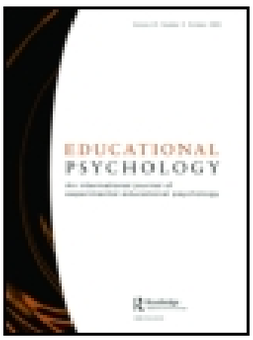

Routledge
Taylor & Francis Group

Educational Psychology

An International Journal of Experimental Educational Psychology

ISSN: 0144-3410 (Print) 1469-5820 (Online) Journal homepage: https://www.tandfonline.com/loi/cedp20

Investigating how errors should be flagged and worked examples structured when providing feedback to novice learners of mathematics

Elisapesi Manson & Paul Ayres

To cite this article: Elisapesi Manson & Paul Ayres (2019): Investigating how errors should be flagged and worked examples structured when providing feedback to novice learners of mathematics, Educational Psychology, DOI: 10.1080/01443410.2019.1650895

To link to this article: https://doi.org/10.1080/01443410.2019.1650895

EDUCATIONAL PSYCHOLOGY
https://doi.org/10.1080/01443410.2019.1650895

Routledge
Taylor & Francis Group

Investigating how errors should be flagged and worked examples structured when providing feedback to novice learners of mathematics

Elisapesi Manson[a] and Paul Ayres[b]

[a]School of Humanities, Education and Theology, Pacific Adventist University, Port Moresby, Papua New Guinea; [b]School of Education, University of New South Wales, Kensington, Australia

ABSTRACT

This study investigated the effectiveness of using a sequence of worked examples as part of the feedback cycle. Worked examples were either presented as full worked examples or partial worked examples (single-step and completion formats). In two experiments, grade 8 students completed a learning phase on a mathematics topic, which was immediately followed by a testing phase. A day later, participants were given feedback on their test papers and provided worked examples to problems where errors were made, and then re-tested. In Experiment 1 ($N = 73$), studying full worked examples led to greater improvement than studying single-step worked examples. In Experiment 2 ($N = 74$), full worked examples led to greater improvement than studying either single-step worked examples or completion worked examples. Furthermore, no learning differences were found when learner errors were directly flagged or otherwise. In conclusion, providing full worked examples as feedback to novice learners was helpful.

ARTICLE HISTORY
Received 29 May 2018
Accepted 29 July 2019

KEYWORDS
Worked examples; direct instruction; mathematical cognition; cognitive load theory; feedback

Introduction

Worked examples provide step-by-step solutions to a problem or task and are a form of direct instruction. They provide an expert problem-solving model, which students can study and learn from (Atkinson, Derry, Renkl, & Wortham, 2000). Rather than acquiring new knowledge through problem-solving or other types of unguided methods (see Kirschner, Sweller, & Clark, 2006), learners are shown worked examples to study. A vast amount of research has shown that for novice learners, in particular, worked examples leads to greater learning outcomes than problem-solving based methods (Ayres & Sweller, 2013).

Despite the success of using working examples, little if at all any, research has been conducted using a sequence of worked examples. Most studies are single learning experiences. Typically, there is the main acquisition phase where a worked example strategy is compared with another instructional method such as problem-solving,

CONTACT Paul Ayres p.ayres@unsw.edu.au School of Education, University of New South Wales, Kensington, Australia

which is then followed by a testing phase to evaluate the effectiveness of the worked examples. Studies into worked examples usually end there with no attempt to provide further worked examples based on learner performance.

Worked examples are a highly effective method of helping students learn about new content, but they do not guarantee that all students master the given content in the time provided. Many students need further instruction and practice. Hence, the current study focused on providing additional learning time through a second set of worked examples. It was assumed that learners could benefit from a sequence of two sets of worked examples. Furthermore, in order to enhance learning further, feedback was provided between the two sets of worked examples. However, in order to explore how to create the most effective learning environment under these conditions, two factors were manipulated. The first factor investigated how to structure the second set of worked examples, in order to deal with the potential influence of developing expertise. The second factor investigated the type of feedback that should be provided to best link together the two sets of worked examples.

Worked examples

The worked example effect

The use of worked examples is not new and have been used extensively as a common teaching strategy in many learning disciplines, and thoroughly investigated within the theoretical underpinnings of cognitive load theory (see Sweller, Ayres, & Kalyuga, 2011). Initial research by Sweller and Cooper (1985) into intermediate school mathematics found that worked examples facilitated superior learning outcomes to conventional problem-solving methods (known as the worked example effect). Building on this initial research, the worked example effect has been replicated in many other mathematics and scientific domains, as well as more non-procedural domains (see Ayres & Sweller, 2013).

The theoretical underpinnings of the worked example effect can be explained by the difficulty posed by learning from minimal guidance strategies that rely on problem-solving (Kirschner et al., 2006). Early research into cognitive load theory demonstrated that learners could solve problems but did not necessarily learn from the process. Faced with novel problems to solve, learners rely on general problem-solving strategies that generate a high cognitive load. As a result, the majority of working memory resources are diverted to solving the problem rather than acquiring new knowledge about the topic (Sweller et al., 2011). In contrast, by providing a solution to a problem, worked examples reduce the number of search processes and problem state manipulations associated with problem-solving, allowing more working memory resources to be devoted to understanding and learning about the problem.

Worked example problem pairs

In their original study into worked examples, Sweller and Cooper (1985) used an alternation strategy of 'study a solution to a problem and solve a similar problem', based on the argument that students needed to solve some problems to avoid a lack of

motivation. The success of this strategy led to its wide-scale adoption in worked examples research. However, more recently, Van Gog, Kester, and Paas (2011) found evidence that there may be no need to actively solve a similar problem, as studying its solution may be just as effective. The key condition is that learners should study a worked example first, rather than initially attempting to solve the problem.

