UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

THE IMPLICATIONS OF FEEDBACK SEQUENCING ON PERFORMANCE AND TRANSFER

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

DUNCAN MCCOLLUM
Norman, Oklahoma
2019

THE IMPLICATIONS OF FEEDBACK SEQUENCING ON PERFORMANCE AND TRANSFER

A THESIS APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY

Dr. Eric Day, Chair

Dr. Daniel Kimball

Dr. Luz-Eugenia Fuenzalida

Dr. Scott Gronlund

Dr. Hairong Song

Acknowledgments

I would like to thank my committee for their endless support and patience. They have provided me with so much encouragement and refused to let me give up. Over the years they have shaped the way I think about cognition and how we can better understand the world through statistics. I truly appreciate the excellent education they have provided me.

I would also like to thank my family, for their unconditional support and love. I know that no matter what I decide to do they will be there for me.

Table of Contents

List of Tables

List of Figures

Abstract

The goal of this study was to investigate the role of sequencing different levels of feedback specificity on performance and solving of novel problems. Prior work has shown a more successful transfer of knowledge when feedback is initially withheld. This is likely due to a result of requiring students to engage in the exploration of problems. In the current study, participants trained on GRE math problems, during which they received either only knowledge of their results or knowledge of their results as well as a hint in the event that they answer incorrectly. The ordering of these treatments varied between groups. Participants ability to transfer their knowledge was only impaired when they encountered a change in the type of feedback they received during training. Otherwise, receiving hints did not affect transfer or performance scores. Additional covariates are explored to explain the pattern of results, and potential improvements to the study are discussed.

The Implications of Feedback Sequencing on Performance and Transfer

A common complaint amongst students is that they will never use what they learn in school (Organisation for Economic Co-Operation and Development, 2004). They report seeing little value in the work they are asked to complete (Yazzie-Mintz, 2007). This is not a new complaint, and the issue has become a prominent topic in cognitive psychology. The value of studying seemingly useless subjects and learning inconsequential material is not only lost on students but is associated with undesirable outcomes including higher dropout rates (Yazzie-Mintz, 2007). However, one key justification for exposing students to a diverse course pallet is that they will apply the problem-solving skills and reasoning abilities they learn to new real-world problems—this is called transfer. The study of transfer has produced some complex and conflicting results, and researchers have been exploring pedagogies that will more readily promote the transferability of knowledge (Bransford, Brown, & Cocking, 2000; Friedlaender, Burns, Lewis-Charp, Cook-Harvey, & Darling-Hammond, 2014). One key element that has been systematically explored is the role feedback plays in the learning and transfer process. There is a common belief that more specific feedback leads to greater performance and learning, and this has been the guidance provided by many instructional texts (see Goodman, Wood, & Hendrickx, 2004, for discussion). Interestingly though, research suggests that feedback does not always appear to help (Kulhavy, White, Topp, Chan, & Adams, 1985; Schmidt, Young, Swinnen, & Shapiro, 1989). In fact, in some cases, it may be preferable to have students attempt to discover solutions on their own without feedback (Schwartz, Chase, Oppezzo, & Chin, 2011; Schwartz & Martin, 2004). This process is thought to promote a deeper level of processing of the content and thus a deeper understanding of the target information. However, this pedagogical approach is not a simple panacea, as there are plenty of nuances that determine when and what feedback will be most effective (Kluger & DeNisi, 1996; Shute, 2008). Given the importance of the applicability to “real world” settings and the implications it can have on student investment in learning, it seems both logical and prudent to systematically explore the circumstances and the extent to which this more demanding approach is appropriate, thus enabling the development of a curriculum that instills more generally useful knowledge.

Transfer

How do we apply problem-solving skills we learn in school to novel problems in the real world? We must have some ability or mechanism that allows us to generalize knowledge. Otherwise, we would be helpless when faced with problems that do not perfectly match problems we learned to solve in school. How similar do problems need to be, and in what ways do they need to be similar for us to transfer stored knowledge effectively to such novel circumstances? A typical transfer study involves an initial training task and a subsequent transfer task. For a training task that involves problem-solving, participants learn how to solve a specific type of problem. Then, for the transfer task, they attempt to apply what they have learned to a novel problem. The effectiveness of a training technique can be determined by comparing performance on a transfer test across conditions that vary as to the use of the technique in the initial training task.

This topic has been researched for over a century with much debate on whether or not transfer is even possible. One of the impetuses for this line of work was testing the benefits of the “doctrine of formal discipline,” which emphasizes the importance of learning subjects such as mathematics and foreign languages, with the purported benefit is that this will exercise the mind, helping it to develop general reasoning and problem-solving skills (Lyans, 1914). However, early research examining the proposed effects of this pedagogical strategy did not support the expected advantages. For instance, Thorndike and Woodworth (1901) failed to demonstrate the transfer of the ability to estimate the magnitude of shapes after having been trained to do so with other shapes. However, Participants were able to improve their ability to estimate the magnitude of shapes they had trained with. They concluded that transfer depended on the number of concrete elements that are identical in the training and transfer task. In another classic study by Judd (1908), evidence of the benefits of training were observed. Participants were trained to hit a target that was submerged underwater at 12 inches and then were tested on their ability to hit a target submerged only 4 inches. Some participants received an explanation of how refraction worked; the other participants received no such explanation and thus had only their experience with the training task from which to learn. The participants with explanation-based learning performed much better in hitting the 4-inch-deep target than the participants with only experience-based learning. Thus, this finding suggests that explanation of a common underlying constraint—refraction in this case—is one aspect of learning that promotes the transfer of problem-solving skills to new problems.

The Judd (1908) study involved quite similar training and transfer tasks and thus tested *near transfer*. By contrast, *far transfer*—involving dissimilar tasks—is of more ultimate importance for the transfer of skills learned in school to real world problems, given the dissimilarity in those two types of task. Broadly, far transfer occurs when knowledge is generalized even when the training and testing problems differ in aspects unrelated to the structure of the problem. This is in contrast to near transfer which occurs when training and testing tasks are very similar (e.g., it could be as similar as solving the same kind of math problem but with different numbers). Near transfer is generally accepted in the scientific literature but there is less support for far transfer (Detterman, 1993). Some argue though that it is possible to observe far transfer between situations that may be different superficially but similar on a structural level. It is speculated that the contradictory findings stem from the failure to specify the ways situations need to be similar in order for transfer to occur (Barnett & Ceci, 2002).

Barnett and Ceci (2002) outline a taxonomy to help researchers provide better operational definitions of the exact kind of far transfer they are attempting to observe. They outline nine dimensions such as knowledge domain, physical context, temporal context, and testing modality to name a few. Each of these dimensions may characterize the relationship between training and transfer and near or far. This allows researchers to better predict under which circumstances transfer will occur.

Another issue that sheds light on the debate is the concept of sequestered problem-solving. Studies that typically fail to demonstrate far transfer require participants to directly apply their learning in environments where they have little opportunity to attempt new approaches and revise their solutions as needed. This is not consistent with how we solve problems and learn new things in real life and may explain why we fail to see transfer in the lab when it seems to exist in reality (Schwartz, Bransford, & Sears, 2005). It is also important to note that transfer may not manifest itself in how people solve new problems but how they learn new things. This is called “transferring in” knowledge in which existing conceptual knowledge is used as a scaffold to assist in learning a new topic (Schwartz et al., 2005). For example, someone may have a good sense of which questions will be most useful for learning about something new.

These new ideas of demonstrating transfer outside of sequestered environments and transferring in knowledge created interest in demonstrating transfer in situations that required invention and discovery. There is evidence that having students try to invent solutions on their own before being given the canonical solution and a lesson can result in better learning and transfer when compared to students who go through worked examples before receiving a lesson (Schwartz & Martin, 2004). This method, called "inventing to prepare for future learning," demonstrates the importance of measuring how well students are prepared to learn (transfer in), as well as use what they learn to solve novel problems (transfer out). Similarly, requiring students to come up with explanations for solutions helps reinforce the correct procedural knowledge while also elucidating the rationale for those procedures (Rittle-Johnson, 2006). Understanding the rationale behind a solution allows students to develop innovative problem-solving approaches when they encounter a modified version of that problem.

In order for transfer to occur, three things must happen: first, the learners must recognize that they have the required knowledge to answer a problem; second, they must be able to recall the necessary procedures and knowledge to solve the problem; and third, they must successfully execute these procedures (Barnett & Ceci, 2002). Gick and Holyoak (1980) make it clear that failure to spontaneously transfer knowledge is often due to failure to recognize that a viable solution exists in memory. In their study, participants could not recognize that they had the correct knowledge on their own because there was a lack of surface similarity between the training and transfer tasks. We tend to be ineffective at searching our memory based on the structural features of our knowledge as opposed to the superficial features. Therefore, if we want to promote transfer, we should seek treatments that make the structural elements more salient when we are searching for them. Vendetti, Wu, and Holyoak (2014) showed that participants can be trained to be more aware of the deep structure of problems. By training participants to solve analogies that were semantically distant, participants adopted a relational mindset when performing an unrelated object matching task. Importantly though, this was only observed when participants generated solutions to the analogies rather than recognize them. This demonstrates that it is possible to develop problem-solving strategies that can then be transferred more effectively to problems which have structural similarities but dissimilar surface features.

Feedback

The role that feedback plays in the learning process has been explored extensively, and the research has yielded many mixed results (for reviews, see Kluger & DeNisi, 1996; Shute, 2008). Feedback has been very well defined in the literature and can be broken down into many different dimensions, all of which interact to help or hurt the learner. Some important factors to consider when designing feedback are specificity, learners' prior knowledge, and frequency.

