e.g. Bereiter, 1980; Flower and Hayes, 1980).

Josephs and his colleagues (Josephs *et al.*, 1994, 1996) have examined how reliance on superficial external cues can lead to poorly calibrated performance judgments. Josephs *et al.* (1994) demonstrated that individuals who produced output that was enhanced to appear physically large (inflated either by large font sizes or by being attached to empty boxes) rated their performance and productivity as greater than those who produced output that was not enhanced in this way. Although this effect was only observed when individuals were allowed to see the physical evidence of their task performance, this work provides evidence that superficial and irrelevant information associated with a task can exert a significant influence on self-evaluations of performance.

Josephs *et al.* (1996) investigated the conditions under which individuals were able to accurately assess their own learning curves (i.e. knowledge about their degree of mastery over a certain type of problem). In addition to direct learning curve access, these researchers introduced a superficial sensory cue called 'problem set size', which is simply the number of problems or items in a set of problems (e.g. the number of carburetors in the assembly room, or the number of calculus problems in a homework assignment). These researchers found that (a) participants who worked on problem sets that were pre-tested to be of relatively uniform difficulty appeared to have access to learning curve information and used this information to judge their performance, but (b) participants who worked on problems sets that were pre-tested to be of variable difficulty were unable to directly access their learning curve and relied on more superficial problem set size information.

The objective of the present studies is to more closely examine the use of the problem set size cue in assessing progress on multiple problem tasks. We propose that problem set size is a readily observable and computationally simple cue that is sometimes used to evaluate progress on a task even when more valid cues are available. We do not suggest that people use the precise ratio of problems completed as a measure of their performance. Instead, we propose that problem set size is used as a reference point against which actual performance is evaluated, and that sometimes it can be an inappropriate reference point. Previous research has shown that inappropriate reference points can lead to errors in a wide variety of contexts, including estimates of one's ability relative to others (Kruger, 1999), determination of how many units of a product to

purchase (Wansink *et al.*, 1998), gambling preferences (e.g. Ganzach, 1996), and even estimates of the number of jelly beans in a jar (Smith, 1999). In short, just as participants inappropriately used the spin of a wheel of fortune as a basis for estimating the percentage of African countries in the United Nations (Tversky and Kahneman, 1974), we propose that test-takers will use an arbitrary problem set size as a basis for estimating how many problems they *should* complete in preparation for an exam, thereby influencing their feelings of preparedness based on the number of problems they actually *do* complete.

OVERVIEW OF EXPERIMENTS

The purpose of the present studies was to demonstrate the influence of problem set size on the amount of work participants would complete before feeling that they were prepared to perform well on a related task in the future and on self-report ratings of performance and preparedness. Specifically, for Studies 1–3 it was predicted that (a) on open-ended tasks where participants could work until they felt well prepared, participants would work longer and complete more problems if they were presented with a large rather than a small problem set, and (b) on tasks where participants' output (i.e. number of problems completed) was held constant, participants presented with large problem sets would evaluate their performance and preparedness less favorably than would participants presented with small problem sets. Study 4 addressed the relationship between cognitive resources and the use of problem set size as a performance cue. Because of the superficial nature of problem set size, it was predicted that the influence of problem set size would be reduced when participants had additional time to reflect on their task performance.

STUDY 1

Method

Overview

Participants were asked to solve a series of anagrams. They were told that the problem set they received was a 'practice' test, and would be followed by a 'real' test using the same type of anagrams. Participants were randomly assigned to one of two conditions: they worked either on a set of 50 anagrams (small problem set condition) or a set of 100 anagrams (large problem set condition). Participants were told to work on the practice test until they felt prepared to perform well on the real test. No real test was actually administered. It was predicted that participants in the large problem set condition would solve more anagrams and work longer than would participants in the small problem set condition.

Participants

Participants were 16 male and 14 female undergraduates at the University of Texas at Austin who participated for course credit.

Procedure

Participants were greeted by an experimenter and led to a laboratory cubicle. After seating the participant at a desk, the experimenter explained that the purpose of the study was to test how various types of environmental conditions (e.g. noise) influence performance on tasks that require a substantial amount of concentration. Participants were told that the experimenter had chosen an anagram test as a typical problem-solving task, and that they would be given the opportunity to prepare for the 'real' anagram test by working on a set of practice problems.

The anagram problems were then described to participants – each anagram consisted of five scrambled letters, and to solve the anagram these letters must be unscrambled to form an English word. Each anagram was presented on a 3-inch by 5-inch index card. After explaining the task, the experimenter left the laboratory briefly to retrieve the anagrams. The experimenter returned to the laboratory with either 50 or 100 anagrams, depending on the experimental condition. Participants were not told how many anagrams were in the problem set. To further indicate the arbitrary nature of the size of the problem set, the experimenter told participants that a virtually unlimited number of practice anagrams was available if the participant completed the first set of practice anagrams.