Worked examples and the expertise reversal effect

As the research into worked examples became more widespread and sophisticated it emerged that the worked example effect was mainly limited to novice learners, with little knowledge on the topic to-be-learned (domain-specific knowledge). As expertise in a domain increases, instructional strategies with less direct guidance become more effective. This effect is called the expertise reversal effect (Kalyuga, Ayres, Chandler, & Sweller, 2003).

To explain the expertise reversal effect, cognitive load theory researchers have argued that as expertise increases, strategies helpful for novices become redundant (Sweller et al., 2011). Once a certain amount of knowledge has been acquired, learners have sufficient knowledge to deal with more complex processing such as that generated by problem-solving. Hence, problem-solving may not increase the cognitive load to such a level that interferes with learning because the information is available in long-term memory (domain-specific knowledge) to deal with its demands. However, requiring learners with some expertise to use worked examples may be redundant leading to unnecessary processing, increases in cognitive load and a subsequent loss in learning. To prevent the expertise reversal effect occurring with worked examples, a number of modifications have been applied to their structure to effectively manage the transition from novice to a more knowledgeable learner.

Restructuring worked examples to deal with the transition to expertise

Most modification strategies have been based on completion problems. A completion problem is a partially worked example, where the learner has to complete some key steps, thus creating a more active role for the learner (see Van Merriënboer, 1990). Completion strategies have been found to be more effective than problem-solving strategies but not necessarily superior to full worked examples (Paas, 1992).

To also assist in the transition from novice to expert, completion problems have also been combined with fading strategies. As domain-specific knowledge increases, worked out steps are gradually faded out until problems can be solved without any guidance (Schwonke et al., 2009). Fading strategies have been found to be more effective than continuing with full worked examples as expertise increases (Salden, Aleven, Schwonke, & Renkl, 2010).

The present study

The present study investigated how to enhance learning by providing a sequence of two sets of worked examples. The first set was used to acquire initial knowledge

about the mathematical domain, and the second set was used to reinforce and improve this knowledge. Crucial to the success of this sequence is the type of feedback provided. In many mathematics classrooms, regardless of the teaching strategies employed, students will receive some form of instruction on a topic, and then complete further tasks related to this topic. Such tasks, whether they are part of formal assessments or not, will be evaluated by the teacher and returned to students. This is likely to occur a day or two later depending upon the type of task, efficiency of the teacher, and the frequency of classes. The current study aimed to replicate this type of teaching practice in an authentic school environment and provide ecological validity, by providing feedback a day after completing the task. The type of feedback given and how often it is provided can make a significant impact on learning as described next.

The importance of feedback

Much is known about the importance of feedback (see Mory, 2004) and under many conditions, it has been found to improve student learning particularly when it is of a formative nature (Shute, 2008). There are a number of important aspects to providing effective feedback. It should provide guidance and opportunities for the learner to improve (Orsmond & Merry, 2011). Effective feedback should allow learners to recognise their next steps and how to take those steps (Mory, 2004). Corrective feedback can help learners improve by identifying errors (Marzano, Pickering, & Pollock, 2001), which should be accompanied by some form of explanations or elaborated information on the reasons responses were correct or incorrect (Hattie & Timperley, 2007). Feedback also needs to be carefully aligned with the student's prior knowledge in order to be effective (Hattie & Timperley, 2007). As frequently shown, learners with low prior knowledge need high instructional support compared to learners with high prior knowledge (Kalyuga et al., 2003).

Feedback in the present study

Worked examples can intrinsically fulfil many of the feedback functions outlined above. They can be a source of error identification, provide guidance and opportunities to improve, provide information on solution steps, and are a form of elaborated information. Worked examples are also an effective way of dealing with differences in prior knowledge as they can be modified a number of different ways. In the present study, two other feedback strategies were directly incorporated into the experimental design. Firstly, feedback on errors made was provided after students had first completed a test on the topic (Shute, 2008). Secondly, the feedback was provided in a timely fashion (the next day) while the mathematics topic was still being studied (see Kulik & Kulik, 1988) and learners had some memory of their previous actions (Mory, 2004).

Study hypotheses

As previously described, worked examples can lead to expertise reversal effects as novice learners transition to more expert learners. This transition can be managed by using fading

strategies (Salden et al., 2010). However, as the target learners in the present study were novice learners in the early stages of knowledge acquisition, different worked example structures were investigated rather than fading strategies. Hence, the first factor manipulated was the structure of the second set of worked examples provided after feedback.

One approach to providing worked examples feedback following a solution error is to provide a complete worked solution to the problem. Providing a full worked example to a problem not only provides a correction to the error, but also may consolidate overall knowledge about that problem. However, if for example, a learner makes only one error or two errors in a multi-step problem solution, providing full worked examples may be redundant (Sweller & Chandler, 1991), as the learner has already demonstrated some knowledge in completing part of the solution successfully. A way to counteract this potential cause of redundancy is to provide a partial worked solution aligned directly with the parts of the solution where errors were made.

However, if expertise effects are not present, a worked example consisting of partial solution steps may lack continuity and fail to show the various connections that need to be made for optimum learning to occur. Furthermore, as worked examples research indicates, learners with low prior-knowledge benefit from full worked examples, it is expected that such learners will benefit from full worked examples during feedback as well. Hence, it was predicted that the low-knowledge learners used in this study would benefit from full worked examples rather than a partial approach. Hence, the first hypothesis tested was:

Hypotheses 1: Full worked examples will lead to higher learning outcomes than partial worked examples when presented as feedback.

