Computers & Education 143 (2020) 103665

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Effects of a review video and practice in video-based statistics training

Hans van der Meij[a,*], Paul Dunkel[b]

[a] University of Twente Faculty of Behavioural, Management and Social Sciences Department of Instructional Technology, P.O. Box 217, 7500, AE Enschede, the Netherlands
[b] Salzstraße 24, 48143 Münster, Germany

ARTICLE INFO

Keywords:
Video
Demonstration-based-training
Review videos
Practice

ABSTRACT

Abstract: This study investigated the effectiveness of video-based statistics training. An experiment is reported in which conditions were systematically varied for the presence or absence of review videos and practice opportunities. Three main types of dependent measures were recorded: engagement, motivation, and knowledge. Seventy students participated in training and were assessed afterwards. Video play measures revealed nearly complete (94.1%) views for original videos, and lower but still substantial views for the review video (60.5%). The results for time spent on the videos were in line with these findings. There was no effect of condition on these engagement measures. Motivation scores were high for task relevance and self-efficacy. Self-efficacy scores were significantly higher when a review video and/or practice opportunity was available. Practice scores were uniformly high in the two practice conditions. The practice conditions scored significantly better than the no-practice conditions on a procedural knowledge post-test. There was no difference between conditions on a declarative knowledge post-test, nor on a transfer test. The conclusion draws attention to the possibility of improving the effectiveness of existing video-based statistics training by adding a complementary review video and arranging for practice with external feedback.

1. Introduction

Statistics is a complex subject matter to learn. It involves acquiring knowledge of concepts, theories and formulas, as well as formulating and testing hypotheses. Another layer of complexity comes from the handling of the statistics software that is often involved (Baglin & Da Costa, 2014). It is, therefore, not surprising that many students perceive statistics courses as highly anxiety-inducing (Chew & Dillon, 2014).

Instructional videos are increasingly being employed in statistics courses to address these complexities. An important advantage of videos is that they can animate calculations and demonstrate how to accomplish computations with statistical software. Another advantage is that students can control their learning pace and view the videos anyplace and anytime. The present study was set up to investigate the effectiveness of two special features in video-based instructions for learning statistics, namely review videos and practice arrangement.

There is a growing number of flipped classroom designs for statistics courses that use videos as the primary instructional resource (e.g., Shinaberger, 2017; Wilson, 2014; Winquist & Carlson, 2014). In these flipped classrooms students must view the videos before

* Corresponding author.
E-mail address: H.vanderMeij@utwente.nl (H. van der Meij).

https://doi.org/10.1016/j.compedu.2019.103665
Received 16 April 2019; Received in revised form 9 August 2019; Accepted 12 August 2019
Available online 21 August 2019
0360-1315/

attending class, in which the focus is on problem-solving and discussion. Empirical studies with flipped classrooms for statistics courses generally report increased motivation, understanding, and task performances (e.g., Burgoyne & Eaton, 2018; Heuett, 2017; Peterson, 2016; Shinaberger, 2017).

Most of the flipped classroom research has been conducted in ecologically valid settings. That is, the research often included numerous videos, intact classes and involved existing courses that ran for weeks or even months, among other things. Such flipped classroom studies therefore do not allow one to draw firm conclusions about just the contribution of the instructional videos (Bishop & Verleger, 2013). In addition, flipped classroom studies often give little, if any, information on the design of the instructional videos. This paper describes a controlled study on videos in which special attention is given to two design features that can potentially raise the effectiveness of video-based instructions for inferential statistics.

A series of instructional videos were developed from scratch, following a worked-example design (see Renkl, 2014b). The videos showed an expert model executing and explaining the actions involved in accomplishing statistical tasks. To increase learning from a (recorded) model, additional design measures are often needed. A Demonstration-Based Training (DBT) approach has been adopted for improving effectiveness (Brar & van der Meij, 2017; Rosen et al., 2010). Essentially, this approach proposes design features that can enhance the observational learning processes of attention, retention, production and motivation. In the reported study, two design features were experimentally manipulated. One concerned the presence or absence of a review video (akin to a summary with text). The other feature concerned the presence or absence of practice with external feedback. So far, both design features have received little attention in video-based training studies.

In the following section we describe the role of review videos (henceforth sometimes simply referred to as a review) and summarize the findings from empirical research. Thereafter, practice is discussed in a similar way. Next, we introduce the experimental design with four conditions (videos only, videos with review, videos with practice, and videos with review and practice). The study investigates three types of dependent variables, namely engagement, motivation and learning. The method section contains a detailed account of the design of the videos, along with the regular descriptions of participants, instruments, procedure and data analysis. The discussion on the findings is structured in line with the dependent variables, namely engagement (video viewing and time), motivation (task relevance and self-efficacy), and learning (declarative and procedural knowledge tests).

2. Research context

2.1. Reviews

A review of an instructional video is similar to a summary in a written text. A text summary conveys a synopsis to the reader. It is a commonly employed means for presenting the main ideas advanced in a text, which empirical evidence has shown to enhance retention (Hartley & Trueman, 1982). The video counterpart for a text summary is a review. A review video provides the viewer with an abstract of the video content. The main aim of a review is to enhance learning.

Very little empirical research on the effects of review videos on learning and motivation appears to have been conducted. A systematic search of the literature with the keywords multimedia or video, in combination with the keywords review or summary yielded a substantial number of hits on automated techniques for creating a video summary. A large number of studies were also found in which professionals (e.g., teachers, medical specialists, athletes) played back videos to find out how they could improve their performance. Empirical studies were rare in which an original video was complemented with a review video to enhance learning, however.

We found only found five empirical studies that investigated the potential contribution of a review video to retention. Four consecutive experiments investigated the effectiveness of video-based training with or without review videos for formatting tasks in Microsoft's Word. Training invariably consisted of repeated cycles of viewing a video on a task followed by practice on that task. The studies investigated the effects of review videos on task performance (i.e., practice, immediate and delayed post-test) and motivation (i.e., task relevance and/or self-efficacy).

In the first study (van der Meij & van der Meij, 2016a), the participants could control video play and task practice. A significant effect of including reviews was found for task performance on an immediate post-test. Also, a significant effect of including reviews was found for self-efficacy.

In the second study (van der Meij & van der Meij, 2016b) the participants were restricted in their video play and time for practice. Video play was experimenter-controlled and a maximum time of 5 min was set for each practice task. This study found a significant benefit from the presence of reviews on practice, and on immediate post-test scores as well as on delayed post-test scores. In this study, the reviews led to higher appraisals for task relevance.

In the third study (van der Meij, 2017), there was again user control over video-play. This study was the first to record engagement with the videos. The findings indicated that over 90% of the content of the instructional (demonstration) videos was viewed both in the video-only and the video with review condition. Engagement with the reviews was relatively low (32%). Even so, the review condition did significantly better on task performances during and after training. No effect of condition was found for self-efficacy.

The fourth study (van der Meij, van der Meij, Voerman, & Duipmans, 2018) compared the effectiveness of reviews with a (no-review) control and second demonstration condition. The latter condition was included to investigated the idea that reviews are better aligned with the retention process than a sheer repetition of content. The participants in the study controlled video play and time for practice. No effects of reviews on task performances were found. This was ascribed to the (added) presence of previews in all videos, causing too much overlap with the reviews. Usage data supported this view, as previews were viewed almost in full while reviews scored less than 50%. Self-efficacy increased most for the review, but a ceiling effect prevented the discovery of an effect of condition.

Another experiment on video-based training with and without reviews revolved around statistics (Brar & van der Meij, 2017). The videos in this study discussed inferential statistics (i.e., t-testing). Participants were students enrolled in a statistics course. Training consisted of user controlled video viewing (no practice). Students were tested after training with a knowledge and performance test. Usage data indicated that 98% of the instructional videos were viewed in both the video-only and video with review condition. The viewing rate for the reviews was very high (90%), but even so it was significantly lower compared to the instructional videos. The findings showed that reviews raised the scores on these tests slightly but not significantly, as was expected. The study did not measure motivation.

Overall, the few studies that exist have found review videos to be viewed less extensively than instructional (demonstration) videos (Brar & van der Meij, 2017; van der Meij, 2017; van der Meij & van der Meij, 2016a,b; van der Meij, van der Meij, et al., 2018). These studies have, nevertheless, tended to find reviews favorable for procedural knowledge development on trained tasks. Only one study examined declarative knowledge development, but found no effect of reviews. Transfer tasks have not been investigated. Effects of reviews on motivation have been mixed. One study found a significant positive effect on self-efficacy, one on task relevance and two studies found no effects on motivation.

