Number and Content Variability of Instructive Examples Promote Structure-based Learning

Faria Sana, Joseph A. Kim
Department of Psychology, Neuroscience & Behaviour, McMaster University

Abstract

For novice learners of statistics, successful recognition of a statistical concept in a problem requires an understanding of abstract rules and principles. Whereas experts focus on structure, novices typically rely on surface features such as the storyline presented in the problem. However, novices can learn to foster expert-like strategies with exposure to examples that vary the surface features to promote structure-based learning. This strategy may be further improved by increasing the number of examples used during initial learning. The purpose of this study was to examine the effects of short-term guided training on problem recognition in novices. Three statistical concepts were each illustrated with two or three examples which had high or low content variability. Results support the use of three high content variable examples as a simple and time-effective implementation to learning.

1. Introduction

During initial learning, studying a source task may improve later performance on a target task. Analogical transfer is the degree to which this prior learning affects later learning. The similarity in content (i.e., storyline, events, names, objects) between the source and target tasks can vary ranging from highly similar (near transfer) to highly dissimilar (far transfer). Importantly, if the tasks are too dissimilar, there may be no transfer at all.

Learning with examples is critical during the initial learning stage; even when rules are provided, learners use examples to understand how concepts are instantiated [1]. Consequently, learning with multiple examples increases the likelihood that transfer will be attempted [2:4]. A single example used during learning may inadvertently lead a learner to focus on irrelevant details specific to the particular example. However, using multiple examples promotes learners to focus on relevant features, and link structural commonalities [5:6]. Gick and Holyoak (1983) [7] argued that multiple source examples promote better encoding and abstraction of generalized schema for a concept category, which in turn encourages structural retrieval of that concept at a later time [8:9]. Accordingly, greater access to a pool of examples will result in better schema abstraction of a given concept.

Prior research has typically involved the use of either one or two examples during training phases, with only a few studies that directly investigated the relationship between number of examples used to learn concepts and transfer performance [10]. Moreover, limited studies have focused on the role of content variability in instructive examples which may better promote attention to structural features of the problem [11:12].

Successful transfer is also dependent on the overall perceived similarity between source and target tasks. Two problems are similar if they share surface features (i.e., storyline, events, names, objects) and/or structural relations (i.e., principles, equations, procedures) [13]. However, perceived similarity is affected by factors such as content, prior knowledge and level of expertise, which does not always correspond to structural similarity. Once perceived similarity is identified, it prompts the retrieval of source examples. However, whether information is accurately transferred largely depends upon the degree of structural similarity between source and target tasks. The effect of perceived similarity on transfer is best demonstrated by research on expert-novice differences. Experts represent problems based on principles and rules, resulting in positive transfer. Novices, unaware of structural similarities and dissimilarities between problems, make use of salient content features to solve later problems, which can lead to negative transfer [14:16]. Importantly, these findings suggest that it is not just the number of examples that influence transfer, but also the representativeness and variability of source examples during initial learning that play a role in the accuracy of transfer. In light of research which suggests that positive transfer is due

to structural rather than surface encoding, we explore conditions under which manipulating exemplar content and number results in long lasting effects on structural learning.

2. Experiment overview

The current study examines problem recognition of statistical concepts in novices using immediate and delayed (24 hour) tests to assess performance on far and near transfer tasks. Four independent groups learned statistical concepts which varied in number of instructive examples used (two or three) and content variability (high or low). Three predictions were made: First, learning with three, rather than two examples may be more beneficial because learners will articulate relevant structural relations with greater accuracy. This abstraction may in turn facilitate better retrieval of the concept representations [17]. Second, high content variability may promote schema construction by providing a more complete representation of concepts as they are not restricted to a limited set of surface features [4]. Conversely, low content variable examples will narrow focus to specific and possibly irrelevant surface features of a concept, resulting in poor performance on far transfer tasks. Third, performance on near transfer tasks will be higher than far transfer tasks, particularly when learning occurs with two examples. This prediction is consistent with Digerjean and Nogry's [17] claim that information learned will likely transfer to another situation if the degree of surface similarity between the source and target tasks is increased.

