International Journal of Instruction *April 2023 ● Vol.16, No.2*
e-ISSN: 1308-1470 ● www.e-iji.net *p-ISSN: 1694-609X*
pp. 229-246

Article submission code: 20220325121502
Received: 25/03/2022 Revision: 27/09/2022
Accepted: 22/10/2022 OnlineFirst: 07/01/2023

Effects of Worked Example on Students' Learning Outcomes in Complex Algebraic Problems

Saidat Morenike Adeniji
University of New England, Australia & University of Ilorin, Nigeria, *sadeniji@myune.edu.au* & *salaudeen.sm@unilorin.edu.ng*

Penelope Baker
University of New England, Australia, *pep.baker@une.edu.au*

High school students have been reported to have difficulties solving complex algebraic problems. This study therefore investigated the effects of worked example instruction on students' learning outcomes in solving complex algebraic problems. The study was a quasi-experiment that involved a pre-test, an intervention, a post-test, and a delay test. The responses of 72 students (aged 14 to 15 years) were scored following the structure of the observed learning outcomes (SOLO) model and analysed using the Rasch model and regression analysis. The results indicated a significant effect of worked examples from the pre-test to the post-test; however, this effect was not completely retained at the delay test. Also, worked examples had a larger effect on the low-ability students than the high-ability students, but student gender neither influenced nor interacted with learning outcomes at the post-test and delay test. Lastly, the results revealed an interaction between the worked example effects and students' expertise level, with the high-ability students experiencing a full reversal of the worked example effect. These results are explained with respect to element interactivity and expertise reversal effects, and inform mathematics educators and teachers of the conditions of the worked example effect and the implications for classroom practices.

Keywords: mathematics learning, cognitive load theory, worked example effect, expertise reversal effect, element interactivity, complex algebra

INTRODUCTION

Like other mathematical topics, proficiency in algebra and particularly solving equations is demonstrated through procedural and conceptual knowledge. While procedural knowledge emphasizes the step-by-step algorithm towards a solution to a task, conceptual knowledge addresses the principles or rich connections among mathematical elements (Hurrell, 2021). Solving equations remains a fundamental gatekeeper to higher level mathematics and has been reported to develop students' critical thinking (Martinez et al., 2016). It is more likely that students who are successful in solving algebraic

Citation: Adeniji, S. M., & Baker, P. (2023). Effects of worked example on students' learning outcomes in complex algebraic problems. *International Journal of Instruction*, *16*(2), 229-246. https://doi.org/10.29333/iji.2023.16214a

problems would cope with higher level mathematics than the reverse. However, despite the different instructional designs that have been identified globally to improve students' mathematics learning, reports continue to indicate that students struggle to solve complex equations using rich procedural and conceptual knowledge (Johari & Shahrill, 2020; Kolawole & Ojo, 2019; Omobude, 2014). Specifically, students have been reported to have difficulties understanding algebraic expressions and equations, conceptualizing the dynamic role of variables, manipulating constants and coefficients to isolate unknowns, and applying algebraic understanding to solve contextualized problems (Johari & Shahrill, 2020; Martinez et al., 2016). Hence, there is a need to focus more on how to help students to develop a deeper understanding of solving complex equations, given its central role in the mathematics curriculum.

Pedagogical practices employed by teachers largely influence students' learning of mathematical ideas (Ayebale et al., 2020; Mosimege & Winnaar, 2021; Moussa-Inaty et al., 2019). The pedagogy of interest in this article is the worked example instructional design, which is in use in many classroom classes around the world. It is expected that the results from this study would inform research and classroom practices. The worked example, which originated from Sweller's (1998) cognitive load theory, is conducted in three forms: problem-example pairs, example-problem pairs, and worked example only. Available empirical studies have established that the example-problem pairs, also known as alternate strategy, is superior to others when solving simple equations (Van Gog et al., 2011). However, Paas and van Merrienboer (2020) recommend the utilization of example-problem pairs to facilitate the understanding of complex mathematical tasks. This study builds on the existing findings from studies on worked example instruction to examine the effectiveness of worked example instruction (example-problem pairs) on how students solve complex algebraic equations. An example of complex algebraic equations is solving simultaneous equations, which is the example used in this study. There are limited studies on simultaneous equations, and several studies report that the majority of students have difficulty in learning to solve simultaneous equations (Johari & Shahrill, 2020; Kolawole & Ojo, 2019; Omobude, 2014). Specifically, while students may have acquired competence in solving linear equations, they tend to have difficulty understanding the relationship between two simultaneous equations. Hence, teaching students how to solve simultaneous equations requires appropriate student-centered pedagogical practices that can facilitate schema construction and the transfer of mathematical knowledge to solving other related problems.

Managing cognitive resources while solving complex algebraic equations is essential because of the presence of highly interacting elements. Previous studies on example-problem pairs have reported its instructional effectiveness for learning simple equations (one-step, two-step and three-step equations) (Alreshidi, 2021; Chen et al., 2019; Ngu & Phan, 2017). However, this study focused on solving complex equations with a minimum of six steps and numerous operational and relational lines. Also, none of the previous studies on worked example instruction evaluated the quality of students' responses as prescribed by the structure of the observed learning outcome (SOLO) model. Rather, in previous studies, a post-interventional quantitative response assessment was utilized for analysis (Chen, 2019; Van Gog et al., 2011). Moreover,

while most of the previous studies focused mainly on the statistical analysis employed to describe the effectiveness of the intervention, none considered the quality of the measurement. Therefore, a genuine non-equal interval measurement model proposed by an item response theory, the Rasch model (Linacre, 2013), was employed in this study.