First, prior knowledge can moderate the effect of a feedback message (Fyfe, Rittle-Johnson, & DeCaro, 2012). If a student is a complete novice, receiving some explanation of how to solve the problem can be useful. However, if a student is familiar with a topic then they may be better off without an explanation and should be allowed to explore the problem on their own (Fyfe et al., 2012). Successful exploration of a problem improves understanding of its structure and its underlying concepts (Schwartz et al., 2011).

Changing the timing of feedback can have interesting effects and should also be considered when designing feedback. Kulhavy and Anderson (1972) demonstrated that receiving correct answers to a multiple-choice test after a delay can improve performance on a future test. It was argued that this was because the delay kept the initial errors from competing with the correct answers. Schmidt et al. (1989) found that decreasing the frequency of feedback on a motor task has also been shown to hurt immediate performance but improves performance on a delayed task. By summarizing performance over many trials feedback was delayed and contained less information.

The dimension of feedback which I am most interested in for this study is specificity. Specificity refers to the amount of information that is conveyed in a feedback message. The simplest form of feedback is knowledge of result (KR), in which a solution is deemed correct or incorrect, and no additional information is given. Students must have at least this much specificity in their feedback if they are expected to learn by doing, and KR can often be achieved implicitly by seeing if an action or response yields an expected result. For instance, they may see that they scored a goal, or that their solution to a math problem can be tested. Immediate performance gains are typically observed when feedback contains more information than just KR (Attali, 2015). This elaborative feedback allows students to recognize that a gap exists in their knowledge and begin to understand how to correct it. For example, a message may include some sort of hint or cue that points the learner in the correct direction without giving away the answer. This may be useful for students who are stuck on a problem and cannot recognize or remember the correct procedures.

On the high end of the spectrum of specificity, feedback may become very complex and customized, providing KR, identifying any errors in their solution, and providing guidance or hints on how to improve their answer. While more specific feedback may improve immediate performance, too much information may be harmful to other skills that may be learned (Kulhavy et al., 1985). As a result, they will be less effective at identifying and correcting errors on their own.

However, there have been other findings that imply that providing more information in a feedback message facilitates transfer to novel problems (Butler, Godbole, & Marsh, 2013). One study demonstrated that learning with feedback of greater specificity can result in better knowledge when test questions require inferences to be made. The reason participants who receive extensive feedback are better at inferences is thought to be that the feedback helps them to develop important connections among critical concepts (Butler et al., 2013). Even though informative feedback can be useful for learning, less informative feedback can result in learning as well (Goodman et al., 2004). It is believed that a lack of specific feedback promotes exploration which leads to the learning of procedural knowledge as well as the underlying concepts (Fyfe et al., 2012; Schwartz et al., 2011). High specificity will lead to high performance and learning; however, it also results in less exploration (Goodman et al., 2004).

One explanation that has been put forth to account for the sometimes detrimental impact of feedback is the guidance hypothesis, which argues that learners develop a dependence on feedback which limits practice in using other processes such as identifying errors and recognizing transferable knowledge (Salmoni, Schmidt, & Walter, 1984; Schmidt et al., 1989; Schmidt & Bjork, 1992). This overreliance on feedback results in subpar performance on other tests of ability. This hypothesis is consistent with the findings of Goodman et al. (2004) who found that fostering exploration by limiting specific feedback is essential to developing a deep understanding of the problem and its potential solutions. While having to figure things out on your own may be less efficient in the short term, it encourages innovation. Developing a deeper understanding of the content is crucial for making that knowledge generalizable by later being able to recognize solutions that are transferable to subsequent novel problems. The guidance hypothesis is also consistent with the results of some of the studies that seek to improve transfer via self-explanation and invention (Schwartz & Martin, 2004; Rittle-Johnson, 2006). When learners have the opportunity and prior knowledge to successfully explore a problem they engage with the content at a deep level. They can gain considerable conceptual knowledge compared to when they are just given specific feedback that only conveys a set of procedures applicable to the training task.

Hypotheses

The purpose of this study was to explore the type and sequencing of feedback that promotes the development of a problem-solving skill that transfers to problems of a different domain. Providing feedback in the form of hints is often beneficial when learning something new but can be detrimental in some cases. One element that determines the efficacy of feedback is how the information is used. It may be acknowledged and help create useful ties between critical concepts. It may also be too helpful, acting as a crutch to learning and decreasing a participant's ability to learn new things later on.

Hypothesis 1.

During training, participants will be more successful at problem-solving for topics that they receive hints for when they need to reattempt an incorrect problem. This is because hints will help them develop topic-specific skills, which can be applied to similar problems. If they do not receive hints when they are incorrect, they will not learn the topic-specific skills as well.

Hypothesis 2.

When participants are tested on problems involving completely new topics without any hints, their performance will be impacted by how well they are able to transfer skills learned during training. Specifically, I am interested in their ability to develop problem-solving skills that can be transferred across GRE topics. If they consistently receive hints, it is predicted that they will fail to develop these general problem-solving skills and perform poorly on the transfer task. Alternatively, if they must solve problems with no hints during training, the process of exploring problems and identifying appropriate problem-solving skills that are useful across topics should help make these skills more transferable.

Hypothesis 3.

Participants who initially train without hints and later train with hints will be prepared to solve problems from novel topics. The experience of attempting to find appropriate strategies on their own will prepare them to learn more from having hints which are eventually provided to them in the second learning phase. Specifically, they should be able to understand what is characteristic of setting up a problem successfully. In contrast, those that initially receive hints and then have them removed will not be as successful at transferring the general problem-solving skills they learn from the hints to new topic areas. This is due to a lack of exploration of potential problem-solving strategies because they have already observed effective approaches in the form of hints.

Methods

Participants.

Undergraduate students (N = 649) were recruited from a participant pool at the University of Oklahoma. Participants received partial course credit for their voluntary participation as one option for satisfying the research participation requirements of their psychology courses.

Materials.

Participants were trained and tested on questions that resemble the questions from the quantitative section of the Graduate Record Examinations (GRE). The quantitative section of the GRE only covers topics which are no higher than what is learned in a second algebra course (Briel & Michel, 2014). In other words, all participants have been exposed to each topic on the GRE. This prior knowledge will ensure that participants were prepared to engage in exploration if needed. However, the test makes simple questions more difficult by placing a greater emphasis on quantitative reasoning skills rather than focusing solely on understanding of basic concepts. This requires participants to abstract mathematical models from questions while disregarding irrelevant information (Morley, Bridgeman, & Lawless, 2004). In addition, all participants should possess the prior knowledge to solve the problems, but the transfer of that knowledge is inhibited by their lack of ability to recognize what concepts to use on the GRE. Although these topics will be independent of each other in terms of their basic concepts, the general problem-solving skills which are needed for GRE questions should overlap between topics. Using the GRE materials allows students to focus on developing recognition (and in turn transfer) of these general problem-solving skills without having to teach optimal problem-solving approaches which are specific only to completely new concepts.

The materials were drawn from test preparation materials (Educational Testing Service, 2012; MG Prep, 2013) and focused on nine core topics that are covered by the GRE (two-variable word problems, divisibility and prime numbers, exponents and roots, functions and formulas, geometry, inequalities and absolute values, measures of central tendency, number properties, and probability). The topics were selected to have as little overlap with each other as possible. This is to ensure that topics are independent. Training on one topic should not prepare participants to answer questions from a new topic simply because they require the knowledge of the same elementary concepts. A pool of five potential questions per topic were selected from test preparation materials. Questions within topics require similar conceptual knowledge but they assess that knowledge with non-isomorphic problems. One hint for each question was developed by examining the canonical solution provided by the source, or the solutions that research assistants came up with during pilot testing of the materials. These solutions outline how the problem should be set up and then the operations or reasoning that must be implemented to get the answer. The hints focus on identifying a first step towards the solution by providing information which can help reframe the problem into a more familiar context. For instance, the hint may highlight that certain information is irrelevant or that the problem includes multiple steps. The hint did not provide any additional information that someone without a hint would not be able to figure out from the problem itself. For example, Steve's property tax is $140 less than Patricia's property tax. If Steve's property tax is $1,960, then Steve's property tax is what percent less than Patricia's property tax? The hint for this question would be to calculate Patricia's property tax first. The hint only provides the first step and does not include a definition of what "percent less" means. A sample question and its hint for each topic may be found in Figure 1.

Design.

The study used a 2 (learning phase 1: hint, no hint) x 2 (learning phase 2: hint, no hint) between-participants design. After attempting each problem, participants received one of two types of feedback: In the no-hint conditions, participants received information stating whether they got the problem correct or incorrect. If they were incorrect, they were asked to reattempt the problem, otherwise, they were given the next question. In the hint conditions, participants followed the same procedure but were provided a hint for how to solve the problem if they were incorrect on their first try. In the first learning phase, half of the participants were randomly assigned to receive a hint or no hint after an initial attempt on each problem, with half of the participants in each of those two conditions then receiving the same or different type of feedback in the second learning phase. After the two learning phases, the transfer phase began, in which participants were tested on their ability to transfer the general problem-solving skills they developed to previously unencountered topics. To control for order effects, items within topics were randomized. Furthermore, topics were counterbalanced across training and transfer phases using a balanced Latin square design.

Procedure.

