Learning from Worked-Out Examples: A Study on Individual Differences

Article *in* Cognitive Science · January 1997

DOI: 10.1207/s15516709cog2101_1 · Source: OAI

1 author:

Alexander Renkl
University of Freiburg
259 PUBLICATIONS **8,118** CITATIONS

SEE PROFILE

COGNITIVE SCIENCE 21 (1), 1997, pp. 1–29 ISSN 0364-0213

Learning from Worked-Out Examples: A Study on Individual Differences

ALEXANDER RENKL

University of Munich

The goal of this study was to investigate individual differences in learning from worked-out examples with respect to the quality of self-explanations. Restrictions of former studies (e.g., lacking control of time-on-task) were avoided and additional research questions (e.g., reliability and dimensionality of self-explanation characteristics) were addressed. An investigation with 36 university freshmen students of education working in individual sessions was conducted. The domain was probability calculation. Prior knowledge and the quality of self-explanations (protocols of the individuals' thinking aloud) were assessed as predictors of learning. A post-test was employed to measure the learning gains as the dependent variable. The following main results were obtained. Most self-explanation characteristics could be regarded as relatively stable person characteristics. The individual differences in the quality of self-explanations were, however, found to be multidimensional. Most important, even when controlling for time-on-task (quantitative aspect), learning gains could be substantially predicted by qualitative differences of self-explanation characteristics. In particular, successful learners tended to employ more principle-based explanations, more explication of operator-goal combinations, and more anticipative reasoning. In addition, there were two types of effective learners, labeled *anticipative reasoners* and *principle-based explainers*.

Worked-out examples consist of the givens of a problem, solution steps, and the final solution itself. Learning from worked-out examples is an important source of learning (VanLehn, 1986, 1996), and it is a learning mode preferred by novices (e.g., Anderson, Farell, & Sauers, 1984; LeFevre & Dixon, 1986; Pirolli & Anderson, 1985; Recker & Pirolli, 1995). Furthermore, research has shown that learning from worked-out examples is typically very effective (e.g., Sweller & Cooper, 1985; Tarmizi & Sweller, 1988; Ward & Sweller, 1990; Zhu & Simon, 1987).

However, in order to successfully learn from these types of examples, the learner has to actively explain the solution steps to himself or herself because not all the information about the rationale of the solution steps, that is necessary for understanding the solution procedure, is included in the examples (cf. Chi, Bassok, Lewis,

Direct all correspondence to: Alexander Renkl, Institut für Pädagogische Psychologie und Empirische Pädagogik, University of Munich, Leopoldstraße 13, 80802 Munich, Germany. E-mail: renkl@mip.paed.uni-muenchen.de.

Reimann, & Glaser, 1989). This incompleteness of worked-out examples is not, however, merely typical of psychological experiments, but also of common textbooks. In this study, individual differences in learning from worked-out examples (domain: probability calculation) were investigated.

INDIVIDUAL DIFFERENCES IN LEARNING FROM WORKED-OUT EXAMPLES

The seminal study of Chi et al. (1989) has shown that there are significant individual differences with respect to the extent to which learners profit from the study of worked-out examples, depending on how well they explain the solutions to themselves. This phenomenon was labeled *self-explanation effect.* The subjects in the study of Chi et al. (1989) had to learn from worked-out examples in the domain of physics (Newton mechanics). First, the subjects read a text on basic concepts and definitions of Newton mechanics. Criterion-referenced testing and, if necessary, remedial learning phases assured that the subjects had comparable background knowledge on these definitions and concepts. Then, they had to study three worked-out examples. Afterwards, the subjects had to solve problems independently. During both the two phases (study of examples and problem solving), the subjects had to think aloud so that the encoding and the use of the worked-out examples could be studied. Finally, a post-test was presented. Chi et al. (1989) divided the subjects into a successful ($n = 4$) and an unsuccessful group ($n = 4$) with respect to the performance in the problem-solving phase (median split). With regard to the study of the worked-out examples, the following main results were obtained. The successful problem solvers assigned more time to the study of the worked-out examples than the unsuccessful ones (on average 13.0 min vs. 7.4 min per example). Accordingly, they produced more task-relevant ideas while trying to explain the example solutions to themselves. There were, however, also qualitative differences between the good and the poor problem solvers. The successful subjects (1) more frequently related solution steps to the domain principles presented in the text (deep-structure self-explanations), (2) more frequently elaborated on the application conditions and goals of operators and (3) more adequately monitored their comprehension, that is, they more frequently diagnosed comprehension failures and less frequently had the illusion of comprehension (for similar results see Ferguson-Hessler & DeJong, 1990). However, as the quantity (time-on-task) and quality of the learning processes were confounded in Chi et al.'s (1989) study, it could not be definitely ruled out that the effective learners were superior merely because they devoted more time to elaborating the worked-out examples.

With respect to the significance of cognitive prerequisites, Chi and VanLehn (1991) found in a re-analysis of Chi et al.'s (1989) data that there were no significant differences between good and poor learners with respect to GPA and on a test on prior domain abilities (Bennett Mechanical Ability Test). However, no strict test was performed regarding the extent to which self-explanation characteristics can explain differential learning gains when prior knowledge is controlled.

Chi, DeLeeuw, Chiu, and LaVancher (1994) analyzed individual differences in an experimental group which was prompted to provide self-explanations while studying a text. The self-explanation effect could be replicated. However, the time-on-task also varied significantly (from 1 hr and 27 min to 2 hr and 53 min; Chi et al., 1994). Thus, in this study, too, the alternative explanation of an underlying time-on-task effect could not be excluded. With respect to the significance of cognitive prerequisites, Chi et al. (1994) found that the size of the self-explanation effect in the prompted group was not dependent on cognitive prerequisites. Due to the interventional nature of the study, the role of cognitive prerequisites for *spontaneous* self-explanations was naturally not investigated.

Pirolli and Recker (1994), who also compared the learning processes of six successful and of six unsuccessful problem solvers during the study of worked-out examples and of instructional text, replicated and extended the results of Chi et al. (1989) in the domain of LISP programming. Basically, the differences between good and poor learners found in the study of Chi et al. (1989) were reproduced. Another very interesting point made by Pirolli and Recker (1994) is that there may be a diminishing return relation between self-explanations and learning. That means, elaborations that are very extensive may become repetitive or distracting from the core principles underlying the solutions. This does not mean that, at a certain point, further elaborations do not foster learning at all. However, the average learning effect of each new elaboration diminishes with each increase in number of elaborations. A diminishing return relation is described by the so-called power function (linear relation between two logarithmic variables) that is often employed to describe the effects of practice on skill (Newell & Rosenbloom, 1981; Rosenbloom & Newell, 1986). Pirolli and Recker (1994) found that, when predicting learning by elaborations, a power function regression explained more variance than an ordinary linear regression (37% vs. 28%). The authors regarded this result as the confirmation of their diminishing return hypothesis. However, the interpretation of this result needs to be qualified because the difference between the explained variances of the linear and of the power function models are rather small given the big sampling errors of correlation coefficients in small samples. On the other hand, no huge difference between the two models is to be expected because they are similar in the sense that a power function can also be fitted to a substantial degree by a linear function. Thus, the issue of diminishing return relations between self-explanation and learning needs further exploration.

There are two more studies that are relevant to the present topic although they were not primarily concerned with individual differences in self-explanations. Recker and Pirolli (1995) reported that the self-explanation effect was replicated, however, no detailed corresponding data were presented. Lovett (1992) analyzed learning by problem solving and by worked-out examples in the same domain as the present study, that is, in probability calculation. She found that learners who elaborated (self-explained) on a central concept (numerator starting value) in the learning materials outperformed in a later test those who did not. These studies thus provide further evidence for the relevance of self-explanation effects, but they did not avoid the restrictions of the investigations discussed above.

RESEARCH QUESTIONS

The present study aims to extend the existing research on individual differences with respect to learning from worked-out examples. In previous studies, the protocols of just a few subjects (typically 4 or 6 good learners vs. 4 or 6 poor learners) were analyzed. This is not surprising given the tremendous efforts required to analyze verbal protocols. These small sample sizes, however, cannot claim to be representative with regard to the total range of individual differences, and it is difficult to determine the reliability and interrelatedness of different indicators of high-quality self-explanations, to employ statistical tests for significance (low-test power), and to compute measures of practical significance. In addition, the median split procedure employed by Chi et al. (1989) and by Pirolli and Recker (1994) has some drawbacks. First, information on interindividual differences is lost through dichotomization of variables. Second, no measures for practical significance are directly available, in comparison with the use of correlations (r or r^2). Third, using median split often leads to the result that subjects near the median can be in different groups although they are actually more similar to each other than to some other members of their respective group. The present study analyzes the associations between continuous variables in a sample of 36 subjects so that the drawbacks mentioned above can be avoided.

The following specific research questions were addressed:

1. *To what extent are individual characteristics of self-explanations generalizable over examples*? This issue can also be regarded as a matter of reliability. Thus, it is examined to what extent learners can be differentiated reliably across different examples with respect to self-explanation characteristics.
2. *To what extent are different self-explanation characteristics correlated*? The interrelations between different aspects of good self-explanations (e.g., monitoring and deep-structure explanations) are analyzed in order to examine whether the construct "quality of self-explanations" can be regarded as unidimensional.
3. *To what extent is the quality of self-explanations associated with cognitive prerequisites*? In order to evaluate the conjecture that the self-explanation effect can be more parsimoniously explained by differences in prior knowledge, the present study attempts to determine the degree to which the quality of self-explanations depends on prior knowledge.
4. *Are the learning results associated with the quality of self-explanations even when time-on-task is controlled*? In previous investigations (Chi et al., 1989; 1994; Pirolli & Recker, 1994), the study time for the worked-out examples was not fixed. It was found that the good learners spent substantially more time-on-task than the poor ones. Time-on-task is, however, a very reliable and powerful predictor of learning (e.g., Helmke & Renkl, 1992). Thus, the question of the extent to which the differences in learning gains are due to quantitative (time-on-task) and to qualitative (structure of self-explanations) aspects remains unanswered. In the present study, the learning time of each learner was strictly limited.