2.2. Practice

As related earlier, the instructional videos in the present study were designed as worked examples. Worked examples can be defined as problem-based instructions in which the instructions begin with a problem statement or definition and then elucidate the problem solving process (Renkl, 2014b). Many worked-example designs have a classic coupling of instruction followed by practice. Students first receive a worked example on a problem and then engage in practice on that problem. This coupling of instruction with practice is also advocated in the DBT-approach (Bandura, 1986). Two main arguments have been raised for the example-practice coupling: consolidation and stimulating deeper processing.

One argument is that practice can help consolidate learning. The worked example provides users with a mental model of the solution process, which can then subsequently be fortified by practicing with a similar problem (e.g., Hodges & Coppola, 2015; van Gog, 2011). Similarly, a testing effect may occur, with practice strengthening the memory trace (e.g., Brewer, Marsh, Meeks, Clark-Foos, & Hicks, 2010).

Another argument why practice may be beneficial for learning is that it can stimulate deeper processing. A moderating factor in the worked examples effect is student effort. Worked examples run the risk of passive and superficial processing (Atkinson, Derry, Renkl, & Wortham, 2000). When students do not actively engage with examples, their effectiveness is seriously threatened. Passive processing is also a risk of video-based instructions (Salomon, 1984). Including practice can have an activating effect. Practice can stimulate students to deepen understanding and thus enhance retention.

To our knowledge, no research appears to have been done on the contribution of practice in video-based learning of statistics. There are, however, a few empirical studies that have investigated the effects of practice on video-based software training. These studies can give some insights into how to arrange such practice in statistics software training and what effects can be expected.

Ertelt in experiment 2 (Ertelt, 2007) compared a video-practice with a video-only condition for learning RagTime, a desktop publishing program. The study did not record engagement, or motivation. A significant but small effect of practice was found on an immediate, delayed and transfer test.

Two recent experiments investigated the effects of practice in a video-based training on formatting tasks in Microsoft's Word. The first study (van der Meij, Rensink, & van der Meij, 2018) consisted of a video-only control and two experimental conditions that varied in the coupling of instruction and practice (i.e., practice-video and video-practice). No engagement data were recorded. Also, effects of conditions on motivation were not measured. The video-practice condition did significantly beter than the other conditions on the practice items included in the training. However, this learning effect did not last. After training there was no difference between conditions on the immediate and delayed performance tests.

A follow-up study (van der Meij, 2018) included a fourth condition that represented a prevalent real-life scenario of software usage, namely practice-video-practice. The study replicated the main findings from the earlier study. That is, the video-practice condition outperformed the other conditions (including practice-video-practice) on the practice items during training. But in this study, too, the learning effects did not last. After training there was again no difference between conditions on an immediate and delayed performance test that assessed retention of how to perform the trained tasks. A significant effect of practice was found only on a transfer test where the outcomes favored the three practice conditions over the control. Motivation was also measured, but no effect of practice on measures of task relevance and self-efficacy was found. The study did not record engagement data.

In summary, empirical research on practice in software training suggests that the best design arrangement is a sequence in which video instruction precedes practice. All in all only modest effects of practice have been reported, however. One possible explanation lies in the specific arrangement of practice vis-à-vis the videos. In the above-mentioned studies, practice followed immediately after a task instruction. Such a direct coupling is referred to as a blocked schedule. An alternative arrangement is a mixed schedule. In such a schedule all task instructions precede practice. Research has shown that a mixed schedule can be more beneficial for learning than a blocked schedule (e.g., Abel & Roediger, 2017; Broadbent, Causer, Williams, & Ford, 2017; Helsdingen, Van Gog, & Van Merriënboer, 2011). The present study therefore arranged for practice after all instructions had been completed.

Another possible reason is that there was only internal feedback to practice. The studies apparently departed from the view that the interaction with the software provided the participants with sufficient intrinsic feedback (i.e., actions yielding visual and auditory information) for consolidating or modifying task performance (see Bandura, 1986). If that had been the case, it would have facilitated software training arrangements. However, the modest findings suggest that, even in the case of human-computer interaction, augmenting, extrinsic feedback may be needed to raise the effectiveness of practice for learning (see Fiorella & Mayer, 2018). The present

study therefore gave extrinsic feedback to practice tasks. Motivation was rarely measured in the reviewed studies. One study that measured motivation found no effect of practice. Not a single practice study reported engagement data.

3. Research design and questions

The research literature has advanced several arguments why reviews and practice should benefit motivation and learning. Empirical research on these features has occasionally, but not always, found positive effects. This study aimed to further explore this issue. It reports on an experimental study with a full factorial design. The four conditions in the study were: control (instructional videos only), practice (instructional videos + practice), review (instructional videos + review video) and review & practice (instructional videos + review video + practice).

Question 1: What is the effect of condition on video engagement? Participants must engage with the videos sufficiently in order for the videos to affect motivation and learning (compare Shinaberger, 2017). To obtain insight into this prerequisite, two engagement measures were recorded: unique play rate and engagement time (see Method). Engagement measures were expected to be equal across conditions.

Question 2: What is the effect of condition on motivation? For motivation the study looked at task relevance and self-efficacy Method). Because the training addressed a prescribed study topic, no effect of condition on task relevance was expected. In accordance with earlier empirical findings, it was expected that reviews would have a positive effect of on self-efficacy. Practice was not expected to influence self-efficacy.

Question 3: What is the effect of condition on learning? Four measures were recorded to assess learning. During training, success on the practice tasks was measured. After training, a knowlegde, performance and transfer test were administered. Based on the reviewed studies, the experimental conditions were expected to yield better scores on all learning outcomes. No differences between experimental conditions were expected on these outcomes.

4. Method

4.1. Participants

The seventy participants were students enrolled in an introductory Research Methods and Descriptive Statistics course offered by the university. The course is a prerequisite for admission to a Master's program in International Business Administration (n = 27), Psychology (n = 18), Educational Science and Technology (n = 12), Communication Science (n = 9), or other (n = 4). The mean age of the 26 male and 44 female students was 24.2 years (SD = 3.9). Participation in the experiment was voluntary and students could opt-out at any time during the study. The participation rate was 78% of all students enrolled in the course. No course credits could be won for participation. All students who completed the experiment received 10 euros. Payment for participation is standard procedure in our university. Because the participants in all conditions were paid, it was not expected to affect the findings on the relative effectiveness of conditions. Students were randomly but evenly assigned to conditions. The number of participants in conditions were: control (n = 16), practice (n = 19), review (n = 18) and review & practice (n = 17).

4.2. Instructional materials

4.2.1. Instructional videos

The instructional videos explained the meaning and calculation of descriptive statistics. In addition, they showed how these could be computed with SPSS statistics software (version 23). There were four instructional video clips. Clip #1 dealt with central tendency. It discussed the mean and median (3 min 58 s). Clip #2 dealt with dispersion. It discussed quartiles and interquartile range (3 min 22 s). Clip #3 dealt with data exploration. It discussed boxplots and outliers (3 min 5 s). Clip #4 dealt with spread. It discussed variance and standard deviation (3 min).

All clips departed from a real-world problem that students could relate to (see Merrill, 2002). Each clip had a 5-part structure. It began with a brief preview and then introduced the main problem. Thereafter, it was animated and explained how to calculate a statistic. This was rounded off with a recap of the meaning of the statistics. Finally, the clip demonstrated and explained the procedure for computation of the statistic in SPSS. The interested reader can find these videos on the YouTube channel from one of the authors.

A Demonstration-Based Training (DBT) approach was followed to create the video clips and the moments of practice (see Brar & van der Meij, 2017; Grossman, Salas, Pavlas, & Rosen, 2013; van der Meij, 2017). In DBT, key observational processes of attention, retention, production (practice) and motivation are linked to supportive design features.

Attention was guided to facilitate selective processing of the videos core content. The preview prepared students for the forthcoming information; it supported top-down attentional processing. Signalling and zooming techniques were used to support bottom-up attentional processes. Signals were in red rather than blue because the former has a stronger attentional value (Kosslyn, Kievit, Russell, & Shephard, 2012).

Retention refers to the process of understanding and storing information for future behavior. An important technique for building understanding is simple-to-complex sequencing of content (van Merriënboer & Kester, 2014). This guideline led us to present the problem of computing the mean and median before discussing dispursion into quartiles and interquartile range, among other such choices.