Data were collected in-lab using E-Prime 2.0 software (Psychology Software Tools, Pittsburgh, PA), The experiment was described to the participants as seeking to improve the usability of their current knowledge. This was meant to convey the message that intelligence is not static and hopefully provided some motivation to participants who do not like math, as suggested by Dweck (1989). Participates then completed a task to measure working memory capacity as a possible covariate of their transfer performance; then the training and transfer tasks; and then a questionnaire to elicit self-reported information about demographics and academic performance, also as possible covariates.

Working Memory Task. Participants first completed an OSPAN task to measure working memory capacity (WMC). This is an abbreviated measure developed by Oswald, McAbee, Redick, & Hambrick (2015) and administered in E-Prime. In this task, participants are shown a simple math problem (e.g., 2*2 + 5 = 9) and must decide if it is true or false. Next, they are shown a letter which they are instructed to remember. After going through a set of these math problems and letter pairs, they will be asked to recall the letters in order. This sequence repeats multiple times with different set lengths and takes approximately 10 minutes to complete.

Training and Transfer Tasks. For this portion of the experiment, there were three major phases, two learning and one transfer. The two learning phases allowed participants to get experience with three topics per phase (presented in blocks) with differing levels of feedback, as noted in the Design section. Each learning phase consisted of a hinting block in which participants did or did not receive hints, depending on their condition. These two learning phases also included assessment questions from each of the trained topic areas. This was designed to measure performance on that topic after receiving different levels of feedback (hints or just reattempts). These learning phases were conducted sequentially with no distractor or interruptions. A diagram depicting the sequence of the learning and transfer tasks can found in Figure 2.

Learning Phase 1. During the first hinting block, participants answered four questions from each of three topic areas. The topic areas were presented sequentially. Participants were encouraged to use scratch paper but were not be allowed to use a calculator. If participants got a question wrong, they received the type of feedback prescribed by their condition and were allowed to reattempt the problem. The combined time of a first and second attempt did not exceed two minutes. After receiving feedback, they were told how much time they have left to reattempt the question. If they were correct on their first attempt, they immediately proceeded to the next question. Once participants completed the first hinting block, they encountered a test question from each of the topics they encountered in that block. The test questions resembled the training questions and they had two minutes to solve it; however, there was no feedback and no reattempts.

Learning Phase 2. As noted in the Design section, in the second learning phase, some participants received the same kind of feedback they had in phase one, while others encountered the other kind of feedback. Phase two included three new topics but otherwise followed the same procedures as phase one.

Transfer Phase. In the final phase, participants encountered two test questions each from three novel topic areas. As with prior test periods, participants only had one try to solve these problems and did not receive feedback.

Demographic and Academic Performance Questionnaire. The demographic and academic questionnaire included basic demographic information (age, ethnic origin, race, and sex), SAT scores, high school grade point average, and current major.

Results

The results are reported separately for each hypothesis; the analysis plan was the same, but the exact variables included in the models differed across hypotheses. I conducted an ANCOVA with treatment condition as the between-subjects factor and WMC included as a covariate to account for differences in participants' ability to deal with the additional load that is required to find solutions without guidance. I compared this base model to competing models based on model fit (i.e. sum of squares explained). The competing models included additional participant ability covariates. Descriptive statistics of key variables are reported in Table 1, and a correlation heatmap of covariates is reported in Figure 3. Plots of mean scores between groups on learning and transfer phases can be found in Figure 4 and Figure 5, respectively.

Hypothesis 1. The first hypothesis was that performance on the assessment for each learning phase would be better in the hint condition than the no-hint condition. To examine this, I looked at only the groups that received the same treatment during a given block. The results are reported separately for each learning phase.

Learning Phase 1 Results. For this analysis, the conditions which received hints during the first learning phase were combined and the conditions which did not receive hints during the first learning phase were combined, yielding two groups to compare. The base model revealed that WMC was a significant predictor of assessment score, $F(1, 645) = 7.53$, $p = .01$. However, contrary to my hypothesis, treatment condition was not a significant predictor, $F(1, 645) = .02$, $p = .89$, nor was the interaction between WMC and condition, $F(1, 645) = .04$, $p = .84$. As evidenced by the statistics in Table 2, adding response time and the number of successful first attempts did improve the model fit significantly. Interestingly, WMC became a non-significant predictor when the number of successful first attempts was added (as in Model 3), so it was removed from subsequent models. The best fitting model was Model 5, $F(3, 645) = 20.9$, $p < .01$, adjusted $R^2 = .08$. This model shows that treatment condition did not have a significant effect on assessment score, $F(1, 645) = .02$, $p = .89$. Number of successful first attempts was a significant predictor, $F(1, 645) = 53.71$, $p < .0001$, as was response time, $F(1, 645) = 8.89$, $p < .001$.

Learning Phase 2 Results. For this analysis the conditions which received hints during the second learning phase were combined and the conditions which did not receive hints during the second learning phase were combined, resulting in two groups to compare. The base model revealed that treatment condition was not a significant predictor of assessment score, $F(1, 645) = .22$, $p = .64$, nor was WMC, $F(1, 645) = 1.02$, $p = .31$, or the treatment by WMC interaction, $F(1, 645) = .74$, $p = .39$. Just like with Learning Phase I, the fit statistics improved when adding other covariates (see Table 3). The best fitting model was Model 5, $F(3, 645) = 23$, $p < .01$, adjusted $R^2 = .09$. This model shows that treatment condition did not have a significant effect on assessment score, $F(1, 645) = .25$, $p = .62$. Number of successful first attempts was a significant predictor, $F(1, 645) = 63.98$, $p < .001$, as was response time, $F(1, 645) = 4.67$, $p = .03$.

Hypothesis 2. The second hypothesis suggests that more exposure to hints will be associated with significantly worse performance on the transfer task. As a reminder, this is because the participants who learn with no hints will develop general problem-solving skills through exploration and deep understanding, which they can then use for novel topics encountered during the transfer phase. The degree of transfer is calculated as the proportion of problems solved in the final transfer phase. To determine if the number of hinting blocks encountered impairs transfer, I conducted an ANCOVA with the number of hinting blocks encountered as a between-subjects factor and WMC as a covariate, as well as their interaction. This analysis combined participants who were in conditions which encountered only one block of hints. This analysis also excluded participants who scored below chance levels (.22) on the assessment questions during the learning phases. Updated descriptive statistics for the modified dataset can be found in Table 4. Differences between groups on transfer scores are shown in Figure 6. The analysis showed that condition was a good predictor of final transfer scores $F(2, 401) = 3.23$, $p = .04$, as well as WMC $F(1, 401) = 8.64$, $p = .003$. As with previous models, this one improved with adding additional covariates (see Table 5). Notably when success on first attempts was added working memory became non-significant (as in Model 3). The best model was one that only included group membership and success on first attempts as predictors $F(3, 403) = 36.2$, $p < .001$, adjusted $R^2 = .21$ (Model 4). Pairwise comparisons revealed that those that receive one block of hints could be expected to score lower on the transfer test than the group that never received hints $t(200) = 2$, $p = .04$.

Hypothesis 3. This hypothesis concerns whether withholding hints until later training promotes greater transfer to new topics than proving hints early on. The two groups which received sequenced feedback were included as two levels of the between-subjects factor in an ANCOVA with WMC as a covariate. An updated correlation heatmap of covariates may be found in Figure 7. As with hypothesis 2, participants who scored below chance levels (.22) during the assessment periods were excluded. This base model revealed that the sequencing of hints had no impact on participants transfer scores $F(1, 196) = .01$, $p = .93$, and working memory capacity failed to predict transfer performance as well $F(1, 196) = .074$, $p = .39$. Additional covariates were examined to see if prediction could be improved by accounting for the number of successful first attempts during the training blocks, number of hints encountered, time spent on problems, and accuracy when given a hint. A series of models were compared to do this (see Table 6). Of these covariates the only one that was shown to be a good predictor was the number of questions correct on the first attempt $F(1, 196) = 34.8$, $p < .001$. As a result the best model included only hint sequence condition and number of questions answered correctly on the first attempt (Model 4), $F(2, 196) = 17.4$, $p < .001$, adjusted $R^2 = .14$.

Discussion

The purpose of this study was to determine the effect of the type and sequencing of hints on performance in a transfer task. Participants completed two learning phases of GRE-type math problems, in which they reattempted problems they answered incorrectly with or without hints (crossed between the two phases) and then completed a series of transfer questions from new topics without any kind of feedback. It was hypothesized that receiving hints would be useful for learning a given topic, but it would impair performance on new topics in the absence of feedback. Furthermore, I expected to see impaired transfer performance by groups that received hints early in the experiment and then had them taken away (relative to the group that received the opposite sequencing of hints).

This study failed to find much support for these hypotheses. Receiving hints during a particular learning phase did not predict participant's scores on assessment questions from that phase. The number of blocks in which hints were encountered was shown to be a predictor of final transfer scores, but it exhibited an interesting trend. It was predicted that more exposure to hints would result in lower transfer scores. However, it was shown that groups that received only one block of hints scored significantly lower on the transfer test than the group that never received hints and the group which consistently received hints. These groups which were exposed to one block of hints experienced a change in treatments during the experiment. It is possible that this change affected their motivation or the strategy they were using, and this had an impact on their transfer scores. Having the hints removed or added may have decreased how much effort was applied to questions in the second learning phase. Causing them to not pick up the problem-solving skills. Additionally, having hints introduced may have led participants to not try on the first attempt so they can receive a hint and eliminate one of the answer choices. It should be noted though that the size of this effect was small ($\omega^2_p = .01$) and may be an artifact of the two sequenced conditions being combined to make one large condition. Finally, the order in which hints were sequenced was not shown to be a good predictor of transfer performance. There is no evidence to indicate that it matters when hints are provided in order to encourage developing transferable problem-solving skills.

