Education Tech Research Dev (2018) 66:1101–1118
https://doi.org/10.1007/s11423-018-9579-9

RESEARCH ARTICLE

Learning from failure: a meta-analysis of the empirical studies

Aubteen Darabi[1] · **Thomas Logan Arrington**[1] · **Erkan Sayilir**[1]

Published online: 22 February 2018

Abstract The authors searched five scholarly databases for a decade of research publications examining learning from failure as an instructional strategy. Out of 187 publications, 62 were found to be relevant to the topic from which only 12 used experimental design to examine the issue and reported statistics appropriate for meta-analysis. The studies also represented only two of our search domains-productive failure and failure-driven memory. The small number of experimental studies on this topic is a telling indication of the state of experimental research in this area. However, they revealed a moderately positive result for the effect of learning from failure. An examination of moderating variables indicated that participants' grade level, subject matter domain, and study's duration, while not significant in explaining the differences across the examined studies, showed positive medium effect sizes. Instructional design implications of our findings and limitations of the study are discussed.

Keywords Learning from failure · Failure-based instructional strategy · Meta-analysis · Learning strategy

Effective methods of instruction and training that lead to learners' productive performance have always been the topic of interest among education scholars. Over a century ago, Morgan (1894), introduced the concept of "trial and error" as a method of learning. Since then, the topic has been explored in different domains in search of efficient methods for facilitating novice learners' promotion to proficient and perhaps expert performers. In recent decades, scholars like Brown et al. (1989) suggested authentic methods and situated

✉ Aubteen Darabi
adarabi@fsu.edu

[1] Department of Educational Psychology and Learning Systems, College of Education, Florida State University, 1114 W. Call Street, Tallahassee, FL 32306-4450, USA

learning where learners face ill-structured problems and learn from the challenges of a real performance environment. Zimmerman (1989) introduced self-regulation method as a meta-cognitive skill that learners can use to learn from their failures. In expertise research, Ericsson et al. (1993) introduced the concept of "deliberate practice" which they suggested will turn a "novice" into an "expert" over time as they recognize their errors and overcome them through deliberate practice. The common theme among these methods, one can argue, is the learners' reflection on their errors, correcting them, and learning from their experience.

The recent advances in instructional design have minimized the learners' errors due to the "extraneous" features of instruction and facilitated their concentration on the "germane" part of learning (Sweller 1994; Paas et al. 2003). This distinction seems to have gone unnoticed by scholars studying learning from failure who argue that the purpose of the instructional design is to eliminate learners' errors. For instance, Kapur (2008) extend this argument to the point that he describes the goal of instructional design as to prevent learners' failure.

The fact is that over the decades, instructional designers have been concentrating on facilitating learners' achievement of learning objectives through formulating design of instruction to minimize the learners' cognitive load caused by confusion and misunderstanding (e.g., Dick et al. 2014; van Merriënboer et al. 2003). These scholars emphasized presentation of instruction in small learning increments and carefully sequenced instructional activities for learners' accomplishment of objectives. They further differentiated among types of cognitive load emanating from instruction (see Sweller 1994; van Merriënboer et al. 2003; Paas et al. 2003; Kalyuga 2011) and suggested that the purpose of instructional design is to eliminate the extraneous load and mediate the germane load of instruction for better accomplishments of learning objectives.

We believe that statements such as Kapur's (2008) that design approaches potentially limit room for learners' errors or prevents learners' failure, confuses the extraneous errors caused by the poor design of instruction with the errors that might be due to germane load of learning tasks or the errors purposefully imbedded in the instruction as instructional strategies. In our view, the latter type of errors are the concerns of learning from failure strategies. We contend that a robust design of instruction that expands the learners' ability to focus on their solutions to the germane challenges of instruction, is a complement to learning from failure strategies.

According to Kapur and Rummel (2012) structured instructional strategies typically produce four types of outcomes: *productive success*, *productive failure*, *unproductive success* and *unproductive failure:*

1. Productive success occurs when short-term performance results in long-term learning. This outcome has been examined through most research in learning and performance. Scafolded practice items imbedded in the performance of an instructional task could be an example of productive success. Here, the learners solve problems within their level of understanding, through some scaffolding. The scaffolding is there to ensure that the learners succeed in application of the instruction. Thus, the learners are succeeding in short-term performance (i.e., task or part-task practices) aimed at long-term learning.
2. Productive failure refers to a learning instance in which short-term performance is not successful, but leads to long-term learning. Unlike other strategies for learning from

failure, the productive failure strategy (Kapur 2008) requires the learners to solve unscaffolded-complex problems *prior* to receiving instruction to consolidate the presented learning concepts (Kapur 2016).

3. The other two approaches, unproductive success and unproductive failure allude to the short-term success or failure in performance where there are no benefits of sustained learning.

In search of a common thread among learning from failure research, based on which they can create a "unified model of failure", Tawfik et al. (2015) identified five prominent failure theories: *cognitive disequilibrium*, *impasse-driven learning*, *productive failure*, *failure-driven memory*, and *negative knowledge*. Tawfik et al. (2015) introduced these theoretical perspectives that refer to the same learning experience in which, "…failure generates an additional inquiry process at the point of failure that may not exist during a successful experience" (p. 977).

Given this discussion, we decided to search for studies conducted in these areas. This expanded the scope of our meta-analysis by including studies using instructional strategies concerned with failure as part of the learning condition. Following is a brief discussion of each of these theoretical perspectives.

Literature Review

Cognitive disequilibrium

Much of the theoretical basis for learning from failure is based on Piaget's cognitive disequilibrium. Cognitive disequilibrium occurs when learners encounter a situation contrary to their current mental model. Learners will be challenged until they either assimilate those differences into their mental model or modify their mental model according to the new situation (Piaget 1952). These challenges can come in various forms such as "…deviations from the norms, obstacles to goals, interruptions of action sequences, contradictions, anomalous information, unexpected feedback, and other forms of uncertainty" (D'Mello and Graesser 2014, p. 106). Disequilibrium is a foundational state of learning as it forces an individual to realize their lack of understanding (Piaget 1977; Tawfik et al. 2015).