Another vital technique for enhancing retention is segmentation, which involves dividing a long video into several clips or creating within-video sections having a clear beginning and end. Segmentation strongly improves the comprehensibility of a video, especially

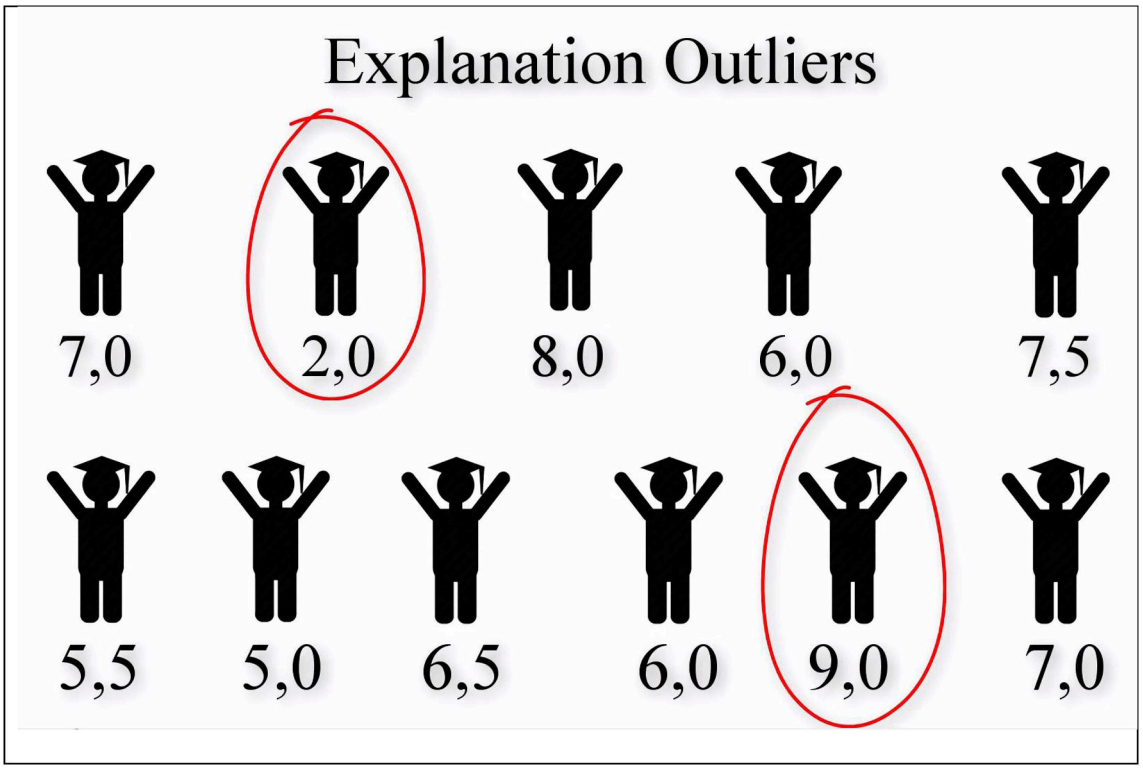

Fig. 1. Screenshot of a labelled video section.

when combined with short 2–5 s pauses (Spanjers, van Gog, & van Merriënboer, 2012; Spanjers, van Gog, Wouters, & van Merriënboer, 2012). Adoption of this technique led to a spit between four separate videos, and to the five-part structure within each video.

Yet another technique to support retention is to present labels. A label summarizes a key point and is often presented as one or two words that appear on the top of a screen. Labels create "desirable difficulty" when combined with the spoken narrative. The small discrepancy between what the narrator says and what is written on the screen can stimulate the user to pay close attention to both sources, resulting in deeper processing (Yue, Bjork, & Bjork, 2013). In the present study, labeling was used as a header for certain video sections. For example, if the narrator explained the calculation of a concept, a header appeared at the top of the screen, thus serving as a label (see Fig. 1).

Finally, all video clips confronted students with a single and simple solution strategy for creating SPSS syntax. In SPSS, the syntax for a statistical procedure can be produced in two ways. Writing one's own syntax is the most efficient and flexible way, but it is also more demanding. Following the design advice to teach only a single method and to select the easiest one (Renkl, 2014a; van der Meij & Carroll, 1998), all videos presented a menu-based method for creating the SPSS syntax.

Practice involves students in exercises that help assess and strengthen the newly acquired knowledge. Supporting practice lies outside the scope of the video itself. Usually, the larger training setting includes an arrangement for practice. This was also the case in the present study. In line with the findings from empirical research, practice (if any) always occurred after the video clips. Practice tasks were problem-centered. Just as for the instructional videos, a real-world problem was chosen that could appeal to the audience (i.e., beer consumption of 50 students at a campus party). During practice, students worked on specially prepared practice file. Besides contributing to the effectiveness and efficiency of this process, use of these files also facilitated giving feedback. The dataset for the practice tasks included the variables age, gender, education, nationality and glasses of beer.

The practice tasks covered all the demonstrated statistics. The students were asked to perform all computations in SPSS. The findings had to be presented in a specially prepared Word file which asked for a screenshot of the pertinent SPSS output. Because SPSS output invariably contains supplementary data, the students also had to write down the found value(s) for the requested statistic to ensure that they had selected the right variable(s) and value(s).

Students could obtain feedback after practice tasks. This feedback was given in a separate file. It consisted of a question, problem definition, written set of action steps in SPSS, and a screenshot of the SPSS output for the statistic for each of the six practice tasks.

Performance success on the practice tasks was measured. For each question, one point was awarded for a screenshot with the correct SPSS output, and one point for stating the correct value(s). The six questions covered topics from all four video clips (two questions related to videos 1 and 2 and one question related to videos 3 and 4). The maximum score for practice was 12 points. Scores were converted to percentages.

Motivation is the process that stimulates students to engage in an activity and to persist in their endeavor (Pintrich & Schunk, 2002). One technique for increasing motivation is task-anchoring. As related earlier, the content was presented to the students as real-

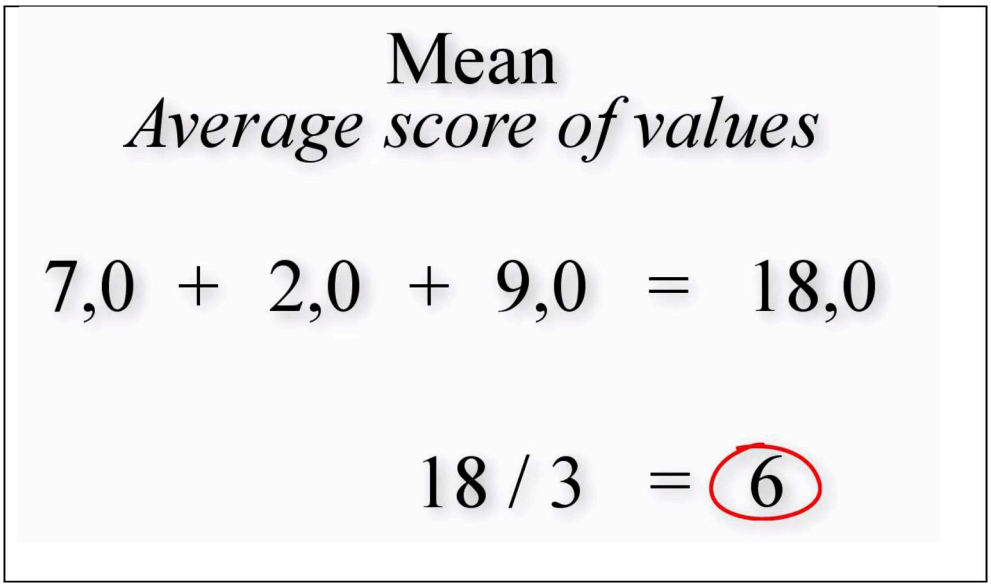

Fig. 2. Screenshot of the expanded slide for calculating the mean in the review video.

world problems that they were likely to value. Another technique for keeping students motivated is by limiting video length. All clips stayed well within the 6 min maximum advised for instructional video (Guo, Kim, & Rubin, 2014).