A number of covariates were considered in this study to help explain performance on assessment and transfer questions. The primary one that was included was WMC. This was of interest because WMC is strongly correlated with fluid intelligence and therefore should be a good predictor of participant's performance and transfer (Shipstead, Harrison, & Engle, 2016). I did find that WMC was usually a good predictor of performance on learning phase assessments and transfer questions, but when the number of successful first attempts was included as an additional predictor the significance of WMC capacity was reduced. This suggests that they share variance, and predict performance in a similar way, such that first attempt success accounts for much of the predictive power of WMC. This would not be surprising given that first attempt success is a measure of performance and WMC is related to performance. This introduces a complication though. The modest relationship between these two predictors (r = .25) means that those who have higher WMC will receive fewer hints. So, the actual strength of the treatment (number of hints or non-hints received) may depend on participant WMC. The finding that first attempt accuracy is a significant predictor in all of these models is to be expected given that the success on the dependent measures is identical to what is needed to get a high score for the number of successful first attempts. In each you have only one chance to answer a question without receiving any feedback. Time spent on training blocks was only a useful predictor for assessment questions from that block. This may be because it is closely associated with how hard participants work on a set of problems. It failed to be a useful predictor of transfer performance as well. Suggesting that the time spent training was not influencing the transferable skills that may have been developing.

This study has a number of limitations that may have resulted in a lack of findings. First and foremost, the hints may not have been useful, causing the treatments to not have an effect. There are two ways the hints were expected to have an effect. First, it should have assisted in the learning of a particular topic. There is evidence that the hints failed to do this given that participants had a 39% chance of answering a question correct after receiving a hint and a 38%

chance of answering a question correctly after not receiving a hint. Secondly, hints were expected to limit the learning of successful approaches to answering GRE style questions. The lack of a difference in transfer performance between groups is good evidence for this. Participants who never received hints and those that received hints in both training blocks did not significantly differ in their performance on final transfer questions (40% versus 41%, respectively). Demand characteristics may have also existed in this study. Participants could have perceived their objective as getting as many questions correct as possible, regardless of how many attempts it takes. This could cause them to adopt the strategy of guessing on their first attempt so that they can receive a hint, or at least reduce the number of answer choices. This would limit the amount of exploration they engage in and would hurt their scores on the test questions when they only get one attempt. Another issue with exploration is that all participants should have engaged in exploration during their first attempt regardless of their condition, effectively reducing the impact of the treatment. The number of topics which were included likely also served as a limitation. GRE test development literature indicates that the test consists of four main topic areas: arithmetic, algebra, geometry, and data analysis (Briel & Michel, 2014). By more than doubling the number of topics participants could encounter in the experiment the likelihood of the topics being truly independent went down. This would limit the effectiveness of the treatments because basic knowledge that was picked up during one training block could be used on a different part of the experiment. For instance, a participant may figure something out about how to approach two variable word problems which could help them on problems from the functions topic. The surplus of topics also hurt the ability to get a reliable measure of performance on any topic because the number of questions per topic had to be limited. Motivation was also likely an issue during this experiment. Participants were made aware of their poor performance by the feedback messages, and this likely hurt their willingness to try because their performance was not tied to any motivational items such as a reward. Approximately 15% of participants used their scratch paper for four questions or fewer. The analyses suffered from limitations as well. The experiment may have been more successful if a simpler and more appropriate design was used. An alternative model which is more consistent with the design of this study has been explored before. It consisted of the main effects of receiving hints in learning phase I, learning phase II, and WMC, as well as the interactions between all three predictors. These were used to predict transfer scores. This model yielded results similar to those seen in hypotheses two and three, there was no effect of hinting status on transfer scores. Furthermore, the interaction between hinting status in learning phase I and II was not significant. All of this is consistent with my conclusion that there is no evidence of sequencing effects on transfer scores. Finally, the instrument that was used for WMC may have malfunctioned. A Shaprio-Wilk test for normality provided evidence that the WMC scores were heavily skewed, $W = 1$, $p < .001$ (a Q-Q plot and histogram of scores for WMC can be found in Figure 8). This was an automated shortened version of the operational span task which was validated by Oswald et al. (2015). Part of the trade off of it being shortened is that it suffers from a ceiling effect. They report a skew parameter of -1.08 which is similar to what was observed in this study (-1.1). WMC likely would have been a more effective covariate if the instrument was able to better discriminate among individuals with high WMC.

While this study failed to do so, it may still be possible for future research to demonstrate that receiving hints will limit the development of transferable problem-solving skills. Researchers should be mindful of the materials they are using and ensure that participants are motivated. These two issues are closely related, the materials should require participants to learn something, but it should not be so hard that it seems impossible. The materials used for training and transfer should also allow for a better understanding of the relationship between what skills are being transferred, rather than having to simply assume that successes are driven by developing general problem-solving skills. Strategy use should also be considered and controlled for if possible. If participants are expecting to receive a hint and not giving their best effort on their first attempt this should be reflected in their reaction time data. Use of a verbal protocol may also be beneficial. Participants may be asked if they adopted any strategies during the study. It would be interesting to find out if participants shift their strategies after the type of feedback changes. To try to limit the use of an answer choice reduction strategy ahead of time participants could be told that they will only receive partial credit for questions answered correctly on the second attempt. Future studies involving this kind of feedback may be unable to get around having different strengths of treatments due to participants answering questions correctly before any feedback is needed. To get around this, researchers may consider excluding participants who do not receive their prescribed feedback on a minimum number of questions. Finally, it would be interesting to explore if participants who never receive hints perform significantly better than those who initially have hints then have them taken away in the second block. This could be due to motivation being lost and could strengthen the case for receiving either one type of feedback consistently or being required to try to solve problems on your own before being given feedback.

Support for this study's hypotheses would provide guidance to educators on how to structure their curriculum so that students are prepared to apply their knowledge to new instantiations of problems they have solved before. Additionally, it could help ensure that students have the skills to learn on their own once they have left the classroom. If limiting feedback had been shown to be an effective approach, educators would still need to be cautious in how they implement it. There is a tradeoff between motivation and self-guided learning. Some students may be enthusiastic about a subject and could be successful at trying to figure out things on their own. Yet, other students will become frustrated or decide they will wait until the teacher gives them a lesson before trying. It is important to be able to recognize when that frustration becomes a barrier and how best to address it.

The traditional “tell-and-practice” approach in which students receive a lesson, and then solve problems on their own may result in efficient and accurate performance on questions that are similar to what they are trained. However, it could limit a student’s ability to learn new things on their own by transferring in problem-solving approaches they developed for similar problems. It may also hinder a student’s capacity to answer questions that require prior knowledge but is unrecognizable on a surface level. Consequently, students may be unable to imagine the myriad ways that they could apply their knowledge in the real world. These issues need to be addressed by developing new pedagogy which prioritizes learning and transfer. Research into new methods can consider the role that feedback plays, and the possibility that teaching yourself may be the most effective at certain times. This may be difficult to do in the lab and in the classroom because of the numerous ideocracies that will influence the effectiveness of a given approach. However, if rules can be found that achieve these goals then perhaps learning and our knowledge will become more valuable.

References

Attali, Y. (2015). Effects of multiple-try feedback and question type during mathematics problem solving on performance in similar problems. *Computers & Education*, *86*, 260–267. https://doi.org/10.1016/j.compedu.2015.08.011

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin*, *128*(4), 612–637. https://doi.org/10.1037//0033-2909.128.4.612

Bransford, J., Brown, A., & Cocking, R. (Eds.). (2000). The design of learning environments. In *How people learn: Brain, mind, experience, and school* (Expanded edition, pp. 131–154). https://doi.org/10.17226/9853

Butler, A. C., Godbole, N., & Marsh, E. J. (2013). Explanation feedback is better than correct answer feedback for promoting transfer of learning. *Journal of Educational Psychology*, *105*(2), 290–298. https://doi.org/10.1037/a0031026

Detterman, D. K. (1993). The case for the prosecution: Transfer as an epiphenomenon. In D. K. Detterman & R. J. Sternberg (Eds.), *Transfer on trial: Intelligence, cognition, and instruction* (pp. 1–24). Westport, CT, US: Ablex Publishing.

Dweck, C. S. (1989). Motivation. In A. Lesgold & R. Glaser (Eds.), *Foundations for a psychology of education* (pp. 87–136). Hillsdale, NJ, US: Erlbaum Associates, Inc.

Educational Testing Service (Ed.). (2012). *Practice book for the paper-based GRE revised general test* (Second). Princeton, NJ: Educational Testing Service.

Friedlaender, D., Burns, D., Lewis-Charp, H., Cook-Harvey, C. M., & Darling-Hammond, L. (2014). *Student-Centered Schools: Closing the Opportunity Gap* [Research Brief]. Stanford, CA: Stanford Center for Opportunity Policy in Education.

Gick, M., & Holyoak, K. (1980). Analogical problem solving. *Cognitive Psychology*, *12*(3), 306–355. https://doi.org/10.1016/0010-0285(80)90013-4

Goodman, J. S., Wood, R. E., & Hendrickx, M. (2004). Feedback specificity, exploration, and learning. *Journal of Applied Psychology*, *89*(2), 248–262. https://doi.org/10.1037/0021-9010.89.2.248

Judd, C. (1908). The relation of special training to general intelligence. *Educational Review*, *36*, 28–42.