5. *Is there a diminishing return relation between self-explanations and learning gains*? It is investigated to what extent the corresponding findings of Pirolli and Recker (1994) can be replicated?
6. *Can different self-explanation styles be identified and, if yes, are they related to learning success*? It is tested whether learner types can be identified which correspond to the prototypes of a good and a poor self-explainer as described by Chi et al. (1989) and Pirolli and Recker (1994).

METHODS

Sample

In university courses, freshmen students of education were invited to participate in the present investigation. In order to find learners with minimal prior domain knowledge, the students were told that persons with prior knowledge of probability calculation should not take part because they might already know all the to-be-learned contents. Thirty-six students volunteered to participate.

Worked-out Examples

A computer monitor was used to present worked-out examples from the domain of probability calculation. The problem specification and the solution steps of each worked-out example were shown on four screen pages. In Figures 1a and b, the four pages of one such worked-out example are presented (see also Figure 2 for the problem specifications of another worked-out example). On the first page, the problem givens were displayed (see the upper section of Figure 1a). The subject could read them and then go to the next page where the first solution step was presented in addition to the problem formulation (see the lower section of Figure 1a). After inspecting this solution step, the subjects continued to proceed to the following page where the next solution step was added (see the upper section of Figure 1b). The whole solution of each problem was presented on the fourth page (see the lower section of Figure 1b). On the next page, a new example was presented. The amount of time spent on each screen and the number of pages and examples inspected were automatically recorded.

The subjects were allowed to regulate the processing speed of the worked-out examples on their own. An external pacing control, for example, by fixing the presentation time for each page, would have interfered with the learners' strategies and would have diminished ecological validity. However, in order to keep the time-on-task for each subject constant, a study time of 25 minutes was fixed. Thus, when 25 minutes were over, the next mouse click on "Next Page" effected a "thank-you screen" to appear. The subjects were informed about this procedure in advance.

Individual differences in processing speed caused the number of examples (pages) inspected by different subjects to vary. In order to preclude the pitfall that the faster subjects acquired a *broader* knowledge base through the inspection of further ex-

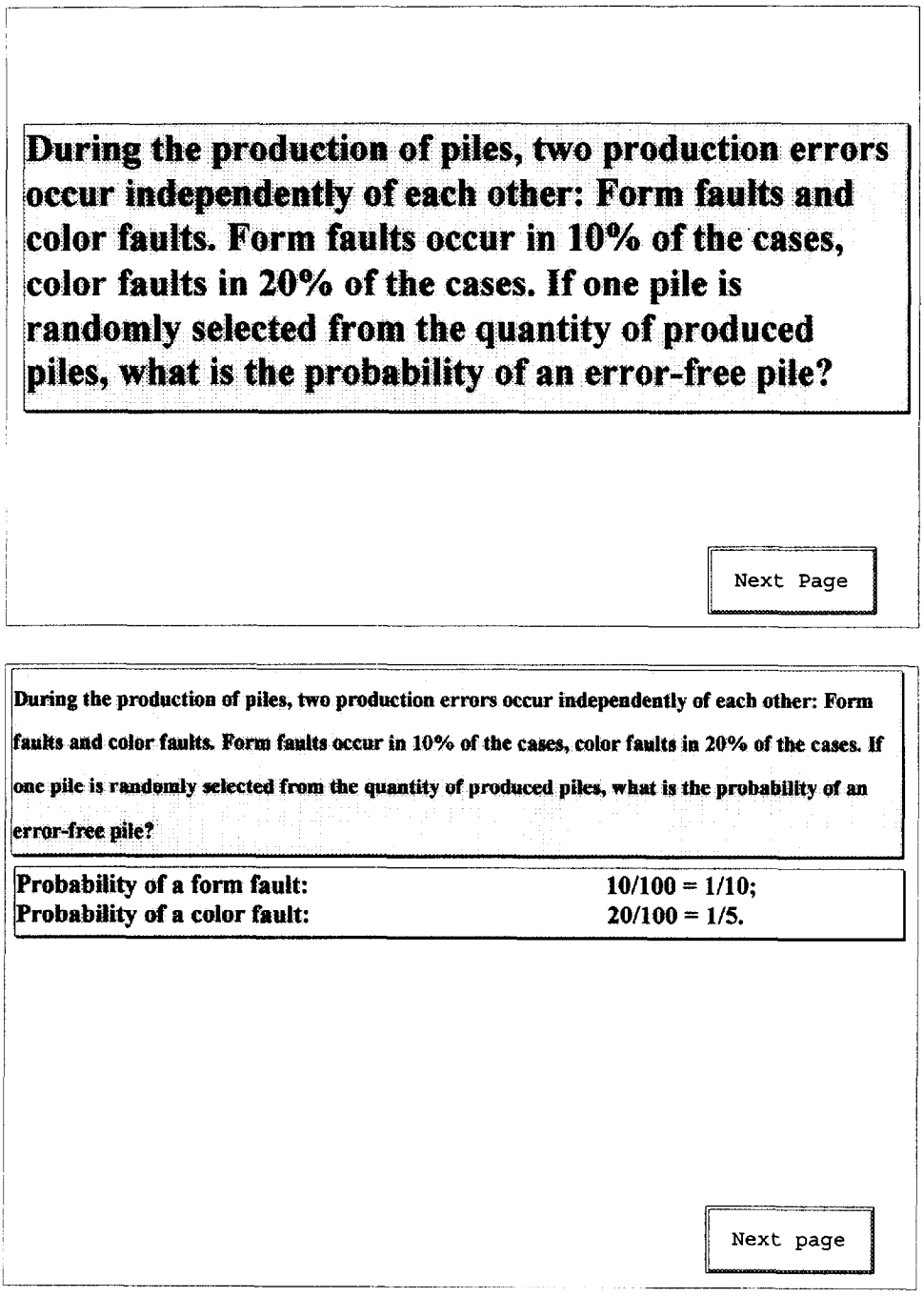
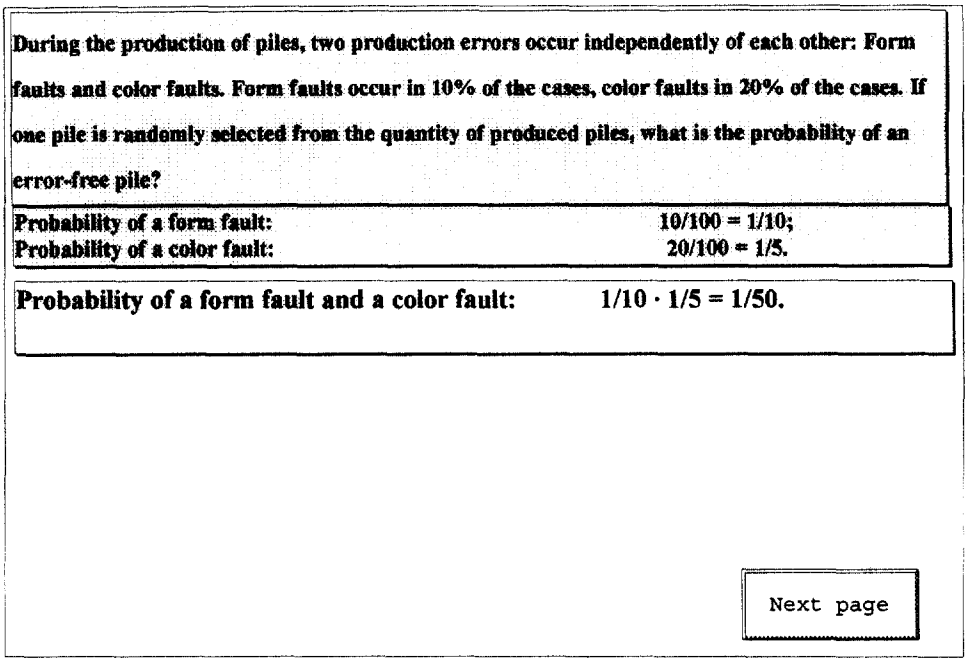


Figure 1a. The first two screens of one worked-out example.

amples with different deep structures, only four types of deep structures were used. Within the available time span of 25 minutes, which was fixed according to the results of pilot studies, every subject processed the first four problems and thereby encountered each type of deep structure. Hence, the faster subjects were confronted