Participants were given an answer sheet and scratch paper to help with their solutions. The experimenter asked participants to work as quickly and accurately as possible, and to continue working until they believed their performance had reached a level that would allow them to do well on a subsequent ‘real’ test that would involve five-letter anagrams virtually identical to the practice problems. Order of anagram presentation was randomized for each participant. Hints at the rate of 1 per minute were provided if the solution time exceeded 1 minute. A hint consisted of providing the participant with the first or next (if not the first hint for a given anagram) successive letter in the solution of the anagram. Participants’ anagram solution times were recorded. Participants were told that their responses would be timed, but that no time limit was involved.

After completing the practice problems, participants were told that there would be no real test. At this point, participants were debriefed and dismissed.

Results

All analyses were first conducted with gender as an additional predictor variable. There were no significant main effects or interactions involving gender, so gender was excluded from the final analyses.

As predicted, one-way (problem set size: small or large) between-subjects ANOVAs revealed that participants in the large problem set condition solved significantly more anagrams than did participants in the small problem set condition, $F(1, 28) = 7.35, p < 0.05$ and that participants in the large problem set condition worked significantly longer than did participants in the small problem set condition, $F(1, 28) = 5.43, p < 0.05$. These results are shown in Exhibit 1.

In addition, a one-way (problem set-size: small or large) ANOVA verified that problem-solving rates (time per anagram) did not differ between the two conditions, $F < 1$. This result suggests that problem set size had no influence on actual performance quality but instead influenced participants’ perceptions of their performance.

Discussion

In Study 1, participants who began the experiment with 100 anagrams solved significantly more anagrams and worked significantly longer than did participants who began with 50 anagrams. Thus, the initial size of a problem set influenced participants’ judgments about their preparedness.

Exhibit 1. Effects of problem set size on number of anagrams solved and time spent working on the anagram task in Study 1

	Problem set size	
	Small	Large
Anagrams solved	21.80 (<i>SD</i> = 11.01, <i>n</i> = 15)	34.27 (<i>SD</i> = 14.00, <i>n</i> = 15)
Working time (seconds)	963.47 (<i>SD</i> = 439.12, <i>n</i> = 15)	1344.73 (<i>SD</i> = 457.23, <i>n</i> = 15)

Although this result is consistent with the hypothesis that even non-diagnostic problem set size information can influence preparedness judgments, it could be the case that participants are attending to problem set size information due to experimental demand characteristics. In the real world, it seems likely that problem set sizes are chosen for a reason – instructors presumably choose homework problems based on their ideas about how much practice is needed to master the problems, and supervisors presumably choose workloads based on what they expect their subordinates to complete. Thus, participants might have inferred that the experimenter selected the size of the practice problem set based on what he believed was necessary to prepare adequately for the real test, and therefore that problem set size was a valid cue for judging their preparedness.

Another potential objection to Study 1 might be that participants weren't attending to the experimental task at all and were actually just using problem set size as a cue for when it would be acceptable for them to quit the practice test and move on to finish the experiment. If this is the case, it is possible that problem set size was used as a stopping cue as it was in Study 1, but that it wasn't really used as a preparedness cue. In other words, even though all participants were instructed to stop when they felt well-prepared for the real test, it is possible that large problem set participants actually felt more prepared (i.e. based on the extra preparatory work they had done) than did small problem set participants, than participants who worked on a smaller problem set.

Finally, it is possible that the results of Study 1 apply only to a limited population (e.g. Americans) or to the specific problem set sizes used in Study 1.

STUDY 2

Method

Overview

The procedure used in Study 1 was modified to address the potential alternative explanations and limitations of the results. Participants determined the number of anagrams they were to work on by drawing numbered slips of paper from a box. They were told that the box contained a random assortment of numbers. In actuality, half of the slips of paper were numbered 15 (small problem set condition) and the remaining slips were numbered 40 (large problem set condition). After selecting a slip of paper from the box, participants were given a number of practice anagrams equal to the number on the slip of paper they had drawn. They were asked to work on these anagrams until they felt prepared to do well on a real test of their anagram-solving ability. After completing the practice problem set, participants were given a short questionnaire asking them to evaluate their preparedness for the real test and their performance on the practice test. It was predicted that participants in the large problem set condition would solve more anagrams and work longer than would participants in the small problem set condition. The post-questionnaire was used as a manipulation check to verify that participants in the two problem set conditions stopped at the same level of subjective preparedness.

Participants

Participants were 17 male and 19 female undergraduate students at the University of Tromsø, Norway. Each participant was paid 50 Norwegian kroner (about \$7) for participating.

Procedure

The procedure was identical to that used in Study 1, except for the random determination of problem set size and the post-questionnaire. Participants were told that this was an experiment simulating a typical office

environment. The experimenter stated that, in office settings, employers often assign their employees seemingly arbitrary amounts of work. Participants were then told that, to simulate the arbitrary nature of employer-assigned tasks, the size of their task would be randomly determined, and that this random determination would be made by having the participant pull a numbered slip of paper from a box. Participants were informed that the random number on the paper they drew from the box would be the number of problems in their initial practice problem set. Unknown to participants, half of the paper slips were numbered 15 and the remaining slips were numbered 40.