Furthermore, despite government policies in every nation around the globe that aim to bridge the gender gap in education, the reality is that gender equality is yet to be achieved in many countries, especially Africa (El Yacoubi, 2015; Leder, 2015); hence an understanding of the influence of gender on students' learning outcomes remains inconclusive among mathematics educators. While some studies report no significant difference in the mathematical performance of male and female students (Ajai & Imoko, 2015; Arnup et al., 2013), others indicate that the achievement gap between male and female students accounts for the gender disparity in STEM enrolments in higher institution (Anaya et al., 2022; OECD, 2019). Factors that indicate the inequality of mathematics learning outcomes based on gender are public perceptions that males are better at mathematics and mathematics-related subjects than females, socio-cultural beliefs, females' low confidence, the mathematical achievements of males and females in internationally recognized examinations, and the low participation of females in higher-level science and technology (Leder, 2015). Continuing research is therefore needed on gender differences in mathematics.

Importantly, the teaching and learning process is not complete until students can demonstrate learning, retain information and recall the information for subsequent use (Adeniji et al., 2018). Therefore, given the importance of minimizing cognitive decay and transferring mathematical ideas to solve problems in other fields, this study examines the lasting effects of the example-problem pairs instruction on students' learning outcomes.

Literature Review

The section provides a brief review of literature on cognitive load theory, which is the theoretical framework for the worked example instruction (Sweller 1998), worked example effect, element interactivity and expertise reversal effect.

Cognitive Load Theory

The cognitive load theory (CLT) is based on the evolution of human knowledge (primary and secondary knowledge) and an understanding of the human cognitive architecture, which is characterized by limited working memory and limitless long-term memory (Sweller, 1998). The working memory processes information and stores the processed information in the long-term memory in the form of a schema. However, the limitation of the working memory becomes obvious when it processes more than four to five pieces of new information at a time (Sweller, 2011). If such new information is not rehearsed within 20 seconds, the information may be lost due to the limited resources available in the working memory (Sweller, 2011). CLT contends that for effective learning to take place, the resources required to process a piece of information must not exceed the capacity of the working memory, otherwise it may lead to unnecessary

cognitive load, which hinders learning (Sweller et al., 2019). Cognitive load is the burden imposed on the working memory during learning. Initially, three forms of cognitive load (extraneous, intrinsic, and germane) were identified to hinder effective learning; however, extensive research findings have proven that the germane load benefits learning while the intrinsic and extraneous loads prevent learning. Therefore, the total cognitive load is the addition of the extraneous load caused by poor instructional design and the intrinsic load derived from the complexity of the learning material (Sweller et al., 2019). Perhaps, the higher the total cognitive load imposed during learning, the fewer the learning outcomes, and vice-versa (Sweller, 1998). While changing the complexity of the mathematical content could alter the intrinsic cognitive load, the extraneous cognitive load can be eliminated by designing effective pedagogies (Sweller, 2019). CLT therefore focuses on ways of managing the cognitive resources during learning and prescribes instructional designs that could effectively lead to schema construction. One of these pedagogies is worked example instruction, which this article focuses on.

Worked Example Effects

The worked example instructional design has been prescribed by CLT as an effective pedagogy that could eradicate or reduce cognitive load during learning. As a way of reducing the cognitive load caused by an inappropriate instructional procedure, worked example instruction guides students' learning to relevant activities that could facilitate the construction of schema and enhance effective learning (Renkl, 2017). Therefore, worked example instruction imposes a relatively low cognitive load on students. The worked example instructional design involves the use of similar and structurally identical solution procedures for solving a problem. For example, worked examples are presented to students to study, and each studied worked example is paired with a similarly structured practice problem. It is expected that students study the worked example and transfer their understanding to solving the similar problem attached to the worked example (Van Gog et al., 2011). The theoretical principle underlying this instructional design within the CLT is the borrowing and reorganizing principle (Chen et al., 2019). This principle contends that human beings borrow information (by imitating, listening, studying, etc.), assimilate the borrowed information, reorganize the information using prior knowledge in their long-term memory, and then store it distinctly for future use.

As mentioned, there is some evidence that supports the effectiveness of the worked example instructional design (Alreshidi, 2021; Barbieri et al., 2021; Chen et al., 2019; Renkl, 2017; Smith et al., 2022). Recent research on this instruction has delved into which format of the worked example is the best of the three available: example-problem pairs, problem-example pairs, and worked example only. The available empirical and theoretical evidence reports the superiority of example-problem pairs over the others (Alreshidi, 2021; Sweller, 2019; Van Gog et al., 2011). Again, cognitive researchers have examined the effectiveness of partially worked example against complete worked examples (with full instructional guidance). The studies have established that providing complete instructional guidance is more beneficial than partial instructional guidance

(Richey & Nokes-Malach, 2013; Sweller, 2011). Therefore, this study examines the short-term and long-term effect of example-problem pairs (with full instructional explanation) on students' learning outcomes in solving simultaneous equations.