Kulhavy, R. W., & Anderson, R. C. (1972). Delay-retention effect with multiple-choice tests. *Journal of Educational Psychology*, *63*(5), 505–512. https://doi.org/10.1037/h0033243

Kulhavy, R. W., White, M. T., Topp, B. W., Chan, A. L., & Adams, J. (1985). Feedback complexity and corrective efficiency. *Contemporary Educational Psychology*, *10*(3), 285–291. https://doi.org/10.1016/0361-476X(85)90025-6

Lyans, C. K. (1914). The doctrine of formal discipline. *The Pedagogical Seminary*, *21*(3), 343–393. https://doi.org/10.1080/08919402.1914.10532694

MG Prep (Ed.). (2013). *5 lb. book of GRE® practice problems*. New York, NY: MG Prep, Inc.

Briel, J., & Michel, R. (2014). Revisiting the GRE general test. In *The research foundation for the GRE revised general test: A compendium of studies,* Wendel, C., & Bridgeman, B. (Eds.), (pp. 1.1.1-1.1.8). Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/s/research/pdf/gre_compendium.pdf

Morley, M. E., Bridgeman, B., & Lawless, R. R. (2004). *Transfer between variants of quantitative items* (No. 00-06R; pp. 1–27). Retrieved from Educational Testing Service website: http://doi.wiley.com/10.1002/j.2333-8504.2004.tb01963.x

Organisation for Economic Co-Operation and Development. (2004). *Learning for the world of tomorrow: First results from PISA 2003*. https://doi.org/10.1787/9789264063556-de

Psychology Software Tools, Inc. [E-Prime 2.0]. (2012). Retrieved from http://www.pstnet.com.

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development*, *77*(1), 1–15. https://doi.org/10.1111/j.1467-8624.2006.00852.x

Salmoni, A. W., Schmidt, R. A., & Walter, C. B. (1984). Knowledge of results and motor learning: A review and critical reappraisal. *Psychological Bulletin*, *95*(3), 355–386.

Schmidt, R. A., & Bjork, R. A. (1992). New Conceptualizations of Practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*(4), 207–218. https://doi.org/10.1111/j.1467-9280.1992.tb00029.x

Schmidt, R. A., Young, D. E., Swinnen, S., & Shapiro, D. C. (1989). Summary knowledge of results for skill acquisition: Support for the guidance hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(2), 352–359. https://doi.org/10.1016/0167-9457(92)90035-A

Schwartz, D. L., Bransford, J. D., & Sears, D. (2005). Efficiency and innovation in transfer. In J. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 1–52). Greenwich, CT: Information Age Publishing. Retrieved from https://aaalab.stanford.edu/assets/papers/2005/Efficiency_and_Innovation_in_Transfer.pdf

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology*, *103*(4), 759–775. https://doi.org/10.1037/a0025140

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, *22*(2), 129–184. https://doi.org/10.1207/s1532690xci2202_1

Shipstead, Z., Harrison, T. L., & Engle, R. W. (2016). Working memory capacity and fluid intelligence: Maintenance and disengagement. *Perspectives on Psychological Science*, *11*(6), 771–799. https://doi.org/10.1177/1745691616650647

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

Thorndike, E. L., & Woodworth, R. S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. II. The estimation of magnitudes. *Psychological Review*, *8*(4), 384–395. https://doi.org/10.1037/h0071280

Vendetti, M. S., Wu, A., & Holyoak, K. J. (2014). Far-out thinking: Generating solutions to distant analogies promotes relational thinking. *Psychological Science*, *25*(4), 928–933. https://doi.org/10.1177/0956797613518079

Yazzie-Mintz, E. (2007). Voices of students on engagement: A report on the 2006 High School Survey of Student Engagement. Bloomington, IN: Center for Evaluation & Education Policy.

Tables

Table 1

Means (SD) of key variables, split by treatment condition.

Phase	Item	Hint, Hint	Hint, No Hint	No Hint, No Hint	No Hint, Hint
Training Phase I	Assessment Score	.46 (.31)	.40 (.29)	.44 (.30)	.43 (.29)
	RT	52.77 (14.83)	54.35 (16.37)	49.24 (15.39)	50.75 (14.91)
Training Phase II	Assessment Score	.43 (.29)	.42 (.29)	.43 (.30)	.39 (.31)
	RT	48.03 (16.14)	43.83 (16.61)	44.93 (16.72)	46.44 (18.55)
Transfer	Transfer Score	.41 (.22)	.39 (.23)	.40 (.26)	.37 (.24)
	Correct First Attempts	10.36 (3.74)	9.73 (3.99)	10.20 (3.54)	9.99 (3.68)

Note. Assessment is calculated as proportion correct. Response time (RT) is measured in seconds.

Correct first attempts is a numeric value between 1 and 24.

Table 2

Model comparisons for Hypothesis 1, Learning Phase 1

Model	Res. *df*	Res. *SS*	*df* Diff	SS	*F*	*p*
1. Condition*WMC	645	56.3	--	--	--	--
2. Condition + WMC	646	56.9	-1	-.33	1.87	0.08
3. Condition + WMC + First Attempts	645	53.2	1	3.81	43.05	<.001
4. Condition + First Attempts	646	53.3	-1	-.08	0.89	.33
5. Condition + First Attempts + RT	645	52.5	1	.73	8.22	<.01

Note. * indicates an interaction and a main effect are included.

Table 3

Model comparisons for Hypothesis 1, Learning Phase 2

Model	Res. *df*	Res. *SS*	*df* Diff	SS	*F*	*p*
1. Condition*WMC	645	56.80	--	--	--	--
2. Condition + WMC	646	56.86	-1	-0.07	.74	.39
3. Condition + WMC + First Attempts	645	51.73	1	5.11	58.1	<.001
4. Condition + First Attempts	646	51.82	-1	-.08	.89	.34
5. Condition + First Attempts + RT	645	51.45	1	.37	4.23	.04

Note. * indicates an interaction and a main effect are included.

Table 4

Means (SD) of key variables, split by treatment condition for hypothesis 2.

Phase	Item	Two Blocks	One Block	No Blocks
Training Phase I	Assessment Score	.58 (.24)	.53 (.23)	.58 (.24)
	RT	53.80 (14.23)	53.33 (15.09)	52.95 (14.23)
Training Phase II	Assessment Score	.55 (.23)	.54 (.23)	.54 (.21)
	RT	50.82 (16.61)	47.03 (17.08)	49.04 (16.44)
Transfer	Transfer Score	.44 (.22)	.38 (.24)	.45 (.27)
	Correct First Attempts	11.19 (3.78)	10.40 (3.87)	10.74 (3.55)

Note. Assessment is calculated as proportion correct. Response time (RT) is measured in seconds.

Correct first attempts is a numeric value between 1 and 24.

Table 5

Model comparisons for Hypothesis 2

Model	Res. *df*	Res. *SS*	*df* Diff	*SS*	*F*	*p*
1. Condition*WMC	401	23.5	--	--		
2. Condition + WMC	403	23.6	-2	-.17	1.44	.24
3. Condition + WMC + First Attempts	402	19.3	1	4.34	74.18	<.001
4. Condition + First Attempts	403	19.3	-1	-.02	.26	.61
5. Condition + First Attempts + RT	402	52.5	1	.13	2.20	.14

Note. * indicates an interaction and a main effect are included.

Table 6

Model comparisons for Hypothesis 3

Model	Res. *df*	Res. *SS*	*df* Diff	SS	*F*	*p*
1. Condition*WMC	195	11.44	--	--	--	--
2. Condition + WMC	196	11.44	-1	0.00	0.00	.99
3. Condition + WMC + First Attempts	195	9.73	1	1.72	34.27	<.001
4. Condition + First Attempts	196	9.75	-1	-.02	.42	.52
5. Condition + First Attempts + RT	195	9.72	1	.02	.49	.49
6. Condition + First Attempts + No. Hints	195	9.73	0	-.01	--	--
7. Condition + First Attempts + No. Hints + Hint Succ.	194	9.71	1	.02	.37	.54

Note. * indicates an interaction and a main effect are included. No. Hints is the number of hints that were received. Hint Succ. is the probability of being correct after receiving a hint. No difference between Model 5 and Model 6 is observed because two predictors are swapped. I am interested in Model 6 compared to Model 4.

Figures

1a.

Working at their respective constant rates machine I makes 240 copies in 8 minutes and machine II makes 240 copies in 5 minutes. At these rates, how many more copies does machine II make in 4 minutes than machine I makes in 6 minutes?

Hint: First determine how many copies are made a second.

- A. 10
- B. 12
- C. 15
- D. 20
- E. 24

1b.

If an integer is divisible by both 8 and 15, then the integer also must be divisible by which of the following?

Hint: First find a number that is divisible by 8 and 15.

- A. 16
- B. 24
- C. 32
- D. 36
- E. 45

1c.

Quantity A	Quantity B
25^7	5^{15}

Hint: Make the base value both 5 or 25.

- A. Quantity A is greater
- B. Quantity B is greater
- C. The two quantities are equal
- D. The relationship cannot be determined from the information given

---

1d.

If $\frac{a-b}{a+b} = 2$ and $b = 1$, what is the value of a?
Hint: In the equation, replace b with 1

- A. 1
- B. 0
- C. -1
- D. -2
- E. -3

1e.

In Triangle ABC, AB = 12, AC = 10, and BC = 5.

Quantity A	Quantity B
The measure of angle A	The measure of angle C

Hint: The longer the side is, the larger its corresponding angle.

- A. Quantity A is greater
- B. Quantity B is greater
- C. The two quantities are equal
- D. The relationship cannot be determined from the information given

---

1f.