The post-questionnaire consisted of two items: (1) How prepared are you for the real test? (2) How well did you perform on the practice test? These items were scored using 9-point rating scales on which a value of 1 indicated 'adequate' performance (e.g. 'performed reasonably well') and a value of 9 indicated exceptional performance (e.g. 'performed extremely well').¹

Results

All analyses were first conducted with gender as an additional predictor variable. There were no significant main effects or interactions involving gender, so gender was excluded from the final analyses.

As in Study 1, one-way (problem set size: small or large) between-subjects ANOVAs revealed that participants in the large problem set condition solved significantly more anagrams than did participants in the small problem set condition, $F(1, 34) = 16.64$, $p < 0.05$ and that participants in the large problem set condition worked significantly longer than did participants in the small problem set condition, $F(1, 34) = 15.82$, $p < 0.05$.² These results are shown in Exhibit 2.³

In addition, one-way (problem set-size: small or large) ANOVAs verified that problem solving rates (time per anagram) and hint rates (average number of hints per anagram) did not differ between the two conditions, F 's < 1 . These results suggest that problem set size had no influence on actual performance quality but instead influenced participants' perceptions of their performance.

As predicted, the post-questionnaire did not show any indication that large problem set participants actually felt better about their performance than small problem set participants. Instead, a non-significant trend for small problem set participants to rate themselves better was observed. One-way (problem set size: small

Exhibit 2. Effects of problem set size on number of anagrams solved and time spent working on the anagram task in Study 2

	Problem set size	
	Small	Large
Anagrams solved	11.06 ($SD = 4.74$, $n = 17$)	28.63 ($SD = 17.16$, $n = 19$)
Working time (seconds)	319.71 ($SD = 193.78$, $n = 17$)	932.79 ($SD = 607.71$, $n = 19$)

¹Pilot testing showed that using extremely negative statements (e.g. 'performed extremely poorly') at the low end of the scale resulted in ceiling effects on the rating items.

²The standard deviations are much larger in the large problem set condition than in the small problem set condition. As such, it is technically appropriate to use nonparametric test statistics. Mann-Whitney U tests was also conducted and revealed the same significant effects as the one-way ANOVAs described in the main text.

³It should perhaps be noted that participants with a problem set size of 40 in Study 2 solved more problems than did participants with a problem set size of 50 in Study 1. Although this appears to contradict the conclusions of this paper, we would suggest that the cultural and linguistic differences between the American (Study 1) and Norwegian (Study 2) samples renders it impossible to make valid comparisons of results across studies.

or large) between-subjects ANOVAs showed a marginally significant advantage for small problem set participants over large problem set participants in terms of perceived preparedness for the real test, $F(1, 33) = 4.15$, $p < 0.10$ (M 's 5.53 and 4.47 respectively), and no difference in terms of perceived performance on the practice test, $F(1, 33) = 1.75$, $p < 0.20$ (M 's 4.47 and 3.74 respectively).

Discussion

Replicating the results obtained in Study 1, participants who began with a large problem set solved significantly more anagrams and worked significantly longer than participants who began with a small problem set before they felt adequately prepared to take the real test. However, this extra work failed to gain them anything in terms of self-perceived performance and preparedness. This study extends and clarifies the results of Study 1 in four important ways: (a) by allowing participants to randomly determine problem set size, any potentially diagnostic properties or demand characteristics associated with problem set size were minimized; (b) the possibility that participants actually felt better about themselves after working longer was ruled out; (c) the findings of Study 1 were replicated using Norwegian participants, demonstrating that the influence of problem set size is not simply an idiosyncratic feature of American culture; and (d) the size of the problem sets was changed to demonstrate the generalizability of the phenomenon beyond the specific problem set sizes used in Study 1.

In both Study 1 and Study 2 the number of problems solved and the amount of time used to solve them were the dependent variables. No direct attempt was made in Study 1 to measure how prepared participants felt after completing the practice problems, and in Study 2 this measurement was only made after participants had received unequal treatments (i.e. they had worked on a different number of problems taken from different sized problem sets).

In a world where people work a relatively fixed number of hours per day, however, the situation is likely to be different. Rather than working until they feel adequately prepared, it is often the case that people put in a fixed amount of effort and accept whatever output they have completed. Study 3 examined the impact of problem set size under these conditions by holding performance constant and using a verbal assessment of performance and preparedness as the dependent measure.

STUDY 3

Method

Overview

Participants were asked to solve a series of practice anagrams, with the objective of preparing themselves for a real test that would ostensibly follow. They drew a number from a box that was used to randomly assign them to one of two conditions: they worked on a practice problem set of either 15 anagrams (small problem set condition) or 40 anagrams (large problem set condition). Participants were asked to work until an unspecified time limit was reached. In fact, all participants were interrupted after completing 12 anagrams. Participants were then asked to evaluate their performance on the practice test and their degree of preparedness for the real test. It was predicted that participants in the small problem set condition would evaluate themselves more favorably than would participants in the large problem set condition.

Participants

Participants were 8 male and 19 female undergraduate students at the University of Tromsø, Norway. Each participant was compensated for participating with an instant payoff lottery ticket worth 20 Norwegian kroner (about \$2.50).