Impasse-driven learning

Impasse-driven learning is a theory that explains how students learn from failure when solving procedural problems. This theory builds upon the notion of Piaget, by specifically looking at how an impasse is beneficial for students to learn new information (Tawfik et al. 2015). According to VanLehn (1988), when solving a problem learners encounter an impasse caused by a lack of knowledge or skills or a level of uncertainty. VanLehn notes two options for the learners to attempt overcome the impasse: repairing or seeking help. Repairing occurs when learners decide to solve the impasse on their own, without any outside support from experts or resources (e.g., texts, tutors, or teachers). When repairing, if learners find the correct solution there will usually be "bugs." These bugs prevent the learners from appropriately encoding the information, which can lead to problems in transferring knowledge. The other option allows learners to seek help identify the correct procedure to follow and overcome the impasse. The help can come in various forms, but

according to VanLehn, it can lead to more accurate performance. However, when learners possess relatable knowledge or skills, they are more likely to recognize and overcome these impasses without help and without creating bugs (Blumberg et al. 2008).

D'Mello et al. (2014) introduced confusion as a means of creating an impasse for learners, though in the same element these confusion instances are creating a cognitive disequilibrium. Those learners that overcame the impasse, or confusion, learned more than those who did not (D'Mello et al. 2014; Lehman et al. 2012). VanLehn et al. (2003) found that students at an impasse were more likely to learn in a tutoring situation than those not facing an impasse. Though, these impasses can be considered as productive or unproductive, as learning is not guaranteed (D'Mello and Graesser 2014).

Productive failure

Kapur (2008) notes that productive failure was built upon research in the areas of impasse-driven learning and cognitive disequilibrium. As an instructional strategy, productive failure focuses on learners solving problems, usually ill-structured or complex, prior to instruction. This method allows for learners to develop multiple problem representations and solutions before receiving some type of reconciliation or consolidation of responses. In a sense, productive failure as a teaching strategy asks students to invent multiple solution to a difficult learning task preceding the instruction for that task. Kapur (2008) found that, those learners who attempt to solve ill-structured problems, even though not outperforming their counterparts solving well-structured problems in their initial solutions, outperformed their counterparts in future application of both well-structured and ill-structured problems. The opportunity to generate solutions to unknown problems was beneficial in preparing the learners for the content of instruction that followed their problem experience.

Productive failure treats the problems encountered by learners as impasses they must overcome or reconcile. In this strategy, the reconciliation or consolidation period is in the form of instructor-led instruction and discussion (Kapur 2014). The instruction has varied in different studies but most recently instruction is shown to focus on common solutions generated in the problem-solving phase comparing those to a canonical solution (Kapur 2013, 2014; Loibl and Rummel 2014b).

The productive failure approach has been compared against solving well-structured problems (Kapur 2008; Kapur and Kinzer 2009), direct instruction or lecture and practice (Kapur 2009, 2010, 2012, 2014; Loibl and Rummel 2014a, b; Westermann and Rummel 2012), facilitated problem solving (Kapur 2010; Loibl and Rummel 2014a), worked examples (Glogger-Frey et al. 2015) and solution evaluation (Kapur 2013, 2014) with most results showing it as a favorable approach towards learning, specifically higher level learning.

Failure-driven memory

Failure-driven memory stems from case-based reasoning (Tawfik et al. 2015). Schank (1999) argued that failure experiences were just as important to success experiences when modifying one's script. Failure experiences occur when learners' expectations are not met. Based on these failure experiences, learners learn to predict failure taking certain actions and avoid failure by taking a different route. However, just failing does not guarantee learners' scripts will change, instead learners must be able to explain and understand why failure occurred (Schank 1999) by deliberately evaluating their own performance, practicing for success (Ericsson et al. 1993), or using self-regulation skill (Zimmerman 1989).

Failure experiences do not necessarily have to come from individual experience to be beneficial as demonstrated by Tawfik and Jonassen (2013). Learners can learn vicariously from the cases of failures of others dealing with a very similar problem within the same domain. The researchers found that students using failure cases performed better than those accessing success cases in an argumentation task. The failure-case group performed better overall, including with better rebuttal and counterclaim scores. However, as Tawfik et al. (2015) notes, the empirical evidence available for this approach is lacking.

Negative knowledge

Negative knowledge is simply defined by Gartmeier et al. (2008) as non-viable knowledge. Gartmeier et al. (2008) continue by highlighting that "…non-viable knowledge is knowledge that somehow stands in contradiction to prior knowledge or is counterproductive regarding a certain goal" (p. 89). Non-viable negative knowledge is beneficial as it focuses learners to see the broader picture to avoid making future errors and to do so with relative confidence (Tawfik et al. 2015).

The impact of negative knowledge was further investigated by Gartmeier et al. (2011) in the career of elderly-care nurses. These researchers found that the most experienced group of nurses had the most specific negative knowledge. They proposed that this was potentially the cause of their ability to avoid errors and perform well. However, Gartmeier et al. (2008, 2011) studies are focused on experiential and workplace learning. Heemsoth and Heinze (2013), focused on classroom learning when exploring negative knowledge. They found that students learning from incorrect examples generated more negative knowledge than those learning from correct examples. This negative knowledge led to improved performance as learners had a more comprehensive understanding of the topic.

Given these relationships, we designed a meta-analysis study to examine the empirical research conducted on these topics for a period of 10 years. Our purpose was to review this body of research and document whether the learners' failure in its different forms contributes to their learning from that experience. In the following sections, we present the methodology and the results of the analysis and continue to discuss our findings. We conclude by highlighting the implications of our findings for designing instructional strategies.