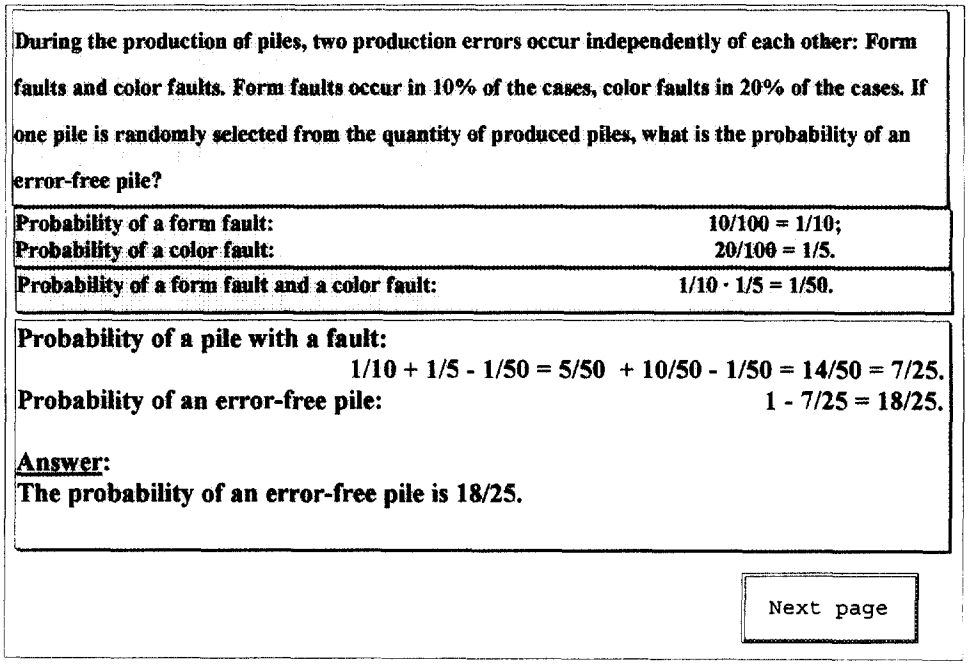


Figure 1b. The last two screens of one worked-out example.

with examples containing new surface features (i.e., new numbers, new objects), but not new deep structures.

The number of inspected examples was introduced as a control variable. The question of the extent to which differences in speed had some effects on learning was

to be explored. Based on the findings of Chi et al. (1989) that good learners studied the individual examples longer than the poor learners did, a negative relation between the number of examples inspected and learning gains could be expected. On the other hand, studying examples with varying surface structures fosters transfer (Catrambone & Holyoak, 1989; Paas & VanMerrienboer, 1994), and thus a positive relation also seems plausible. Renkl (1995) found a nonsignificant positive association between number of examples and learning gains. In the present study, the extent to which the control variable "number of examples inspected" is related to learning is examined for the case where indicators of the superficiality of processing taken from the verbal protocols are controlled.

Thinking Aloud Procedure

The subjects were asked to verbalize their thoughts concurrently with the study of examples. The corresponding instruction was structured according to the guidelines of Ericsson and Simon (1993). The subjects were told to talk aloud and verbalize anything that comes to their mind. They were not instructed to provide special information. Thus, the subjects' spontaneous self-explanations were assessed.

Before the study of the examples, the thinking (talking) aloud procedure was trained using a warm-up problem (word problem in arithmetic). When the subject did not talk for more than 15 seconds the experimenter said to him or her: "(Please) keep talking".

Coding of Verbal Protocols

The original plan was to use the coding system from Chi et al. (1989) with minor adaptations in order to analyze the thinking aloud protocols of the subjects while they were studying the worked-out examples. The inspection of the verbal protocols of pilot subjects showed that the deviation of the present investigation from Chi et al.'s (1989) study with respect to the chosen domain, the presentation mode of the examples, the type and complexity of examples, and the number of examples studied on average by each person made major adaptations necessary.

The protocols were thoroughly examined for content segments that correspond to the following categories:

1. *Principle-based explanation*. The number of times that subjects referred to the principles of probability calculation was counted. However, if a principle was merely mentioned without any elaboration (e.g., "This is the multiplication rule"), this category was *not* scored. There had to be some elaboration of a principle (e.g., "It gets multiplied, because the events are independent from each other;" this statement referred to the meaning of the multiplication rule). This category corresponds to the Chi et al.'s (1989) codings of the learners' references to Newton's Laws (the underlying domain principles in that study).
2. *Goal-operator combinations*. This category was scored if a (sub-)goal and an operator that led to this (sub-)goal was explicitly mentioned (e.g., "Through this multiplication we get the probability of tiles with color and form faults").

This category corresponds to the code "Impose a goal or purpose for an action" in the Chi et al. (1989) study.

3. *Anticipative reasoning*. If a subject computed a probability in advance, that is, without looking at the worked-out solutions, this category was coded (e.g., "Then the probability of tiles with color and form faults is 1/50"). This category is not directly analogous to anything in previous studies. This may be due to the fact that the presentation mode of the worked-out examples in the present study differed from other investigations in that the solution steps were presented in a step-by-step mode. Hence, the subjects could, instead of having a look at the next page with a further solution step, anticipate on their own (i.e., compute to-be-found probabilities) and compare their predicted probabilities with the ones presented in the next solution steps provided by the program.
4. *Elaboration of problem situation*. This category referred to information about the situation that the subjects inferred from the givens of a problem (e.g., "If the first ball is drawn, the overall number of balls is reduced by one"). This category has some similarity to the category "Refine or expand the conditions of an action" used by Chi et al. (1989). However, statements that were assigned to the category "elaboration of problem situation" did not typically refer to specific actions or operators. Rather they more globally indicated the construction of a situation model (Kintsch, 1986), that is, a mental model of the initial situation and of the flow of events described in the problem statement.
5. *Noticing coherence*. This category was employed to test the extent to which the perception of coherence between examples can foster the induction of abstract schemata and, consequently, problem-solving performance (cf. Catrambone & Holyoak, 1989; Gick & Holyoak, 1983). In this category, each statement was coded in which the worked-out example presently being studied was related to an earlier one (e.g., "This is the same problem as the one with the aircraft pilots"). It is, however, important to note that the present category is not equivalent to processes during problem solving (i.e., while using examples) such as noticing analogies as described by Holyoak and colleagues (e.g., Catrambone & Holyoak, 1989; Gick & Holyoak, 1983). The present category refers to processes during self-explanation activities.
6. *Monitoring-negative*. All indicators of non-understanding were scored into this category (e.g., "Now I don't understand it any more").
7. *Monitoring-positive*. If a subject indicated that s/he understood a solution step, this category was scored (e.g., "Oh yeah, I see"). Both monitoring categories were identical to the ones used in the study by Chi et al. (1989).

The analysis of pilot subjects showed that the frequently employed method of independently first segmenting and then coding thinking-aloud protocols did not make sense. This was due to the fact that the "size of the units" varied too strongly from category to category so that no reasonable common grain-size of segmentation could be found. Thus, the protocols were segmented with the coding categories in mind. In the cases of monitoring statements, often a single word (e.g., "clear") was regarded as segment. When coding explanations of a solution step in terms of the

underlying probability principle, longer statements were regarded as a unit. However, the coding categories were distinct and there were no inclusions of segments. For example, a monitoring statement always indicated the end of a previous segment.[1]

The protocols were independently coded by both the author and a research assistant. The interrater agreement with respect to assigning the protocol segments to the coding categories was 89.3%, or expressed as Cohen's (1960) Kappa which corrects for chance agreement .87. This amount of interrater agreement can be regarded as satisfactory. In cases of divergence, the author re-examined the protocols and made the final decision.

In addition to qualitative analyses, the length of the verbal protocols (number of words) was determined. This measure served as a control variable to exclude the possibility that self-explanation effects are, in reality, due to some kind of verbosity or word fluency effects.

Instruments

Pretests

A subsample of items used in the algebra test by Lienert and Hofer (1977) called the *Mathematiktest für Abiturienten und Studienanfänger* [Mathematics test for 13th graders and university freshmen] was used to measure performance in algebra problems as the indicator of mathematics ability (10 items). Furthermore, six relatively simple probability calculation problems were employed as a pretest (e.g., "If you play the dice twice, what is the probability of two sixes?"). In both pretests, one point was awarded for each correct item solution.

Post-test

The post-test consisted of 15 items. Three items were relatively simple problems such as those employed in the pretest. The other 12 items were constructed according to the following rationale (see Figure 2): four items were identical to the first four worked-out examples, except that some irrelevant information was inserted (i.e., same deep structure, similar surface structure, irrelevant information); four items had the same deep structure, but a different surface structure; four items had a similar surface structure, but the deep structure was changed. For the correct solution of a post-test item, two points were awarded. If at least half of the solution was correct, one point was dispensed. Computational errors which very occasionally occurred were ignored.

Instructional Text

The instructional text, which provided basic knowledge for the study of the worked-out examples, contained about 700 words (including formulas) and a diagram to illustrate the *addition principle* in probability calculation (see below). The following principles were explained in a rather abstract manner: *definition of probability* (p[target events] = n[target events]) / n[all possible events]), *multiplication principle* for in-

Givens of a worked-out example

In an aptitude test for aircraft pilots, 40% of the applicants do not pass the physical examination and 60% do not pass the psychological tests. 20% of the applicants fail because of the physical and the psychological examination. What is the probability that two randomly selected applicants fit the job?

Same deep structure · similar surface structure · irrelevant information

In an aptitude test for aircraft pilots, 40% of the applicants do not pass the physical examination and 60% do not pass the psychological tests. 20% of the applicants fail because of the physical and the psychological examination. 40% merely failed to pass the psychological tests. What is the probability that two randomly selected applicants fit the job?

Same deep structure · different surface structure

Production errors cause 15% of pencils to be of an incorrect length and 10% of the incorrect diameter. In 5% of the cases, both faults are present. If two pencils are randomly selected, what is the probability that neither has an error?