Element Interactivity

Element interactivity is the relationship that exists between elements in a learning material. Elements could be numbers, concepts, symbols or procedures. The cognitive load imposed during learning is determined by the rate of element interactivity in the learning material (Chen et al., 2015). Therefore, from the CLT perspective, element interactivity influences effective learning: a low level of element interactivity is where the individual elements of a learning material can be learned in isolation without referring to the other elements (Chen et al., 2015) while a high level of element interactivity is when a material cannot be learned without referring to other elements. A high level of element interactivity may impose a high cognitive load on the students' learning. Solving simultaneous equations have highly interacting elements and they are therefore expected to impose a high cognitive load. For example, solving $2x - y = 3$ and $x + y = 3$ simultaneously requires the students to process the individual elements (such as '2' as a coefficient of x; 'y' and 'x' as variables; '3' as a constant, and '-' and '+' as operators), the relationship between the elements of each equation (i.e., the left side equals the right side) and the relationship between the two simultaneous equations (i.e., the variables x and y in the two equations are equal). This requirement imposes a high intrinsic cognitive load on the students' learning; however, this type of intrinsic cognitive load can be reduced by sequencing, which means breaking the learning materials down into bits such that at the initial stage, fewer elements are interacting. The students can then carefully progress to the full element interactivity material (Chen et al., 2017). Therefore, in this study, the teaching activities were sequenced into initial and subsequent lessons. Sequencing is beneficial because it emphasizes building on learners' prior knowledge (Kalyuga & Renkl, 2010; Sweller, 2011), which CLT refers to as the expertise reversal effect.

Expertise Reversal Effect

Schema is a cognitive structure that represents human knowledge about something (Sweller, 1998). A high-ability student has a schema related to the subject domain (Kalyuga & Renkl, 2010), while a low-ability student lacks schema connected to the task at hand, and they therefore require effective instructional guidance to build schemata. The presence or absence of a schema impacts on the effectiveness of the worked examples and may determine students' learning outcomes. The presence of a schema means that the same materials can represent high element interactivity for a novice or low-ability student but represent a single element for high-ability or experienced students (Chen et al., 2017; Sweller et al., 2019). The differential requirements of the low-ability and high-ability students form the basis of measuring the expertise reversal effect. The expertise reversal effect occurs when an effective instruction becomes less effective as students gain expertise in the domain (Kalyuga & Renkl, 2010; Sweller, 2011), which means the effectiveness of instruction may not be continuous, and students' expertise levels will influence the instruction's effectiveness.

Therefore, this study examined the influence of students' expertise levels on the main effect of the example-problem pairs.

The specific research questions are:

1. What effects do worked examples have on students learning outcomes in complex algebraic problems?
2. Do students expertise levels influence their learning outcomes?
3. What differential learning outcomes are observed based on students' gender?
4. Does students' gender or expertise level interact with the effectiveness of the worked examples?

METHOD

The study reports the effects of the worked example instructional design on students' leaning outcomes in complex equations. This investigation followed a one-group, within-subject, quasi-experimental design that involved a pre-test, a post-test and a delay test. It focused on measuring students' learning outcomes at the three time points to infer causal influences, knowledge retention and the conditions facilitating the learning.

Sample

The samples for the experiment were first-year senior secondary school students (aged 14 to 15 years) in Nigeria. A multi-stage sampling technique was used to select 72 students and their regular mathematics teacher. In Nigeria, schools are either private or public and single-gender or mixed. Participants were drawn from a public school because there are four times more public schools' students than private school students (Nigerian Federal Ministry of Education, 2017). Also, mixed schools are appropriate for studying the impact of gender. In all, students from a school that has had 10 years' continuous participation in external examinations and had a mathematics teacher with at least five years' teaching experience were recruited. There was a relatively equal number of males (N = 38) and females (N = 34). There were 29 low-ability students and 43 high-ability students identified in the pre-test measures (in logits). A median value of 0.11 was selected for categorization, with students who obtained a score below 0.11 being categorized as low-ability students and students who obtained a score of 0.11 and above being categorized as high-ability students. Since English is the official language in schools, this study was conducted in English. The students have similar cultural backgrounds and training in relation to learning mathematics. Following the Nigerian Ministry of Education mathematics syllabus, participating students had basic linear algebra skills at the time of the study (i.e., $y + 3y = 4y$), which served as a pre-requisite for learning how to solve simultaneous equations. Hence, it was assumed that all students had received equivalent mathematics training before the study was conducted.

Research Instrument

The pre-test, post-test and delay test questions were similar and consisted of nine open-ended questions: five targeted the procedural knowledge of solving simultaneous

equations and four focused on conceptual knowledge. The research instruments were validated by a panel of experts that comprised a professor of mathematics education, a professor of psychology education, two senior lecturers in mathematics education, two mathematics teachers and the ethics committee of the authors' affiliated university. All comments and suggestions from the panel relating to the instruments were incorporated before final approval. The methods for solving simultaneous equations tested in this study were substitution, elimination, and graphical and word problem methods. Students were required to provide step-by-step answers to the questions and explain each procedural step. Thus, students are required to follow a sequence of actions to achieve the final answer and also demonstrate rich connections between discrete pieces of elements in the equations.