Methodology

Data collection

Given their familiarity with the relevant literature, the authors met in a discussion session to identify the searching, screening, and selecting criteria for the investigation. Considering the salient topics of learning from failure represented in the instructional model synthesized by Tawfik et al. (2015), the authors decided on a search procedure to include the following keywords: *cognitive disequilibrium, impasse-driven learning, productive failure,* and *failure-driven memory,* and *negative knowledge*. We agreed to use Tawfik et al. (2015) comprehensive set of keywords, which seemed to be constructed based on the prior literature concerning the concepts that lead to research on learning from failure.

We realized that most scholarly investigations looking at *learning from failure* as an instructional strategy appeared in recent years, even though the root sources of this

construct were introduced long before (i.e., cognitive disequilibrium and impasse-driven learning). For this reason, we decided to search only for studies that concentrated on this construct in the literature during the past 10 years.

The authors identified five scholarly databases (i.e., *ERIC, PsychInfo, Google Scholar, Web of Science,* and *Dissertation Abstracts*) to search for the period of 2005–2015 using the identified keywords. All the publications drawn from these five databases were written in English. Besides the authors, a graduate student and a research associate were recruited for assistance with the search. They were trained in an orientation session by the leading author and instructed about the studies' purpose and the search criteria.

The studies identified in our search covered a variety of domains such as: mathematics (Kapur 2012; Loibl and Rummel 2014a, b) science (Kapur 2009; Pathak et al. 2008), basic device interaction (D'Mello et al. 2014) or other complex fields such as business, education, or nursing (Gartmeier et al. 2008; Glogger-Frey et al. 2015; Tawfik and Jonassen 2013). Many of the articles took place across varying time frames and locations including: Singapore, India, Canada, Germany, and the United States of America. Lastly, the educational level of participants ranged from junior high to professionals. The enumerated variation in the studies provided us with a general framework for understanding the key concepts to be coded and examined in these articles. From our point of view, these codes mediate the effectiveness of learning from failure based on which the framework of this meta-analysis was built.

Our search resulted in 187 studies across the five databases conducted within the 10-year period. A copy of each publication was saved on a Blackboard organization page to be accessed by the reviewing team. We initially coded the 187 publications with the following identifiers in preparation for our review process:

- Publication author(s)
- First author's affiliate institution
- First author's country
- Year published
- Publication title
- Type of publication (e.g. article, dissertation, book chapter, presentation, etc.)
- Publication/presentation venue

Inclusion and exclusion criteria

The authors reviewed the 187 publications using the first criteria demonstrated in Fig. 1. The first selection criterion was to determine whether the publications were directly examining learning from failure. In this review, for a publication to be selected for further examination, it should have included an examination of *a clear learning experience where learners were likely to fail, encounter failure of their own, or be exposed to failure of others and use that experience for better learning performance.* Through this process, 62 publications relevant to failure-based learning were identified. The remaining 125 publications were picked up in our search because they had mentioned one or more of the keywords in their title, abstract, or in their review of literature but they were not specifically examining our topic of interest. The authors conducting this review and selection process, discussed the cases individually and came to full agreement on selecting the 62 publications for further analysis.

In the second review and selection round, the same two authors applied our second selection criteria—*whether the studies compared a failure condition with a control*

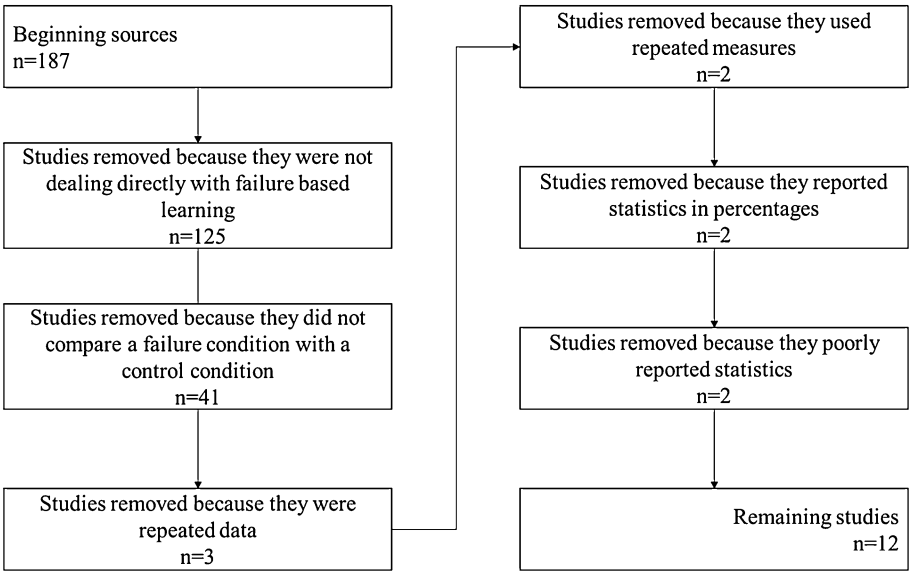

Fig. 1 Inclusion and exclusion flowchart. A visual depiction of the studies removed and their reason for removal

condition (i.e., a treatment where failure was not expected or where the process is scaffolded to prevent failure). This review resulted in identifying 17 publications and four presentation manuscripts that included group comparison or within-subjects design. However, we found three of the presentation manuscripts to have been already included in three of the identified published articles and thus they were dropped from our analysis. This left us with a total of 18 publications that basically represented the five research domains of learning from failure as our search keywords indicated.

Our next step was to scrutinize the selected studies in terms of specifics of their research design, group comparison procedures, and statistical analysis. This time six of the studies were found not to fit the parameters set for our investigation and had to be excluded. Two of the studies used repeated measures of learning from failure strategy versus conventional strategy within the same group of subjects that led to the fluctuation of effect sizes due to the brief time between the strategies. Another pair of articles reported post-test scores and standard deviations in percentages which would have slanted the results due to extremely high variability had we included them in our analysis. Lastly, another two articles did not provide sufficient information in the sense that the stats were poorly reported (i.e., some comparisons were mentioned as non-significant but not reported in the article).