Different deep structure · similar surface structure

In an aptitude test for aircraft pilots, 40% of the applicants do not pass the physical examination and 60% do not pass the psychological tests. 20% of the applicants fail because of the physical and the psychological examination. What is the probability that at least one out of two randomly selected applicants fits the job?

Figure 2. Types of post-test items and corresponding examples (cf. also Renkl, 1995).

dependent events (p[A and B] = p[A] * p[B]), *addition principle* (p[A and/or B] = p[A] + p[B] − p[A and B]), *principle of complementarity* (p[non A] = 1 − p[A]). The worked-out examples and the test items were based on these principles of probability calculation. The average study time of this text was 11.5 minutes (*SD*: 3.4 minutes). Individual differences in study time were not significantly related to post-test performance (r = −.02), and thus not discussed further.[2]

Procedure

The subjects worked in individual sessions of about two hours. First, the mathematical pretests were presented. In order to provide or re-activate basic knowledge that allowed the subjects to understand the worked-out examples, an instructional text

TABLE 1
Scores for Pretests and Post-tests

	M	SD	Reliability
Pretest-probability (6)[a]	1.08	1.36	.69
Pretest-algebra (10)[a]	5.03	2.50	.76
Post-test (30)[a]	11.25	6.83	.84
Simple problems (6)[a]	3.92	1.78	.42
Irrelevant information (8)[a]	3.33	2.26	.58
Different surface (8)[a]	2.39	2.48	.73
Different deep structure (8)[a]	1.61	2.02	.65

Note. [a] Theoretical maximum.

on basic principles of probability calculation was given to the subjects. The comprehension of these basic concepts was assessed by a criterion-referenced test which was evaluated immediately. If there was a wrong answer, the experimenter gave a semi-standardized explanation and had the subject re-read the corresponding text passage. After this procedure, the subjects were informed that they had to study the worked-out examples for 25 minutes. They were instructed to think aloud during this period. Finally, the subjects worked on the post-test.

RESULTS

Descriptive Statistics, Reliabilities, and Intercorrelations of Pretest and Post-test Scores

As Table 1 shows, both pretests showed satisfactory reliabilities (Cronbach alpha coefficients). As there were 10 items in the algebra test, the resulting mean of about 5 indicated that the test items were of medium difficulty. The mean of the probability pretest (about 1 with a theoretical maximum of 6) indicated that subjects with low prior knowledge levels were indeed selected for the present study.

The post-test, taken as a whole, proved to be reliable. It can, however, be divided into subscores according to different types of post-test items (simple problems; problem with irrelevant information; problems with different surface structure; problems with different deep structure). An insufficient reliability was obtained for the simple items. Thus, no separate analyses were made with this subscore. For the scale of items with different surface structure and the scale of problems with different deep structure, satisfactory internal consistencies were determined. The reliability estimate for items with irrelevant information was relatively low (.58) but acceptable for group analyses (Table 1). The mean total posttest score was about 11 (theoretical maximum: 30) which signified that it was relatively difficult. As can be seen from the means of the three reliable subscores (Table 1), the item difficulty increased from problems with irrelevant information over problems with different surface structure to different deep structure problems. The relatively low rates of transfer from the learning materials to problems with different surface structure (e.g., Gick & Holyoak, 1980) and with different deep structure (see Lovett, 1992, for analogous results in the domain of probability calculation) is in accord with literature on analogical transfer.

TABLE 2
Intercorrelations Between Test Scores

	(2)	(3)	(4)	(5)	(6)
(1) Pretest-probability	−.06	.53*	.42*	.37*	.57*
(2) Pretest-algebra		.03	.00	.01	.04
(3) Post-test			.84[a]	.79[a]	.87[a]
(4) Irrelevant Information				.46*	.64*
(5) Different surface					.71*
(6) Different deep structure					

Notes. * $p < .05$ (two-tailed test of significance).
[a]Tests of significance did not make sense because the variables were arithmetically dependent.

TABLE 3
Number of Self-explanations, Protocol Length, and Number of Examples

	M	*SD*	Reliability
Self-explanation variables			
Principle-based explanations	4.89	5.01	.80
Goal-operation combinations	1.53	2.25	.80
Anticipative reasoning	2.53	3.42	.81
Elaboration of situation	1.61	1.50	.58
Noticing coherence	1.44	1.25	.06
Monitoring-negative	8.08	6.10	.59
Monitoring-positive	7.03	5.37	.74
Protocol length	1633.44	305.58	—
Number of examples	9.49	3.51	—

Table 2 shows the intercorrelations between pretest and post-test scores. Surprisingly, the pretest in algebra did not significantly correlate with any other test score. The more proximal probability pretest was substantially related to the post-test, taken as a whole, as well as to its subscales (Table 2). The post-test subscales were all associated with one another. The performance on problems with different surface and different deep structure was strongly correlated ($r = .71$). Thus, these two subscales seemed to measure the same dimension of individual differences. Hence, no discriminant associations with predictors was to be expected, and a separate treatment of these two subscales did not make sense. The achievement on items with irrelevant information was also substantially associated with performance on both other subscales (different surface: .46; different deep structure: .64), but not to an extent that a separate view on this aspect would not have made sense. As a consequence of this pattern of results, in addition to the post-test as a whole, two subscales were kept for further analyses: near transfer (items with irrelevant information) and medium transfer (items with different surface and with different deep structure).[3]

Number of Self-Explanations, Protocol Length, and Number of Examples

Table 3 shows that, on average, about 5 principle-based explanations were given during the 25-minute period of example study. Other types of elaborations that were supposed to foster learning did not occur, at least on the average, very frequently.

During the study of examples, on average, to-be-computed probabilities were anticipated about 2.5 times, goal-operator combinations were explicated about 1.5 times, the situation of the problem was elaborated about 1.6 times, and coherence between examples was noticed about 1.4 times. Metacognitive statements concerning an individual's understanding were quite frequent. On average, the subjects stated a lack of understanding about 8 times and understanding about 7 times.

The descriptive statistics of the control variables show that the subjects' verbal protocols had an average length of about 1600 words. The mean number of examples inspected was 9.49. This indicates that the subjects studied 9 to 10 worked-out examples on average.

To What Extent Are Individual Characteristics of Self-explanations Generalizable Over Examples?

It was investigated whether persons can be differentiated reliably across different examples with respect to self-explanation variables or whether the quality of self-explanations was, instead, a function of specific person-example interactions. For this purpose, each person's scores were computed for the examples 1,3,...and for the examples 2,4,...respectively (odd-even method). Then they were correlated with each other and finally corrected by the Spearman-Brown formula. As Table 3 shows, the reliability estimates for principle-based explanations, for explication of goal-operator combinations, for anticipative reasoning, and for positive monitoring statements were all above .70 and thus sufficient. The reliability of negative monitoring statements and of elaboration of the problem situation were .59 and .58 respectively. This was not very satisfactory, but still acceptable. For noticing coherence between examples, the correlations between the protocol halves were very low and consequently a reliability estimate of .06 was obtained. This means that the subjects could not be reliably differentiated with respect to their tendency to notice coherence. The very low reliability coefficient also indicated that even when extending the observation period (in order to increase the reliability in analogy to test prolongation), no reliable individual differences were to be expected. Thus, this variable was excluded from further analyses.[4]

To What Extent Are Different Self-explanation Characteristics Correlated?

Of the 15 correlations between the self-explanation variables (without control variables), only one association reached the 5%-level of significance (two reached the 10%-level of significance; Table 4). As will be shown in a later section, even variables that proved to be correlated with learning gains were not significantly intercorrelated. The correlation between explication of goal-operator-combination and principle-based explanations was, however, quite strong ($r = .64$). Thus, it does not seem to be reasonable to interpret it as coincidental covariation due to sampling errors. This association means that learners who tended to assign meaning to operators did this in two ways: relating operators to domain principles and to goals.

TABLE 4
Interrelations Between Self-explanation Variables, Protocol Length, and Number of Examples

	(2)	(3)	(4)	(5)	6)	(7)	(8)
(1) Principle-based explanations	.64*	−.01	.27	−.23	−.10	.47*	−.18
(2) Goal-operation combinations		−.01	−.03	−.33#	.05	.39*	.11
(3) Anticipative reasoning			.20	−.24	−.20	−.12	.03
(4) Elaboration of situation				.10	−.17	.03	−.29#
(5) Monitoring-negative					.25	.12	−.18
(6) Monitoring-positive						.40*	.23
(7) Protocol length							.22
(8) Number of examples							

Note. # $p < .10$; * $p < .05$ (two-tailed test of significance).

TABLE 5
Interrelations Between Prior Knowledge and Self-explanation Variables

	Pretest Algebra	Pretest Probability
Principle-based explanations	.28#	.00
Goal-operation combinations	.32#	.24
Anticipative reasoning	−.31#	.43*
Elaboration of situation	.10	−.03
Monitoring-negative	.19	−.26
Monitoring-positive	.06	−.30#

Note. # $p < .10$; * $p < .05$ (two-tailed test of significance).

On the whole, the present data do not suggest that quality of self-explanations is a unidimensional construct. This means that learners who are good self-explainers with respect to certain aspects are not necessarily good self-explainers in all respects.

The following findings were obtained with regard to the control variables (Table 4). The number of worked-out examples studied was not, at the 5%-level, significantly associated with any other self-explanation variable. In contrast, protocol length was significantly correlated with principle-based explanations, the explication of goal-operator combinations, and positive monitoring. Thus, it was reasonable to include the control variable "protocol length" in further analyses.

To What Extent Is the Quality of Self-explanations Associated with Cognitive Prerequisites?