3. Method

3.1. Participants

A total of 84 first year psychology undergraduate students (mean age = 20.53, males = 39, female = 45) from McMaster University, Hamilton, Ontario, Canada enrolled in the study for course credit or CAD $10 payment. Data of 3 participants were discarded due to failure to follow instructions (1) or prior exposure to non-parametric statistics (2).

3.2. Experimental-design

The between- and within-subjects experimental design involved learning three non-parametric statistical tests/concepts, each paired with either two or three illustrative examples. The experiment consisted of a 2 (number of examples: two or three) x 2 (content of exemplars: surface or structure emphasizing) x 2 (testing delay: immediate or 24 hour delayed) mixed-factorial design.

The factors number of examples and content of exemplars were manipulated between-groups. Number of examples was operationalized by varying the number of examples used to illustrate each of the three concepts. This factor consisted of two levels: (1) two-examples condition, in which each concept was demonstrated by a set of two word-problems, and (2) three-examples condition in which each concept was paired with a set of three word-problems. Content of exemplars was operationalized by varying the content (i.e., storyline and details of the context of examples) within and across concepts. This factor was comprised of two levels: (1) surface-emphasizing examples: all examples of a given concept shared similar storylines which were different from the storylines used for the other two concepts, and (2) structure-emphasizing examples: different exemplar storylines were used to illustrate a concept, but the same storylines were used for all three concepts. The factor of testing delay was manipulated within-groups. It was operationalized by varying the time when learning was tested: (1) immediate testing occurred at the end of the first session, and (2) delayed testing occurred 23 – 25 hours following the first session.

3.3. Materials

Materials consisted of a questionnaire, descriptions of three non-parametric tests, two and three example-sets illustrating each of the three statistical concepts, six multiple-choice questions to screen for prior learning, and a series of eighteen problem-recognition tasks, nine of which were far transfer and nine near transfer tasks.

The assessment questionnaire was used to collect information on participant's age, gender, academic year and program, and number of mathematics and statistics courses taken since grade 11.

Participants learned about three statistical tests in this experiment: Chi-squared test, Kruskal-wallis test, and Wilcoxon signed-rank test. Half-page descriptions, taken from an introductory statistics book [18], were used to explain characteristics of the concepts. Table 1 highlights the characteristics of interest.

A set of two or three word-problems were used to illustrate each concept following its description. The examples were primarily used to show participants a typical word problem for each statistical test and were not worked-out examples with solutions.

Adopting Quilici and Mayer's [12] methodology, we refer to sets as 'surface-emphasizing' (i.e., low content variable) for

Table 1. Structural features of three statistical concepts used during the learning phase

Statistical test	Structural features			
	IV (# of groups)	Sample	DV	Main question
Kruskal-wallis	3+	independent	quantitative	are there any differences between three or more conditions?
Wilcoxon signed-rank	2	dependent	quantitative	is there a significant change in a condition after some treatment?
Chi-squared test of independence	2	-	categorical	is there a relationship between two variables?

conditions in which examples of a given statistical concept shared similar versions of a particular storyline, and different storylines were used for each statistical concept (see Table 2). 'Structure-emphasizing examples' (i.e., high content variable) were used in conditions where each example consisted of a different storyline for a given concept, but the same exemplar storylines were used to demonstrate all three concepts (see Table 3).

The distinct storylines were kept consistent across all conditions. Storyline 1 was always about gentlemen's hair preference. Storyline 2 consisted of grannies catching colds, and storyline 3 (which was used in the three-example conditions) contained content on fruit population.

An initial screening test consisting of six multiple-choice questions was used to gauge initial understanding of the statistical concepts. It included basic recall questions such as matching the type of sample (e.g., paired, independent) and data (e.g., categorical, quantitative) with the corresponding concept. For each participant, the total score from this screening test was used to measure below chance performance; four participants were removed from the study as a result of this exclusion criterion.