Because of our detailed and systematic screening and scrutiny looking for the best-fit empirical publication, we selected 12 articles that used an experimental design to measure learning from failure and reported statistics that lend themselves to the analytical purpose of this study. We calculated Cohen's kappa (κ) to examine the authors' agreement on inclusion and exclusion of the publications. The test yielded 0.89 which indicates strong agreement among the authors.

Coding procedures

Meta-analysis methodology typically requires establishing a framework for identifying the key characteristics of the topic of study to be coded. As discussed above, selecting and categorizing the publications provided us with information that we needed to use in conducting our analysis. We used that information to create a database and store the design characteristics and the reported statistics of the 12 studies. The lead author conducted a short coding orientation and practice meeting for the two coauthors who then proceeded to retrieve and code the following information from each article to be stored in our data base:

- Means, and standard deviations for the experimental and control groups
- Number of subjects in treatment group
- Number of subjects in control group
- Total number of subjects
- Study's treatment instructional strategy
- Study's reported outcome
- Study's research design
- Sampling method
- Group assignment method
- Study's lead author
- Study's location by country
- Participants' grade level
- Study's subject matter
- Study's duration

Indicators of study quality (Valentine and Cooper 2008) such as sampling and group assignment methods were included in the list even though quality of study was not used as a moderator due to low variation among the 12 studies on these variables. The publications reported mostly a convenience sampling method and group assignments were mostly based on the available grouping (i.e., quasi-experimental). These studies represented only two of the domains, productive failure and failure-driven memory, that we searched as keywords (see Table 1).

Average interrater agreement among coders was calculated at 93%. Disagreements between coders were resolved by discussion between coders and consultation with the project lead when needed. At the completion of coding process, we reviewed the coded variables and identified six of them that could differentiate among the studies in terms of their findings in line with our stated purpose. Table 2 presents a list of these "moderating" variables and their coding values as entered in the database.

Table 1 Selected publications and domains they represented

Research domain/keyword	Publications
Productive failure	Glogger-Frey et al. (2015) Kapur (2009, 2010, 2012, 2014) Kapur and Bielaczyc (2012) Kapur and Lee (2009) Loibl and Rummel (2014a, b) Pathak et al. (2008) Westermann and Rummel (2012)
Failure-driven memory	Tawfik and Jonassen (2013)

Effect size calculation and analysis

We used means, standard deviations, and sample sizes for the experimental and control groups to calculate the Cohen's d (Cohen 1988). Specifically, we divided the post-test mean score differences between the experimental and control groups by the pooled standard deviation

$$d = \frac{\bar{x}_T - \bar{x}_C}{s_p}$$

where $\bar{x}_T$ is the mean of the treatment group, $\bar{x}_C$ is the mean of the control group, and s_p is the pooled standard deviation obtained as follows:

$$s_p = \sqrt{\frac{(n_T - 1)s_T^2 + (n_C - 1)s_C^2)}{n_T + n_C}}$$

Here, n_T and n_C represent sample sizes, and s_T and s_C are standard deviations for the treatment and control groups, respectively. To avoid the small sample bias, we then transformed Cohen's d to Hedges' g effect size (Hedges 1981) where a positive standardized mean difference favors the "learning from failure" strategy and negative one favors the control strategies.

For the studies that provided multiple outcomes either because of various outcome measures or multiple comparisons with the same control group, we computed variance of the composite score and combined the effect sizes across outcomes into one to ensure the independence of effect sizes for overall analysis. Studies that reported effect sizes from independent groups were treated as independent (Borenstein et al. 2009) resulting in multiple effect sizes for some of the articles.

Data analysis

Given the calculated effect sizes, we used the Comprehensive Meta-Analysis software, version 2 (Borenstein et al. 2005) and R studio to do the following:

1. Calculate standardized mean differences and transform them into Hedge's *g*.

Table 2 Moderating variables and their coded values

Variables	Code values
Study's location by country	1. Germany 2. Singapore 3. United States of America
Participants' grade level	1. Junior high (grades 6–8) 2. High (grades 9–12) 3. College (undergraduates)
Study's subject matter domain	1. Science/mathematics 2. Other (education, business)
Reported means	1. Pretest/posttest 2. Adjusted posttest
Study's duration	1 = within 1 week 2 = within more than 1 week
Study's lead author	Authors name

2. Test the homogeneity of standardized mean differences under random-effects model to examine the overall differences.
3. Draw funnel plot and forest plots and calculate all relevant statistics.
4. Conduct subgroup analysis with a mixed-effects model to explore the potential effects of moderators on the differences between learning from failure groups and control group.

Results modeling

Random-effects model assumes that each study differs from population effect by subject-level sampling error and a random component (Hedges and Olkin 1985; Lipsey and Wilson 2001). In mixed-effects model the errors term is treated as random and the moderators treated as fixed effects. In other words, the fixed-effects model is adopted across subgroups and the random-effects model is adopted within groups. In the present study, we used the random-effects model to explore the variability of the effect sizes and used the mixed-effect (similar to analysis of variance for categorical moderators) to examine our categorical moderating variables.

Homogeneity tests

We adopted the random-effects model to synthesize overall mean effect size. We examined Q test, a Chi square test of homogeneity of effect sizes (Hedges and Olkin 1985) and I^2 (Higgins et al. 2003) to assess the presence of heterogeneity. A significant Q test would indicate a significant variation of the effect size, and a non-significant Q test would indicate a non-significant variation (Borenstein et al. 2009).