The algebra pretest as the indicator of general mathematics ability was not significantly (5%-level of significance) associated with any self-explanation variable. It showed a tendency (10%-level) to be positively correlated with principle-based explanations and with the explication of goal-operator combinations and to be negatively associated with anticipative reasoning (Table 5). Even the relatively proximal pretest in probability calculation was only weakly associated with self-explanation variables, except for the significantly positive correlation with anticipative reasoning. Merely a tendency (10%-level) to be negatively correlated with prior

TABLE 6
Relations of Quality of Self-explanations, Protocol Length, and Number of Examples to Learning: Zero-order Correlations (in front of the slash) and Partial Correlations (behind the slash)

	Post-test Total	Near Transfer	Medium Transfer
Principle-based explanations[a]	.38* / .44*	.35* / .38*	−.31* / −.36*
Goal-operation combinations[a]	.37* / .30*	.25 / .17	.43* / .37*
Anticipative reasoning[a]	.49* / .34*	.39* / .25	.48* / .34*
Elaboration of situation[a]	.12 / .16	.10 / .12	.06 / .09
Monitoring-negative	−.46* / −.39*	−.36* / −.30#	−.44* / −.36*
Monitoring-positive	−.19 / −.04	−.27 / −.17	−.08 / .09
Protocol length	.09 / .18	.17 / .24	.10 / .19
Number of examples	.23 / .15	.05 / −.05	.40* / .35*

Note. # $p < .10$; * $p < .05$; [a]one-tailed tests of significance because of theoretical expectations.

knowledge in probability calculation was found for positive monitoring statements (Table 5). In sum, there were relatively weak relations between prior knowledge and self-explanation characteristics.

A possible factor that might account, however, for these results is the restricted variance of prior knowledge in probability calculation in the present sample. Although the subjects could reliably be differentiated with respect to their level of prior knowledge, solely subjects with minimal preknowledge in probability calculation (see descriptive pretest results) were selected for this study. Thus, if subjects with a broader range of prior knowledge levels were included, other results might have been obtained.

Are the Learning Results Associated with the Quality of Self-explanations Even When Time-on-task Is Controlled?

As Table 6 shows, all post-test scores were significantly positively associated with principle-based explanations and anticipative reasoning. The explication of goal-operator combinations was significantly correlated with the post-test as a whole and with medium transfer performance. For near transfer, the association failed to reach the level of significance. The elaboration of the problem situation was not significantly related to post-test scores.

In contrast to previous studies, it was found that the more the subjects felt that they did not understand, the worse they performed on the post-test. This was indicated by the strong negative correlation between negative monitoring statements and the post-test. Positive monitoring statements were unrelated to the post-test achievement which was also in contrast to previous research.

The protocol length was not substantially related to any post-test score. Thus, the relations of the self-explanation variables to the post-test cannot be explained by some kind of verbosity or word fluency effect. The number of examples which were inspected was significantly related to medium transfer, but not to near transfer achievement. This leads to a positive, but not significant correlation between the number of inspected examples and the post-test as a whole.

Partialing out the pretest scores in probability calculation did not substantially change the pattern of results, as shown in Table 6 (the algebra pretest was not used a control variable, because it was unrelated to pretest and post-test scores; see above).[5]

In summary, successful learners tended to provide many principle-based explanations, to frequently anticipate to-be-computed probabilities, and to seldom state lack of comprehension. The explication of goal-operator combinations and the inspection of a relatively large number of examples seemed to be especially relevant for the medium transfer performance.

For exploratory reasons, hierarchical multiple regression analyses were performed. Thus, preliminary hypotheses can be generated about which variables could, beyond confounded effects, explain unique variance in learning gains. The dependent variables were the post-test scores. The predictors included the pretest, the self-explanation, and the control variables. First, the strongest predictor of the dependent variable was included. Additional predictors were entered into the regression equation if they significantly increased the proportion of explained variance. The resulting final regression models are shown in Table 7.

Taking the post-test as a whole, approximately 50% of the variance could be explained. The pretest, principle-based explanations, anticipative reasoning were significant predictors explaining unique variance proportions. In addition, the number of examples inspected, which, on the bivariate level, was not significantly associated with the post-test as a whole, was included in the regression equation. This finding may possibly be explained by the assumption that inspecting multiple examples within a given time span can foster learning, if the single examples are not processed too quickly and superficially, that is, for example, without principle-based explanations and without anticipative reasoning. Thus, when controlling for high-quality self-explanations, which necessarily precludes from moving too quickly through the examples, the number of examples inspected is a positive predictor of learning.

With regard to near transfer performance, solely principle-based explanations and the pretest were significant predictors. The proportion of explained variance (estimation for the population) was just about 25%. This was comparatively low (Table 7). In the case of medium transfer, it was quite astonishing that, when taking self-explanation characteristics into account, the pretest could not explain substantial unique variance. The explication of goal-operator combinations, anticipative reasoning, and the number of examples inspected can account for about half of the variance (Table 7). Thus, the quality of self-explanations seemed to be of major importance for the acquisition of well transferable knowledge.

It is interesting to note that the assessed aspects of the quality of self-explanations had a greater impact on the medium transfer than on the near transfer performance. This may be due to the fact that the aspects coded in this study were especially selected under the perspective of identifying factors that produce high-quality learning (i.e., transferable knowledge). For near transfer tasks, a more passive or rote learning style focussing on "syntactic" aspects may be quite successful (cf. the distinction of transformational analogy and analogical search control by VanLehn & Jones, 1993; VanLehn, Jones, & Chi, 1992). As the latter aspect was not represented as well in the

TABLE 7
Multiple Regressions for the Prediction of Post-test Scores by Self-explanation Variables: Statistically Significant Standardized Regressions Weights

	Post-test Total	Near Transfer	Medium Transfer
Pretest	.33	.42	n.s.
Principle-based explanations	.42	.35	n.s.
Goal-operation combinations	n.s.	n.s.	.40
Anticipative reasoning	.34	n.s.	.47
Number of examples	.23	n.s.	.34
R^{2a}	.55	.30	.53
Adjusted R^{2b}	.49	.26	.49

Notes. Only those self-explanation variables were included in this table that were a significant predictor in at least one regression equation.
[a]All R^2 were statistically significant at the 5% level;
[b]Estimation of the proportion of explained variance in the population.

present coding system, the near transfer could not be predicted to the same extent as the medium transfer.

Is There a Diminishing Return Relation Between Self-explanations and Learning Gains?

One major problem in testing diminishing return relations in form of power law functions is that, taken seriously, the variables have to be measured at the level of rational scales because both variables in a regression equation have to be logarithmically transformed. However, the post-test and the frequencies in verbal protocols can (at best) be regarded as interval scales that accordingly can be subjected to any linear transformation. The results of regression analyses with logarithmic variables are not, however, invariant to linear transformation. On the other hand, if there is some kind of diminishing return relation, regression with logarithmic variables should result in higher proportions of explained variance than traditional linear regression. The exact amount of explained variance of regressions with logarithmic variables should, however, not be interpreted. Hence, the results from Pirolli and Recker (1994) and the following results have to be interpreted with the discussed restriction in mind.[6]

Regressions in which the logarithmic post-test scores were predicted by the logarithmic self-explanation scores were performed. Solely the four self-explanation variables in the narrow sense (principle-based explanations, goal-operator combinations, anticipative reasoning, elaboration of problem situation) were included.[7] As Table 8 shows, no consistent pattern was obtained. In more than half of the cases, the ordinary linear regression could explain higher proportions of variance than the power function regression, and even when the power function lead to higher proportions of explained variance these increases were not substantial. Thus, given the measurement problems (see above) and taking into account the principle of parsimony, it is sensible to further assume linear relations between self-explanation variables and learning (at least on the interindividual plane).

TABLE 8
Post-test Variance Explained by Conventional Linear Regression and by Regression with Logarithmic Variables (Power Function)

	Post-Test		Near Transfer		Medium Transfer	
	Linear Function	Power Function	Linear Function	Power Function	Linear Function	Power Function
Principle-based explanations	14.4	16.6	12.3	12.8	9.6	8.3
Goal-operation combinations	13.7	16.2	6.3	6.3	18.5	19.8
Anticipative reasoning	23.8	23.4	15.2	13.1	23.0	22.7
Elaboration of situation	1.5	1.0	1.0	.4	.4	.2

Can Different Self-explanation Styles Be Identified and, If Yes, Are They Related to Learning Success?

In order to explore whether there are different self-explanation styles, a cluster analysis was performed. This means that the learners were grouped according to their similarity with respect to the self-explanation variables. For this purpose, those self-explanation variables were selected that proved to be related to learning gains (indicated by significant bivariate correlations or multiple regression weights): principle-based explanations, explication of goal-operator combinations, anticipative reasoning, negative monitoring statements, and number of examples. In order to prevent the situation that certain variables determine the cluster solution more than others due to larger variances, z-standardized variables were used. Achievement data (pretest and post-test) were not included, because it was an investigation of whether different self-explanation styles could be identified independently of achievement. In a second step, the question of whether the diverse styles differ with respect to achievement was to be tested.

The cluster analysis method which was employed was the Ward procedure with squared Euclidian distances. The resulting dendrogram (Figure 3) favored a four cluster solution, because aggregating the cases in three (or less) clusters increased sharply the residual variance (intracluster variance). Although the Ward procedure usually tends to result in equal size clusters, in this case relatively unequal group sizes were obtained (cluster 1: $n = 4$; cluster 2: $n = 8$; cluster 3: $n = 19$; cluster 4: $n = 5$).