During the testing phase, participants responded to problem recognition tasks, which entailed a series of eighteen word-problems with varying content. For a given word-problem, participants were asked to select the correct statistical test used to solve the problem from one of three options. Word-problems in the testing phase consisted of nine far transfer questions and nine near transfer questions.

Far transfer occurs when learned information is transferred to a context dissimilar to the one in which information was initially learned. These word-problems, taken from various statistics books, were completely novel in content (i.e., the storylines used in these questions were not similar to the storylines used in the source examples). Near transfer occurs when some learned information is carried over to a context similar to the one in which information was initially processed. Therefore, these target word-problems shared similar surface features (i.e., storylines) with their source examples. See Tables 4 and 5 for a summary of operationalizing near and far transfer recognition tasks in surface- and structure-emphasizing conditions.

3.4. Reliability of word-problems

Far transfer word-problems were taken from statistics textbooks and near transfer problems were created by the experimenter. To ensure that the latter set of word-problems were not influenced by experimenter bias, two raters (PhD candidates from the Department of Statistics, McMaster University) were asked to independently record which non-parametric test would be used to solve each problem. Cronbach's α (r = .848) was used to calculate the inter-rater agreement. Experimenter made revisions to the word-problems with low item agreement.

3.5. Counterbalancing

The experimental materials were counterbalanced such that participants were assigned to conditions according to a fixed rotation based on the time of their arrival, the order of statistical tests presented to the participants was rotated throughout the conditions, and the order of problem-recognition tasks was randomized.

3.6. Procedure

Participants were tested in groups of 1-5 in an hour long session. All experimental conditions were presented in an online survey (Lime Survey) using a laptop in a separate cubicle. Participants were instructed to learn the concepts and answer a series of questions about the concepts as accurately as possible. The experiment started with the assessment questionnaire, followed by the learning phase, in which participants read descriptions of three non-parametric tests, with two or three example word problems paired with each concept. For a given condition, the examples were either surface or structure-emphasizing. The initial screening test concluded the learning phase in which participants answered six multiple choice questions about the concepts. This was followed by the testing phase where participants responded to problem recognition

Table 2. Design for the surface-emphasizing conditions

Statistical concept	Example 1	Example 2	Example 3
Kruskal-wallis	storyline 1	storyline 1	storyline 1
Chi-squared	storyline 2	storyline 2	storyline 2
Wilcoxon signed-rank	storyline 3	storyline 3	storyline 3

Table 3. Design for the structure-emphasizing conditions

Statistical concept	Example 1	Example 2	Example 3
Kruskal-wallis	storyline 1	storyline 2	storyline 3
Chi-squared	storyline 1	storyline 2	storyline 3
Wilcoxon signed-rank	storyline 1	storyline 2	storyline 3

Table 4. Summary of how we operationalized near and far transfer questions in the surface-emphasizing condition

	Kruskal-wallis test	Chi-squared test	Wilcoxon signed-rank test
Content of source examples	storyline 1	storyline 2	storyline 3
Near transfer questions	share storyline 1	share storyline 2	share storyline 3
Far transfer questions	novel storyline	novel storyline	novel storyline

Table 5. Summary of how we operationalized near and far transfer questions in the structure-emphasizing condition

	Kruskal-wallis test	Chi-squared test	Wilcoxon signed-rank test
Content of source examples	storyline 1,2,3	storyline 1,2,3	storyline 1,2,3
Near transfer questions	share storyline 1,2,3	share storyline 1,2,3	share storyline 1,2,3
Far transfer questions	novel storyline	novel storyline	novel storyline

tasks which entailed a series of word problems with varying content. For example, content of a Chi-squared problem could be novel (i.e., testing transfer in far contexts) or share surface features with source examples of Chi-squared (i.e., transfer in near contexts). Participants recorded which test they would use to solve a word-problem. Problem recognition tasks were given immediately after the learning phase and again 24 hours later. All tasks were presented one at a time. Participants were debriefed and dismissed at the end of the second session.

3.7. Scoring

Responses on the initial learning test and problem-recognition tasks were scored as either correct (1 point) or incorrect (0 points), with averages reported in the paper.