Results

Our analysis of the 12 studies resulted in 23 effect sizes corresponding to varying research settings reported by these studies. For instance, Kapur and Lee (2009) presented results from three different settings within one study which resulted in three effect sizes. The average effect size across all the studies with 95% confidence intervals (g is 0.43; 95% CI is 0.19 to 0.68) is presented in Fig. 2. The average effect size, albeit modestly, indicates superiority of learning from failure compared to other instructional strategies used in the selected studies. The Q test shows statistically significant heterogeneity (Q (22) = 77.5, $p < 0.001$) which supported the choice of random-effects model. Also, the high heterogeneity ($I^2 = 73.5\%$; $T^2 = 0.22$) indicates that the variability across studies is beyond sampling error. It shows that 73.5% of the total variation in effect size is due to true between-study variability, rather than sampling error (Borenstein et al. 2009; Lipsey and Wilson 2001). To examine this high variability, we analyzed moderating variables.

Moderator analyses

Among the 6 identified moderating factors (see Table 2), *Study's location* contained the same information as the *Study's lead author* thus it was precluded from the final analysis. The five-remaining categorical moderating variables were examined for explaining the variability in effect sizes. We used a mixed-effects model by implementing subgroup

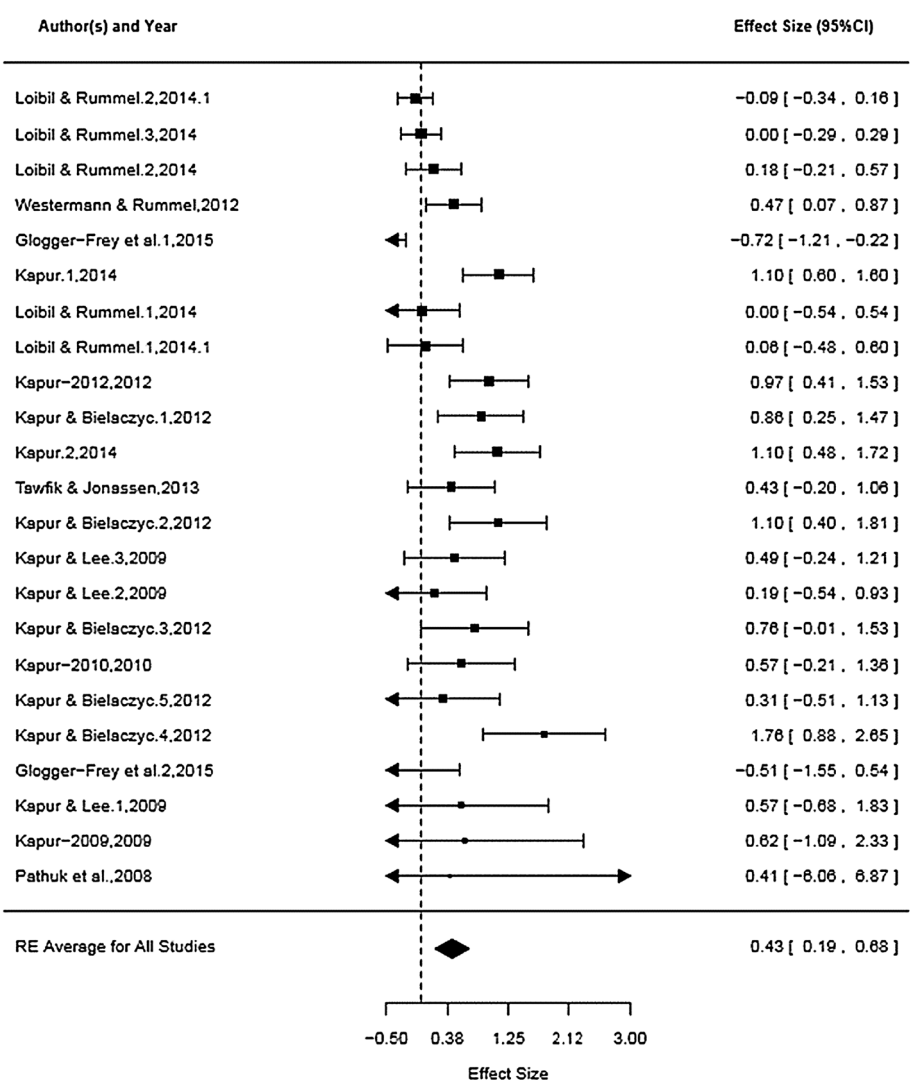

Fig. 2 Forest plot. The average effect size (g) across studies and their 95% confidence interval

analysis (treating the error as random and the moderators as fixed effects). This was followed by an investigation of the between-group heterogeneity (Q_B). The results of these analyses are displayed in Table 3 followed by a description of findings on moderating variables.

Education

Participants' education ranged from junior high school to college. Studies with subjects from junior high school and high school showed a higher effect ($g = 0.65$ and $g = 0.38$ respectively) compared to studies with college students that produced a smaller effect size

Table 3 Modeling results for overall and by moderating variables

Moderating variables	*n*	Effect size		95% confidence interval		Test of null	Heterogeneity		
		g	SE	LL	UL	*z*-value	Qb	*df*	*p*
Fixed-effects	23.00	0.28	0.06	0.17	0.39	4.89			
Random-effects	23.00	0.43	0.12	0.19	0.67	3.61			
Education									
Junior high	11.00	0.65	0.16	0.33	0.96	4.05			
High school	9.00	0.38	0.17	0.04	0.71	2.21	2.6	2.00	0.27
College	3.00	0.06	0.40	− 0.73	0.84	0.14			
Domain									
Science/math	20.00	0.50	0.13	0.25	0.75	3.95	1.1	1.00	0.29
Other	3.00	0.06	0.40	− 0.73	0.84	0.14			
Reported means									
Pre/post test	12.00	0.19	0.15	− 0.10	0.49	1.28	7.4	1.00	0.01
Adjusted post test	11.00	0.72	0.12	0.48	0.96	5.87			
Duration									
< 1 week	16.00	0.43	0.16	0.12	0.74	2.75	0.01	1.00	0.93
> 1 week	7.00	0.45	0.13	0.19	0.71	3.43			
Researcher									
Kapur	13.00	0.85	0.11	0.65	0.11	8.06	30.9	1.00	< 0.001
Others	10.00	0.02	0.11	− 0.20	0.23	0.14			

($g = 0.06$). The Q test showed no significance of between-level variance (Q_B (2) = 2.6, $p = 0.27$) indicating that subjects' education level is not a significant moderator in explaining the heterogeneity between the studies.