Also of critical interest in this study was how learner errors are flagged. Flagging errors are a form of corrective feedback, which in its basic form indicates whether a response is correct or not. Research suggests that errors should be indicated and not left for the learner to discover (Van Beuningen, de Jong, & Kuiken, 2012). Hence, indicating an error followed by a worked example on the problem provides both corrective and elaborative feedback. However, if worked examples are provided to learners when errors are made then learners are able to locate their errors themselves. Subsequently, worked examples represent a form of indirect flagging. Hence, an important question arises in that, does directly flagging errors enhance learning compared to indirect flagging errors when using worked examples as part of the feedback cycle? This form of indirect flagging may be superior to direct flagging but in the case of low expertise, learners may be an extra burden, raising cognitive load sufficiently to interfere in learning. Therefore the following prediction was made:

Hypothesis 2: Directly flagging errors will lead to higher learning outcomes than not explicitly flagging errors.

Experiment 1

The first experiment was conducted to not only directly test hypotheses 1 and 2, but also to explore some of the factors that may impact on the testing of both

hypotheses. Two worked example conditions were compared: full worked examples and single-step worked examples (a form of partial worked example).

As reported in the literature review, worked examples have taken different forms. The most frequent format is to use the study-solve problem pairing originally used by Sweller and Cooper (1985). However, Van Gog et al. (2011) have shown that it is not always necessary to solve a problem and worked examples can be effective by simply studying worked examples. Therefore in this study to explore whether it is necessary to use problem pairs as part of the feedback, a study-only option was used. Furthermore, to test whether effects could be found without directly indicating errors, learner errors were not directly flagged and had to be found indirectly through the worked examples.

Method

Participants

Seventy-three eighth-grade (median age of 13 years) students (40 boys, 33 girls) from high schools in New South Wales, Australia, initially participated in this experiment. Nineteen participants were later excluded because they either did not participate in all phases of the experiment and/or scored above 80% in the pre-test and hence, were considered to have little potential to improve in their performance in the post-test. As previously indicated, we were specifically interested in learners who were in need of further instruction rather than those who had shown a high degree of mastery of the topic.

Participants were then randomly assigned to one of two conditions after the initial testing phase: A single-step worked example group ($N = 27$) or a full worked example group ($N = 27$). All students had recently gained some initial prior learning experience with the content covered in this experiment from their normal mathematics classes but were still considered novice learners in the domain.

Design

The experiment consisted of 4 stages. In the first stage (acquisition), all participants received instruction on the given topic through a set of worked examples. During the second stage, participants were tested (pre-test) on the topic studied during the acquisition phase. For the third stage, participants received feedback according to the group they were assigned to and given more study time. In the final fourth stage, they were re-tested (post-test) on the same topic.

Learning topic

The experimental materials were based on the topic *linear equations* (e.g. solve $2x + 10 = 18$). The expected knowledge and skills outcomes stipulated that learners would be able to solve equations using algebraic methods including simplification of algebraic expressions that involve addition, subtraction, multiplication, division, fractions and expansions of algebraic expressions by removing grouping symbols.

Acquisition stage

The instructional materials consisted of six pairs of linear equation problems consistent with the paired problem format of study a solution to a problem and solve a similar problem (see Cooper, & Sweller, 1987). Each pair consisted of a complete solution to a problem (e.g. solve $2x + 10 = 18$) to be studied, and a structurally similar problem to be solved (e.g. solve $3x + 8 = 23$). These materials were presented on A4 (21 cm $\times$ 29.7 cm) sheets of white paper. The worked out solution was given on the left side of the sheet, and the problem to be solved on the right side, with sufficient space for students to complete their solution steps and answers. A list of correct worked out solutions for all six problems to be solved was available during the instructional phase on a separate sheet of paper. Participants were instructed to refer to each solution after solving a problem and make appropriate corrections before moving to the next pair.

Test materials

The pre-test consisted of 10 problems to solve. The first six problems were similar problems (similar test) to those experienced during acquisition; while the remaining four problems were considered transfer problems (transfer test), as they required more mathematical manipulation [e.g. solve $3\ (7x - 2) = 3\ (2x + 3)$] from the problems encountered during the acquisition phase, but were still a form of linear equations but more complex. This test was presented on sheets of A4 paper with sufficient space for students to show their solution steps and answers. The post-test had an identical format to the pre-test (six similar problems and four transfer problems) with structurally similar problems but not identical to the pre-test.

Feedback phase

Pre-test answer sheets for each participant were first photocopied. On the photocopies test answers were scored, tallied and kept by the researcher. On the originals, feedback was provided according to the conditions, and returned to the participants next day. No numerical marks were indicated on the original test papers to avoid potential negative effects of receiving such marks.

All photocopied test papers were marked with the same marking criteria according to a strict rubric. Each problem in the pre- and post-tests were assigned three marks. One mark was deducted for each error made or missing information. Therefore, 3 marks were awarded when no error was made, 2 marks were awarded for 1 error made, 1 mark when 2 errors were made, and 0 marks for 3 or more errors. If an error was made, then subsequent solutions steps that were correct based on that error were not penalised unless the level of problem complexity was diminished.

Feedback for each test paper was provided based on the assigned feedback group and the mark for each question on the photocopied test paper. For the single-step worked example group, the feedback consisted of the correct worked example step for each line where an error was made, positioned in the space provided on the right of the test paper in line with the error. If no errors were made no feedback was given. If one error was made then one solution step was provided, if two errors were made then two solutions steps were shown, and so on.

Table 1. Example of types of feedback given in Experiment 1.

	Correction feedback	
Participant's error	One-step worked example	Full worked example
$5x - 8 = 27$		$5x - 8 = 27$
$5x - 8 + 8 = 27 + 8$		$5x - 8 + 8 = 27 + 8$
$5x = 45$ (error made here)	$5x = 35$	$5x = 35$
$x = 45/5 = 9$		$x = 35/5 = 7$

Before each participant in the single-error worked example group received their feedback, participants were told that a solution step was given alongside each incorrect line on the space provided on the right of the test paper. They were also told study the correct solution and identify where they had made mistake(s). From this perspective, errors were not directly flagged but could be found by comparing learner solutions with the worked examples. Participants were also required to study the solutions.