4.2.2. Review video

The review video was a stand-alone, abbreviated version of the four video clips (2 min 48 s). It gave a condensed account of the main conceptual and procedural information presented in the original clips, following the same design techniques. The review discussed the topics in the same sequence as did the video clips. For each topic, a 3-part structure was followed. First, the statistic was defined, then the calculation was demonstrated. Next, the solution steps were shown (see Fig. 2). Finally, the computation in SPSS was shown. The narrative was personalized. It presented action information in the "I" form (conversational style) to better align it with the viewer's possible mental rehearsal (e.g., "To calculate the mean, I add up all scores and divide the sum by the number of scores"). There was no display or discussion of SPSS output.

4.3. Instruments

4.3.1. Website

A dedicated website guided the participants through the experiment and gave access to all resources and programs (e.g., videos, SPSS, Word files, questionnaires). Fig. 3 shows the webpage with the videos. The left side, which was permanently visible, gave the table of contents that linked to the videos. After selecting a video, a still with the title and subtitle of the clip appeared. Students could then start the video by selecting the play button from a standard toolbar. This toolbar also enabled other types of *user control* such as (temporary) stops and rewinding.

4.3.2. User logs

All students' interactions with the videos were unobtrusively recorded. This record giving the status of each second of a video (e.g., play, replay, pause and skip) yielded the input for the engagement measures. The unique play rate was the number of seconds of a video that were presented at least once in play mode. It was expressed as a percentage of the total number of seconds in a video. For example, if a student watched the first 100 and last 10 s of Clip #4 (180 s), unique play rate was (100 + 10)/180 = 61%. The minimum score of 0% indicated that a video was never set into play mode. The maximum score of 100% meant that every video second was played at least once. Unique play rate is a proxy for viewing.

Engagement time is the mean total amount of time participants spent on each video. The measure (in seconds) includes replays and pauses. Replays may be caused by participants noticing that they missed certain information, need better understanding, or want to check for confirmation, among other things. Pauses too can be a sign of a meaningful action on a video, as they enable participants to reflect on, or to study, a particular frame or slide for a longer period. Engagement time was expressed as a percentage of the total number of seconds in a video. For example, if a student was engaged for 200 s with Clip #4 (180 s), engagement time was (200/180 = 111%.

4.3.3. Motivation questionnaire

The motivation questionnaire concentrated on the two key constructs from expectancy-value theory, namely task relevance and self-efficacy (Eccles & Wigfield, 2002). Task relevance refers to the present and future utility of an activity. It indicates the importance

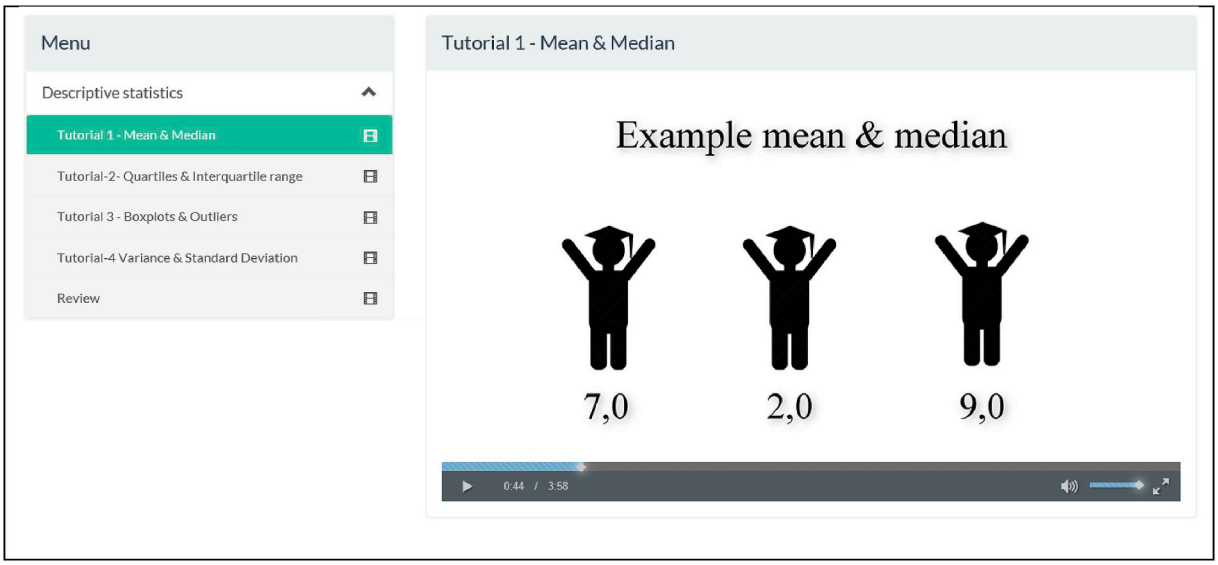

Fig. 3. Screenshot of the website (Review condition) with links to video clips. Clip #1 has been activated. The displayed slide is from the introduction to mean and median.

of a task to the user's goals or concerns (van der Meij, 2007). Self-efficacy refers to the user's expectancy for success in novel tasks in the same domain (Bandura, 1997). The items for the questionnaire were adapted from the Motivated Strategies for Learning Questionnaire (MSLQ) (Pintrich, Smith, Garcia, & McKeachie, 1991). There were 4 questions on task relevance (e.g., "I think that SPSS is relevant for my study", "I found the tasks important") and 5 questions on self-efficacy (e.g., "I can now compute a mean in SPSS", and "The tasks were easy"). Responses were given on a 7-point Likert scale with the response anchors *strongly disagree* (1) – *strongly agree* (7). The MSLQ is widely used in educational research where it has been found to have high reliability and good predictive validity (Duncan & McKeachie, 2005; Pintrich, Smith, Garcia, & McKeachie, 1993). In the present study the reliability scores, Cronbach α, for task relevance and self-efficacy were satisfactory, respectively 0.67 and 0.69.

4.3.4. Performance test

This test resembled the practice. It was problem-centered and revolved around a similar set of variables in the database (50 participants with age, gender, education, nationality and facebook friends). Participants also had to report with SPSS screenshots and statistic(s) value(s). In short the test differed in surface level features but had the same underlying deep level features as practice. Scoring was identical for practice tasks and performance tests, with one point for a correct answer and one point for a screenshot of the correct SPSS output, resulting in a maximum value of 12 points. Scores were converted to percentages.

4.3.5. Transfer test

The transfer test consisted of three items on untrained topics (i.e., mode, range, and histogram). The Facebook database was used for the SPSS computations that were requested. One point was awarded for a correct answer and the test maximum was 6 points. Scores were converted to percentages.

4.3.6. Knowledge test

Declarative knowledge was tested with 11 multiple-choice items with four alternatives for each question. The test covered content from all four video clips. An example of an item was "Deviation is …" (Alternatives: (a) the same as variance, (b) a score far from the median, (c) the difference between each score and the mean, (d) the difference between each score and the median). One point was awarded for a correct answer and the test maximum was 11 points. Scores were converted to percentages.

4.4. Procedure

The topics in the videos were created together with the teacher of the Research Methods and Descriptive Statistics course. The experiment was conducted after the students had received several lectures about general research methodology in class. The students were told that the experiment prepared them for the forthcoming statistics lecture in the course and that it would take about 1 h to complete. They were informed that the set-up consisted of pre-training with videos and a number of statistical assignments. Participation in the pre-training was possible only before its content was taught in class. Students were told that there was a password-protected website with videos and other resources. Students could engage in the pre-training at home or at the university. During pre-training, a logging program recorded the students' actions on the website. After typing their code number and password, students were directed to the website that instructed them to view the videos in the indicated sequence (ending with the review video

in the Review conditions). Also, they were told that they could process the videos as long as they wanted. Next, the students in the practice conditions engaged in the mandatory hands-on SPSS practice, and they could consult the feedback for each task. Thereafter, these students were tested. For the students in the other conditions, testing followed immediately after video viewing was completed. Testing started with the performance test, followed by the motivation questionnaire, transfer test, and knowledge test. After the experiment was completed students were thanked, debriefed and paid for participation.

4.5. Data analysis

A check on the random distribution of participants for age and gender revealed no significant difference between conditions. Likewise, as a proxy check on prior knowledge distribution, there was no difference between conditions on scores obtained in the methods course that immediately preceded the Research Methods and Descriptive Statistics course.