In order to determine whether the resulting cluster solution yielded a grouping of learners that has some relevance with respect to learning success, the extent to which the achievement scores varied between clusters was examined. An analysis of variance showed that the groups did not differ significantly with respect to their prior knowledge in probability calculation (Table 9). For all post-test scores, except for the unadjusted near transfer performance, significant group differences ($p < .05$) were obtained (Table 9). The clusters 1 and 2 showed a mean above 0.45 (z-score) in every post-test measure meaning that they were (about) half a standard deviation or more above the grand mean in each case. Cluster 1 and cluster 2 did not differ significantly in any measure, although there was a tendency for unadjusted post-test scores to indicate greater success for cluster 1 than for cluster 2, whereas the reverse was true

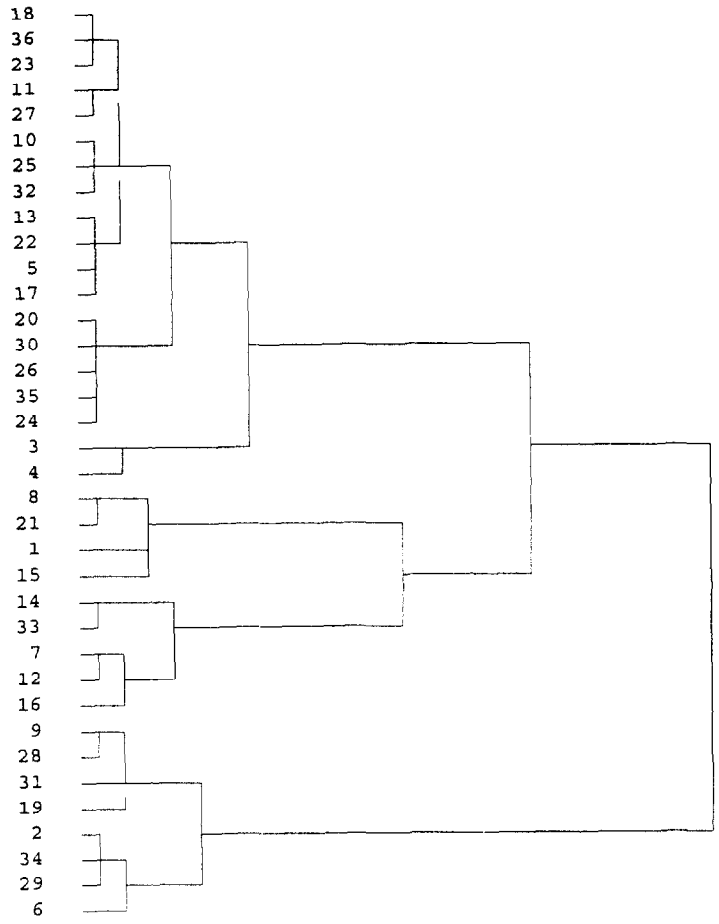


Figure 3. Cluster analysis of self-explanation characteristics: Dendrogram.

TABLE 9
Self-explanation Styles: Means (Standard Deviations in Parentheses) of Ordinary and of Adjusted (Covariate Pretest-Probability) Achievement Scores

	Cluster 1 $n = 4$	Cluster 2 $n = 8$	Cluster 3 $n = 19$	Cluster 4 $n = 5$	F^a p	Post-hoc Comparisons
Pretest-probability	.67 (1.34)	−.34 (1.04)	−.18 (.79)	.67 (1.16)	2.03 > .10	—
Post-test	.95 (.62)	.57 (.90)	−.50 (.72)	.23 (1.39)	4.99 < .05	C2, C1 > C3
Adj. post-test	.69 (.73)	.86 (.96)	−.47 (.74)	−.15 (.98)	5.76 < .05	C2, C1 > C3; C2 > C4
Near transfer	.74 (.51)	.57 (1.16)	−.36 (.93)	−.15 (.70)	2.84 < .10	—
Adj. near transfer	.50 (.19)	.78 (1.12)	−.31 (.94)	−.47 (.37)	3.57 < .05	C2 > C3, C4
Medium transfer	.90 (.72)	.45 (.90)	−.48 (.60)	.39 (1.67)	4.29 < .05	C1, C2 > C3
Adj. medium transfer	.64 (1.11)	.72 (.98)	−.45 (.59)	04 (1.39)	4.11 < .05	C1, C2 > C3

Notes. For better comparability, all variables were z-standardized;
[a]df = 3,32 for ordinary scores; df = 3,31 for adjusted scores (analysis of covariance).

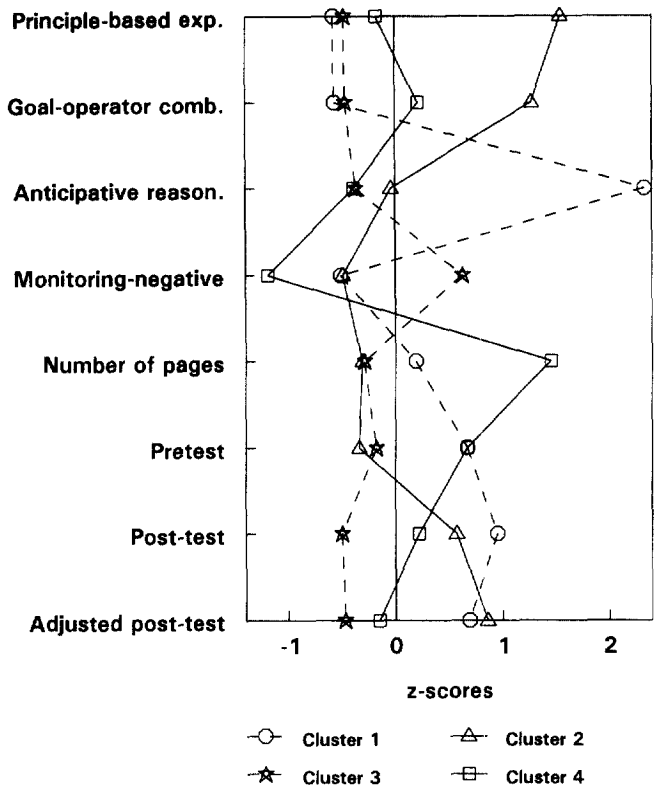


Figure 4. Profiles of the self-explanation clusters.

for the adjusted scores. This was due to the fact that cluster 1 started at a slightly higher prior knowledge level. Cluster 3 seems to have been the least successful group. In each case where the group differences were significant, post-hoc comparisons (Duncan) showed that cluster 3 performed more poorly than clusters 1 and 2 (except for the adjusted near transfer score where no significant difference between cluster 1 and 3 was found). Cluster 4 performed between the successful and the unsuccessful groups, so that most post-hoc comparisons did not reach the level of significance (Table 9).

It could be shown that, although the clusters were determined independently from achievement data, they differed substantially in this respect. Hence, it was worthwhile to inspect the self-explanation styles of the different clusters (Figure 4). Cluster 1 concentrated their efforts on the anticipative computation of to-be-found probabilities and did not provide either many principle-based explanations nor explicated many goal-operator combinations. This group self-diagnosed just a few comprehension impasses and inspected a medium number of examples. The post-test performance was high. However, this group started at a relatively high level of prior knowledge. Members of this group could be labeled as *anticipative reasoners*.

Cluster 2 was—according to the adjusted means—the most successful group. It could be characterized by a self-explanation style that emphasized the assignment of

"meaning" to the operators, both by explicating the underlying principle and the corresponding subgoal. Anticipative reasoning was performed merely at a medium level. The learners in this group infrequently noticed comprehension failures and inspected a slightly below-average number of examples. They started at a relatively low prior knowledge level, but reached a high level of learning success. This self-explanation style could be characterized as data-driven, but nevertheless active (cf. Reimann, 1994; Reimann & Schult, 1996). It was named *principle-based*.

Cluster 3 was the comparatively large group of 19 learners that could be described as unsuccessful. This relative failure to profit from studying worked-out examples obviously resulted from the poor quality of self-explanations: There were few principle-based explanations, few nominations of goal-operator combinations, and a low level of anticipative reasoning. In addition, this group also noticed many comprehension impasses and did not inspect many examples. This self-explanation style was labeled *passive*.

The individuals in cluster 4 engaged in an average amount of principled-based explanations and explications of goal-operator combinations. Anticipative reasoning was relatively infrequent. Interestingly, although they were merely medium successful learners, they very seldom noticed comprehension problems and assigned relatively little time to each example (i.e., inspected many examples). There is some similarity of this style to the unsuccessful learners described by Chi et al. (1989). Those subjects also inspected the examples for a relatively brief time and their learning success contrasted with the low extent of self-diagnosed comprehension difficulties. This self-explanation style is named *superficial*.

DISCUSSION

To What Extent Are Individual Characteristics of Self-explanations Generalizable Over Examples?

Individual differences in the quality of self-explanations are consistent over different worked-out examples. The self-explanation variables focussed on in this study could be reliably measured (with the exception of noticing coherence between examples). This means that these variables can be viewed as an expression of a person characteristic that is relatively stable over different worked-out examples (at least in one domain). The fact that for noticing coherence between examples no sufficient reliability (i.e., differentiability of subjects) was obtained indicates that it is not trivial to assume that there are reliable individual differences. Obviously, individuals do not consistently differ from each other with respect to any self-explanation variable that seems to be important from a theoretical point of view.