For the full worked example group, a full worked solution was positioned alongside (space provided on the right of the test paper) each answer where an error was made. Participants were told that a full worked solution was given alongside their answers that contained any errors on the space provided on the right of the test paper. They were also told to study the correct solution and identify where they have made mistakes.

Each worked example correction for both groups was digitally inserted using Microsoft Word in the space provided to maintain efficiency and clarity. Table 1 provides an example of one error in the calculation (left column) with feedback for the single-step worked examples group (middle column) and the full worked example group (right column).

Procedure

The experiment was conducted in the students' mathematics classes over two lessons with the assistance of the classroom teacher. The first lesson started with the signing of the ethics consent form, followed by an outline of the experiment and relevant instructions given by the researcher. Participants were then given the instructional worksheet of six-paired worked examples to complete (20 min). All participants used the same instructional material. At the end of the instructional phase, all worksheets were collected before participants were given the common pre-test (20 min). After lesson one, participants were randomly assigned to the two conditions, either the single-step worked examples or the full worked examples. A photocopy of the pre-test was then scored, and feedback inserted on the original pre-test according to the conditions of the two feedback methods. In the second lesson (Day 2) students were handed back their pre-tests. At the start of the second lesson, participants were given instructions to study the feedback inserted into their test papers (10 min). Immediately after participants studied their respective feedback, both groups were given the common post-test (20 min). Table 2 provides a summary of the procedures.

Table 2. Summary of experiment procedures in Experiment 1.

Lesson outline	Description	Time duration
Day 1: Lesson 1		
Introduction	Participants were briefed and completed ethics consent form.	10 min
Acquisition stage	Instructional worksheet containing six-paired worked examples were completed.	20 min
Pre-test stage	Similar and transfer test questions answered	20 min
Day 2: Lesson 2		
Introduction	Participants were briefed and instructed on the feedback given.	10 min
Feedback	Participants studied their feedback from the pre-test.	10 min
Post-test	Similar and transfer questions answered	20 min
Debriefing	Words of thanks to the participants from the investigator and class teacher. Students were also allowed to ask questions.	10 min

Table 3. Group means (and SDs) of overall test scores in Experiment 1.

	Pre-test		Post-test	
Type of worked example	Similar	Transfer	Similar	Transfer
Single-step	8.93 (4.30)	0.96 (1.99)	9.29 (5.62)	1.07 (2.04)
Full	9.86 (4.49)	1.36 (1.91)	10.96 (5.01)	1.43 (1.89)
Combined groups	9.39 (4.38)	1.16 (1.94)	10.13 (5.34)	1.25 (1.96)

Results

Scoring of tests

As described above, each problem in the pre- and post-tests were assigned 3 marks, and 1 mark was deducted for every error made. Both similar (maximum score of 18) and transfer problems (maximum score of 12) were marked using the same marking criteria. The means and standard deviations for all test dependent variables are presented in Table 3. The similar test (Questions 1–6) and transfer test (Questions 7–10) were analysed separately to investigate potential differences between the different types of questions. Cronbach alpha tests on retention and transfer scores for the pre-test were 0.82 and 0.54, respectively, and 0.76 and 0.47 for the post-test. Whereas the retention test had a high degree of reliability, the transfer test had low reliability, suggesting a lack of a single transfer construct. It was notable that transfer scores were extremely low barely reaching an overall accuracy rate of 10% on the post-test (see Table 3), due to many non-attempts by the participants. It is highly likely that lack of attempts had a significant impact on reliability, and therefore no further analysis was conducted on the transfer data.

Initial analysis of overall test scores

An initial analysis was conducted on the total test data (Table 3) that included problem answers to test questions that received no feedback. Pre-test scores were used as a covariate to control for potential prior-knowledge differences. For similar test scores, the ANCOVA revealed no significant group differences ($F < 1$, ns.).

Feedback analysis

To examine the impact on the problems that directly received feedback an additional analysis was completed. The following filtering procedure was first conducted.

Table 4. Group means (and SDs) of direct feedback test scores in Experiment 1.

	Pre-test		Post-test	
Type of worked example	Similar	Transfer	Similar	Transfer
Single-step	2.39 (1.89)	0.32 (0.90)	2.46 (2.80)	0.21 (0.79)
Full	2.36 (2.11)	0.18 (0.55)	4.43 (2.62)	0.68 (1.49)
Combined groups	2.38 (1.99)	0.25 (0.75)	3.45 (2.86)	0.45 (1.21)

Questions in the pre-test that were awarded 0 marks in the single-step worked examples group received a full worked example as feedback because each solution step was incorrect and therefore a single error correction was provided for each step, resulting in the same number of steps as a full worked examples correction. Therefore, for scores of 0, there was no difference between this experimental condition and the full worked example group. Similarly, if an answer to a question was totally correct (score of 3), then no differences occurred between conditions, as no feedback was provided at all. Therefore in order to identify real differences between the conditions, all questions that were awarded 0 or 3 marks in both groups were omitted from the data. Furthermore, all marks in the post-test that corresponded to the omitted questions in the pre-test were also omitted from the data. For example, if Question 1 for participant A was omitted from the pre-test data, Question 1 for participant A was also omitted from the post-test data.