Assumption testing revealed non-normal distributions for all data. Therefore, comparisons involved non-parametric tests (i.e., Kruskal Wallis, Wilcoxon, Mann-Whitney). For the Mann-Whitney *U* test, the exact significance is reported. Due to missing data, the degrees of freedom slightly vary across measures. Testing was two-tailed with alpha set at 0.01 to correct for repeated testing. For effect size the *r*-statistic is reported (Field, 2013). This statistic tends to be qualified as small for $r = 0.10$, medium for $r = 0.30$, and large for $r = 0.50$ (Cohen, 1988). For a sample of 70 participants, and an alpha of 0.05, the power exceeds 0.80 when a large effect is obtained (Cohen, 1992).

5. Results

5.1. Engagement

Table 1 presents the data for *unique play*. This engagement measure represents the percentage of a video that was set in play-mode, a proxy for viewing. The table shows that overall, students had viewed over 90% of the *instructional videos*. There was no difference between conditions on unique play rates for these videos, $H(3, N = 69) = 0.48$, $p = 0.92$.

Table 1 also shows that the unique play rate for the *review video* was about 60%. This means that just under two-thirds of the review may have been viewed. There was no difference between conditions on unique play rates for the review video, $H(1, N = 31) = 0.42$, $p = 0.52$. A repeated measures analysis showed that there was a large, significant difference on unique play rates by video type, $T(31) = 69$, $z = 3.51$, $p < 0.001$, $r = 0.63$. A much smaller part of the review video was played and hence potentially viewed.

Table 2 presents the data for *engagement time*. This measure records the mean amount of time participants engaged with a video. The table shows that overall, students took 40% longer to process the *instructional videos* than a straightforward play would have taken them. There was no difference between conditions on engagement time for these videos, $H(3, N = 69) = 2.60$, $p = 0.46$.

Table 2 further shows that the engagement time for the *review video* was about 95%. This means that, on average, the time spent on this video was equal to its length. The standard deviations again point to considerable differences between participants. There was no difference between conditions on engagement time for the review video, $H(1, N = 31) = 1.50$, $p = 0.22$. A repeated measures analysis showed that there was a large, significant difference for engagement time by video type, $T(31) = 54$, $z = 3.80$, $p < 0.001$, $r = 0.68$. Participants spent relatively more time processing the instructional videos than the review video.

5.2. Motivation

Table 3 presents the data for motivation. For both task relevance and self-efficacy the overall mean scores are above the scale mean of 4. This indicates that students gave a positive appraisal for the value of the training for their studies, and also expressed confidence in their knowledge of the statistics discussed in the videos.

There was no difference between conditions on *task relevance*, $H(3, N = 67) = 3.10$, $p = 0.38$, but conditions did differ for *self-efficacy*, $H(3, N = 67) = 14.49$, $p = 0.002$. Detailed analyses showed that there were no differences between the experimental conditions. In contrast, each experimental condition differed significantly from the control condition. The data from the comparisons for Review & Practice, Review, and Practice, were $U(33) = 58.5$, $z = 2.81$, $p = 0.004$, $r = 0.50$; $U(33) = 62.5$, $z = 2.66$, $p = 0.007$, $r = 0.46$; $U(33) = 40.0$, $z = 3.48$, $p < 0.001$, $r = 0.61$, respectively. The effect size values pointed to a large effect.

Table 1
Mean unique play rates (standard deviation) per condition and video type.

	Instructional Video M SD	Review Video M SD
Review & Practice ($n = 17$; 15)[a]	97.8% (3.6)	62.7% (41.9)
Review ($n = 18$; 16)[a]	93.5% (12.7)	58.4% (43.2)
Practice ($n = 18$)	92.8% (13.7)	n.a.
Control ($n = 16$)	92.2% (23.7)	n.a.
Total ($n = 69$; $n = 31$)[a]	*94.1% (14.8)*	*60.5% (41.9)*

[a] The number of subjects for, respectively, original video clip, and review video.

Table 2
Mean engagement time (standard deviation) per condition and video type.

	Instructional Video M SD	Review Video M SD
Review & Practice (n = 17; 15)[a]	170.2% (125.5)	120.3% (140.9)
Review (n = 18; 16)[a]	127.5% (44.2)	72.3% (64.7)
Practice (n = 18)	127.0% (47.8)	n.a.
Control (n = 16)	142.5% (66.4)	n.a.
Total (n = 69; n = 31)[a]	*141.3% (77.8)*	*95.5% (109.3)*

[a] The number of subjects for, respectively, instructional video, and review video.

5.3. Learning

The overall mean score for *practice* was 93.8% (see Table 4). This finding indicates that students were capable of successfully completing, nearly all of the hands-on assignments during training. There were no differences between conditions for practice, $U(35) = 122$, $z = 1.15$, $p = 0.25$.

The overall mean score on the *performance test* was 89%. This finding indicates that students successfully completed nearly all hands-on tasks after training. There was a significant difference between conditions on this test, $H(3, N = 70) = 18.07$, $p < 0.001$. Table 3 shows that the two practice conditions had nearly identical scores on this test. Compared with the non-practice conditions, the difference was statistically significant, with a medium effect size, $U(70) = 919.5$, $z = 4.02$, $p < 0.001$, $r = 0.48$. Detailed analyses showed that the practice conditions differed from the control condition, $U(52) = 400.5$, $z = 2.68$, $p = 0.007$, $r = 0.37$, and also that the practice conditions did better than the review condition, $U(54) = 519$, $z = 4.08$, $p < 0.001$, $r = 0.55$.

The 91% mean score on the *transfer test* indicated that students did very well on this test. There was no difference between conditions on this test, $H(3, N = 70) = 3.99$, $p = 0.26$.

The 76% mean score on the *knowledge test* indicated that students did moderately well on this test. There was no difference between conditions on this test, $H(3, N = 67) = 3.42$, $p = 0.33$.

6. Discussion

6.1. Engagement

Possibly the most frequently used measure for video engagement is dropout rate. Video hosting websites such as Wistia and YouTube commonly report this statistic as a signal of whether a video did a good job capturing and holding the attention of its audience. Dropout rate can be defined as the percentage of viewers who navigate away from a video before its completion (Kim et al., 2014). The present study measured two different engagement statistics to obtain a slighty more detailed perspective on the issue whether the participants interacted enough with the videos. That is, the study recorded unique play and engagement time as each measure gives different insights in participants' video processing activities.

Unique play reports the percentage of a video that was set in play mode at least once. It indicates how much of a video could have been watched, a prerequisite for learning from a video. The data showed that nearly all instructional videos were viewed in full in all conditions. The finding compares favorably with the average of 52% reported by Kim et al. (2014) for videos in MOOCS. Admittedly, those videos appear in a different context than the experiment reported here. But even so, the unique play rate can be considered high. The relatively low standard deviations further indicate that most participants had similar rates. This satisfies an important prerequisite for video-based training. The unique play rates for the review video were considerably lower, namely around 60%. This limits its effectiveness.

Unique play is similar to the audience retention measure in YouTube Analytics (Google, 2018). Audience retention shows how much of a video has been viewed. Like unique play, the goal is to get audience retention as close to 100% as possible, because it signals that viewers have processed the whole video. Audience retention is a more sophisticated metric than unique play, however, in that it also shows *when* viewers drop off. Information about drop-off points is especially useful as a cue for where a video might benefit from redesign.

Table 3
Means (standard deviation) for task relevance* and self-efficacy* per condition.

	Task relevance M SD	Self-efficacy M SD
Review & Practice (n = 17)	5.97 (0.64)	5.62 (0.62)
Review (n = 17)	5.88 (0.60)	5.61 (0.68)
Practice (n = 17)	6.09 (0.67)	5.85 (0.53)
Control (n = 16)	5.72 (0.63)	4.91 (0.70)
Total (n = 67)	*5.92 (0.64)*	*5.51 (0.71)*

* Scale values range from 1 to 7, with higher values meaning a more positive rating.

Table 4
Means (standard deviation) for knowledge assessment per condition.

	Practice M SD	Performance Test M SD	Declarative Knowledge Test M SD	Transfer Test M SD
Review & Practice (n = 16; 17; 17; 17)[a]	94.3% (13.5)	96.6% (8.4)	73.3% (15.6)	83.3% (26.4)
Review (n = −; 18; 17; 18)[a]	n.a.	76.4% (20.5)	81.8% (15.1)	92.6% (17.4)
Practice (n = 19; 19; 17; 19)[a]	93.4% (9.0)	96.1% (8.0)	75.4% (17.5)	94.7% (12.5)
Control (n = −; 16; 16; 16)[a]	n.a.	87.0% (13.6)	72.7% (15.2)	92.7% (10.5)
Total (n = *35; 70; 70; 67*)[a]	*93.8% (11.1)*	*89.1% (15.7)*	*75.9% (15.9)*	*91.0% (17.9)*

[a] The number of subjects for, respectively, practice, performance test, declarative knowledge test, and transfer test.