4. Results

Data were analyzed with a 2 (number of examples: two, three) x 2 (exemplar content: surface, structure) x 2 (time delay: immediate, 24 hours later) fixed-factorial ANOVA. All results deemed significant were reliable at $p < .05$. Post-hoc pairwise comparisons were Bonferroni-corrected to the .05 level. Table 6 provides the means and n for all conditions.

4.1. Performance on near problem-recognition tasks

Figure 1 displays the proportion of correct responses on near transfer problem-recognition tasks as a function of the number of examples paired with each concept, content of examples and testing delay. A marginal main effect of examples was observed [$F(1,76) = 3.74$, MSe = .10, $p = .052$, $\eta^2_p = .047$], in which learning with three examples led to higher proportion of correct responses than did learning with two examples). Pairwise comparisons revealed that performance on immediate test was not significantly different when learning occurred with two and three structure-emphasizing examples; however, performance on tasks for the two structure-emphasizing condition marginally decreased 24 hours later, [$t(1, 18) = 1.831$, $p = .07$, $d = .54$].

4.1.1. Discussion

Marginally significant differences in near transfer analyses suggest that learning with three examples may be more advantageous for problem recognition compared to two examples. Moreover, lasting results, as measured by delayed testing, diminish when learning occurs with two high variable examples compared to the other conditions.

4.2. Performance on far problem-recognition tasks

Figure 2 displays the proportion of correct responses on far transfer problem-recognition tasks as a function of the number of examples paired with each concept, content of exemplar content, and testing delay. First, a main effect of examples was observed [$F(1,76) = 3.69$, MSe = .10, $\eta^2_p = .046$] in which using three examples to study a concept led to higher proportion of correct responses on far transfer tasks that did studying with two examples. Second, there was a marginal significant main effect of content [$F(1,76) = 3.28$, p = .056, $\eta^2_p = .041$], in which structure-driven learning led to higher proportion of correct responses than did surface-driven learning. Finally, examples x content x time delay interaction was marginally significant at [$F(1,76) = 3.60$, p = .06, MSe =.05, $\eta^2_p = .045$]. We proceeded to perform planned multiple comparisons without the analysis of simple main effects.

Pairwise comparisons showed three significant differences: The effect of exemplar content is observed only when you learn concepts using two examples; in this condition, structure-driven learning led to higher proportion of correct responses on the immediate test than did surface-driven learning, [$t(1,38) = 2.01$, $d = .644$]. In comparison, when using three examples, proportion of correct responses on the immediate test was similar whether learning was surface-driven or structure-driven. Furthermore, proportion correct on the immediate test was higher for three- vs. two-examples condition, but only when learning was surface-driven [$t(1,40) = 2.35$, $d = .729$]. When learning was structure-driven, immediate test performance was similar for both two-example and three-example conditions. However, paired t-tests between immediate and delayed performance for both conditions, two and three examples, showed a decrease in proportion correct 24 hours post initial learning for the former

Table 6. Proportion correct on problem-recognition tasks as a function of the number of examples used, exemplar content and time delay

content of exemplars	problem-recognition tasks	number of examples paired with each concept	
		two examples	three examples
Immediate testing			
structure-emphasizing	far transfer tasks	.61 (n = 20)	.62 (n = 22)
	near transfer tasks	.69	.70
surface-emphasizing	far transfer tasks	.43 (n = 20)	.61 (n = 22)
	near transfer tasks	.58	.67
delayed testing (24 hours post initial test)			
structure-emphasizing	far transfer tasks	.47 (n = 19)	.63 (n = 21)
	near transfer tasks	.51	.69
surface-emphasizing	far transfer tasks	.44 (n = 20)	.52 (n = 20)
	near transfer tasks	.57	.69