Following this filtering that removed all scores of 0 and 3 in the pre-test (and corresponding post-tests) the analysis was conducted on questions that only received scores of 1 and 2 marks in the pre-test. Hence, feedback scores were computed for each participant by adding together their marks for the question(s) that only received feedback minus the described omissions. For example, participant A scored 1 mark in question 1, and 2 marks in question 5 in the pre-test; and then scored 3 marks in question 2 and 3 marks in question 5 in the post-test. The modified pre-test score for participant A was therefore 3 $(1 + 2)$ and the post-test score was 6 $(3 + 3)$. Hence, 3 for the pre-test can be validly compared with 6 for the post-test. Hence, direct improvements according to the feedback conditions were measured.

Mean group scores are shown in Table 4. As in the previous analysis, the pre-test scores were used as a covariate in ANCOVA to control for potential prior-knowledge differences before the feedback was given. The ANCOVA revealed significant group difference for similar test scores, F (1, 53) = 13.47, $MSe = 55.798$, p = .001, partial $\eta^2 = 0.203$, where the full worked example group (adj $M = 4.45$, $SE = 0.385$) scored significantly higher than the single-step worked example group (adj $M = 2.45$, $SE = 0.39$). The large effect size of .203 (see Cohen, 1988) indicates a large difference between the groups.

Differences between the pre- and post-tests for similar problems. The above analysis examined differences between the feedback strategies; whereas the following analysis examined overall (both groups combined) differences between the pre- and post-tests (see Table 4). For the similar test problems, a paired t-test showed the post-test score was significantly greater than the pre-test score $t(55) = 2.55$, $p = .014$, Cohen's $d = 1.1$, indicating that overall, the combined feedback strategies had a positive effect.

Discussion

The results on the similar test problems that received feedback indicated that providing full worked examples was significantly more effective than providing single-step worked examples (partial worked examples) for similar (to acquisition) test problems and, therefore, Hypothesis 1 was supported for these type of problems. The low Cronbach alpha for the transfer test suggested a source of unreliable data, and therefore no conclusions were drawn from this data, other than the participants found the problems very difficult.

Although significant group differences were found when the analysis was restricted to problems where feedback had been provided, no overall significant effects were found when the whole data set was considered. Two deliberate design conditions could have contributed towards this lack of overall effect. Firstly, errors were not directly indicated but had to be found by examining the worked examples. Simply asking participants to study may have led to a focus on locating errors rather than gaining a deeper understanding of the problem solution. Secondly, worked examples during feedback were only studied and not paired in a study-solve format as originally designed by Sweller and Cooper (1985) and have been used successfully since. As a consequence, a number of changes were made in the next experiment, as well as increasing the number of interventions investigated.

Experiment 2

A number of changes were made in this experiment. Firstly, during feedback worked example problem pairs were provided rather than just studying a worked example. For the first group (full worked example), for each incorrect answer participants were required to study a full worked example correction followed by a similar problem to solve (see Sweller & Cooper, 1985). For the second group (single-step worked example), participants were required to study single line error corrections for only the lines where an error was made (identical to Experiment 1), and then complete single-line solutions to similar problems.

Secondly, a third group was introduced in the form of a complete worked example. The results from the first experiment indicated that under the given conditions a single-step worked example was inferior to a full worked example. If, as argued above, partial solutions may lack continuity and fail to show all the connections needed to understand and learn about a problem, a single-step solution is the most radical form of a partial worked example. In contrast, completions problems have more continuity as they involve an ordered sequence of solution steps (Van Merriënboer & Krammer, 1987). Hence, a completion worked example group was introduced, where participants were shown a completed worked example correction, which started from the first line an error was identified and ended with the last line of the solution. Participants were then required to solve a similar problem starting from the same point where the first error was identified through to final completion.

Thirdly, the first experiment did not indicate where errors were made but relied on participants to locate them by referring to the worked examples provided. Because some research suggests that indicating errors (corrective feedback) is an important

component of effective feedback (Marzano et al., 2001), more effective feedback may be provided by the instructor/teacher directly indicating errors. Hence, for these three groups, errors were directly flagged.

Fourthly, a fourth group (no error indicated) was included that took the form of a full worked example problem pair, identical to group 1, but any errors made by the participants were not directly flagged. Participants could deduce that an error had been made because a worked example was provided, but the exact error(s) could only be located if their answer was checked against the correct solution given in the worked example. A broader examination of the first hypothesis was conducted by comparing the four groups. Furthermore, by comparing the two full worked example groups alone, Hypothesis 2 could be tested as flagging errors or not was the only variable manipulated.

Finally, the number of problems tested in the transfer set was reduced to enable participants more time to complete the problems set.

Method

Participants

Seventy-four grade 8 (median age of 13) students (38 boys, 36 girls) from high schools in New South Wales, Australia, initially participated in this experiment. The same inclusion criteria applied in the first previous experiment was applied in this experiment, hence, 25 participants were excluded from this experiment because they either did not participate in both lessons and/or scored above 80% in the pre-test. Participants were randomly assigned to the four groups after the first pre-test phase accordingly: full worked example ($N = 12$), completion worked example ($N = 14$); single-step worked example ($N = 11$), and no-error indicated ($N = 12$). All students had recently gained some prior learning experience with the content material covered in this experiment from their normal mathematics classes.

Design

Similar to the first experiment, there were 4 stages. In the first stage (Acquisition), all participants received instruction on the given topic through a set of paired worked examples. During the second stage, participants were tested (pre-test) on the topic studied during the acquisition phase. For the third stage, participants received feedback according to the group they were assigned to and given more study time. In the final fourth stage, they were re-tested (post-test) on the topic.

Learning topic

The experimental materials were based on the same topic covered in the previous experiment (solving linear equations).

Acquisition stage

The instructional material consisted of the same problems used in the previous experiment and followed the same-paired format for the worked examples as previously outlined.