The other measure of video processing registered in the present study was engagement time. It is a record of the total amount of time spent on a video. When unique play is close to its maximum (100%), engagement time signals the presence of replays and/or pauses. Both are deliberate actions of the participant that can contribute to better understanding. Indeed, research advocates the inclusion of pauses to increase the effectiveness of instructional videos (Spanjers, van Gog, Wouters, et al., 2012). The findings showed that the engagement time for the instructional videos exceeded the duration of the videos by no less than 40% on average. Just as for unique play, the highest score was found for the review & practice condition, but statistical testing revealed no difference between conditions. A similar outcome emerged for the review video. Here too the review & practice condition took longer than the length of the video and had the highest score of the review conditions.

Engagement time is commonly reported in large-scale automated data analyses. This measure of video-watching session length includes plays, replays and pauses (Guo et al., 2014). It is a time-stamped record that begins when a video opens and ends when it closes or another event occurs (e.g., the viewer navigates to another page). When audience retention is at its maximum and engagement time is over 100%, it is a useful metric for gauging the comprehensibility of a video. The presence of rewinds and pauses may signal that viewers have experienced difficulties understanding the video after a first round of video-play. It is a token that the pace may be too fast and design adjustments are called for.

Jointly, the two engagement findings in the present study give off the strong impression that the audience found the instructional videos informative enough to play these nearly in full. In addition, the participants spent a considerable amount of 'extra' time for pauses and replays which could signal that they took deliberate actions to further their comprehension of the videos.

The engagement data can be used for practical purposes. As mentioned earlier, video processing is important in flipped classroom courses, which depend heavily on students coming to classes well-prepared (Shinaberger, 2017). Metrics such as unique play and engagement time could therefore serve as a check or gatekeeper (with or without implications) in such courses. On the other hand, viewing data need to give valid insights into video processing and this validity can be questioned. For instance, the two measures recorded in the present study cannot reveal whether the viewer is actively paying attention to the video and how information is processed. If the research requires more detailed process information, logging is not an option and other, more time-consuming measures are needed, such as eye-movement registrations, observations, and think-aloud protocols.

6.2. Motivation

Even though statistical findings abound in daily life, students often fail to notice the value of statistics. For instance, Ruggeri, Dempster, Hanna, and Cleary (2008) found that a majority of psychology students whom they questioned failed to see the relevance of understanding statistics. In addition, one of the reasons why statistics courses are widely considered to be highly anxiety-evoking stems from (low) self-efficacy, as illustrated in the research by Payne and Israel (2010). More generally, these and other studies have given evidence that the two central motivational constructs from expectancy-value theory, task relevance and self-efficacy, affect motivation and learning in a statistics course (e.g., González, Rodríguez, Faílde, & Carrera, 2016; Hood, Creed, & Neumann, 2012).

To contribute to this literature, the present study measured whether review and practice affected task relevance and self-efficacy. Before discussing these findings, it is noteworthy to look at the high appraisals given by the students for these constructs. The overall scores for task relevance and self-efficacy were very positive being considerably higher than the neutral scale value.

We believe that these favorable scores were at least partly due to the Demonstration-Based Training approach that was adopted to create the videos (see Brar & van der Meij, 2017; Grossman et al., 2013; van der Meij, 2017). In this approach, the key design guideline for task relevance is anchorage in the task domain of the audience. It calls for a selection of problems that an audience can easily recognize as real and relevant. Two guidelines that are likely to have strongly contributed to self-efficacy are segmentation and simple-to-complex sequencing. These guidelines argue for respectively a meaningful division of subject matter into manageable problems, and a presentation order of gradually increasing task complexity.

There was no effect of condition on task relevance. Apparently, the basic design of the instructional videos (more than) sufficiently conveyed the value of the statistics that were taught. There were also no differences for self-efficacy between the three experimental conditions. However, each and every experimental condition contrasted significantly with the control, and the effect sizes pointed to the existence of a large difference in self-efficacy.

The finding of a positive contribution of an included review to self-efficacy extends the outcome obtained in one of the earlier reported studies on reviews in software training (van der Meij & van der Meij, 2016a) that involved a different audience (primary and secondary school students) and another domain and software (text processing in Microsoft Word). Perhaps the review served as a

favorable criterion or benchmark for the students' own appraisals of what they had learned from the videos. The review may have increased self-efficacy because it helped reduce students' anxiety about statistics. In Bandura's (1997) theory, anxiety is one of four main sources of self-efficacy information.

Earlier studies had not found an effect for practice. One possible reason for the significant effect of practice on self-efficacy in the present study is the provision of feedback. Where other studies merely enabled practice, the present study also offered feedback on demand. In Bandura's (1997) theoretical framework, feedback plays a key role in mastery experience which is another primary source of self-efficacy information. In addition, various empirical studies have given proof of the effect of feedback for increasing self-efficacy (e.g., Aro et al., 2018; Beatson, Berg, & Smith, 2018; Dimotakis, Mitchell, & Maurer, 2017; van Dinther, Dochy, & Segers, 2011).

6.3. Learning

A mean score of 93.8% was found for practice tasks. The finding indicates not merely that students could complete nearly all training tasks successfully, but also that they were willing to engage in performing these tasks during training. This suggests that the students found practice a valuable training component.

The mean score of 89.1% on the performance test indicates that the students generally achieved a success rate above a mastery-level criterion of 80%. Because no pretest was administered, it is impossible to tell how much each student already knew or learned. The two practice conditions, which did not differ between themselves, accomplished a near perfect score (96%) on the performance test. These conditions also achieved a significantly higher score on the performance test than did the non-practice conditions (81%). The effect size pointed to a moderate difference. In other words, the present study found that the inclusion of practice helped increase the scores on the performance test.

This finding is not in line with video-based software training research that found no advantage of practice on an immediate and delayed performance test (van der Meij, 2018; van der Meij, Rensink, et al., 2018). One possible explanation lies in the mixed practice schedule that was followed to accommodate practice. In a mixed practice schedule instructions about different tasks are presented before students can engage in task practice. The arrangement is more challenging than the classic task instruction – task practice pairing and there is empirical research that suggests that this can increase learning outcomes (e.g., Brady, 2008; Helsdingen et al., 2011; Merbah & Meulemans, 2011).

Another possible explanation lies in the provision of external feedback. In the above-mentioned studies, no such feedback was given. Student had to rely on their own appraisals and/or on internal feedback that is part and parcel of human-computer-interaction. In the present study, students could access external feedback after engaging in practice.

There is a huge literature on the positive effects of feedback on learning (e.g., Clark, 2012; Evans, 2013; Fletcher, 2018; Shute, 2008; van der Kleij, Feskens, & Eggen, 2015). There is an important distinction between formative and summative feedback and between various types of feedback. In the present study, the feedback was formative, meaning that students could use it during training to improve their learning. Generally, three main types of feedback are distinguished (Shute, 2008). The first and most basic form is knowledge of results, which informs the student about the correctness of his or her answer. The second type is knowledge of the correct answer, in which the feedback provides the right answer to the student. The third type, elaborated feedback, can entail a variety of complementary information. For this type, the distinction between information about the correct answer and (new) instruction is blurred. Elaborated feedback may, among others things, consist of a worked example, a description of the correct solution steps and a discussion of prevalent mistakes for the task or problem. Elaborated feedback is considered to be the most effective type (Shute, 2008; van der Kleij et al., 2015). The present study offered elaborated external feedback. It was chosen over other forms of feedback because it was deemed best for stimulating student reflection and deeper processing of task practice.

In software training, users must acquaint themselves with the handling of the program. In such a training direct instructions on how to accomplish tasks tend to be followed by moments of practice to enhance learning. The interface scaffolds practice as it provides users with signs of task progression. There is considerable internal feedback (Bandura, 1986) that can help users in finding the right trajectory through the interface. In other words, users are called upon more for their recognition of the proper sequence of menu options, than for unprompted recall of a series of steps in a solution path.