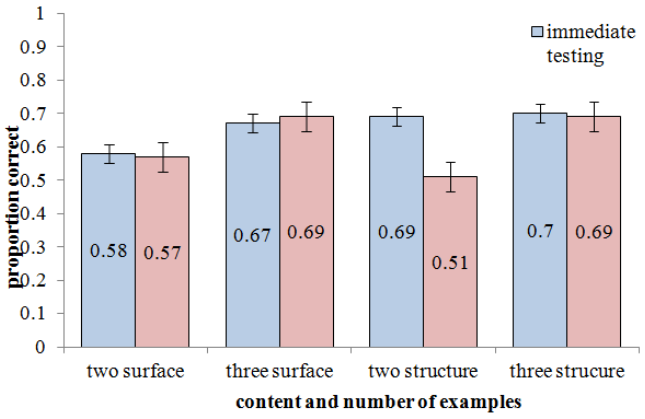

Figure 1. Proportion correct on near transfer problem-recognition tasks as a function of the number of examples used and the type of exemplar set during initial learning

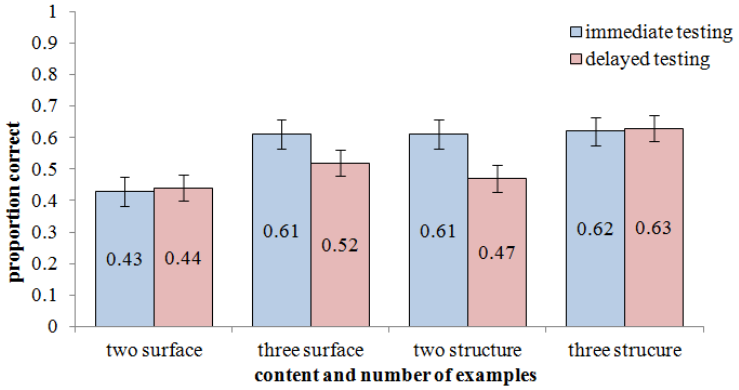

Figure 2. Proportion correct on far transfer problem-recognition tasks as a function of the number of examples used and the type of exemplar set during initial learning

condition, [$t(1,18) = 2.39$, $d = .547$][1] but not for the latter condition.

4.2.1. Discussion

Results from this experiment clearly indicate that the use of three structure-emphasizing examples fosters accurate problem recognition immediately after learning, and 24 hours later. Moreover, the use of two examples to study concepts is beneficial only when learning occurs with structure-emphasizing content and in immediate testing situations. This heightened performance decreases 24 hours later.

5. Discussion and conclusion

When participants learned with structure-emphasizing examples, performance on immediate far transfer tasks was no higher in the three-example condition than in the two-example condition. However, learning with three examples appeared to produce a better retrieval probe for source examples 24 hours later. This effect is consistent with our predictions and other findings in problem-solving literature: concept representation affects source access and largely determines problem-solving success [19] [4]. If the primary concern for a learner is immediate performance or low-level learning, then studying with two or three structure-emphasizing examples will produce comparable results. However, if the concern is to improve encoding of memory probes and promote high-level learning, then using three structure emphasizing examples yields substantial long-term gains; participants who learned using three structure emphasizing examples performed equally well on the immediate and delayed far transfer tests, whereas performance for those who learned using two structure-emphasizing examples decreased in the delayed test.

Learning with three structure-emphasizing examples was critical, since participants who used three surface-emphasizing examples demonstrated a decrease in performance 24 hours later. Although this decline was not statistically significant, we speculate that the reduced benefits were likely due to partial structural inconsistencies in retrieval. In addition, since surface-driven learners did not receive multiple cover stories, they were less likely to generate structure-based concept schemas. Among the participants who received surface-emphasizing examples, those who learned with three, rather than two examples, were more likely to retrieve relevant rules to solve far transfer tasks 24 hours later.

The central finding is that highly focused short-term interventions can make novices sensitive to structural features of problems. However, unless a more extensive intervention with three structure-emphasizing examples is used, the effects of successful transfer tend to fade with dissimilar content and time. This study has practical and theoretical implications for classroom instruction and learning mechanisms. The primary theoretical implication is that high variable content with three examples is effective because it fosters structural representations of concepts; with dissimilar surface features, this intervention allows abstraction of increasingly accurate rules to determine category membership on the basis of structure. From a pedagogical standpoint, the issue may be of how teaching can be tailored to promote far transfer in a way that results in better initial learning. Further research is under way to explore the roles of supporting guidance and interventions, such as self-

[1] Given that this comparison is within-subjects, the effect size d is corrected for dependence between responses using Gravetter and Wallnau's (2009) equation (11.3) [18].

explanation prompts and feedback, to further optimize near and far transfer performance.