Test materials

To enable more time for providing test answers, the first test (pre-test) was reduced to eight problems (six similar and two transfer) identical to the first eight problems used in Experiment 1. Data provided from the first experiment indicated that most students had no time to attempt the last two transfer problems, which contributed towards a low Cronbach alpha score. Hence, eliminating these problems was expected to enhance the reliability of the transfer test, as well as reduce the overall time demands of the tests. The post-test presented after feedback consisted of the first eight problems of the previous post-test.

Feedback phase

Consistent with the first experiment, the pre-test for each participant was photocopied and scored and kept by the researcher, while the original pre-test paper included the feedback and was returned to the participants. On three of the conditions (full worked example, single-step worked example, and completion worked example) any error made was flagged by positioning a red cross (X) at the end of the line of the solution step where errors were made. In contrast, errors for the no-error indicated group were not flagged at all.

For each incorrect answer to a problem, participants received a worked example problem pair consisting of a solution to the problem followed by a similar problem (the same for all 4 groups) to be solved, structured according to the group condition. For both the full-worked example group and the no-indicator group, a full worked example was given with a similar problem to solve (see Table 1). For the completion worked example group, participants were required to study a completion worked example correction, which started from the first line an error was identified and ended with the last line of the solution. In the example, in Table 1 the completion worked example would consist of the following lines ($5x = 35$, $x = 35/5$, $x = 7$). Participants were then required to solve a similar problem (e.g. solve $7x = 28$) starting from the same point where the first error was identified. For the single-step worked example group, participants were required to study a single-step solution to the error made (see Table 1), and then complete a single-step solution to a similar problem.

Procedure

The procedure and study/test times were exactly the same as in Experiment 1.

Results

Scoring of tests

Scoring of tests was exactly the same as in Experiment 1 with maximum scores of 18 (similar test) and 6 (transfer test). Cronbach alpha scores for the similar pre- and post-tests were 0.89 and 0.88 respectively, providing a satisfactory level of reliability. Because the transfer test consisted of only two items (problems to be solved), Cronbach alphas were not completed (see Eisinga, Te Grotenhuis, & Pelzer, 2013), instead, Person correlations were calculated. A weak correlation of 0.33 was found on

Table 5. Group means (and SD) of overall test scores in Experiment 2.

	Pre-test		Post-test	
Type of worked example	Similar	Transfer	Similar	Transfer
Full	6.07 (3.32)	0.54 (0.84)	11.21 (6.66)	1.79 (2.33)
Completion	10.20 (4.31)	0.57 (0.53)	12.53 (7.25)	1.93 (1.87)
Single-step	7.55 (2.91)	0.36 (0.39)	9.73 (8.13)	1.46 (1.70)
No error indicated	7.54 (4.14)	0.54 (0.56)	13.08 (7.17)	1.69 (1.97)
Combined groups	7.91 (3.97)	0.51 (0.60)	11.74 (7.17)	1.74 (1.94)

Table 6. Group means (and SDs) of direct feedback test scores in Experiment 2.

	Pre-test		Post-test	
Type of worked example	Similar	Transfer	Similar	Transfer
Full	1.15 (0.64)	0.54 (0.69)	4.14 (4.42)	0.65 (1.65)
Completion	1.48 (0.56)	0.90 (0.85)	3.00 (2.59)	1.20 (1.47)
Single-step	1.44 (0.42)	0.73 (0.79)	3.00 (3.46)	0.64 (0.81)
No error indicated	1.36 (0.53)	0.85 (0.78)	6.39 (3.75)	1.08 (1.55)
Combined groups	1.36 (0.55)	0.76 (0.77)	4.13 (3.76)	0.91 (1.42)

the transfer pre-test and a stronger correlation of 0.67 on the post-test. Hence, only a reasonable degree of reliability for a transfer construct was found on the post-test.

Initial analysis of overall test scores

The means and standard deviations for all test dependent variables are presented in Table 5. The post-test data reported in Table 5, which includes total test scores, were analysed using ANCOVA where pre-test scores were used as the covariate to control for potential prior-knowledge differences. For both similar and transfer test scores, the ANCOVA revealed no significant group differences (both $F < 1$, ns.).

Feedback analysis

The feedback analysis was conducted for problems in the post-test that corresponded to pre-test questions that received feedback with scores of 1 mark or 2 marks as described in Experiment 1 (see Table 6). Pre-test scores were used as a covariate in the ANCOVAs.

The ANCOVA revealed significant group differences for similar test scores, F (3, 44) $= 4.70$, $MSe = 46.03$, $p = .006$, partial $\eta^2 = 0.243$. This large effect size (see Cohen, 1988) indicates a large difference between the groups. Post-hoc tests with Bonferroni corrections revealed that participants in the full worked example group (adj $M = 5.31$, $SE = 0.19$) group and the no error indicated group (adj $M = 6.87$, $SE = 0.90$) scored significantly higher than participants in the completion worked example group (adj $M = 2.76$, $SE = 0.85$) and the single-step worked example group (adj $M = 3.10$, $SE = 0.94$). There were no other significant differences between groups.

There were no significant group differences on the transfer questions ($F < 1$, ns.).

Differences between pre- and post-tests. A paired t-test on the combined data for all 4 groups (see Table 6) showed a significant improvement between the pre- and post-tests, $t(48) = 3.37$, $p = .001$, $d = 1.6$. As can be seen from Table 6 each feedback

strategy generated improvements; whereas for the transfer set no significant overall improvements were found ($t < 1$, ns.).

Discussion

This experiment tested two hypotheses. The feedback analysis, which examined only questions that received correctional feedback, indicated that the groups with full worked examples (full worked example group and the no-error indicated group) made significantly higher improvements than the completion worked example and single-step worked example groups for similar problems. Hence support was found for Hypothesis 1.Participants scored higher after receiving full worked examples than receiving completion worked examples or single-step worked examples.