For this reason it makes sense to investigate whether practice without feedback works in software training. But there are also practical considerations. If internal feedback already does a proper job, or users can replay videos when needed, designers save time and effort by relying on that feedback alone. Indeed, software makers such as Adobe and TechSmith that accommodate practice with their training videos do not provide external feedback presumably for these reasons. However, these companies may also do so because it is not self-evident what external feedback should look like if it must be different from the demonstration video. Another reason why one might refrain from giving external feedback to practice, is that it can help in preventing users from being seduced into short-circuiting their activity and thinking. That is, adjunct question research points out that the provision of external feedback may induce more passive processing (Hamaker, 1986; Roelle, Rahimkhani-Sagvand, & Berthold, 2017). The students in the present study could have refrained from active practice and simply requested feedback. The findings tentatively suggest that this did not happen.

7. Conclusion

The findings from the present study give a promising prospect for the effectiveness of video-based learning of statistics. There was a high level of engagement with the instructional videos, and lower but still substantial engagement with a review video. Also, the students made extensive use of the opportunity to practice during training. Furthermore, the results after training were good in the control condition, and even better in the experimental conditions. We believe that these findings attest to the general design qualities of

the videos that were constructed, and to the practice arrangement. As such, this study extends a growing set of studies that support the effectiveness of the DBT approach for design of video-based software training (e.g., Brar & van der Meij, 2017; van der Meij, 2017).

Both manipulated features yielded a significant effect on one or more of the dependent variables. The review positively affected appraisals for self-efficacy, as did practice. Practice also yielded higher scores on a performance test after training. This extends previous research on software training that failed to find such an effect (van der Meij, 2018; van der Meij, Rensink, et al., 2018). The present study tentatively indicates that practice in software training benefits from a specific (i.e., mixed) arrangement in combination with external feedback. More research is needed to disentangle the contribution of either factor in learning. A noteworthy characteristic of both features is that they do not require a redesign of the existing video. The design of a review video and an arrangement for practice are possible while leaving existing instructional videos intact.

A limitation of the study is that there was no pre-test. A pre-test was omitted because we wanted to avoid the risk of an interaction effect of testing and treatment (Campbell & Stanley, 1963). That is, we were afraid that a pre-test would predispose the participants to selective viewing, akin to the effect found for adjunct questions before text reading. As a result, the study could yield only information about the relative effectiveness of conditions.

In short, the present study invites further research on the effectiveness of reviews and practice in video-based statistics training, and it invites a more frequent application in existing flipped classroom designs for statistics courses. The suggestion to include practice with external feedback, and possibly also course credit, accords with a recent recommedation for mathematics flipped classrooms (Lo, Hew, & Chen, 2017).

Conflicts of interest

The authors declare that they have *no conflict of interest*.

Acknowledgements

Compliance with *Ethical Standards*: A research proposal describing the study has been approved by the ethics committee of the University. Participants gave informed consent.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compedu.2019.103665.

References

Abel, M., & Roediger, H. L. (2017). Comparing the testing effect under blocked and mixed practice: The mnemonic benefits of retrieval practice are not affected by practice format. *Memory & Cognition, 45*(1), 81–92. https://doi.org/10.3758/s13421-016-0641-8.

Aro, T., Viholainen, H., Koponen, T., Peurac, P., Räikkönen, E., Salmi, P., et al. (2018). Can reading fluency and self-efficacy of reading fluency be enhanced with an intervention targeting the sources of self-efficacy? *Learning and Individual Differences, 67*, 53–66. https://doi.org/10.1016/j.lindif.2018.06.009.

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. https://doi.org/10.3102/00346543070002181.

Baglin, J., & Da Costa, C. (2014). How do students learn statistical packages? A qualitative study. In H. MacGillivray, B. Phillips, & M. Martin (Vol. Eds.), *Topics from Australian conferences on teaching statistics springer proceedings in mathematics & statistics: Vol. 81*, (pp. 169–187). New York, NY: Springer.

Bandura, A. (1986). *Social foundations of thought and actions: A social cognitive theory*. Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (1997). *Self-efficacy. The exercise of control.* New York, NY: Freeman and Company.

Beatson, N. J., Berg, D. A. G., & Smith, J. K. (2018). The impact of mastery feedback on undergraduate students' self-efficacy beliefs. *Studies In Educational Evaluation, 59*, 58–66. https://doi.org/10.1016/j.stueduc.2018.03.002.

Bishop, J. L., & Verleger, M. A. (2013). The flipped classroom: A survey of the research. *Paper presented at the 120th American society of engineering education (ASEE) annual conference & exposition, atlanta, GA*.

Brady, F. (2008). Contextual interference effect and sport skills. *Perceptual & Motor Skills, 106*, 461–472.

Brar, J., & van der Meij, H. (2017). Complex software training: Harnessing and optimizing video instructions. *Computers in Human Behavior, 70*, 1–11. https://doi.org/10.1016/j.chb.2017.01.014.

Brewer, G. A., Marsh, R. L., Meeks, J. T., Clark-Foos, A., & Hicks, J. L. (2010). The effects of free recall testing on subsequent source memory. *Memory, 18*(4), 385–393. https://doi.org/10.1080/09658211003702163.

Broadbent, D. P., Causer, J., Williams, A. M., & Ford, P. R. (2017). The role of error processing in the contextual interference effect during the training of perceptual-cognitive skills. *Journal of Experimental Psychology: Human Perception and Performance, 43*(7), 1329–1342. https://doi.org/10.1037/xhp0000375.

Burgoyne, S., & Eaton, J. (2018). The partially flipped classroom: The effects of flipping a module on “Junk Science” in a large methods course. *Teaching of Psychology, 45*(2), 154–157. https://doi.org/10.1177/0098628318762894.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago, IL: Rand McNally & Company.

Chew, P. K. H., & Dillon, D. B. (2014). Statistics anxiety update: Refining the construct and recommendations for a new research agenda. *Perspectives on Psychological Science, 9*(2), 196–208. https://doi.org/10.1177/1745691613518077.

Clark, I. (2012). Formative assessment; Assessment is for self-regulated learning. *Educational Psychology Review, 24*(2), 205–249. https://doi.org/10.1007/s10648-011-9191-6.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155.

Dimotakis, N., Mitchell, D., & Maurer, T. (2017). Positive and negative assessment center feedback in relation to development self-efficacy, feedback seeking, and promotion. *Journal of Applied Psychology, 102*(11), 1514–1527. https://doi.org/10.1037/apl0000228.

Duncan, T. G., & McKeachie, W. J. (2005). The making of the motivated strategies for learning questionnaire. *Educational Psychologist, 40*(2), 117–128.

Eccles, J. S., & Wigfield, A. (2002). Motivation beliefs, values, and goals. *Annual Review of Psychology, 53*(1), 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153.