Procedure

The procedure was identical to that used in Study 2, with three exceptions. First, to avoid the possibility that participants might use randomly drawn problem set sizes as a basis for inferring the distribution of numbers they could have drawn and thereby the experimenter's expectations concerning the amount of work they should complete, participants were given false information about the distribution of the random numbers. Specifically, participants were told that the slips of paper they could draw contained the values 10, 15, 20, 25, 30, 35, 40, 45, and 50, with an equal frequency for each value. Second, rather than being instructed to work until they felt prepared for the real test, participants were instructed to work until the experimenter stopped them after an unspecified time limit was reached. Finally, before being informed that there would be no real test, participants were given a self-evaluation questionnaire. This questionnaire asked participants to assess: (a) how well they thought they had performed on the practice test, (b) how good they were at solving anagrams, (c) how prepared they felt to take the real test, (d) how well they thought they would perform on the real test, (e) how many practice anagrams they felt the average person would have solved under the same testing conditions, and (f) how many anagrams they could complete on the real test within a 15-minute time limit. Items (a) through (d) used 9-point rating scales, where a value of 1 indicated 'adequate' performance (e.g. 'performed reasonably well') and a value of 9 indicated exceptional performance (e.g. 'performed extremely well'). Items (e) and (f) were free response measures. After completing the questionnaire, participants were debriefed and dismissed.

Results

All analyses were first conducted with gender as an additional predictor variable. There were no significant main effects or interactions involving gender, so gender was excluded from the final analyses. Preliminary analyses indicated that participants in the small and large problem set conditions performed equally well in terms of both time taken to solve the anagrams and number of hints required, F 's < 1.

Because the six self-evaluative items used different measurement scales, participants' responses to all of these items were standardized. The item assessing performance of the average person was reverse scored, because lower values reflect positively on the participant (i.e. a low value for the average person implies that the participant performed better than average). A Cronbach's alpha of 0.80 indicated a high degree of internal reliability among self-evaluative items (a) through (e), but including item (f) resulted in a substantial reduction in reliability (Cronbach's alpha = 0.72). Consequently, the standardized scores on items (a) through (e) were combined to form a single self-evaluation index, which was computed as the simple sum of the five standardized scores. The raw value for item (f) was analyzed separately.

The self-evaluation index was used as the dependent measure in a one-way (problem set size: small or large) between-subjects ANOVA. As predicted, participants in the small problem set condition ($M = 1.63$, $SD = 3.75$, $n = 13$) evaluated themselves significantly more favorably than did participants in the large problem set condition ($M = -1.51$, $SD = 3.15$, $n = 14$), $F(1, 25) = 5.58$, $p < 0.05$. A one-way (problem set size: small or large) between-subjects ANOVA also revealed that participants in the small problem set condition ($M = 26.31$, $SD = 9.20$, $n = 13$) believed they could complete significantly more problems on the real test than their large problem set counterparts ($M = 16.79$, $SD = 11.42$, $n = 14$), $F(1, 25) = 5.64$, $p < 0.05$.

Discussion

Study 3 demonstrates that problem set size influences not only how much work one does before feeling prepared, but also how prepared one feels after doing a fixed amount of work. This influence occurs despite the fact that problem set size is clearly an invalid performance cue. What, then, are the limitations of this influence? Josephs *et al.* (1996) demonstrated one situation in which the influence of problem set size on judgments of preparedness was attenuated: when these researchers made a valid performance cue (i.e. learning

curve information) more accessible, the valid cue was used in preference to problem set size. This suggests that the accessibility advantage of problem set size information might be an essential determinant of when this information will be used.

But why do people prefer to use easily accessible cues even when they are not particularly valid? One explanation might be the chronic dilemma faced by anyone trying to evaluate their performance on a task: the person must evaluate herself at the same time she is performing the task. At least when the task is a demanding mental task (e.g. anagrams or math problems), both the evaluation and the performance itself can require a great deal of cognitive resources. Thus, it is possible that the performance of such a task interferes with the accurate assessment of task performance. In fact, many experimenters have used manipulations very similar to the anagram task in the present studies to impair participants' judgment processes. Such 'cognitive load' manipulations have been shown to interfere with a variety of mental processes, including impression formation (e.g. Newman *et al.*, 1996), use of stereotypes (e.g. Gordon and Anderson, 1995), and mood control (e.g. Wegner *et al.*, 1993).

In effect, the anagram task in Studies 1–3 might have placed participants under cognitive load. This contention is supported by the behavior of the participants in the first three studies: (a) none of these participants showed any indication of evaluating their performance at any point during the practice test; and (b) during debriefing interviews after completing the procedures, many participants commented on the high degree of concentration required to solve the anagrams. This suggests the possibility that the cognitive demands of solving the anagram task resulted in an overreliance on easily accessible evaluative cues such as problem set size. Study 4 attempted to test this hypothesis by reducing the mental interference created by the anagram task.