Which of the following inequalities have at least one positive solution and at least one negative solution?

Hint: Sketch out the graphs for the equations on each side of the inequalities.

- A. $\frac{5}{3}x < x$
- B. $x^3 < x$
- C. $x - 6 < x - 7$

1g.

Of 30 theater tickets sold, 20 tickets were sold at prices between $10 and $30 each and 10 tickets were sold at prices between $40 and $60 each.

Quantity A	**Quantity B**
The average (arithmetic mean) of the prices of the 30 tickets	$50

Hint: Try to find the average if all tickets were the most expensive and the average if all tickets were the cheapest.

- A. Quantity A
- B. Quantity B
- C. The two quantities are equal
- D, The relationship cannot be determined from the information given

---

1h.

If $a > b > c > d$ and $a = 2$, which of the following must be negative?

Hint: Does the question state that the values must be integers?

- A. ab
- B. ac
- C. bd
- D. None of the above

1i.

Among the people attending a convention in Europe, 32 percent traveled from Asia and 45 percent of those who traveled from Asia are women. What percent of the people at the convention are women who traveled from Asia?

Hint: Try to working through this problem by saying there are 100 people at the convention.

- A. 14.4%
- B. 73.1%
- C. 27.4%
- D. 9.8%
- E. 44.7%

Figure 1a-i. One sample question and its corresponding hint from each topic. (a) Two-variable word problems; (b) Divisibility and prime numbers; (c) Exponents and roots; (d) Functions and formulas; (e) Geometry; (f) Inequalities and absolute value; (g) Measures of central tendency; (h) Number properties; (i) Probability.

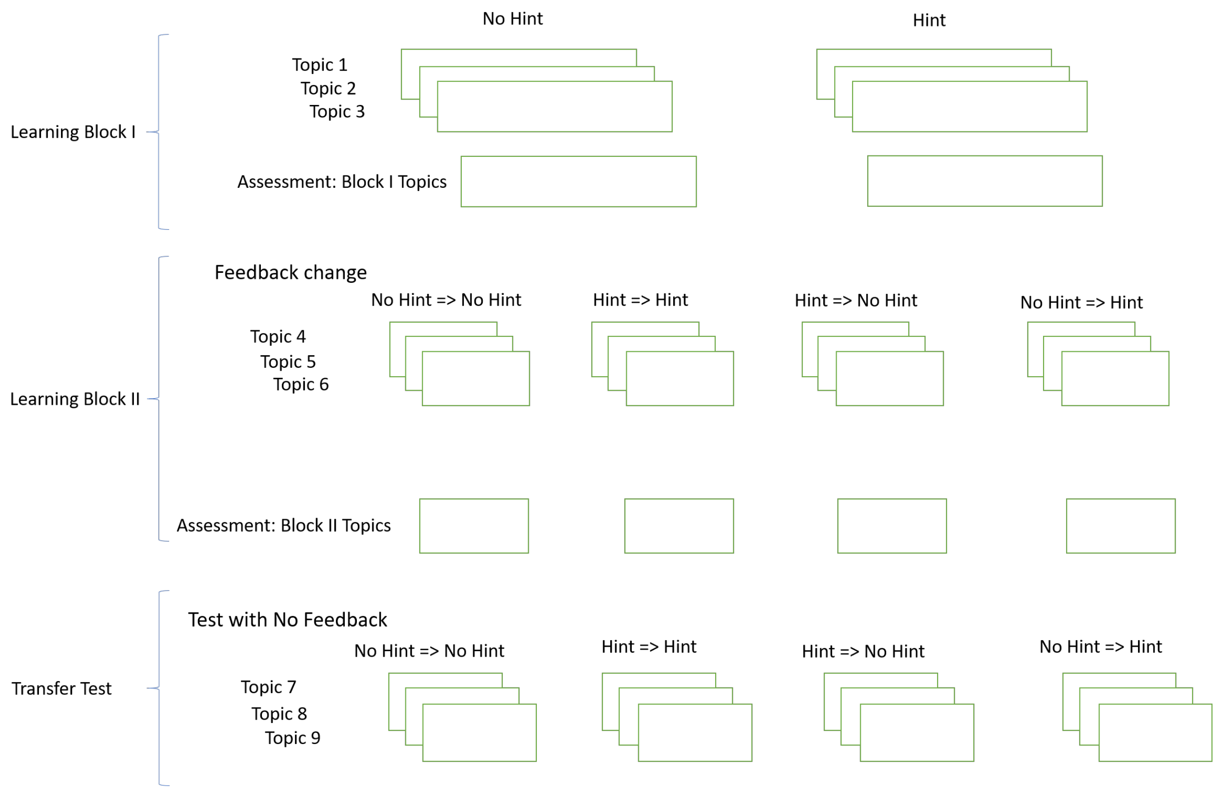


Figure 2. Diagram depicting the training and transfer task. Participants first train on three topics which have four questions each. When they answer a question incorrectly, they either are given a hint and told to try again, or they are simply told that they were incorrect and to try again. Participants then receive one assessment question from each topic, they are not allowed reattempts on these assessment questions. Learning block II follows the same procedures except each of the groups is divided in half so that they will either continue receiving the same kind of feedback (hints or no hints) or they will receive a different kind of hint. During this phase they train on three new topics. Finally, participants encounter two questions from three new topics in the transfer phase. They do these questions without feedback or reattempts.

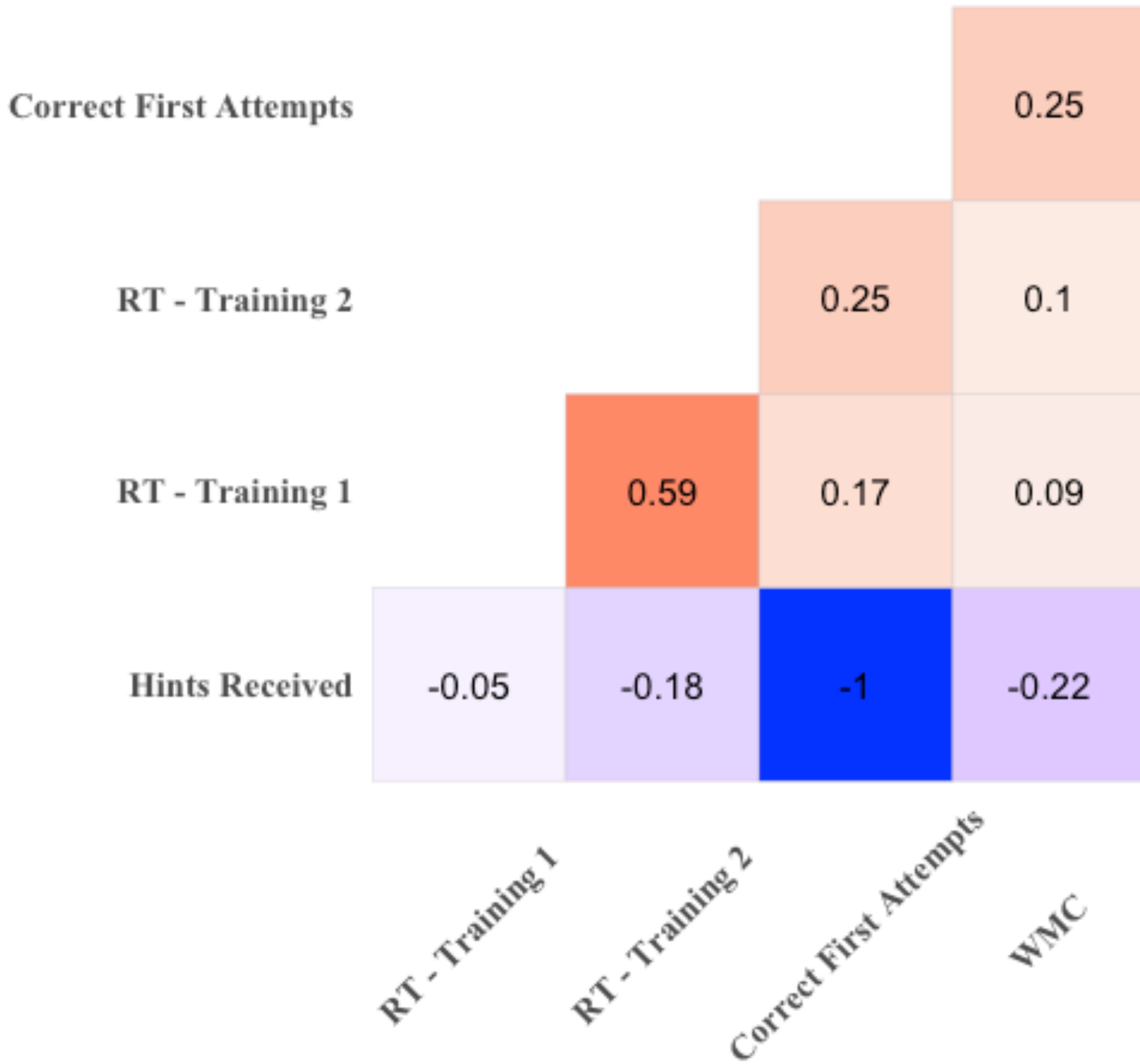


Figure 3. Correlation heatmap for all covariates considered across all conditions. Note, hints received is artificially low due to it being tied to treatment (one condition will never receive hints). $|r| > .07$ is significant at $\alpha = .05$.