Subject matter domain

Studies' subject matter domain were coded into two categories of science and math versus other fields. We found a small effect size for the other fields category ($g = 0.06$) but the science and math category produced a relatively larger effect size ($g = 0.5$). However, the low Q value (Q_B (1) = 1.1, $p = 0.29$) of the variance between the domain categories suggesting there is not a substantial variability among studies in terms of this moderator.

Duration

Duration of the studies' treatments were reported in various units (e.g., sessions, lessons units, periods, weeks, etc.). We converted the studies duration into a dichotomous variable by determining whether the duration would fit into a week or would extend further. Effect size for the studies conducted within one-week period was moderate ($g = 0.44$) and for the studies that lasted more than a week was slightly higher ($g = 0.46$). Although both duration categories produced medium effect sizes, the between-level variance for this variable was not statistically significant (Q_B (1) = 0.01, $p = 0.93$).

Researcher

We created a dummy variable based on the study's lead author and used it as a moderator to investigate whether the effect of learning from failure was different in terms of who conducted the studies. Kapur's significant share of the selected studies and the number of effect sizes calculated for his studies (13 of the 23) provided a sensible base for comparing his studies with all others. Our examination showed a large mean effect sizes ($g = 0.86$) for Kapur's studies compared to other authors that had smaller mean effect size ($g = 0.02$) with a detectable between-level variance (Q_B (1) = 30.95, $p < 0.01$).

We noticed that Kapur reported adjusted mean scores for post-test scores where the pre-test scores were used as covariate for both control and treatment groups. Considering that this might be the reason for the difference between his studies and others, we created another moderator called "reported means" for further examination.

Reported means

We divided the studies into two categories, one that reported mean score for treatment and control groups and other that reported adjusted posttest scores for the treatment and control groups. Studies reported adjusted mean value for posttest score had a relatively large effect ($g = 0.72$) and those that reported unadjusted post-test scores had smaller effect ($g = 0.19$) with significant between-level variability (Q_B (1) = 7.4, $p < 0.01$). This finding supported our speculation about the difference between Kapur's studies and others discussed above.

Publication bias

Meta-analytic results were accompanied by the inspection of publication bias through visual inspection of the funnel plot, and the Duval and Tweedie (2000) "trim and fill" methods to find out if any "missing studies" were needed. In the present study, we reported a funnel plot (Fig. 3) which plots inverse standard error (SE) against Hedge's g. Studies with larger standard errors will be close to bottom of the plot and vice versa. The funnel plot of SE by Hedges' g appears to be asymmetric with an unequal proportion of studies in both sides of the plot suggesting that publication bias may exist within our selected studies due to the lack of sufficient number of studies with negative effect. Figure 3 displays the trim and fill results, which indicates an imbalance presented by eight imputed studies (white circles) on the left of the funnel plot. This means that our overall effect size ($g = 0.43$, 95% CI 0.17, 0.66), represented by the black circles, was reduced to 0.11 (− 0.14, 0.36) when considering the eight imputed studies. The results of the "trim and fill" method for this funnel plot (Fig. 3) detects a possible bias in the results of this analysis.

Discussion

The fact that out of 62 publications examining the learning-from-failure strategy we ended up with only 12 articles with experimental design, is a telling indication of the state of experimental research in this area. However, even though small in numbers, these studies still revealed a moderately positive result for the effect of learning from failure as an

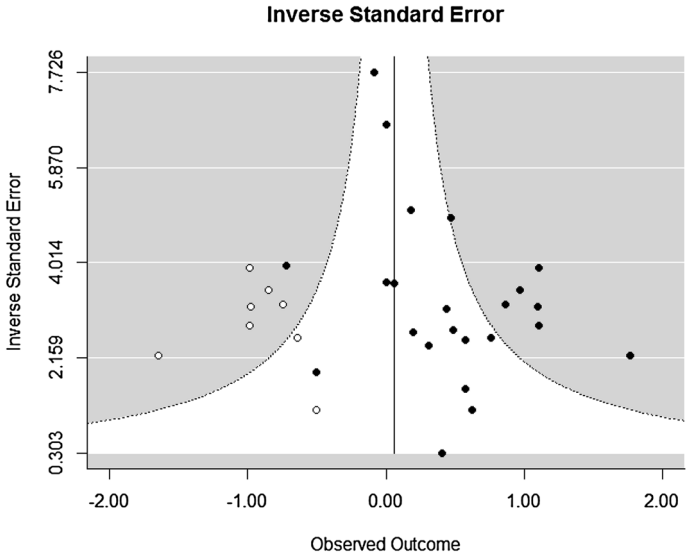

Fig. 3 Funnel plot. This figure depicts publication bias in our results

instructional strategy. Further test results revealed a substantial variability among studies which justified examining several moderating variables to explain the issue. The results of this analysis indicated that *participants' grade level*, *study's subject matter domain*, and *study's duration* had no significant role in the variability among the studies because the effect sizes were similar across the categories within these moderators.

However, these moderators yielded positive medium effect sizes (see Cohen 1988). Given these effect sizes, one can speculate that a well-designed experiment in this field should use sufficient time for application of the strategy in terms of *duration of instruction*. In the studies we examined, the duration did not exceed 2 weeks. We contend that longer exposure to this learning strategy might result in better gain.

In terms of subject matter domain, the majority of studies we examined dealt with math and science. One can speculate that these authors found math or science to be more appropriate domains for application of this strategy due to their problem-solving properties. Further research in areas other than math and science will provide information on strategy's effectiveness across subject domains (Kapur 2015).

For grade level, the moderate effect size for junior high and high school participants over the college level participants might be explained by citing the difficulties of controlling conditions when conducting experiments in college level classes. Further research of the failure-based learning strategy at the college level may prove to the contrary.