Procedure for Data Collection

The procedure for data collection was in line with the standard ethics for research, and was approved by the Human Research Ethics Committee of the University of New England, Australia, approval number HE20-224. Following recruitment, a pre-test was administered to the students by their regular mathematics teacher. The data were collected seven months after the COVID-19 pandemic lockdown and there was a need to reduce the intrinsic cognitive load for solving complex equations. Hence, the worked example intervention was sequenced into initial and main lessons. In the initial lesson, the teacher revised solving linear equations, and the students were then exposed to the main lesson through the instruction and acquisition sheets. The instruction sheet explained what simultaneous equations are and the different methods for solving them, and also provided a worked example (with explanations) for each method for solving simultaneous equations. Students were required to study the instruction sheet and were free to request further explanation from the teacher. Then, the students were presented with an acquisition sheet that contained eight worked examples. Each of the worked examples was paired with a similar practice problem, and the students were required to carefully study the worked examples, understand them, and then transfer their understanding to solving the paired problem. For example, the solution procedure for $x = 2 + y$ and $x + y = 8$ was provided for the students to study and then they needed to transfer their understanding to solve a similarly structured problem ($x = 1 + y$ and $x + y = 7$). It was expected that the continuous practice of the paired problems would facilitate the construction of schema that would help students to solve the questions in the post-test at the initial stage and the delay test at a later stage. The post-test was administered to the students immediately following the acquisition phase in order to measure the effectiveness of the instructional intervention. To ascertain whether the instructional effects observed at the post-test were retained for a longer period, a delay test was administered three weeks after the post-test. The three-week span was as recommended by Cohen et al. (2018). Any delay test effects could be attributed to both the intervention, the revision effect, and students' retention abilities.

Data Analysis

Students' responses to the three tests were scored by employing the structure of the observed learning outcomes (SOLO) model (Biggs & Collis, 2014). The SOLO model

provides a rubric for categorising students' responses into increasing levels of reasoning by considering the quantity and quality of the responses to each task. The model consists of five modes of functioning and five levels of responses in each mode. The responses from the tests carried out in this study fell within the concrete symbolic mode (reflecting declarative knowledge of symbol system in the empirical world) and formal mode (corresponding to the demonstration of abstract concepts) only. This study identified five levels of response to the conceptual questions (prestructural, unistructural, multistructural, formal mode 1 and formal mode 2) and four levels of response to the procedural questions (prestructural, unistructural, multistructural, and relational). The scoring was carried out such that the prestructural = 0, unistructural = 1, multistructural = 2, relational = 3, formal mode 1 = 4, and formal mode 2 = 5. An intra-rater assessment of the scoring process yielded a reliability index of 0.93, which indicated that the scoring of the responses was consistent.

Students' scores were imported into Rasch Winstep software for analysis. This was used to generate initial estimates of the students' learning outcomes. This is important because unlike other statistical analyses, the Rasch model does not assume an equal interval among test items. Specifically, Rasch analysis is significant for providing estimates of students' abilities (in logits), item difficulties, and measures of model fit. Table 1 provides statistical estimates of the reliability and how well the data fit the Rasch model. The fitness of the model was determined using infit and outfit measures. The ideal values for the productive measurement range were between 0.5 and 1.5 (Linacre, 2013). Almost all of the mean square fit statistics (infit and outfit) in Table 1 were within this range, except for the outfit of Test 1, which may be a result of random responses from low-performing students. Therefore, these data could be claimed to fit the Rasch model. Moreover, the high item separation indices (>3) indicated that the sample size was large enough to establish a reproducible item difficulties hierarchy of the instrument. Additionally, the students' reliability of more than 0.5 means the existence of more than one ability level; therefore, students were grouped into two ability levels (low and high) based on their pre-test scores.

Table 1
Rasch summary statistics for items (I) and students (S) estimates

Tests	Rasch separation index (I)	Rasch separation index(S)	Infit (I)	Infit (S)	Outfit (I)	Outfit (S)	Reliability (I)	Reliability (S)
Test1	7.08	1.31	0.93	0.83	1.70	1.23	0.98	0.63
Test2	5.79	1.52	1.04	0.98	0.99	0.99	0.97	0.70
Test3	4.98	0.92	1.02	1.04	0.95	0.95	0.96	0.46

Students' abilities and item difficulties on the Rasch measurement scale are shown in Figure 1 (pre-test and post-test). The map represents the hypothetical distribution of items (Question 1 to 9 represented by Q1-Q9) and students (2 students represented by '#' or 'X' and one student represented by '.') along the same variable. The straight line (in the upward direction) on the map represents the variable of measurement. The '#' or 'x' at the top of the variable line represents the most able students, while those down the line are the least able students. Similarly, the questions at the top of the line are the most

difficult items (Q5 and Q8 for pre-test, Q4 for post-test), while those down the variable line are the least difficult items. Furthermore, Figure 1 shows that the highest performing students at the pre-test operated at 1 logit, and the lowest performing students at the pre-test operated at -4.5 logits. After the intervention, the performance improved to 2.7 logits and -2.3 logits, respectively.

Lastly, the student measures from the Rasch software were imported into Statistical Package for Social Sciences (SPSS), where initial correlations of the students' outcomes were determined across the three time points. Also, regression analyses, such as student *t*-test and repeated measures analysis of variance, were used to answer the research questions and test the associated hypotheses. The results from these statistics were used to infer causal relationships between the independent and dependent variables.

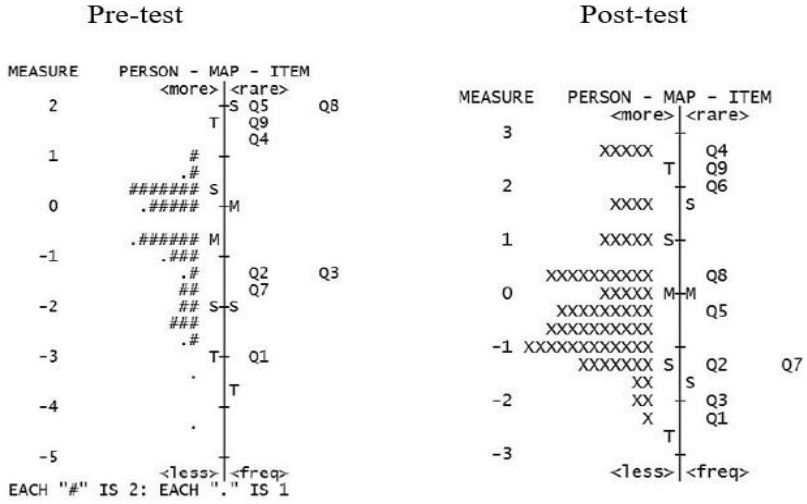

Figure 1
Wright map of students' abilities and item difficulties at the pre-test and post-test

FINDINGS

This section presents the results of the analyses and answers the research questions.