6. References

[1] J.R. Anderson, J.M. Fincham, and S. Douglass, "The role of examples and rules in the acquisition of a cognitive skill", *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23(4), 1997, pp. 932-945.

[2] D. Homa, and J. Cultice, "Role of feedback, category size, and stimulus distortion on acquisition and utilization of ill-defined categories", *Journal of Experimental Psychology: Learming, Memory and Cognition,* 10(1), 1984, pp. 83-94.

[3] Tulving, E., and F.I.M. Craik, *The Oxford Handbook of Memory,* Oxford University Press, United Kingdom, 2000, pp. 109–122.

[4] K.J. Kurtz, and J. Loewenstein, "Converging on a new role for analogy in problem solving and retrieval: When two problems are better than one", *Memory & Cognition,* 35(2), 2007, pp. 334–341.

[5] K. Forbus, D. Gentner, and K. Law, "MAC/FAC: A model of Similarity-based Retrieval", *Cognitive Science*, 19(2), 1995, pp. 141-205.

[6] J.E. Hummel, and K.J. Holyoak, "Distributed representations of structure: A theory of analogical access and mapping", *Psychological Review,* 104(3), 1997, pp. 427– 466.

[7] M.L. Gick, and K.J. Holyoak, "Schema induction and analogical transfer", *Cognitive Psychology,* 15(1), 1983, pp. 1-38.

[8] B.H. Ross, and P.T. Kennedy, "Generalizing from the use of earlier examples in problem solving", *Journal of Experimental Psychology: Learning, Memory and Cognition,* 16(1), 1990, pp. 42-55.

[9] J. Loewenstein, L. Thompson, and D. Gentner, "Analogical encoding facilitates knowledge transfer in negotiation", *Psychonomic Bulletin & Review*, 6(4), 1999, pp. 586-597.

[10] S.K. Reed, and C.A. Bolstad, "Use of examples and procedures in problem solving", *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 17(4), 1991, pp. 753-766.

[11] J.L. Quilici, and R.E. Mayer, "Teaching students to recognize structural similarities between statistics word problems", *Applied Cognitive Psychology,* 16(3), 2000, pp. 325-342.

[12] J.L. Quilici, and R.E. Mayer, "Role of examples in how students learn to categorize statistics word problems", *Journal of Educational Psychology, 88*(1), 1996, pp. 144-161.

[13] D.L., Medin, R.L. Goldstone, and D. Gentner, "Respects for similarity", *Psychological Review*, 100(2), 1993, pp. 254-278.

[14] M.T.H. Chi, P.J. Feltovich, and R. Glaser, "Categorization and representation of physics problems by experts and novices", *Cognitive Science,* 5(2), 1981, pp. 121-152.

[15] A.D. Bruin, "*Fostering expert learning strategies in novices",* 2006. Retrieved from http://publishing.eur.nl/ir/repub/asset/8144/proefschrift_de_bruin_final2.pdf

[16] D. Billing, "Teaching for Transfer of Core/Key Skills in Higher Education: Cognitive Skills", *Higher Education: The International Journal of Higher Education and Educational Planning, 53*(4), 2007, pp. 483-516.

[17] A. Didierjean, and S. Nogry, "Reducing structural-element salience on a source problem produces later success in analogical transfer: What role does source difficulty play?" *Memory and Cognition,* 32(7), 2004, pp. 1053–1064.

[18] Gravetter, F.J., and L.B. Wallnau. *Statistics for the Behavioral Sciences*, Eighth Edition. Cengage Learning, 2008.

[19] D. Gentner, J. Loewenstein, and L. Thompson, "Learning and transfer: A general role for analogical encoding", *Journal of Educational Psychology,* 95(2), 2003, pp. 393-405.