The second hypothesis predicted that directly flagging errors would lead to higher learning outcomes than not explicitly flagging errors. This was tested by comparing the two full worked example conditions only (full worked example and no error indicated groups), as these had identical designs apart from errors being shown or not. It should be noted that the other two groups were not included because a controlled comparison would not be possible due to a lack of varying design features. No significant difference was found between the two full worked example groups suggesting that it did not matter whether errors were directly indicated or not, when using full worked examples. Hence Hypothesis 2 was not supported. No evidence was available to confirm that participants whose errors were not flagged (no error indicated group) actually pinpointed specific errors using the worked examples. However, it is feasible that participants may have just studied the worked examples rather than specifically locating their original errors. Nevertheless, as the no-error-indicated group outperformed the two completion worked examples, where errors were flagged, locating actual errors, may not be that important if full worked examples are provided.

General discussion

The study investigated the sequencing of two sets of worked examples with individual feedback. One manipulation of conditions compared full worked examples with partial worked examples. In Experiment 1, full worked examples were compared with single-step worked examples, and in Experiment 2, compared with single-step worked examples and completion worked examples. In both experiments, full worked examples led to significant improvements compared to partial worked examples on problems where feedback was provided. Improvements were found when errors were flagged (Experiment 2) or not flagged (Experiments 1 and 2), and when worked examples were studied only (Experiment 1) or presented in pairs, where the paired problem had to be solved (Experiment 2). These results suggest that for these learners, full worked examples were significantly helpful and not redundant. Clearly, these learners had not transitioned far towards expertise and therefore did not experience expertise reversal effects caused by processing redundant information (Kalyuga et al., 2003). Mastery levels of the mathematics topic were sufficiently low to benefit from a second round of full worked examples following individual feedback.

Consistent with research into novice learning strategies (Kirschner et al., 2006), full guidance in the form of fully explained worked examples was most effective. For these learners, partial worked examples did not provide sufficient direction, feasibly because they lack overall coherence, failing to show important connections between the various solution steps. As pointed out by Hattie and Timperley (2007) feedback needs to match the student's prior knowledge in order to be most effective. In this case, these fairly novice learners needed full guidance during feedback and further study time.

Evidence was also found that locating one's own errors was neither an advantage nor disadvantage when full worked examples were used. Findings from other research suggest that identifying errors and misconceptions are important but needs to be accompanied by more elaborated information (Marzano et al., 2001). In this study when errors were not flagged, full worked examples were provided and as a consequence further elaboration was possible. If learners did identify the errors themselves deeper processing and more active learning may have occurred.

Overall, the results on problems that received feedback were found only on similar (to acquisition) problems. In the first experiment transfer data was not analysed due to the poor reliability of the data as a transfer construct. In the second experiment, greater reliability was found but no difference was found between the strategy groups. In both cases, transfer scores were very low, which not only impacted on reliability but also suggested that the problems were too difficult for the majority of students in the study. Accordingly, these problems are more consistent with far-transfer rather than near-transfer problems.

Previous research has shown that transfer knowledge is difficult to achieve. Perkins and Salomon (1989) argued that for transfer to occur, learners must have a well automated skill that can be applied to a similar situation or abstracted a principle that can be applied to a new situation. It is therefore likely that the practice provided in our study was insufficient to enable automation to occur to transfer to more difficult problems (see Cooper & Sweller, 1987). Furthermore, to achieve significant transfer, other strategies such as variability of worked examples (Paas & van Merriënboer, 1994), specific strategies that flag transfer links (see Bassok, 1990), or the use of directed general problem-solving strategies (see Youssef-Shalala, Ayres, Schubert, & Sweller, 2014), are often required. Hence, it can be concluded that much more practice and intervention was required to produce meaningful transfer effects in this domain with the given learners.

The positive effect on similar problems suggests that full worked examples are a good feedback strategy for learners with this level of mathematical knowledge, but may only extend to problems similar to the initial acquisition problems. It is known that as expertise increases, worked examples during acquisition need to be modified. A fading strategy (Salden et al., 2010) is one such method. Finally, as domain-specific knowledge increases significantly, problem-solving without any guidance can be used (Schwonke et al., 2009). Such methods avoid redundant strategies than can lead to the expertise reversal effect (Kalyuga et al., 2003). However, these strategies have been completed during initial acquisition and it is uncertain what the best-worked examples feedback (post initial acquisition) strategies are when expertise increases. For example, if learners are using fading examples during

acquisition, should fading examples be used during feedback as well? Should there always be a match between the type of worked examples used in acquisition and feedback? Clearly, future research is required to answer these important questions.

The present study had a number of unique features. As described above, some studies into worked examples have assessed student learning and then tailored instruction accordingly. Rapid assessment techniques and intelligent tutors have been used to assess learners' knowledge and then apply the most appropriate learning tasks. But such processes are system controlled and do not necessarily provide direct or delayed feedback. They may, therefore, lack opportunities to engage in important feedback processes and reflections. The present study provided such opportunities, as well as giving the learner more responsibility for locating their errors. It also provided feedback on the next day (delayed feedback), more consistent with everyday mathematics classrooms (non-computerised). This feedback also matched some best-practice principles because it was timely (Kulik & Kulik, 1988; Mory, 2004), conducted after a test (Shute, 2008), and provided opportunities for improvement (Orsmond & Merry, 2011).

