Allen, P. J., Fielding, J. L., East, E. C., Kay, R. H. S., Steele, C. S., & Breen, L. J. (2021). Using StatHand to Train Structural Awareness and Promote the Development of Statistic Selection Skills. *Scholarship of Teaching and Learning in Psychology*, 7(1), 53-67. <https://doi.org/10.1037/stl0000177>

Peer reviewed version

Link to published version (if available):
[10.1037/stl0000177](https://doi.org/10.1037/stl0000177)

[Link to publication record in Explore Bristol Research](#)
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via American Psychological Association at <https://psycnet.apa.org/fulltext/2019-75005-001.html>. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: <http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

Using StatHand to Train Structural Awareness and Promote the Development of Statistic Selection Skills

Peter J. Allen, Jessica L. Fielding, Elizabeth C. East, and Ryan H. S. Kay

University of Bristol

Chloe S. Steele and Lauren J. Breen

Curtin University

Peter J. Allen, School of Psychological Science, University of Bristol, United Kingdom; Jessica L. Fielding, School of Psychological Science, University of Bristol, United Kingdom; Elizabeth C. East, School of Psychological Science, University of Bristol, United Kingdom; Ryan H. S. Kay, School of Psychological Science, University of Bristol, United Kingdom; Chloe S. Steele, School of Psychology, Curtin University, Australia; Lauren J. Breen, School of Psychology, Curtin University, Australia.

Support for this project has been provided by the Australian Government Department of Education and Training (ID13-2954). The views expressed in this project do not necessarily reflect the views of the Australian Government Department of Education and Training.

The results reported in this article were previously presented at the 2019 British Psychological Society Division of Academics, Researchers, and Teachers in Psychology (BPS DART-P) and European Society for Psychology Learning and Teaching (ESPLAT) conferences.

Correspondence concerning this article should be addressed to Peter Allen, School of Psychological Science, University of Bristol, 12a Priory Rd, Bristol, UK, BS8 1TU. Email: p.allen@bristol.ac.uk

Abstract

Psychology students struggle to recall, recognise or explain how they would select appropriate statistics for common research designs. These selection skills are underpinned by structural awareness, which is the ability to look past the surface (or topic) features of a research design and focus instead on its deep structural characteristics. Although most psychology undergraduates display limited structural awareness, it can be trained. In this pre-registered experiment, we designed and evaluated a novel method of training structural awareness. This training method made use of StatHand, a free iOS and web application, in scaffolded activities designed to highlight how the structural (but not surface) characteristics of a research design determine the selection of an appropriate statistical analysis. Bayesian analyses clearly indicated that this training was effective. Specifically, trained undergraduate psychology students ($n = 50$) out-performed an un-trained control group ($n = 52$) on five measures of structural awareness (performance on two sets of triad judgement tasks, two sets of explanation tasks and a scenario generation task) immediately following training, and again one week later ($\delta = 0.71$ to 1.60). At both time points, the trained students also showed greater selection skills than the un-trained control students ($\delta = 0.52$ and 0.57). Finally, on five of these six outcome measures, the trained students showed no decrease in performance between the two time points. Educators are encouraged to consider how they can adapt our methods and materials for deployment in a classroom context or online activities.

Keywords: mobile learning app; statistics education; decision tree; Scholarship of Teaching and Learning (SoTL); iPad; iPhone

Using StatHand to Train Structural Awareness and Promote the Development of Statistic Selection Skills

In 2014, the Statistical Literacy Taskforce of the Society for the Teaching of Psychology (Division 2 of the American Psychological Association) specified five learning goals for an undergraduate psychology degree. The second of these was, “apply appropriate statistical strategies to test hypotheses”. Psychology students who meet this goal are, amongst other things, able to “select ...an appropriate statistical analysis for a given research design, problem, or hypothesis” (p. 2). Research indicates that many psychology students struggle to recall, recognise and explain how they would select appropriate statistics for common research designs.