1. What effects do worked examples have on students learning outcomes in complex algebraic problems?

An initial examination of the overall changes in performance over time was explored. A correlation between pre-test and post-test yielded a low correlation coefficient χ^2 (72) = 0.35, while a moderate correlation was obtained for a correlation between the post-test and delay test $\chi^2(72) = 0.50$ at $p < 0.01$. This indicated a weak correlation in students' learning outcomes between the pre-test and post-test and a moderate correlation between the post-test and delay test. On average, students' learning outcomes at the pre-test were relatively low (M = -0.75, SD = 1.15), and increased significantly at the post-test (M = -0.14, SD = 1.19, $t_{(71)}$ = -3.88, p = 0.00, d = 0.46). The learning outcomes slightly

declined at the delay test; however, the decline was not significant (M = -0.31, SD = 1.37, $t_{(71)}$ = -3.88, p = 0.17, d = 0.16). These results show a relatively moderate worked example effect size at the post-test and a weak effect size at the delay test.

To test the hypothesis of equal means across the three tests, a repeated measure analysis of variance was conducted to explore the within-subject worked example effects on students' learning outcomes. Mauchly's test of sphericity was significant; hence, sphericity was not assumed [χ^2 (2) = 6.50, p = 0.04]. The repeated measure analysis indicated that there was a significant effect of the worked examples on students' learning outcomes across the three tests (F(1.84, 130.45) = 8.88, MSE = 0.89, p = 0.00, η_p^2 = 0.11) (as shown on Table 2). A post hoc comparison using the Bonferroni adjustment also revealed that a significant difference existed between the pre-test and post-test but there was no significant difference in the means of the post-test and delay test. This means that the worked example effect was obtained at the post-test but the effect was not completely retained at the delay test.

Table 2
Mean, standard deviation, and repeated measures analysis of variance for example-problem pairs on students' learning outcomes

	Pre-test		Post-test		Delay test			
Variables	*M*	*SD*	*M*	*SD*	*M*	*SD*	*F*(1.84,130.45)	η_p^2
Learning Outcomes	-0.75	1.15	-0.14	1.19	-0.31	0.87	8.88**	0.11

**p<0.05

Do students' expertise levels influence their learning outcomes?

To answer this question, an independent t-test was conducted to examine the differences between low-ability and high-ability students across the three tests. As expected, there was a significant large difference between the low-ability and high-ability students in the pre-test ($t_{(70)}$ = -11.91, p = 0.00, d = 2.86). In the post-test, the significant difference between the low-ability and high-ability students became moderate ($t_{(69.93)}$ = -3.20, p = 0.00, d = 0.72), and interestingly, in the delay test, no significant difference was found between the learning outcomes of the low-ability and high-ability students. This means that students' expertise levels influenced their learning outcomes. The worked example effects appeared to bridge the gap between the low-ability and high-ability students at the delay test and favoured the low-performing students more than the high-ability students. A repeated measure analysis on the low-ability and high-ability students across the three tests yielded F(1.60, 44.56) = 19.45, MSE = 0.97, p = 0.00, η_p^2 = 0.41 and F(2, 84) = 4.60, MSE = 0.61, p = 0.01, η_p^2 = 0.01, respectively. This result also showed a large worked example effects on the low-ability students (0.41) and a small effect on the high-ability students (0.01).

Table 3
Mean, standard deviation, and repeated measures analysis of variance on students' learning outcomes based on expertise levels

Variables (No.)	Pre-test *M*	Post-test *M*	Delay test *M*		Sig	η_p^2
Low (25)	-1.89	-0.62	-0.32	$F(1.60,44.56) = 19.45$	0.00*	0.41
High (43)	0.01	0.19	-0.30	$F(2,84) = 4.60$	0.01*	0.01

*$p<0.05$

What differential learning outcomes are observed based on students' gender?

In relation to the influence of gender on students' learning outcomes, the mean gain of female students (0.66) from the pre-test to post-test was higher than that of their male counterparts (0.57). However, an independent *t*-test analysis showed that there was no significant difference in the learning outcomes of male and female students at each of the time points, which meant that students' gender does not influence their learning outcomes [$t_{(70)} = -0.14$, $p = 0.89$, $d = 0.03$; $t_{(60.01)} = -0.43$, $p = 0.67$, $d = 0.10$ and $t_{(70)} = -0.26$, $p = 0.79$, $d = 0.06$]. Moreover, an analysis that explored how male and female students' learning outcomes improved across the three time points yielded equal effects [$F(1.69, 62.57) = 4.66$, $MSE = 0.86$, $p = 0.02$, $\eta_p^2 = 0.11$ and $F(2, 66) = 4.14$, $MSE = 0.93$, $p = 0.02$, $\eta_p^2 = 0.11$, respectively]. Hence, worked example instruction had the same effect on male and female students at the post-test and delay test.