STUDY 4

Method

Overview

Participants drew numbered slips of paper from a box. Half of the paper slips were numbered 15 (small problem set condition) and half were numbered 40 (large problem set condition). After drawing a slip of paper, participants were given a number of practice anagrams equal to the number they had drawn. They were asked to work on these practice anagrams until they felt prepared to do well on a real test of their anagram solving ability. In addition to the manipulation of problem set size, cognitive load was manipulated. In the cognitive load condition, participants worked as in the first three experiments – they simply informed the experimenter when they reached an adequate state of preparedness. In the no load condition, however, participants were asked to stop after each anagram and consider whether they needed further preparation, thus interrupting the flow of the anagram task to give them a better opportunity to evaluate their progress. We anticipated that this interruption was likely to have side effects beyond reducing cognitive load (e.g. boredom with repeatedly asking for more anagrams and increased feelings of pressure to stop the anagram task), and therefore expected a main effect such that participants in the no load condition would solve fewer anagrams than participants in the cognitive load condition. The critical prediction, however, was that problem set size would have a greater influence on participants in the cognitive load condition than on those in the no load condition.

Participants

Participants were 35 male and 28 female undergraduate students at the University of Tromsø, Norway. Each participant was compensated for participating with an instant payoff lottery ticket worth 20 Norwegian kroner (about \$2.50)

Procedure

For participants in the cognitive load condition, the procedure was identical to that used in Study 2. Half of participants were randomly assigned to the no-load condition. There were two differences between the no-load condition and the cognitive load condition: (a) prior to beginning the anagram task, the experimenter instructed no load participants to consider whether they had reached an adequate state of preparation after they finished each practice anagram, and (b) although the pile of practice problems was placed in front of the participant as in previous experiments, the experimenter (rather than the participant) turned over each card to reveal the next problem if the participant indicated that he or she needed to do more problems to reach an adequate level of preparation.

Results

All analyses were first conducted with gender as an additional predictor variable. There were no significant main effects or interactions involving gender, so gender was excluded from the final analyses.

A 2 (problem set size: small or large) \times 2 (cognitive load or no cognitive load) between-subjects ANOVA was conducted with number of practice anagrams completed as the dependent variable. The predicted cognitive load by problem set size interaction was significant, $F(1, 59) = 10.80, p < 0.05$ (see Exhibit 3), such that problem set size had a greater effect on the performance of participants in the cognitive load condition than on the performance of participants in the no-load condition. This analysis also revealed the anticipated main effect for cognitive load, $F(1, 59) = 10.51, p < 0.05$, such that no-load participants solved fewer anagrams than their counterparts under cognitive load, and a main effect for problem set size, $F(1, 59) = 10.42, p < 0.05$, such that participants with large problem sets solved more anagrams than did participants with small problem sets.

Planned comparisons showed that participants in the cognitive load condition solved significantly more anagrams in the large problem set condition ($M = 29.61, SD = 17.11, n = 18$) than in the small problem set condition ($M = 10.81, SD = 4.78, n = 16$), $F(1, 32) = 18.02, p < 0.05$, but that participants in the no-load condition did not solve a significantly different number of anagrams in the large problem set ($M = 10.77, SD = 9.89, n = 13$) and small problem set ($M = 10.94, SD = 8.74, n = 16$) conditions, $F < 1$.⁴

A second 2 (problem set size: small or large) \times 2 (cognitive load or nocognitive load) between-subjects ANOVA was conducted with time spent working on the practice problems as the dependent variable. Again, the predicted cognitive load by problem set size interaction was significant, $F(1, 59) = 12.56, p < 0.05$, such that problem set size had a greater influence in the cognitive load condition than in the no load condition (see Exhibit 4). This analysis also revealed the anticipated main effect for cognitive load, $F(1, 59) = 7.85,$

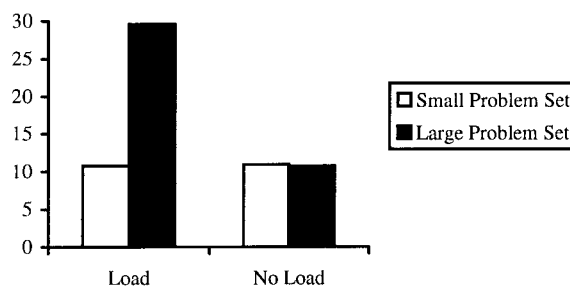


Exhibit 3. Effects of cognitive load and problem set size on number of anagrams solved in Study 4

⁴In the cognitive load condition, the standard deviations are much larger in the large problem set condition than in the small problem set condition, making it appropriate to use nonparametric test statistics. For both number of anagrams solved and time spent solving anagrams, Mann-Whitney U tests were also conducted and revealed the same effects as the planned comparisons shown in the main text.