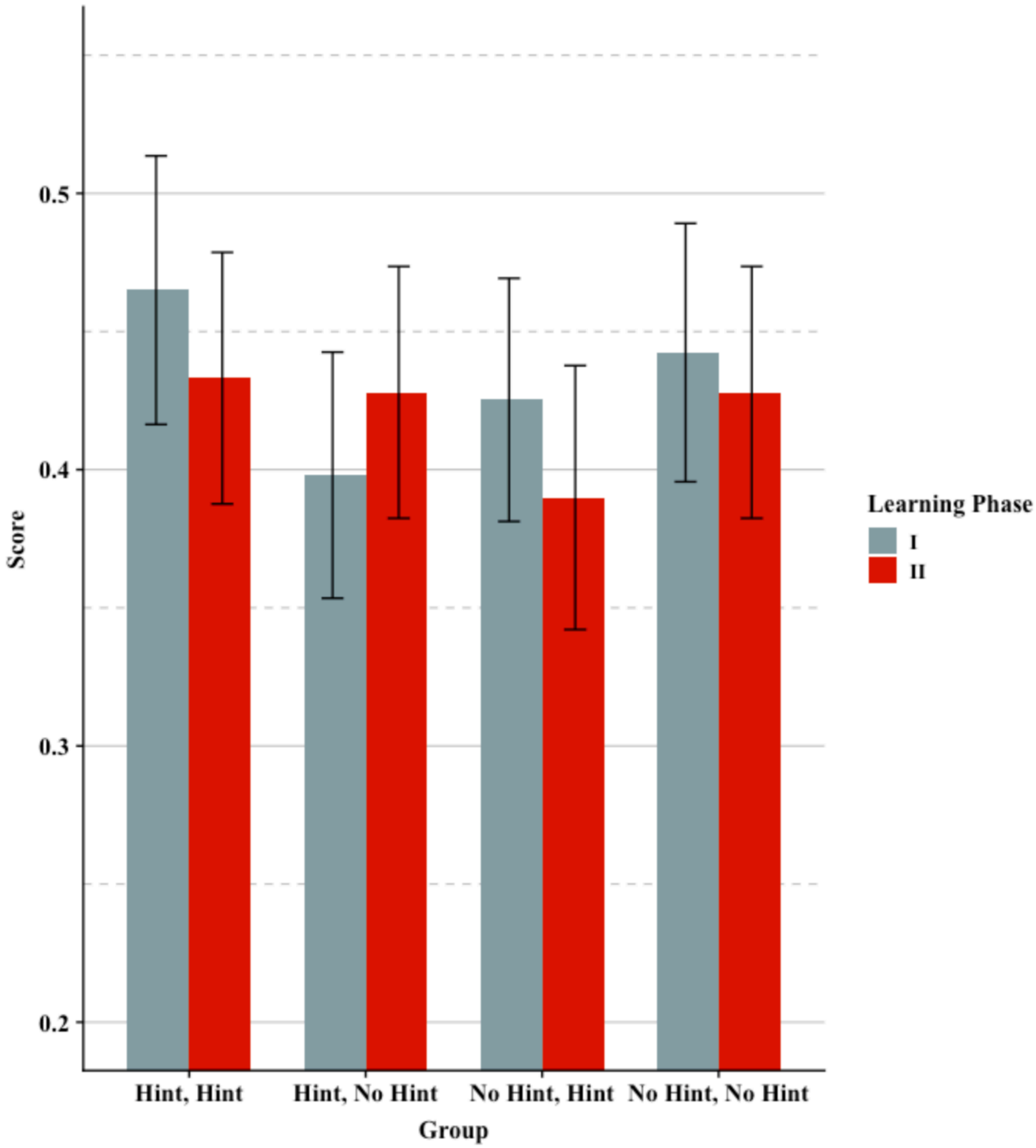


Figure 4. Performance of groups on learning phase assessments. 95% confidence intervals are included.

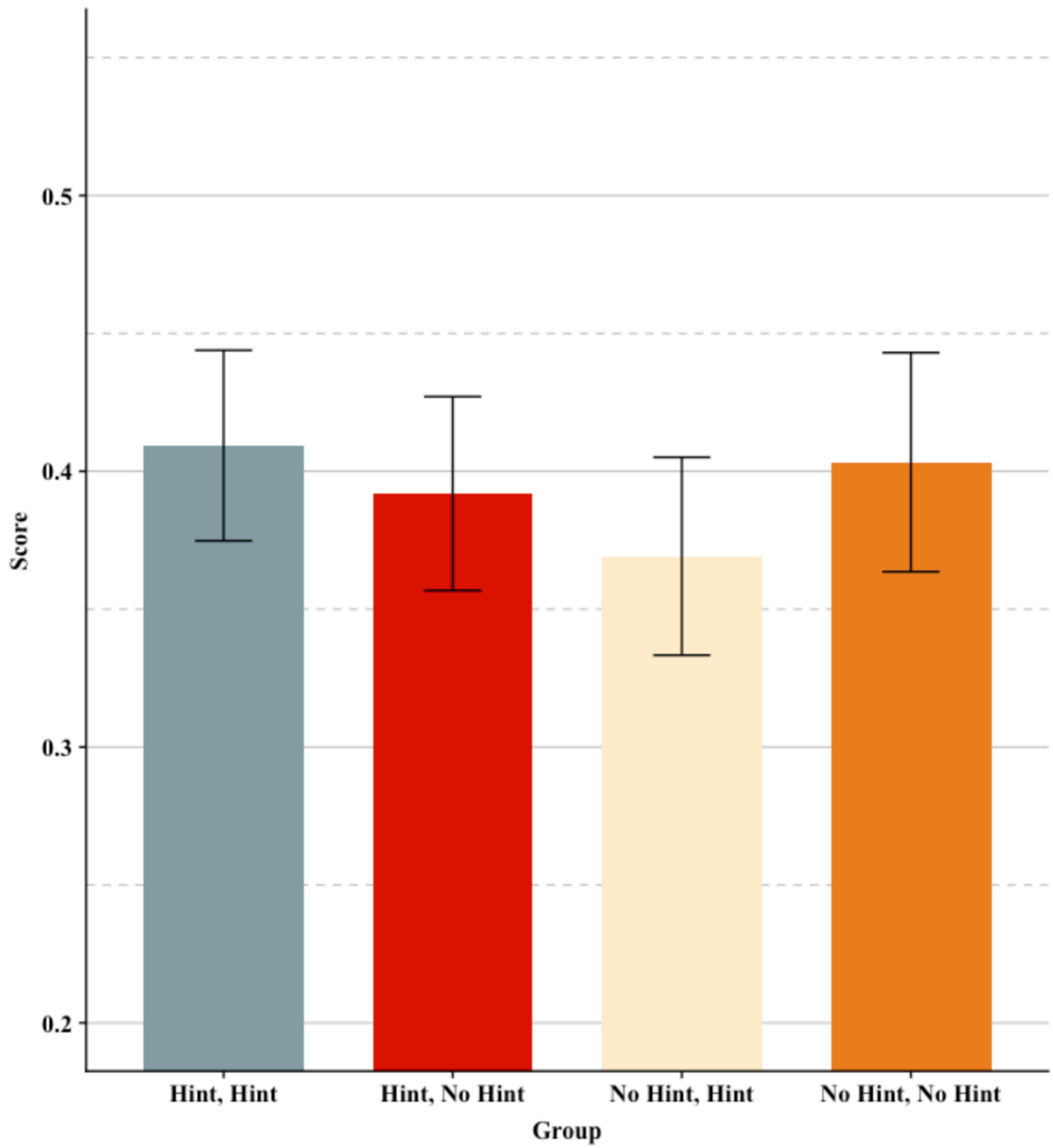


Figure 5. Performance of groups on transfer test. 95% confidence intervals are included.