A large number studies included in our analysis were conducted by one scholar by himself or in collaboration with others. This behooved us to examine the studies in terms of the *researcher*. We found a significant difference between the studies of this one scholar and others in terms of the effect size reported. Examining the results further, we realized that the difference was due to this author's reporting of adjusted means for his studies. Focusing on adjusted means as a moderator, we further analyzed the studies and found that studies that reported unadjusted means showed lower effect sizes. On the other hand, studies with adjusted means reported significantly higher effect sizes. These findings prompted us to speculate that by reporting adjusted means one practically account for or

control some of the variables that may contribute to the lower effect size of the treatment thus making the analysis more robust in terms of reporting the impact of the strategy used with treatment groups.

Another issue worthy of discussion is the fact that we computed a combined effect across outcomes for the studies with multiple outcomes and multiple comparisons. Just like calculating means for any sets of data, one may misrepresent the size of the actual reported effect sizes individually for each outcome. For example, the effect sizes for Loibl and Rummel (2014a, b) studies were reported close to zero while the authors reported a positive effect size for their treatment group dealing with conceptual knowledge (e.g., $d = 1.35$) and a negative effect size for the same subjects in procedural knowledge (e.g., $d = -0.68$) favoring the control group's participants.

The implications of these findings for designing and developing instruction includes using the strategy because of its effectiveness, especially as it seems to be more malleable to systematic design of instruction. Compared to other strategies used in the studies included in this analysis the failure-based strategy showed a positive and significant result. This indicates that the instructional designers and instructors may include the strategy in the courses they design and teach. However, the work of Tawfik et al. (2015) demonstrate the design complexity of using the strategy. This implies that in designing instruction, one need to consider the components of the strategy in relation to the instructional goals. We suggest complementing the strategy with ones that have more empirical basis. The systematic features of instructional design will lend themselves to more easily employing this strategy.

Limitations

A major limitation in our study was the lack of sufficient number of studies reporting negative effect sizes due to the limited number of group comparison studies in this area. This limitation presents the possibility of publication bias. Even though the reported funnel plot (Fig. 3) was almost symmetric, we still need more studies with negative effect sizes to eliminate a potential bias. The insufficient number of studies in this area presented another problem which could be a limitation for gaining a better understanding of the impact of learning from failure as an instructional strategy. Even though we used several keywords in our search for related studies from multiple sources, we ended up with a number of studies that represented only two of the keywords: *Productive Failure* and *Failure-Driven Memory*. This again, is another indication of lack of empirical studies dealing with this topic. We suggest that a better analysis such as ours would benefit from having more studies available.

The complexity of this topic is demonstrated by the number of constructs related to the different dimensions of the topic (see Tawfik et al. 2015). Failure-based learning, while a topic of interest for scholars, seems to present a multi-faceted area of research which limits conducting experimental studies in which one can control or account for this multi-dimensionality. Thus, this presents a limitation for future studies interested in exploring this topic.

Acknowledgements We would like to acknowledge the contributions of Hulya Yurekli and Allison Born for their assistance with the search for literature.

Funding The authors received no funding for this project.

Compliance with ethical standards

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

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

Blumberg, F. C., Rosenthal, S. F., & Randall, J. D. (2008). Impasse-driven learning in the context of video games. *Computers in Human Behavior, 24*(4), 1530–1541. https://doi.org/10.1016/j.chb.2007.05.010.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). *Comprehensive meta-analysis version 2* (p. 104). Englewood, NJ: Biostat.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley. https://doi.org/10.1003/9780470743386.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32–42. https://doi.org/10.3102/0013189x018001032.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillside, NJ: Lawrence Earlbaum Associates.

D'Mello, S., & Graesser, A. (2014). Confusion and its dynamics during device comprehension with breakdown scenarios. *Acta Psychologica, 151,* 106–116. https://doi.org/10.1016/j.actpsy.2014.06.005.

D'Mello, S., Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction, 29,* 153–170. https://doi.org/10.1016/j.learninstruc.2012.05.003.

Dick, W., Carey, L., & Carey, J. O. (2014). *The systematic design of instruction* (8th ed.). Saddle River, NJ: Pearson.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455–463.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363–406. https://doi.org/10.1037/0033-295x.100.3.363.

Gartmeier, M., Bauer, J., Gruber, H., & Heid, H. (2008). Negative knowledge: Understanding professional learning and expertise. *Vocations and Learning, 1*(2), 87–103. https://doi.org/10.1007/s12186-008-9006-1.

Gartmeier, M., Lehtinen, E., Gruber, H., & Heid, H. (2011). Negative expertise: Comparing differently tenured elder care nurses' negative knowledge. *European Journal of Psychology of Education, 26*(2), 273–300. https://doi.org/10.1007/s10212-010-0042-5.

Glogger-Frey, I., Fleischer, C., Grüny, L., Kappich, J., & Renkl, A. (2015). Inventing a solution and studying a worked solution prepare differently for learning from direct instruction. *Learning and Instruction, 39,* 72–87. https://doi.org/10.1016/j.learninstruc.2015.05.001.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics, 6*(2), 107–128. https://doi.org/10.2307/1164588.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York, NY: Academic Press.

Heemsoth, T., & Heinze, A. (2013). The impact of incorrect examples on learning fractions: A field experiment with 6th grade students. *Instructional Science, 42*(4), 639–657. https://doi.org/10.1007/s11251-013-9302-5.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). *Measuring inconsistency in meta-analyses. Bmj, 327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557.

Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need? *Educational Psychology Review, 23*(1), 1–19.

Kapur, M. (2008). Productive failure. *Cognition and Instruction, 26*(3), 379–424. https://doi.org/10.1111/cogs.12107.

Kapur, M. (2009). Productive failure in mathematical problem solving. *Instructional Science, 38*(6), 523–550. https://doi.org/10.1007/s11251-009-9093-x.