Table 4
Mean, standard deviation, and repeated measures analysis of variance on students' learning outcomes based on gender

Variables (No.)	Pre-test *M*	Post-test *M*	Delay test *M*		Sig	η_p^2
Male (38)	-0.77	-0.19	-0.34	$F(1.69,62.57) = 4.66$	0.02*	0.11
Female (34)	0.73	0.07	-0.28	$F(2,66) = 4.14$	0.02*	0.11

*$p<0.05$

Does students' gender or expertise level interact with the effectiveness of the worked examples?

To determine whether students' gender or ability levels significantly interacted with the effectiveness of the worked examples, a mixed design regression analysis was performed. The result indicated no interaction between gender and the worked example effects [$F(1.84, 128.58) = 0.05$, MSE = 0.90, $p = 0.95$, $\eta_p^2 = 0.00$]. However, there was a significant interaction between students' ability levels and the worked example effect, with a large effect size [$F(2, 140) = 25.82$, MSE = 0.60, $p = 0.00$, $\eta_p^2 = 0.27$]. This means that students' ability level influenced the strength of the relationship between their learning outcomes and the worked example effects across the three time points. Figure 2 shows the ordinal interaction of students' ability levels with the worked

example effect. The 1, 2, and 3 in the *x*-axis represent the pre-test, post-test, and delay test, respectively. Again, the worked example effects were higher for the low-ability students than the high-ability students.

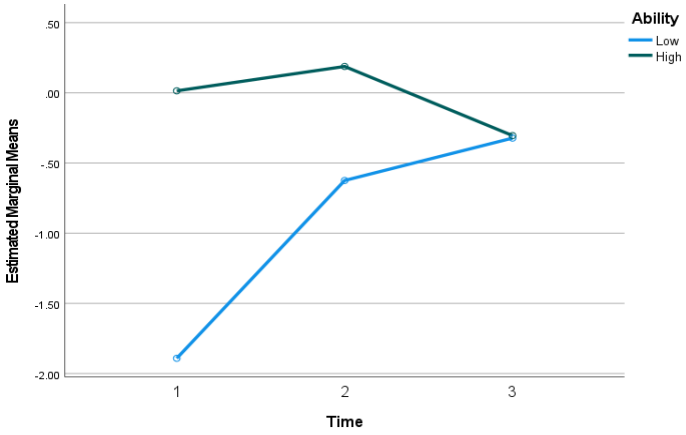

Figure 2
A line graph showing the interaction of students' ability with the worked example effect

DISCUSSION

This study investigated the short-term and long-term impacts of worked examples (example-problem pairs) on students' learning outcomes in complex algebraic problems. The results indicated a significant worked example effect on students' learning outcomes at the post-test (short-term); however, this effect slightly decreased over a period of time (delay test), but the decrease was not significant. This study found a medium effect size of worked examples at the post-test and a weak effect size at the delay test, suggesting that the worked example effect is more beneficial for improving students' learning outcomes in the short term. This result agrees with the findings of Alreshidi (2021), Berbieri et al. (2021) and Chen et al. (2019), who found a significant effect of the worked examples at the post-test. However, the weak worked example effect obtained at the delay test is not consistent with the findings of Chen et al. (2016) who suggest that the worked example effect is more obvious at the delay test than at the post-test.

An explanation for the significant worked example effect reported at the post-test could be that students are successful in borrowing schema from the worked examples, and are able to transfer them to the post-test using the borrowing and reorganising principle of the human cognitive architecture (Paas & van Merrienboer, 2020). Similarly, students were observed to solve familiar routine questions at the post-test in less time but struggle with solving non-routine questions, which may mean that students mastered the procedural steps but acquired little conceptual knowledge to deal with unfamiliar problems. Moreover, the majority of the students could not provide comprehensive explanations for their procedures. Thus, it appears that a few students memorized the

sequential steps to the solution, while others consciously selected unautomated procedures, knowledge of sequencing of actions, and a meaningful reflection of the procedures to provide a solution to the problems, which corresponds with the demonstration of procedural knowledge as explained by Hurrell (2021).

The slight decrease at the delay test may have been due to forgetfulness, which may arise from lack of revision or the non-challenging and direct instruction offered by the worked example. Again, as a way of managing the cognitive resources, the worked example does not allow students to develop their own way of finding solutions to the problem, which limits their experiences during schema construction. Likewise, Sweller et al. (2019) acknowledge that in worked examples, students may only learn appropriate moves for solving problems without carefully studying them as expected. In this scenario, students' knowledge of solving algebraic problems will experience a waning effect. Therefore, worked examples should not only emphasise learning the appropriate moves for solving a problem but also the consequences of the moves.

Furthermore, no significant difference was found in the worked examples effects according to gender, which indicates that the worked example effects on both males and females was relatively similar. This result contradicts Abott's (2021) finding that revealed a significant difference in worked example effects in favour of females. The contradictory results may be due to the form of worked example used. While Abott's study utilized the faded worked example – with gradual reduction of instructional guidance – this study provided full guidance in its worked examples.

Conversely, the worked example effect was influenced by students' level of expertise in favour of the low-ability students. This means the worked examples had a greater effect on the low-ability students than the high-ability students, both in the short term and the long term. In fact, in the long term, the worked example effect was capable of levelling the gap between the low-ability and high-ability students. This pattern of results is consistent with the expertise reversal perspective of the CLT (Chen et al., 2017; Kalyuga & Renkl, 2010) and the findings of Chen et al. (2016). The results from this study may explain the role of schema in learning. The schema present in the long-term memory of the high-ability students provides initial guidance, and the additional and full instructional guidance provided in the worked example instruction may therefore tend to overlap the learning components, resulting in redundancy (Kalyuga & Renkl, 2010). Processing and integrating the overlapping components (schema-based and instructional-based guidance) requires more cognitive resources, which results in a higher cognitive load and hence reduced effectiveness of the instruction. On the other hand, the-low ability students, who have little or no schema, tend to make use of all the instructional guidance to optimize their learning with minimal cognitive resources. Thus, this result supports the previous findings that low-ability students require full instructional guidance for high element interactivity materials (Chen, 2017).