Ertelt, A. (2007). *On-screen videos as an effective learning tool. The effect of instructional design variants and practice on learning achievements, retention, transfer, and motivation*Unpublished doctoral dissertation. Freiburg, Germany: Albert-Ludwigs Universität.
Evans, C. (2013). Making sense of assessment feedback in higher education. *Review of Educational Research, 83*(1), 70–120. https://doi.org/10.3102/0034654312474350.
Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). London, UK: Sage.
Fiorella, L., & Mayer, R. E. (2018). What works and doesn't work with instructional video. *Computers in Human Behavior, 89*, 465–470. https://doi.org/10.1016/j.chb.2018.07.015.
Fletcher, A. K. (2018). Help seeking: Agentic learners initiating feedback. *Educational Review, 70*(4), 389–408. https://doi.org/10.1080/00131911.2017.1340871.
González, A., Rodríguez, Y., Faílde, J. M., & Carrera, M. V. (2016). Anxiety in the statistics class: Structural relations with self-concept, intrinsic value, and engagement in two samples of undergraduates. *Learning and Individual Differences, 45*, 214–221. https://doi.org/10.1016/j.lindif.2015.12.019.
Google (2018). *YouTube Analytics.* Retrieved August 2019 from https://creatoracademy.youtube.com/page/lesson/engagement-analytics.
Grossman, R., Salas, E., Pavlas, D., & Rosen, M. A. (2013). Using instructional features to enhance demonstration-based training in management education. *The Academy of Management Learning and Education, 12*(2), 219–243. https://doi.org/10.5465/amle.2011.0527.
Guo, P. J., Kim, J., & Rubin, R. (2014). March). *How video production affects student engagement: An empirical study of MOOC videos. Paper presented at the L@S '14, atlanta, GA*.
Hamaker, C. (1986). The effects of adjunct questions on prose learning. *Review of Educational Research, 56*(2), 212–242. https://doi.org/10.2307/1170376.
Hartley, J., & Trueman, M. (1982). The effects of summaries on the recall of information from prose: Five experimental studies. *Human Learning, 1*, 63–82.
Helsdingen, A., Van Gog, T., & Van Merriënboer, J. J. G. (2011). The effects of practice schedule and critical thinking prompts on learning and transfer of a complex judgment task. *Journal of Educational Psychology, 103*(2), 383–398. https://doi.org/10.1037/a0022370.
Heuett, W. J. (2017). Flipping the math classroom for non-math majors to enrich their learning experience. *Primus, 27*(10), 889–907. https://doi.org/10.1080/10511970.2016.1256925.
Hodges, N. J., & Coppola, T. (2015). What we think we learn from watching others: The moderating role of ability on perceptions of learning from observation. *Psychological Research, 79*, 609–620. https://doi.org/10.1007/s00426-014-0588-y.
Hood, M., Creed, P. A., & Neumann, D. L. (2012). Using the expectancy value model of motivation to understand the relationship between student attitudes and achievement in statistics. *Statistics Education Research Journal, 11*(2), 72–85.
Kim, J., Guo, P. J., Seaton, D. T., Mitros, P., Gajos, K. Z., & Miller, R. C. (2014). Understanding in-video dropouts and interaction peaks in online lecture videos. *Paper presented at the L@S '14, Atlanta, GA* March 4-5.
Kosslyn, S. M., Kievit, R. A., Russell, A. G., & Shephard, J. M. (2012). PowerPoint presentation flaws and failures: A psychological analysis. *Frontiers in Psychology, 3*, 1–22. https://doi.org/10.3389/fpsyg.2012.00230.
Lo, C. K., Hew, K. F., & Chen, G. (2017). Toward a set of design principles for mathematics flipped classrooms: A synthesis of research in mathematics education. *Educational Research Review, 22*, 50–73. https://doi.org/10.1016/j.edurev.2017.08.002.
van der Meij, H. (2007). Goal-orientation, goal-setting and goal-driven behavior in (minimalist) user instructions. *IEEE Transactions on Professional Communications, 50*(4), 295–305.
Merbah, S., & Meulemans, T. (2011). Learning a motor skill: Effects of blocked versus random practice a review. *Psychologica Belgica, 51*(1), 15–48. https://doi.org/10.5334/pb-51-1-15.
Merrill, M. D. (2002). First principles of instruction. *Educational Technology Research & Development, 50*(3), 43–59. https://doi.org/10.1007/BF02505024.
Payne, J., & Israel, N. (2010). Beyond teaching practice: Exploring individual determinants of student performance on a research skills module. *Learning and Individual Differences, 20*, 260–264. https://doi.org/10.1016/j.lindif.2010.02.005.
Peterson, D. J. (2016). The flipped classroom improves student achievement and course satisfaction in a statistics course: A quasi-experimental study. *Teaching of Psychology, 43*(1), 10–15. https://doi.org/10.1177/0098628315620063.
Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education. Theory, research, and applications* (2nd ed.). Upper Saddle River, NJ: Merrill Prentice Hall.
Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the motivated Strategies for learning questionnaire (MSLQ).* Ann Arbor, MI: University of Michigan. National Center for Research to Improve Postsecondary Teaching and Learning.
Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ). *Educational and Psychological Measurement, 53*, 801–813. https://doi.org/10.1177/0013164493053003024.
Renkl, A. (2014a). Toward an instructionally oriented theory of example-based learning. *Cognitive Science, 38*, 1–37. https://doi.org/10.1111/cogs.12086.
Renkl, A. (2014b). The worked examples principle in multimedia learning. In R. E. Mayer (Ed.). *The Cambridge handbook of multimedia learning* (pp. 391–412). (2nd ed.). New York, NY: Cambridge University Press.
Roelle, J., Rahimkhani-Sagvand, N., & Berthold, K. (2017). Detrimental effects of immediate explanation feedback. *European Journal of Psychology of Education, 32*, 367–384. https://doi.org/10.1007/s10212-016-0317-6.
Rosen, M. A., Salas, E., Pavlas, D., Jensen, R., Fu, D., & Lampton, D. (2010). Demonstration-based training: A review of instructional features. *Human Factors, 52*(5), 596–609. https://doi.org/10.1177/0018720810381071.
Ruggeri, K., Dempster, M., Hanna, D., & Cleary, C. (2008). Experiences and expectations: The real reason nobody likes stats. *Psychology Teaching Review, 14*(2), 75–83.
Salomon, G. (1984). Television is "easy" and print is "tough": The differential investment of mental effort in learning as a function of perceptions and attributions. *Journal of Educational Psychology, 76*(4), 647–658.
Shinaberger, L. (2017). Components of a flipped classroom influencing student success in an undergraduate business statistics course. *Journal of Statistics Education, 25*(3), 122–130. https://doi.org/10.1080/10691898.2017.1381056.
Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795.
Spanjers, I. A. E., van Gog, T., & van Merriënboer, J. J. G. (2012a). Segmentation of worked examples: Effects on cognitive load and learning. *Applied Cognitive Psychology, 26*, 352–358. https://doi.org/10.1002/acp.1832.
Spanjers, I. A. E., van Gog, T., Wouters, P., & van Merriënboer, J. J. G. (2012b). Explaining the segmentation effect in learning from animations: The role of pausing and temporal cueing. *Computers & Education, 59*, 274–280. https://doi.org/10.1016/j.compedu.2011.12.024.
van Dinther, M., Dochy, F., & Segers, M. (2011). Factors affecting students' self-efficacy in higher education. *Educatiional Research Review, 6*, 95–108. https://doi.org/10.1016/j.edurev.2010.10.003.
van Gog, T. (2011). Effects of identical example-problem and problem-example pairs on learning. *Computers & Education, 57*, 1775–1779. https://doi.org/10.1016/j.compedu.2011.03.019.
van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research, 85*(4), 475–511. https://doi.org/10.3102/0034654314564881.
van der Meij, H. (2017). Reviews in instructional video. *Computers & Education, 114*, 164–174. https://doi.org/10.1016/j.compedu.2017.07.002.
van der Meij, H. (2018). Cognitive and motivational effects of practice with videos for software training. *Technical Communication, 65*(3), 265–279.
van der Meij, H., & Carroll, J. M. (1998). Principles and heuristics for designing minimalist instruction. In J. M. Carroll (Ed.). *Minimalism beyond the nurnberg funnel* (pp. 19–53). Cambridge, MA: MIT Press.
van der Meij, H., Rensink, I., & van der Meij, J. (2018a). Effects of practice with videos for software training. *Computers in Human Behavior, 88*, 439–445. https://doi.org/10.1016/j.chb.2017.11.029.
van der Meij, H., & van der Meij, J. (2016a). The effects of reviews in video tutorials. *Journal of Computer Assisted Learning, 32*, 332–344. https://doi.org/10.1111/jcal.12136.
van der Meij, H., & van der Meij, J. (2016b). Demonstration-Based Training (DBT) for the design of a video tutorial for software instructions. *Instructional Science, 44*, 527–542. https://doi.org/10.1007/s11251-016-9394-9.
van der Meij, H., van der Meij, J., Voerman, T., & Duipmans, E. (2018b). Supporting motivation, task performance and retention in video tutorials for software

training. *Educational Technology Research & Development, 66*(3), 597–614. https://doi.org/10.1007/s11423-017-9560-z.
van Merriënboer, J. J. G., & Kester, L. (2014). The four-component instructional design model: Multimeda principles in environments for complex learning. In R. E. Mayer (Ed.). *The Cambridge handbook of multimedia learning* (pp. 104–148). (2nd ed.). New York, NY: Cambridge University Press.
Wilson, S. R. (2014). The flipped class: A method to address the challenges of an undergraduate statistics course. *Teaching of Psychology, 40*(3), 193–199. https://doi.org/10.1177/0098628313487461.
Winquist, J. R., & Carlson, K. A. (2014). Flipped statistics class results: Better performance than lecture over one year later. *Journal of Statistics Education, 22*(3), 1–10.
Yue, C. L., Bjork, E. L., & Bjork, R. A. (2013). Reducing verbal redundancy in multimedia learning: An undesired desirable difficulty? *Journal of Educational Psychology, 105*(2), 266–277. https://doi.org/10.1037/a0031971.