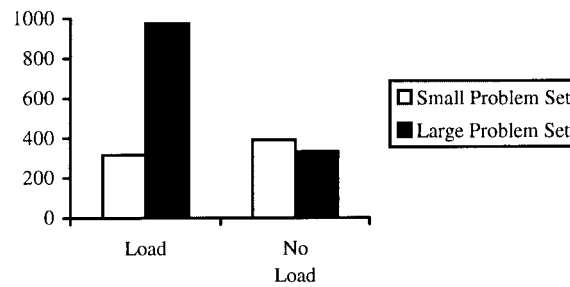


Exhibit 4. Effects of cognitive load and problem set size on time spent solving anagrams in Study 4

$p < 0.05$, such that no-load participants spent less time working on the problems than their counterparts under cognitive load, as well as a significant main effect for problem set size, $F(1, 59) = 8.75$, $p < 0.05$, such that participants with large problem sets spent more time working on the problems than did participants with small problem sets.

Planned comparisons showed that participants in the cognitive load condition spent significantly more time (in seconds) working on the anagram task in the large problem set condition ($M = 974.61$, $SD = 596.53$, $n = 18$) than in the small problem set condition ($M = 316.81$, $SD = 199.76$, $n = 16$), $F(1, 32) = 17.64$, $p < 0.05$, but that participants in the no-load condition did not spend significantly different amounts of time working on the anagram task in the large problem set ($M = 332.69$, $SD = 335.15$, $n = 13$) and small problem set ($M = 392.00$, $SD = 303.61$, $n = 16$) conditions, $F < 1$.

Discussion

The results of Study 4 are consistent with the interpretation that working on a task can interfere with the judgment processes involved in evaluating one's mastery over that task. The cognitive load group replicated the results of the first three studies. These participants used the size of their practice problem set as a cue for evaluating their performance, as demonstrated by the fact that they worked more problems and spent more time when they had a large problem set than when they had a small problem set.

Conversely, participants in the no-load condition demonstrated that the invalid influence of problem set size can be negated by periodically 'stepping back' and evaluating one's progress on a task. These participants were not influenced by problem set size in terms of either the number of anagrams solved or the amount of time spent working on the problem set. Importantly, the manipulation of cognitive load had no effect on the performance of participants with small problem sets, ruling out the possible alternative explanation that the load manipulation was such a distraction that it inhibited the performance of all participants.

Although the anagram task is very similar to tasks that have been used to manipulate cognitive load in previous research, it might still seem a bit unorthodox to have used a manipulation of cognitive 'un-load' in the present study. Some might suggest that what was really manipulated in the present study was reflection time, and we wouldn't disagree with this characterization. The point remains that participants who didn't spend time considering their progress on the experimental task relied heavily on the invalid problem set size cue, whereas participants who did spend time considering their progress were not influenced by problem set size, presumably abandoning it in favor of a more appropriate performance cue (cf. Josephs *et al.*, 1996).

GENERAL DISCUSSION

Four experiments provided evidence for a metacognitive learning strategy based on problem set size. Participants took less time and effort to reach a subjective state of preparedness, and rated their performance and

preparedness more favorably when they worked on small problem sets rather than large problem sets. This effect was observed cross-culturally, and across different combinations of problem set sizes. The influence of problem set size was eliminated when participants evaluated their progress after each item in the problem set.

The importance of these findings is perhaps best evaluated in the context provided by the literatures associated with metacognitive skill acquisition and with human judgment and decision processes. With a few exceptions (e.g. Bryson *et al.*, 1991), metacognition research tends to portray people as skilled at identifying and using good learning strategies while abandoning poor learning strategies. This portrayal has appeared in a variety of learning domains, including reading (e.g. Daneman, 1991), writing (e.g. Bryson *et al.*, 1991), memory (e.g. Pressley *et al.*, 1985), and general problem-solving (e.g. Kluwe, 1987). In contrast, the judgment and decision-making literature describes a multitude of errors, biases, and shortcomings associated with human decision-making (see Kahneman *et al.*, 1982, for a review).

The research presented in the current paper can be viewed as a bridge between the metacognitive literature and the judgment and decision-making literature, rather than clearly fitting into either of these domains. The problem being studied, the use of a particular metacognitive stopping strategy in skill acquisition, falls within the domain of metacognition. On the other hand, the heuristic nature of the use of problem set size as a stopping cue is more typical of judgment and decision-making research. Thus, we believe that the present research contributes both to the judgment and decision-making literature, by demonstrating a novel judgment heuristic and its application to skill acquisition, and to the metacognition literature, by documenting a metacognitive stopping strategy that sometimes interferes with efficient skill acquisition.

The process of self-evaluation

The ultimate goal of research such as this is to fully understand the process of self-evaluation in a learning context (e.g. the determination of when one has done enough to consider oneself adequately prepared). This process, in general terms, has two components: (a) the set of strategies or cues used to make self-evaluative judgments, and (b) the rules used to determine which of these strategies will be used in a particular instance.