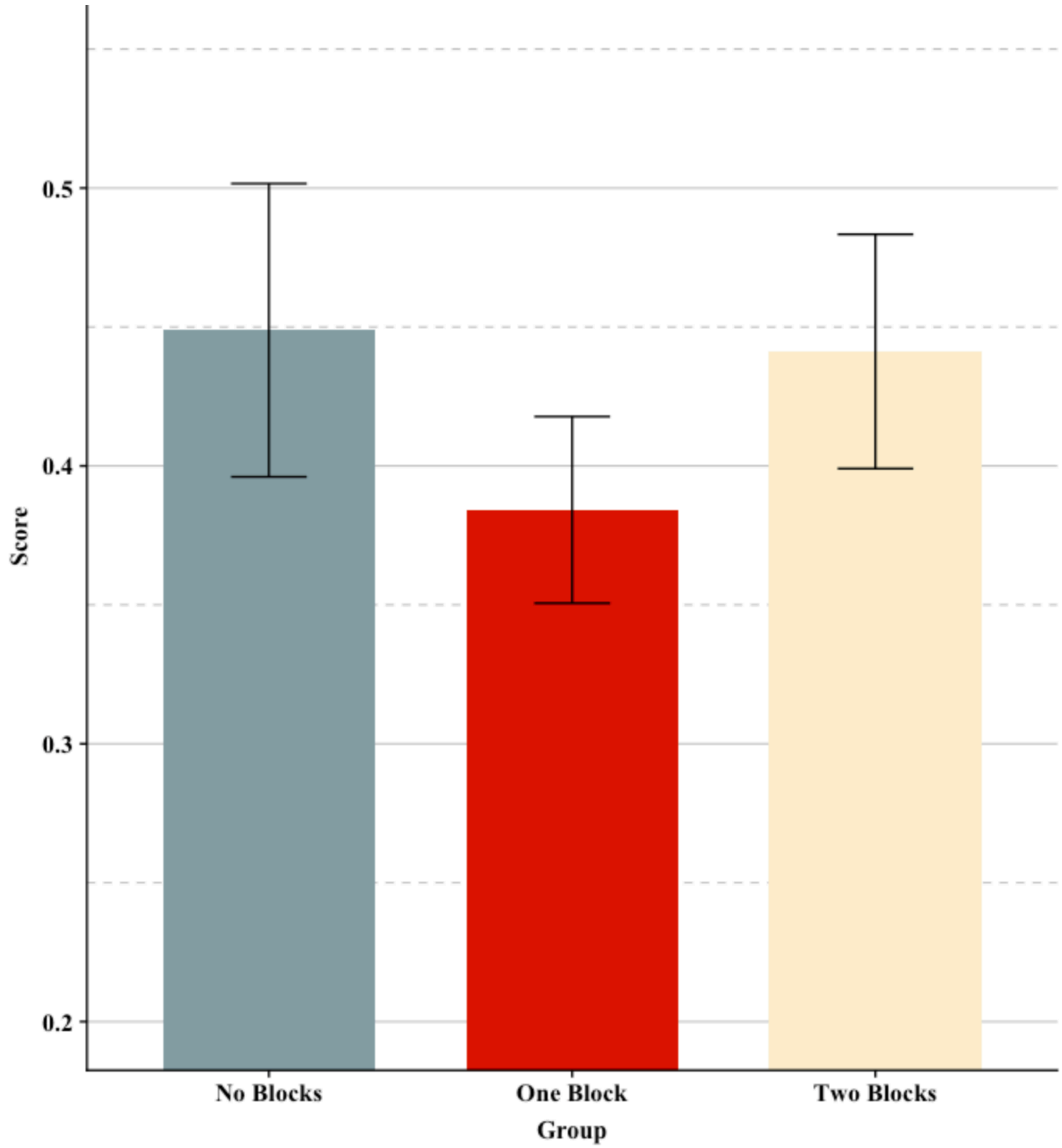


Figure 6. Mean transfer scores between groups depending on how many blocks of hints are provided. 95% confidence intervals are included.

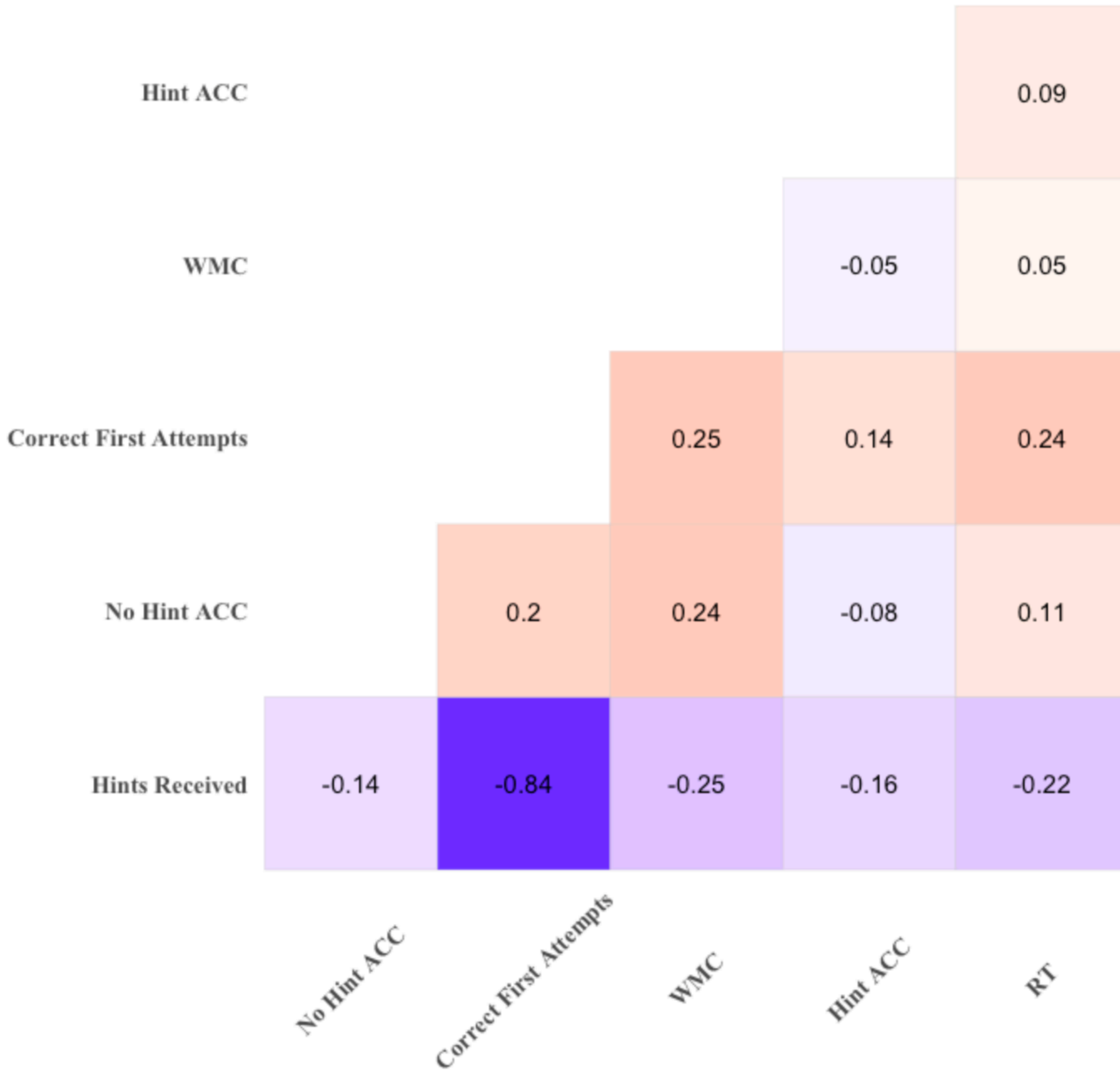
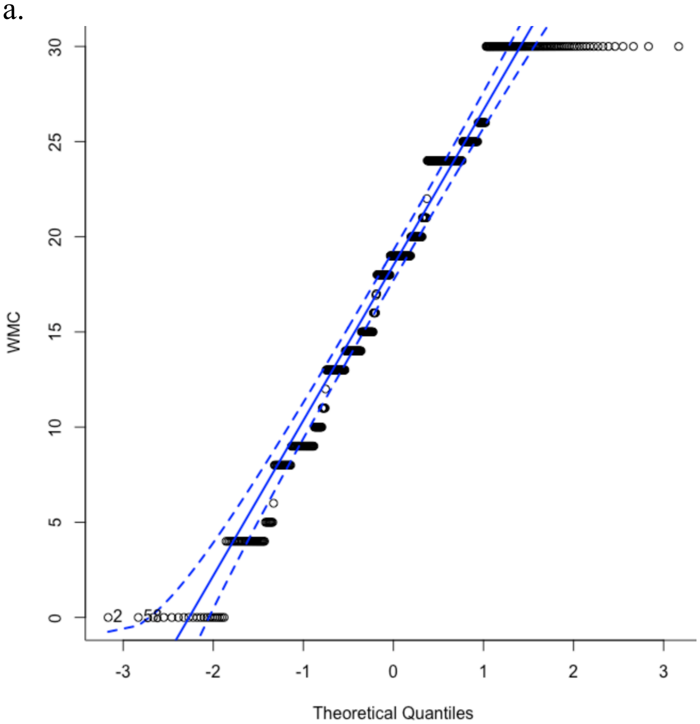


Figure 7. Updated correlation heatmap of covariates for the two groups which received sequenced feedback. $|r| > .15$ is significant at $\alpha = .05$.

a.



b.

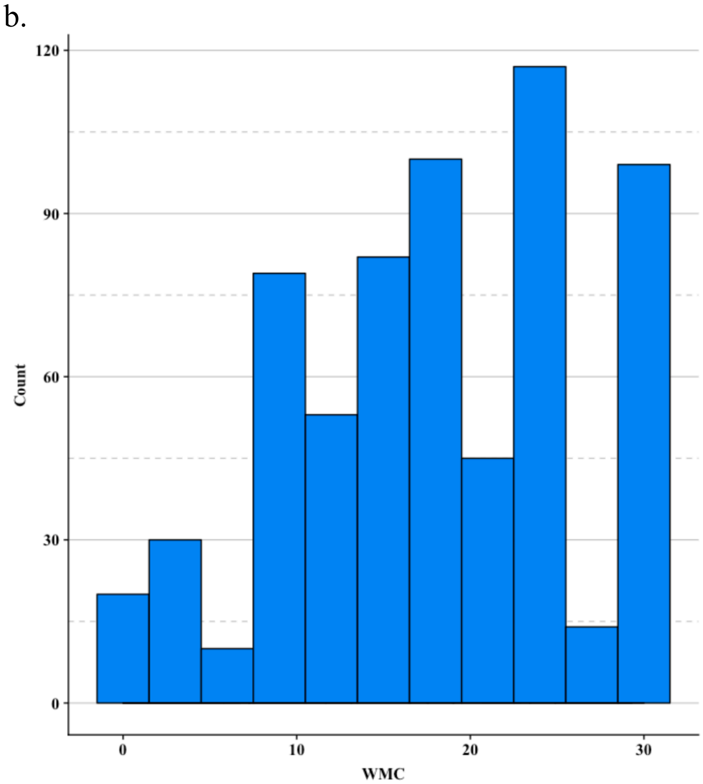


Figure 8. Q-Q plot of WMC scores (a), and a histogram of WMC scores (b).