Kapur, M. (2010). A further study of productive failure in mathematical problem solving: Unpacking the design components. *Instructional Science, 39*(4), 561–579. https://doi.org/10.1007/s11251-010-9144-3.

Kapur, M. (2012). Productive failure in learning the concept of variance. *Instructional Science, 40*(4), 651–672. https://doi.org/10.1007/s11251-012-9209-6.

Kapur, M. (2013). Comparing learning from productive failure and vicarious failure. *Journal of the Learning Sciences, 23*(4), 651–677. https://doi.org/10.1080/10508406.2013.819000.
Kapur, M. (2014). Productive failure in learning math. *Cognitive Science, 38*(5), 1008–1022. https://doi.org/10.1111/cogs.12107.
Kapur, M. (2015). Learning from productive failure. *Learning: Research and Practice, 1*(1), 51–65. https://doi.org/10.1080/23735082.2015.1002195.
Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist, 51*(2), 289–299. https://doi.org/10.1080/00461520.2016.1155457.
Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *Journal of the Learning Sciences, 21*(1), 45–83. https://doi.org/10.1080/10508406.2011.591717.
Kapur, M., & Kinzer, C. K. (2009). Productive failure in CSCL groups. *International Journal of Computer-Supported Collaborative Learning, 4*(1), 21–46. https://doi.org/10.1007/s11412-008-9059-z.
Kapur, M., & Lee, J. (2009). Designing for productive failure in mathematical problem solving. In N. Taatgen, & V. R. Hedderick (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2632–2637). Austin, TX: Cognitive Science Society.
Kapur, M., & Rummel, N. (2012). Productive failure in learning from generation and invention activities. *Instructional Science, 40*(4), 645–650. https://doi.org/10.1007/s11251-012-9235-4.
Lehman, B., D'Mello, S., & Graesser, A. (2012). Confusion and complex learning during interactions with computer learning environments. *The Internet and Higher Education, 15*(3), 184–194. https://doi.org/10.1016/j.iheduc.2012.01.002.
Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* (Vol. 49). Thousand Oaks, CA: Sage Publications.
Loibl, K., & Rummel, N. (2014a). The impact of guidance during problem-solving prior to instruction on students' inventions and learning outcomes. *Instructional Science, 42*(3), 305–326. https://doi.org/10.1007/s11251-013-9282-5.
Loibl, K., & Rummel, N. (2014b). Knowing what you don't know makes failure productive. *Learning and Instruction, 34,* 74–85. https://doi.org/10.1016/j.learninstruc.2014.08.004.
Morgan, C. L. (1894). *An introduction to comparative psychology*. Boston, MA: Adamant Media Corporation.
Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*(1), 1–4.
Pathak, S. A., Jacobson, M. J., Kim, B., Zhang, B., & Deng, F. (2008). Learning the physics of electricity with agent-based models: The paradox of productive failure. In T.-W. Chan, G. Biswas, F.-C. Chen, S. Chen, C. Chou, M. Jacobson, Kinshuk, F. Klett, C.-K. Looi, T. Mitrovic, R. Mizoguchi, K. Nakabayashi, P. Reimann, D. Suthers, S. Yang, & J.-C. Yang (Eds.), *Proceedings of the 17th International Conference on Computers in Education* (pp. 221–228). Taipei, Taiwan: Asia-Pacific Society for Computers in Education.
Piaget, J. (1952). *The origins of intelligence in children*. New York, NY: WW Norton & Co.
Piaget, J. (1977). *The development of thought: Equilibration of cognitive structures* (Vol. viii). Oxford, UK: Viking.
Schank, R. (1999). *Dynamic memory revisited* (2nd ed.). Cambridge, UK: Cambridge University Press.
Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction, 4*(4), 295–312.
Tawfik, A., & Jonassen, D. (2013). The effects of successful versus failure-based cases on argumentation while solving decision-making problems. *Educational Technology Research and Development, 61*(3), 385–406. https://doi.org/10.1007/s11423-013-9294-5.
Tawfik, A. A., Rong, H., & Choi, I. (2015). Failing to learn: Towards a unified design approach for failure-based learning. *Educational Technology Research and Development, 63*(6), 975–994. https://doi.org/10.1007/s11423-015-9399-0.
Valentine, J. C., & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The Study Design and Implementation Assessment Device (Study DIAD). *Psychological Methods, 13*(2), 130–149. https://doi.org/10.1037/1082-989x.13.2.130.
van Merriënboer, J. G., Kirschner, P. A., & Kester, L. (2003). Taking the load off a learners' mind: Instructional design for complex learning. *Educational Psychologist, 38*(1), 5–13. https://doi.org/10.1207/s15326985ep3801_2.
VanLehn, K. (1988). Toward a theory of impasse-driven learning. In D. H. Mandl & D. A. Lesgold (Eds.), *Learning issues for intelligent tutoring systems* (pp. 19–41). New York, NY: Springer. https://doi.org/10.1007/978-1-4684-6350-7_2.

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction, 21*(3), 209–249. https://doi.org/10.1207/s1532690xci2103_01.

Westermann, K., & Rummel, N. (2012). Delaying instruction: Evidence from a study in a university relearning setting. *Instructional Science, 40*(4), 673–689. https://doi.org/10.1007/s11251-012-9207-8.

Zimmerman, B. J. (1989). Models of self-regulated learning and academic achievement. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement* (pp. 1–25). New York, NY: Springer.

Aubteen Darabi is an Associate Professor of the Instructional Systems and Learning Technologies and Director of Center for Learning and Performance Systems of the Learning Systems Institute at Florida State University. Darabi teaches graduate courses in instructional systems and manages research and development projects funded by state and federal agencies.

Thomas Logan Arrington is a doctoral candidate in the Instructional Systems and Learning Technologies program in the College of Education at Florida State University.

Erkan Sayilir is a doctoral candidate of Measurement and Statistics in the College of Education at Florida State University.