To What Extent Are Different Self-explanation Characteristics Correlated?

The intercorrelations between the self-explanation characteristics were rather low. Thus, the question whether the quality of self-explanations is a unidimensional construct is answered in the negative. This means that if a person employs certain

effective self-explanation strategies such as principle-based explanations, it does not imply that this person also shows a great amount of anticipative reasoning. This finding does not, however, signify that one cannot find individuals who fit the prototypes of a good or poor self-explainer with respect to several aspects.

The multidimensionality of self-explanation characteristics also has a consequence for strategy training. If different learners lack different effective self-explanation strategies, an effective intervention procedure should involve a proper diagnosis of individual deficits. Hence, training efforts can be specifically tailored to individual needs and need not train some strategies already employed by a learner. Nevertheless, there may, of course, be some poor learners who lack most of the effective strategies and actually need some type of "overall" training.

To What Extent Is the Quality of Self-explanations Associated with Cognitive Prerequisites?

Although there were reliable differences in prior knowledge that did predict final performance in probability calculation, self-explanation characteristics do not seem to significantly depend on prior knowledge (except for anticipative reasoning). This finding is consistent with the results of Chi and VanLehn (1991). This can be regarded as good news because higher levels of prior knowledge do not seem to be a major prerequisite for the employment of the types of self-explanation strategies used by the successful learners in the present study. The present findings do not, however, exclude the possibility that, when higher standards for evaluating the quality of self-explanations (e.g., similarity to explanations given by mathematics teachers) are employed, substantial prior knowledge may be necessary to meet these standards.

Are the Learning Results Associated with the Quality of Self-explanations Even When Time-on-task Is Controlled?

To date, there have been no strong counter-arguments to the potential objection that the empirically found self-explanation effects may actually be time-on-task effects. This study has shown that even when controlling for time-on-task self-explanation effects are found. Thus, it is shown that the association of an active self-explanation style with learning gains, which was found by Chi et al. (1989; 1994) and by Pirolli and Recker (1994), cannot be explained merely by the increased time-on-task that the successful subjects in those studies spent in studying examples. Hence, apart from time-on-task effects, qualitative differences of self-explanation play a major role in successful learning.

The positive association between principle-based explanations and learning is not only in agreement with Chi et al.'s (1989) study, but also with research in mathematics learning which stresses the importance of principle-based knowledge for effective problem solving (e.g., Hiebert, 1986). The relation of the explication of goal-operator combinations to learning confirms, in particular, the assumptions made by Catrambone and Holyoak (1990; Catrambone, 1994; 1995) that the explicit encoding of subgoal-operator connections fosters transfer to new tasks.

The positive effect of anticipative reasoning to learning has not been the focus of previous empirical research on individual differences in learning from examples. The theoretical assumptions made by Reimann and his colleagues (Reimann, 1994; Reimann, Schult, & Wichmann, 1993) about the importance of an expectation-driven example processing style are, however, confirmed. Anticipative reasoning is an especially important variable of individual differences because it is rather uncorrelated to other self-explanation characteristics and thus a nonredundant predictor of learning (except for near transfer), as was shown by the results of the multiple regression analyses. Anticipative reasoning may be of significance because it prevents some type of illusion of competence. As the goal of studying examples is to be able to later produce solutions, it is reasonable for the subjects to check their own present competence level by anticipating solution steps and then comparing them with the worked-out solutions.

In sharp contrast to the results of Chi et al. (1989) and of Ferguson-Hessler and DeJong (1990), it was found in this study that the poor learners frequently diagnosed their comprehension failures and did not tend to have the illusion of comprehension. Chi et al. (1989) interpreted their results as follows. The poor students do not realize that they do not understand or even think that they do understand although they do not; the good students, in contrast, diagnose their comprehension failures and can then initiate working towards a better understanding. The usefulness of the self-diagnosis of comprehension failures depends, however, on the possibility or ability to resolve them effectively. If the instructional materials are relatively difficult and there are no external support devices (e.g., availability of a tutor or a computer-based help system), efforts to improve comprehension may be unsuccessful. Possibly, the worked-out examples used in the present study were of a kind that comprehension failures frequently could not be rectified by the students who were left to their own devices during their study of the worked-out examples. In addition, the worked-out examples to be studied were not very susceptible to illusion of understanding. Poor learners were not characterized by high frequencies of (non-veridical) positive monitoring statements. On the contrary, they diagnosed numerous comprehension failures. Apparently, they were not, however, very successful in resolving them. Informal inspection of the protocols confirmed this assumption. Thus, in the present study, the number of negative monitoring statements was more or less an indicator of difficulties in the learning process and not of effective metacognitive control. Taken together, many negative monitoring statements cannot be generally viewed as characteristic of good learners. Depending on context variables, such as the quality and difficulty of the instructional materials or the availability of help devices, negative monitoring statements may also be a characteristic of poor learners.

With respect to the control variables, the following findings were obtained. The protocol length, as the indicator of verbosity and word fluency, was not associated with learning. The number of examples inspected was predictive of learning, especially of the medium transfer performance. This finding indicates that the inspection of multiple examples can obviously foster the acquisition of transferable knowledge (Catrambone & Holyoak, 1989; Paas & VanMerrienboer, 1994).

The discussion about the influence of self-explanation characteristics must be qualified in that the present study was correlational in nature. Thus, the possibility cannot be excluded that other variables not included may be responsible for the associations obtained. There is, however, evidence from experimental studies in which self-explanations were either trained (e.g., Bielaczyc, Pirolli, & Brown, 1995) or prompted (Chi et al., 1994) that indicates that self-explanations do, in fact, influence learning gains.

Is There a Diminishing Return Relation Between Self-explanations and Learning Gains?

The present results do not support the assumption of diminishing return relations between self-explanations and learning gains. The corresponding hypotheses by Pirolli and Recker (1994) had been inspired by the power law of practice (Rosenbloom & Newell, 1986). Originally, the power law of practice described an intraindividual relation, that is, the relation between practice trials and time to complete a task. In Pirolli and Recker's (1994) study and in the present investigation, however, relations on the interindividual level were investigated. It is important not to mix up the two perspectives (Renkl, 1993; Valsiner, 1986). Thus, although a diminishing return function on the interindividual level was not found, this does not necessarily indicate that there is no diminishing return relation on the intraindividual level. Perhaps the power law relation was not confirmed on the interindividual plane, because there were no subjects that explained the worked-out examples to themselves in a redundant, "over"-thoroughly manner so that additional explanations were of diminishing use.

Can Different Self-explanation Styles Be Identified?

Different self-explanation styles could be identified that varied in their success in learning. Maybe the most interesting finding is that there are two distinct (relatively) successful styles (i.e., the *anticipative reasoners* and the *principle-based explainers*). This finding confirms claims made by several authors (Mandl & Renkl, 1992; Renkl & Mandl, 1995; Weinert, 1988) that theoretical models (as well as empirical studies) should take into account more frequently that there may exist substitute mechanisms. This means that in the absence of a favorable condition (e.g., many principle based-explanations) another condition (e.g., frequent anticipative reasoning) can be substituted. Hence, there are multiple effective ways of learning.

The finding that more than half of the subjects had to be assigned to the group of unsuccessful learners, reaffirms research findings that learners, left to their own devices, typically fail to show effective learning behaviors when no external support (e.g., teacher guidance or scaffolding) is present (Njoo & DeJong, 1993; Stark, Graf, Renkl, Mandl, & Gruber, 1995).

It is interesting to compare the present results with the clusters of learners found by Recker and Pirolli (1995) in their study on self-explanation effects. Although there is no direct correspondence, there is some similarity between the clusters. Recker and Pirolli (1995) identified clusters that differed from each other with respect

to their balancing of gains and costs associated with efforts in self-explaining instructional material. The group that maximized effort during the instructional period in order to prepare for problem solving seems to be comparable to the groups of anticipative reasoners and principle-based explainers. The individuals in the study by Recker and Pirolli (1995) who minimized costs by minimizing cognitive effort resemble the passive learners in the present study. Finally, the learners who were labeled as superficial correspond to the subjects who, according to Recker and Pirolli (1995), try to balance costs and gains. They do this by first skimming through the instructional materials and then restudying the instructional material more deeply when (during problem solving) task goals have become more concrete. However, in the present investigation, restudying was not possible so that the extent to which the superficial learners would have behaved in the manner described by Recker and Pirolli remains open.

OUTLOOK

Of course, the present study not only provides answers to research questions, but also leaves open many significant queries to be addressed in future research. At this point, however, merely three important issues should be discussed.

First, for self-explanation characteristics such as principle-based explanations or explication goal-operator combinations, substantial empirical evidence has accumulated so that their importance for learning can be taken for granted. Anticipative reasoning, in contrast, has not been focussed on in the previous research. The extent to which this self-explanation characteristic is also of significance in the context of other presentation modes than step-by-step procedures, of other domains, etc. should be investigated in further studies.

Second, the divergent findings with respect to monitoring in the present study and in previous investigations could only speculatively be explained. Systematic experimentation on potential moderating conditions would, however, be necessary in order to obtain more conclusive explanations of these divergent findings.

Finally, the issue of diminishing returns deserves further investigation. An analysis of the relation of the number of explanations and learning on the intraindividual plane would appear to be fruitful. Thus, more direct evidence on the question of the degree to which—beyond a certain point—additional self-explanations are of diminishing use can be obtained.

Acknowledgments

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, Re 1040/1-1).

I thank M. Chi, H. Gruber, K. VanLehn, and an anonymous reviewer for valuable comments and suggestions on earlier drafts.