In terms of the interaction effect, the expertise levels of students interacted with the worked example effects. This result corresponds with the findings of previous studies (Chen et al., 2019; Kalyuga, 2007). However, this finding extends the previous studies by providing the short-term and long-term interactions. The post-test results indicated a

full expertise reversal, which suggests an interaction effect with significant differences between the low-ability and high-ability students. Similarly, at the delay test, there was a significant ordinal interaction between worked example instruction and students' expertise, but there was no difference between the low-ability and high-ability students. This indicates a partial reversal as students' expertise level changes.

Accordingly, the practical implication of these findings is that the schema generated from the worked example instruction needs continuous usage for it to be adequately retained for a longer period. Therefore, teachers should provide more practice questions to the students. Also, teachers need to adjust the pedagogical approaches used in the classroom so that as students gain more expertise in the domain, less guidance is provided to high-ability students. This is in line with the recommendation of Martin and Evans (2020) on load reduction instruction (LRI), where high explicit instruction is provided to students at the initial stage, and as students advance in knowledge, there is a gradual reduction in the guidance provided.

Lastly, in relation to the different methods for solving simultaneous equations, it was observed that students found word problems (non-routine questions) the most difficult. This may be because word problems require students to understand individual concepts in the question and identify the necessary information required to form equations before using the procedural knowledge gained from routine questions to solve the equations simultaneously. This observation appears to limit the benefits of the worked example instruction not to be too reliable when the requirement is for students to transfer mathematical understanding to solving real-world problems.

CONCLUSION

This study investigated the effects of worked example instruction on students' learning outcomes in high element interacting material and considered students' expertise and knowledge retention. The results indicated a significant medium sized effect of worked example in the short term and a weak effect in the long term. Also, worked example instruction was more beneficial for the low-ability students than the high-ability students, both in the immediate and long term, and high-ability students experienced a full reversal effect of the worked example instruction. Furthermore, there was no difference in student outcomes from the worked example effect based on gender, and there were significant interactions of worked example effect and students' ability level both in the short term and long term. These results imply that learning highly interacting materials through worked examples may not help to retain acquired schema, and that worked example instruction is not appropriate for high-ability students.

A limitation to this study was that it was conducted during the COVID-19 pandemic, which had a global negative effect on populations and activities. In order to minimize the spread of the virus, the students' regular mathematics teacher carried out the intervention. Given this situation, it was not possible in this study to observe the effect of worked examples on low-ability and high-ability students' responses to routine and non-routine questions. A further line of investigation would be to compare the effects of worked examples on routine and non-routine algebraic questions, and determine the

ability level that benefits the most for each of these types of questions. Another future study may consider ways of improving students' conceptual understanding through worked examples. It would be good to also examine students' perceptions of the effectiveness of the worked examples. This study has contributed to the boundary conditions for the effect of worked examples by emphasizing the expertise reversal effect and element interactivity.

ACKNOWLEDGEMENTS

This article is part of the doctoral thesis of the first author, which was supported by a University of New England Postgraduate Research Award.

REFERENCES

Abbot, A. (2021). Gender differences in perceptions of the use of faded worked example in mathematics. In Marks, R. (Ed.), *Proceeding of the British Society for Research into Learning Mathematics, 41*(1), 1-6. https://bsrlm.org.uk/wp-content/uploads/2021/05/BSRLM-CP-41-1-01.pdf

Adeniji, S. M., Ameen, S. K., Dambatta, B., & Orilonise, R. (2018). Effect of mastery learning approach on senior school students' academic performance and retention in circle geometry. *International Journal of Instruction, 11*(4), 951-962. http://dx.doi.org/10.12973/iji.2018.11460a

Ajai, J. T., & Imoko, B. I. (2015). Gender differences in mathematics achievement and retention scores: A case of problem-based learning method. *International Journal of Research in Education and Science, 1*, 45-50. https://doi.org/10.21890/ijres.76785

Alreshidi, N. A. K. (2021). Effects of example-problem pairs on students' mathematics achievements: A mixed-method study. *International Education Studies, 14*(5), 8-18. https://doi.org/10.5539/ies.v14n5p8

Anaya, L., Stafford, F, & Zamaro, G. (2022). Gender gaps in math performance, perceived mathematical ability and college STEM education: The role of parental occupation. *Education Economics.* 30(2), 113-128. https://doi.org/10.1080/09645292.2021.1974344

Arnup, J. L., Murrihy, C., Roodenburg, J., & McLean, L. A. (2013). Cognitive style and gender differences in children's mathematics achievement. *Educational Studies, 39*(3), 355-368. https://doi.org/10.1080/03055698.2013.767184

Ayebale, L., Habaasa, G., & Tweheyo, S. (2020). Factors affecting students' achievement in mathematics in secondary schools in developing countries: A rapid systematic review. *Statistical Journal of the IAOS, 36*(S1), 73-76. https://doi.org/10.3233/SJI-200713

Barbieri, C. A., Booth, J. L., Begolli, K. N., & McCann, N. (2021). The effect of worked example on student learning and error anticipation in algebra. *Instructional Science, 49*(4), 419-439. https://doi.org/10.1007/s11251-021-09545-6.