The present studies suggest that problem set size is one cue that can contribute to self-evaluations, but when will this cue be used? It seems clear that problem set size is not used simply because it is perceived to be diagnostic—the random draw procedure should have made the non-diagnosticity of problem set size evident, yet participants continued to use problem set size as an indicator of their performance. Instead, we believe that problem set size information is used primarily because it is easy to apply. In other words, because problem set size (a) is based on simple, easily observable characteristics of the environment and (b) requires essentially no mental effort to apply to judgments, it seems reasonable to postulate that ease of application is an important factor leading to the use of problem set size information. As such, we propose that problem set size is likely to be used as a metacognitive cue whenever other, more diagnostic cues cannot be processed within the cognitive and motivational limitations of the performer. This interpretation is consistent with the results of Study 4, in which participants were shown to be less reliant on problem set size information when they were given more time to think about their progress. It is also consistent with Josephs *et al.*'s (1996) finding that when participants were given easier access to a better learning cue (learning curve information), they abandoned problem set size information in favor of that cue. Further research is needed, however, both to further clarify the conditions under which more or less diagnostic cues are accessed by learners, and to identify exactly what these cues are.

Applications of judgment by problem set size

The judgment by problem set size strategy has straightforward applications to both business (i.e. employee productivity) and education (i.e. test preparation) settings. However, the present research suggests that

attempts at application might be more hazardous than they appear. The results of Studies 1, 2, and 4 indicate that one way to increase performance on open-ended tasks is to increase the size of the problem set (e.g. by assigning more tasks to a worker). However, the results of Study 3 suggest that, to the extent that time or effort is held constant (e.g. an 8-hour work day or a 1-hour study period), problem set size is *negatively* related to feelings of performance and preparedness. As such, it is possible that repeated exposure to increased problem set sizes could lead to chronic reductions in feelings of self-efficacy and self-esteem. Such potential long-term effects should be carefully considered in any attempted application of these findings.

ACKNOWLEDGEMENTS

We would like to thank Dan Gilbert, Art Glenberg, Doug Krull, Ruth Maki, Bill Swann, Karl Halvor Teigen, Bas Verplanken, and Bill von Hippel for their comments on earlier versions of this manuscript. This research was supported by a predoctoral fellowship from the National Science Foundation Research Fellowship Program (GER-9253937). The support of this institution is gratefully acknowledged.

REFERENCES

- Bereiter C. 1980. Development in writing, In *Cognitive Processes in Writing*, Gregg, L. W. and Steinberg, E. R. (eds). Erlbaum: Hillsdale, NJ; 73–93.
- Bryson M, Bereiter C, Scardamalia M, Joram E. 1991. Going beyond the problem as given: Problem solving in expert and novice writers, In *Complex Problem Solving: Principles and Mechanisms*, Sternberg, R. J. and Frensch, P. A. (eds). Erlbaum: Hillsdale, NJ; 61–84.
- Campione JC. 1987. Metacognitive components of instructional research with problem learners, In *Metacognition, Motivation, and Understanding*, Weinert FE, Kluwe RH. (eds). Erlbaum: Hillsdale, NJ; 117–140.
- Collins A, Gentner DA. 1980. Framework for a cognitive theory of writing, In *Cognitive Processes in Writing*, Gregg LW, Steinberg ER. (eds). Erlbaum: Hillsdale, NJ; 61–72.
- Crosby F, Bromley S, Saxe L. 1980. Recent unobtrusive studies of black and white discrimination and prejudice: A literature review, *Psychological Bulletin*, **87**: 546–563.
- Daneman M. 1991. Individual differences in reading skills, In *The Handbook of Reading Research*, (Vol. 2), Barr R, Kamil ML, Mosenthal PB, Pearson PD. (eds). Longman: White Plains, NY.
- Farh J, Dobbins GH. 1989. Effects of comparative performance information on the accuracy of self-ratings and agreement between self- and supervisor ratings, *Journal of Applied Psychology*, **74**: 606–610.
- Festinger LA. 1954. A theory of social comparison processes, *Human Relations*, **7**: 117–140.
- Flower LS, Hayes JR. 1980. The dynamics of composing: Making plans and juggling constraints, In *Cognitive Processes in Writing*, Gregg LW, Steinberg ER. (eds). Erlbaum: Hillsdale, NJ; 31–50.
- Funke J. 1991. Solving complex problems: Exploration and control of complex social systems, In *Complex Problem Solving: Principles and Mechanisms*, Sternberg RJ, Frensch PA. (eds). Erlbaum: Hillsdale, NJ; 185–222.
- Ganzach Y. 1996. Preference reversals in equal-probability gambles: A case for anchoring and adjustment, *Journal of Behavioral Decision Making*, **9**: 95–109.
- Gioia DA, Sims HP. 1985. Self-serving bias and actor observer differences in organizations: An empirical analysis, *Journal of Applied Social Psychology*, **15**: 547–563.
- Glaser R, Pellegrino JW. 1987. Aptitudes for learning and cognitive processes, In *Metacognition, Motivation, and Understanding*, Weinert FE, Kluwe RH. (eds). Erlbaum: Hillsdale, NJ; 267–288.
- Glenberg AM, Sanocki T, Epstein W, Morris C. 1987. Enhancing calibration of comprehension, *Journal of Experimental Psychology: General*, **116**: 119–136.
- Gordon RA, Anderson KS. 1995. Perceptions of race-stereotypic and race-nonstereotypic crimes: The impact of response time instructions on attributions and judgments, *Basic and Applied Social Psychology*, **16**: 455–470.
- Greeno JG. 1983. Forms of understanding in mathematical problem solving, In *Learning and Motivation in the Classroom*, Paris SG, Olson CM, Stevenson, HW. (eds). Erlbaum: Hillsdale, NJ.