Notes

1. Segmentation and coding (of all categories) were performed in a single step.
2. As the purpose of the text on probability principles was to equalize individual prior knowledge differences with respect to these principles, no associations with text study variables and later learning were expected. In other words, the nonsignificant correlation of the study time and later learning is compatible with the assumption of a successful equalization of prior knowledge differences in these principles.
3. A measure of far transfer (e.g., problems based on the same probability principles but with different deep *and* different surface structures) was not available.
4. Even when the relative frequency of noticing coherence (absolute frequency divided by the number of inspected worked-out examples having an isomorphic predecessor) was regarded, no sufficient reliability score was obtained.
5. At first glance, it may be interesting to also operationalize the self-explanation characteristics by relative frequencies (absolute frequencies divided by the number of examples inspected) when determining their relation to learning gains. Thus, the extent to which the *inspected* examples were deeply processed could be analyzed and whether this extent influenced learning could be determined. However, using relative frequencies in determining the quality of self-explanations placed those subjects at disadvantage who inspected relatively many examples. This procedure may, for example, result in a person with many principle-based explanations receiving a low score for the corresponding variable because he or she inspected many examples. Since there is no theoretical rationale that would justify this, the corresponding data should not be discussed in detail. It should be solely mentioned that the pattern of results remains unchanged, except that the positive relations of principle-based explanations and of the explication of operator-goal combinations to most achievement scores fall below the level of significance.
6. The problem of scale quality is not relevant when using the number of learning trials as predictor and time for task completion as dependent variable, as in the original formulation of the power law by Newell and Rosenbloom (1981).
7. Since there were scores of zero and the logarithm of zero is undefined, a constant of 1 was added to each variable.

REFERENCES

Anderson, J.R., Farrell, R., & Sauers, R. (1984). Learning to program in LISP. *Cognitive Science, 8*, 87–129.

Bielaczyc, K., Pirolli, P., & Brown, A.L. (1995). Training in self-explanation and self-regulation strategies: Investigating the effects of knowledge acquisition activities on problem solving. *Cognition and Instruction, 13*, 221–252.

Catrambone, R. (1994). Improving examples to improve transfer to novel problems. *Memory and Cognition, 22*, 606–615.

Catrambone, R. (1995). Aiding subgoal learning: Effects on transfer. *Journal of Educational Psychology, 87*, 5–17.

Catrambone, R., & Holyoak, K.J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 1147–1156.

Catrambone, R., & Holyoak, K.J. (1990). Learning subgoals and methods for solving probability problems. *Memory and Cognition, 18*, 593–603.

Chi, M.T.H., Bassok, M., Lewis, M.W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 18*, 145–182.

Chi, M.T.H., DeLeeuw, N., Chiu, M.H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439–477.

Chi, M.T.H., & VanLehn, K.A. (1991). The content of physics self-explanations. *The Journal of the Learning Sciences, 1*, 69–105.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Ericsson, K.A., & Simon, H.A. (1993). *Protocol analysis* (Revised edition). Cambridge, MA: MIT Press/Bradford.

Ferguson-Hessler, M.G.M., & DeJong, T. (1990). Studying physics texts: Differences in study processes between good and poor performers. *Cognition and Instruction, 7*, 41–54.

Gick, M.L., & Holyoak, K.J. (1980). Analogical problem solving. *Cognitive Psychology, 12*, 306–355.

Gick, M.L., & Holyoak, K.J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1–38.

Helmke, A., & Renkl, A. (1992). Das Münchener Aufmerksamkeitsinventar (MAI): Ein Instrument zur systematischen Verhaltensbeobachtung der Schüleraufmerksamkeit im Unterricht [The Munich Attention Inventory (MAI): An instrument for the systematic observation of students' on-task-behavior during instruction]. *Diagnostica, 38*, 130–141.

Hiebert, J. (Ed.) (1986). *Conceptual and procedural knowledge: The case of mathematics*. Hillsdale, N.J.: Erlbaum.

Kintsch, W. (1986). Learning from text. *Cognition and Instruction, 3*, 87–108.

LeFevre, J.-A., & Dixon, P. (1986). Do written instructions need examples? *Cognition and Instruction, 3*, 1–30.

Lienert, G.A., & Hofer, M. (1977). *Mathematiktest für Abiturienten und Studienanfänger. M-T-A-S* [Mathematics test for 13th graders and university freshmen]. Göttingen: Hogrefe.

Lovett, M.C. (1992). Learning by problem solving versus by examples: The benefits of generating and receiving information. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society 1992*. Hillsdale, NJ: Erlbaum.

Mandl, H., & Renkl, A. (1992). A plea for "more local" theories of cooperative learning. *Learning and Instruction, 2*, 281–285.

Newell, A., & Rosenbloom, P.S. (1981). Mechanisms of skill acquisition and the law of practice. In J.R. Anderson (Ed.), *Cognitive skills and their acquisition*. Hillsdale, NJ: Erlbaum.

Njoo, M.K.H., & DeJong, T. (1993). Exploratory learning with a computer simulation for control theory: Learning processes and instructional support. *Journal of Research in Science Teaching, 30*, 821–844.

Paas, F.G.W.C., & VanMerrienboer, 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.

Pirolli, P., & Anderson, J.R. (1985). The role of learning from examples in the acquisition of recursive programming skills. *Canadian Journal of Psychology, 39*, 240–272.

Pirolli, P., & Recker, M. (1994). Learning strategies and transfer in the domain of programming. *Cognition and Instruction, 12*, 235–275.

Recker, M.M., & Pirolli, P. (1995). Modelling individual differences in students' learning strategies. *The Journal of the Learning Sciences, 4*, 1–38.

Reimann, P. (1994). *Lernprozesse beim Wissenserwerb aus Beispielen: Analyse, Modellierung, Förderung* [Learning processes in knowledge acquisition from examples: Analyses, modelling, fostering] (Habilitationsschrift). Freiburg: Universität Freiburg.

Reimann, P., & Schult, T.J. (1996). Turning examples into cases: Acquiring knowledge structures for analogical problem solving. *Educational Psychologist, 31,* 123–132.

Reimann, P., Schult, T.J., & Wichmann, S. (1993). Understanding and using worked-out examples: A computational model. In G. Strube & K.F. Wender (Eds.), *The cognitive psychology of knowledge*. Amsterdam: North-Holland.

Renkl, A. (1993). Kovariation und Kausalität: Ein ausreichend durchdachtes Problem in der pädagogisch-psychologischen Forschung? [Covariation and causality: A sufficiently reflected problem in psychological and educational research?] In C. Tarnai (Ed.), *Beiträge zur empirischen pädagogischen Forschung*. Münster: Waxman.

Renkl, A. (1995). Learning for later teaching: An exploration of mediational links between teaching expectancy and learning results. *Learning and Instruction, 5*, 21–36.

Renkl, A., & Mandl, H. (1995). Kooperatives Lernen: Die Frage nach dem Notwendigen und dem Ersetzbaren [Cooperative learning: The question of necessary and replaceable conditions], *Unterrichtswissenschaft, 23*, 292–300.

Rosenbloom, P.S., & Newell, A. (1986). The chunking of goal hierarchies: A generalized model of practice. In R.S. Michalski, J.G. Carbonell, & T.M. Mitchell (Eds.), *Machine learning* (Vol. 2). Los Altos, CA: Kaufman.

Stark, R., Graf, M., Renkl, A., Gruber, H., & Mandl, H. (1995). Förderung von Handlungskompetenz durch geleitetes Problemlösen und multiple Lernkontexte [Fostering action competence by guided problem solving and multiple learning contexts]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 27*, 289–312.

Sweller, J., & Cooper, G.A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction, 2*, 59–89.

Tarmizi, R.A., & Sweller, J. (1988). Guidance during mathematical problem solving. *Journal of Educational Psychology, 80*, 424–436.

Valsiner, J. (1986). Between groups and individuals: Psychologists' and laypersons' interpretation of correlational findings. In J. Valsiner (Ed.), *The individual subject and scientific psychology*. New York: Plenum.

VanLehn, K. (1986). Arithmetic procedures are induced from examples. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics*. Hillsdale, NJ: Erlbaum.

VanLehn, K. (1996). Cognitive skill acquisition. In J. Spence, J. Darly, & D.J. Foss (Eds.), *Annual Review of Psychology*. Palo Alto, CA: Annual Reviews.

VanLehn, K., & Jones, R.M. (1993). Learning by explaining examples to oneself: A computational model. In S. Chipman & A.L. Meyrowitz (Eds.), *Foundations of knowledge acquisition: Cognitive models of complex learning*. Boston, MA: Kluwer.

VanLehn, K., Jones, R.M., & Chi, M.T.H. (1992). A model of the self-explanation effect. *The Journal of the Learning Sciences, 2*, 1–59.

Ward, M., & Sweller, J. (1990). Structuring effective worked examples. *Cognition and Instruction, 7*, 1–39.

Weinert, F.E. (1988). Jenseits des Glaubens an notwendige und hinreichende Bedingungen schulischen Lernens [Beyond the belief in necessary and sufficient condition of school learning]. In J. Lompscher, W. Jantos, & S. Schönian (Eds.), *Psychologische Methoden der Analyse und Ausbildung der Lernfähigkeit*. Berlin: Gesellschaft für Psychologie der DDR.

Zhu, X., & Simon, H.A. (1987). Learning mathematics from examples and by doing. *Cognition and Instruction, 4*, 137–166.