Biggs, J. B., & Collis, K. F. (2014). *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome)*. Academic Press.

Chen, O., Kalyuga, S., & Sweller, J. (2015). The worked example effect, the generation effect, and element interactivity. *Journal of Educational Psychology, 107*(3), 689. http://dx.doi.org/10.1037/edu0000018

Chen, O., Kalyuga, S., & Sweller, J. (2016). Relations between the worked example and generation effects on immediate and delayed test. *Learning and Intruction.* 45, 20-30. https://doi.org/10.1016/j.learninstruc.2016.06.007

Chen, O., Kalyuga, S., & Sweller, J. (2017). The expertise reversal effect is a variant of the more general element interactivity effect. *Educational Psychology Review, 29*(2), 393-405. https://doi.org/10.1007/s10648-016-9359-1

Chen, O., Retnowati, E., & Kalyuga, S. (2019). Effects of worked example on step performance in solving complex problems. *Educational Psychology, 39*(2), 188-202. https://doi.org/10.1080/01443410.2018.1515891

Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education* (8th ed.). Routledge.

Hurrell, D. (2021). Conceptual knowledge or procedural knowledge or conceptual knowledge and procedural knowledge: Why the conjunction is important to teachers. *Australian Journal of Teacher Education, 46*(2), 57-71. https://doi.org/10.14221/ajte.2021v46n2.4

Johari, P. M. A. R. P., & Shahrill, M. (2020). The common errors in the learning of the simultaneous equations. *Infinity Journal, 9*(2), 263-274.

Kalyuga, S., & Renkl, A. (2010). Expertise reversal effect and its instructional implications: Introduction to the special issue. *Instructional Science, 38*(3), 209-215.

Kolawole, E. B., & Ojo, O. F. (2019). Effects of two problem solving methods on senior secondary school students' performance in simultaneous equations in Ekiti state. *International Journal of Current Research and Academic Review, 6*(11), 155-161. https://doi.org/doi:10.20546/ijcrar.2019.701.003

Leder, G. C. (2015). *Gender and mathematics education revisited*. The Proceedings of the 12th International Congress on Mathematical Education.

Linacre, J. M. (2013). DIF sample size nomogram. *Rasch Measurement Transactions, 26*(4), 1392-1402.

Martin, A. J., & Evans, P. (2020). Load reduction instruction (LRI): Sequencing explicit instruction and guided discovery to enhance students' motivation, engagement, learning, and achievement. In S. Tindall-Ford, S. Agostinho, & J. Ford (Eds.), *Advances in cognitive load theory: Rethinking teaching* (pp. 15-29). Routledge.

Martinez, M. V., Bragelman, J., Stoelinga, T. (2016). Underprepared students perfomance in algebra in a double-period high school mathematics program. *The Mathematics Educator, 25*(1), 3-31. https://files.eric.ed.gov/fulltext/EJ1110635.pdf

Mosimege, M., & Winnaar, L. (2021). Teachers' instructional strategies and their impact on learner performance in Grade 9 mathematics: Findings from TIMSS 2015 in South Africa. *Perspectives in Education, 39*(2), 324-338.

Moussa-Inaty, J., Atallah, F., & Causapin, M. (2019). Instructional mode: A better predictor of performance than students' preferred learning styles. *International Journal of Instruction, 12*(3), 17-34. https://doi.org/10.29333/iji.2019.1232a

Ngu, B. H., & Phan, H. P. (2017). Will learning to solve one-step equations pose a challenge to 8th grade students? *International Journal of Mathematical Education in Science and Technology, 48*(6), 876-894.

Nigerian Federal Ministry of Education (2017). *The Nigerian digest of education statistics 2014-2016*. https://www.education.gov.ng/images/docs/news/digest2017pdf

Omobude, E. O. (2014). *Learning mathematics through mathematical modelling: A study of secondary school students in Nigeria* [Unpublished master's thesis]. University of Agder.

Paas, F. & van Merriënboer, J. J. G. (2020). Cognitive-Load Theory: Methods to Manage Working Memory Load in the Learning of Complex Tasks. Current Directions in Psychological Science, 29(4). https://doi.org/10.1177%2F0963721420922183

Renkl, A. (2017). Learning from worked-examples in mathematics: Students relate procedures to principles. *ZDM, 49*(4), 571-584. https://doi.org/10.1007/s11858-017-0859-3

Richey, J. E., & Nokes-Malach, T. J. (2013). How much is too much? Learning and motivation effects of adding instructional explanations to worked example. *Learning and Instruction, 25*, 104-124.

Smith, H., Closser, A.H., Ottmar, E. & Chan, J. Y. (2022). The impact of algebra worked-example presentations on students learning. *Applied Cognitive Psychology, 36*(2), 363-377. https://doi.org/10.1002/acp.3925

Sweller, J. (2011). Cognitive load theory. In B. Ross (Ed.), *Psychology of learning and motivation* (Vol. 55, pp. 37-76). Elsevier.

Sweller, J., van Merrienboer, J. J. G. & Paas, F. (1998) Cognitive architecture and instructional design. *Educational Psychology Review* 10, 251–296. https://doi.org/10.1023/A:1022193728205

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261-292. https://doi.org/10.1007/s10648-019-09465-5

Van Gog, T., Kester, L., & Paas, F. (2011). Effects of worked example, example-problem, and problem-example pairs on novices' learning. *Contemporary Educational Psychology*, *36*(3), 212-218.