- Josephs RA, Giesler RB, Silvera DH. 1994. Judgment by quantity, *Journal of Experimental Psychology: General*, **123**: 21–32.
- Josephs RA, Hahn ED. 1995. Bias and accuracy in estimates of task duration, *Organizational Behavior and Human Decision Processes*, **61**: 202–213.
- Josephs RA, Silvera DH, Giesler RB. 1996. The learning curve as a metacognitive tool, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **22**: 510–524.
- Kahneman D, Slovic P, Tversky A. 1982. *Judgment under Uncertainty: Heuristics and Biases*, Cambridge University Press: Cambridge.
- Kluwe RH. 1987. Executive decisions and regulation of problem solving behavior, In *Metacognition, Motivation, and Understanding*, Weinert FE, Kluwe RH. (eds). Erlbaum: Hillsdale, NJ; 31–64.
- Kruger J. 1999. Lake Wobegon be gone! The below-average effect and the egocentric nature of comparative ability judgments, *Journal of Personality and Social Psychology*, **77**: 221–232.
- Maki RH, Berry SL. 1984. Metacomprehension of text material, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **10**: 663–679.
- Maki RH, Serra M. 1992. The basis of test predictions for text material, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **18**: 116–126.
- Mazzoni G, Cornoldi C. 1993. Strategies in study time allocation: Why is study time sometimes not effective?, *Journal of Experimental Psychology: General*, **122**: 47–60.
- Metcalfe J. 1986. Feeling of knowing in memory and problem solving, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **12**: 288–294.
- Nelson TO, Leonesio RJ, Landwehr RS, Narens L. 1986. A comparison of three predictors of an individuals memory performance: the individuals feeling of knowing versus the normative feeling of knowing versus base-rate item difficulty, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **12**: 279–287.
- Newman LS, Duff KJ, Hedberg DA, Blitstein J. 1996. Rebound effects in impression formation: Assimilation and contrast effects following thought suppression, *Journal of Experimental Social Psychology*, **32**: 460–483.
- Pelham BW, Sumarta TT, Myaskovsky L. 1994. The easy path from many to much: The numerosity heuristic, *Cognitive Psychology*, **26**: 103–133.
- Pressley M, Borkowski JG, OSullivan J. 1985. Childrens metamemory and the teaching of memory strategies, In *Metacognition, Cognition, and Human Performance*, (Vol. 1), Forrest-Pressley DL, MacKinnon GE, Waller TG (eds). Academic Press: New York; 111–153.
- Schooler JW, Ohlsson S, Brooks K. 1993. Thoughts beyond words: When language overshadows insight, *Journal of Experimental Psychology: General*, **122**: 166–183.
- Schrader BW, Steiner DD. 1996. Common comparison standards: An approach to improving agreement between self and supervisory performance ratings, *Journal of Applied Psychology*, **81**: 813–820.
- Smith H. 1999. Use of the anchoring and adjustment heuristic by children, *Current Psychology: Developmental, Learning, Personality, Social*, **18**: 294–300.
- Sternberg RJ, Frensch PA. 1991. *Complex Problem Solving: Principles and Mechanisms*, Erlbaum: Hillsdale, NJ.
- Tversky A, Kahneman D. 1974. Judgment under uncertainty: Heuristics and biases, *Science*, **185**: 1124–1131.
- Wansink B, Kent RJ, Hoch SJ. 1998. An anchoring and adjustment model of purchase quantity decisions, *Journal of Marketing Research*, **35**: 71–81.
- Weaver CA. 1990. Constraining factors in calibration of comprehension, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **16**: 214–222.
- Wegner DM, Erber R, Zanakos S. 1993. Ironic processes in the mental control of mood and mood-related thought, *Journal of Personality and Social Psychology*, **65**: 1093–1104.
- Wilson TD. 1985. Strangers to ourselves: The origins and accuracy of beliefs about ones own mental state, In *Attribution: Basic Issues and Applications*, Harvey JH, Weary G (eds). Academic Press: New York; 9–36.

Authors' biographies:

David Silvera is an Associate Professor for the Institute of Psychology at the University of Tromsø, Norway. His primary research interests are self-esteem, social judgment, and decision making.

Robert Josephs is an Associate Professor of Psychology in the Department of Psychology at the University of Texas at Austin. His research focuses on sex differences in intellectual functioning.

R. Brian Giesler, PhD, is an Assistant Research Scientist at the Indiana University Schools of Medicine and Nursing and an investigator with the Mary Margaret Walther Program for Cancer Care Research. His research interests include judgement and decision making in medical contexts.

Authors' addresses:

David Silvera, Institute of Psychology, University of Tromsø, 9037 Tromsø, Norway.

Robert Josephs, Department of Psychology, University of Texas at Austin, TX, USA.

R. Brian Giesler, Indiana University Schools of Medicine and Nursing, Indianapolis, IN, USA.