Three limitations are noted that provide possible directions for future research. Firstly, the study time when feedback was provided was rather brief in both experiments. Clearly, learners could have benefitted from more time to reflect on their errors and to study the worked examples. Nevertheless, the time available was sufficient to generate significant differences between the conditions. However, it is unknown whether more feedback time would have made further differences or not. So future research could investigate the impact of feedback time. Secondly, the far transfer problems were too difficult for the given students in the sample. To investigate transfer effects within the given design, near-transfer problems could have provided more insights as well as greater practice time. Thirdly, although it was beyond the scope of the present study, a number of other possible boundary conditions could be investigated. For example, instead of targeting problems where errors were made, the original full set of worked examples used during acquisition, or a similar set could be provided, especially if more feedback time was available. A control group could also have been included in order to compare the feedback conditions when no feedback was given. Also, a problem-solving group could have been included as another control condition, where no worked examples are provided. Hence, future research could include more feedback time and different boundary conditions, as well as a greater range of learners' prior knowledge.

The results of this study have some clear educational implications for classroom practice. The findings in Experiment 2 showed that timely and focused feedback led to gains in performance, highlighting the importance of feedback itself. Consistent with a whole body of research (see Mory, 2004) feedback improves learning.

In particular, whole worked examples can be used as an effective form of feedback, especially for learners encountering the topics in the early stages of knowledge acquisition. However, until more research is completed, some caution must be shown in generalising this result to more knowledgeable learners.

When using whole worked examples it may not be necessary to specifically point out errors. Learners can match the worked examples with their own attempts, and accordingly, use the worked examples for further learning.

In conclusion, the results of this study suggest that providing worked examples as part of the feedback cycle can lead to learning improvements, particularly for learners with low levels of prior knowledge. Feedback is widely accepted to be a powerful learning tool (Hattie & Timperley, 2007) and worked examples provide a form of essential elaboration (Shute, 2008). For the learners of this study, full worked examples led to higher learning outcomes than partial worked examples.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*, *70*(2), 181–214. doi:10.2307/1170661

Ayres, P., & Sweller, J. (2013). The worked example effect. In J. A. C. Hattie & E. M. Anderman (Eds.). *International Guide to Student Achievement* (pp. 408–410). Oxford, UK: Routledge.

Bassok, M. (1990). Transfer of domain-specific problem solving procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(3), 522–533. doi:10.1037//0278-7393.16.3.522

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cooper, G., & Sweller, J. (1987). The effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology*, *79*(4), 347–362. doi:10.1037//0022-0663.79.4.347

Eisinga, R., Te Grotenhuis, M., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach or Spearman-Brown? *International Journal of Public Health*, *58*(4), 637–642. doi:10.1007/s00038-012-0416-3

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. doi:10.3102/003465430298487

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, *38*(1), 23–31. doi:10.1207/S15326985EP3801_4

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75–86. doi:10.1207/s15326985ep4102_1

Kulik, J. A., & Kulik, C. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, *58*(1), 79–97. doi:10.2307/1170349

Marzano, R., Pickering, D., & Pollock, J. (2001). *Classroom instruction that works. Research-based strategies for increasing student achievement*. Alexandria, VA: Association for Supervision and Curriculum Development.

Mory, E. H. (2004). Feedback research review. In D. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 745–783). Mahwah, NJ: Lawrence Erlbaum.

Orsmond, P., & Merry, S. (2011). Feedback alignment: Effective and ineffective links between tutors' and students' understanding of coursework feedback. *Assessment & Evaluation in Higher Education*, *36*(2), 125–136. doi:10.1080/02602930903201651

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context-bound? *Educational Researcher*, *18*(1), 16–25. doi:10.2307/1176006

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, *84*(4), 429–434. doi:10.1037/0022-0663.84.4.429

Paas, F. G. W. C., & van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology*, *86*(1), 122–133. doi:10.1037//0022-0663.86.1.122

Salden, R. J. C. M., Aleven, V., Schwonke, R., & Renkl, A. (2010). The expertise reversal effect and worked examples in tutored problem solving. *Instructional Science*, *38*(3), 289–307. doi:10.1007/s11251-009-9107-8

Schwonke, R., Renkl, A., Krieg, C., Wittwer, J., Aleven, V., & Salden, R. (2009). The worked-example effect: Not an artefact of lousy control conditions. *Computers in Human Behavior*, *25*(2), 258–266. doi:10.1016/j.chb.2008.12.011

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. doi:10.3102/0034654307313795

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. New York, NY: Springer.

Sweller, J., & Chandler, P. (1991). Evidence for cognitive load theory. *Cognition and Instruction*, *8*(4), 351–362. doi:10.1207/s1532690xci0804_5

Sweller, J., & Cooper, G. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, *2*(1), 59–89. doi:10.1207/s1532690xci0201_3

Van Beuningen, C. G., de Jong, N. H., & Kuiken, F. (2012). Evidence on the effectiveness of comprehensive error correction in second language writing. *Language Learning*, *62*(1), 1–41. doi:10.1111/j.1467-9922.2011.00674.x

Van Gog, T., Kester, L., & Paas, F. (2011). Effects of worked examples, example-problem, and problem-example pairs on novices' learning. *Contemporary Educational Psychology*, *36*(3), 212–218. doi:10.1016/j.cedpsych.2010.10.004

Van Merriënboer, J. J. G. (1990). Strategies for programming instruction in high school: Program completion vs. program generation. *Journal of Educational Computing Research*, *6*(3), 265–285. doi:10.2190/4NK5-17L7-TWQV-1EHL

Van Merriënboer, J. J. G., & Krammer, H. P. (1987). Instructional strategies and tactics for the design of introductory computer programming courses in high school. *Instructional Science*, *16*(3), 251–285. doi:10.1007/BF00120253

Youssef-Shalala, A., Ayres, P., Schubert, C., & Sweller, J. (2014). Using a general-problem solving strategy to promote transfer. *Journal of Experimental Psychology: Applied*, *20*(3), 215–231. doi:10.1037/xap0000021