To illustrate weaknesses in recall, Gardner and Hudson (1999) asked psychology and education students to specify appropriate statistical analyses for a range of typical research scenarios. Despite all having completed at least two years of study, the students were only able to name appropriate analyses for around 25% of the scenarios to which they were exposed. Even the highest performing student, a fourth-year undergraduate, had an accuracy level of just 56%. To illustrate weaknesses in recognition, Ware and Chastain (1989) administered a multiple-choice statistic selection test at the end of a first-year psychology statistics module. When they wrote the test, Ware and Chastain (1989) believed it would be easy enough for the students to master. In actuality, the students averaged less than 45%. Finally, to illustrate weaknesses in explanation, Allen, Dorozenko, and Roberts (2016) asked psychology undergraduates to describe how they would select appropriate statistics for several typical research scenarios. At the time these students were interviewed, they had completed an average of three research methods and statistics modules. Despite their experience, the students described selection strategies that were haphazard and inefficient, and unlikely to reliably lead them to appropriate statistical analyses.

These weaknesses have prompted educators to develop various tools to guide students through the statistic selection process. The most common of these are decision trees, which are widely included in introductory statistics textbooks (e.g., Allen, Bennett, & Heritage, 2019; Nolan & Heinzen, 2017). The inclusion of decision trees in textbooks is supported by research indicating that they promote timely and accurate statistic selection, and that they are popular with students (Carlson, Protsman, & Tomaka, 2005; Protsman & Carlson, 2008). However, despite their efficacy and popularity, paper-based decision trees are not without limitations. Foremost amongst these is brevity. They usually need to fit onto a single sheet of paper. Consequently, information that would assist navigation (e.g., definitions of key terms, examples etc.) is either spatially separated from paper decision trees or is missing entirely.

To overcome these physical constraints, decision trees to guide statistic selection have been developed for digital media (e.g., Koch & Gobell, 1999). One recent example is StatHand (Allen et al., 2016, 2017), a free iOS (available from the App Store) and web (see <https://stathand.net>) application that asks users a series of questions about the nature of their research design. These questions are annotated with definitions and examples, which means that even relative novices can use StatHand without needing to consult further afield. In answering the application's questions, the user progressively narrows in on, and ultimately identifies, a statistic appropriate to their circumstances. In a recent experimental evaluation, 217 psychology students were randomised to four different decision-making guides and asked to select an appropriate statistical analysis for each of five research scenarios (Allen, Finlay, Roberts, & Baughman, 2019). The decision-making aids were StatHand, a familiar paper decision tree, a familiar textbook, and the decision tree and textbook combined. The students in the StatHand condition demonstrated higher statistic selection accuracy than students in the other three conditions ($\delta = 0.50$ to 0.64).

Ostensibly, tools like StatHand promote statistic selection accuracy because they encourage ‘structural awareness’ (Allen et al., 2017). Structural awareness reflects the ability to look past the surface (or topic) level features of a research design and instead focus on its deep structural characteristics (i.e., the number and nature of its independent and dependent variables) and the associations between them (Quilici & Mayer, 2002). Although most undergraduate psychology students display limited structural awareness (Allen, Dorozenko, & Roberts, 2016; Rabinowitz & Hogan, 2008), it can be trained. Such training can take multiple forms. For example, Quilici and Mayer (1996, 2002) and Yan and Lavigne (2014) exposed students to worked examples, research scenarios grouped to highlight structural similarities and differences, and activities designed to focus students’ attention on these structural characteristics. Relative to untrained students, the trained students in these studies were more likely to group new research scenarios according to their structural characteristics (rather than surface or topic characteristics), correctly identify the structural characteristics defining each group, and write new scenarios that also shared these structural characteristics. Furthermore, the trained students were more likely to select and apply appropriate statistics to novel research scenarios. Each of these behaviours reflects structural awareness.

The current study took a different approach to training structural awareness. This approach was inspired by the success of ‘wise’ psychological interventions across a range of contexts, including higher education (Walton, 2014). Wise interventions are brief and targeted. They are designed to change specific behaviours (in both the short and longer term) by exploiting specific psychological processes. In our study, those processes were meta-cognition and structural awareness. Our training made use of StatHand in scaffolded activities designed to guide students to appropriate statistical analyses for four simple research designs. We did this in a way that explicitly highlighted the deep structural characteristics of each design and repeatedly prompted students to reflect on *how* and *why* the structural (but not surface)

characteristics of a study determine the selection of an appropriate statistical analysis. We hypothesised that, compared to un-trained students, trained students would perform better on five tasks reflecting structural awareness. These five tasks were Surface vs. Deep Structural (S-D) and Deep Structural vs. Neither (D-N) triad judgement and explanation tasks (Rabinowitz & Hogan, 2008; Yan & Lavigne, 2014), and a scenario generation task (Quilici & Mayer, 2002). Furthermore, we hypothesised that, compared to un-trained students, trained students would demonstrate stronger statistic selection skills (Ware & Chastain, 1989, 1991). Finally, we hypothesised that these effects would be evident immediately following training (or a control task) and again following a one-week delay.

Method

Design

This was a pre-registered (see <https://osf.io/gjptz>) mixed factorial experiment with one randomised between subjects independent variable (IV; condition: training or control), one within subjects IV (time: time 1 immediately following training and time 2 one week later) and seven dependent variables (DVs). Five of these DVs reflect structural awareness (the S-D and D-N triad judgement and explanation tasks, and the scenario generation task). The sixth, selection skills, reflects the ability to correctly identify appropriate statistical analyses for familiar research scenarios. The interpretation of the final DV, performance on the S-N triad judgement tasks, is ambiguous. This variable has been analysed for exploratory purposes only.

Participants

A total of 102 undergraduate psychology students at the University of Bristol in the United Kingdom participated in this study for either course credit (68.62%) or £25 cash. The rationale for $N = 102$ reflects our judgements about anticipated effect sizes, the smallest effect size likely to be of interest to educators and the number of psychology undergraduates at the University of Bristol that we could feasibly test in an academic year. A frequentist sensitivity

power analysis using G*Power 3.1.4 (Faul, Erdfelder, Lang, & Buchner, 2007) indicated that with $N = 102$ we had power of .80 for detecting an effect of at least $d = 0.64$ in a one-sided independent samples t -test evaluated for significance at $\alpha = .01$. This effect size was used as the basis for a fixed- n Bayes Factor Design Analysis (BFDA; Schönbrodt & Wagenmakers, 2018), which indicated a .86 probability of observing Bayes Factors (BFs) > 3 (i.e., at least moderate evidence in favour of the research hypothesis; Wagenmakers et al., 2018) in a one-sided Bayesian independent samples t -test. The probability of inconclusive or anecdotal evidence (BFs between 3 and .33) was estimated to be .14 whilst the probability of false negatives (BFs $< .33$) was $< .01$.

As illustrated in Table 1, the training and control groups shared similar demographic characteristics. Three students (two from the training group) chose not to return for the second wave of testing.

[INSERT TABLE 1 ABOUT HERE]

Training Materials

The materials used during training were StatHand (Allen, Roberts, et al., 2016; Allen et al., 2017) installed on an iPad, training booklets and practice booklets.

Training booklets. At the top of each page of each four-page training booklet was a research scenario (see section S1 of the online supplement at <https://osf.io/24v7c/>) with structural characteristics corresponding to an independent samples t -test, a paired-samples t -test, a chi-square test of contingencies or a McNemar test of change. The scenario was followed by the question, “what statistical test should [the protagonist in the scenario] use?” Below this were two columns headed by “why did you choose this test?” (on the left) and “mapping story onto StatHand example” (on the right). The eight possible versions of this booklet contained the same four research scenarios in different orders. These orders ensured that the experimenter

exposed each participant to both within and between subjects designs, and to nominal and continuous data before handing over the iPad during training.

Practice booklets. The eight versions of this booklet contained the four research scenarios in section S2 of the online supplement and were otherwise identical to the training booklets described above.

Measures

The following measures were packaged into an online questionnaire deployed using Qualtrics (Qualtrics, Provo, UT).

Triad judgement tasks. The two sets of 16 research scenarios developed for three sets of triad judgement tasks were inspired by Quilici and Mayer (1996) and are presented in sections S3 and S4 of the online supplement. The scenarios vary on surface and deep structural characteristics. The surface characteristics are defined by narrative content, whereas the deep structural characteristics are defined by statistical analyses. The four statistical analyses represented in the scenarios are the independent samples *t*-test, paired-samples *t*-test, chi-square test of contingencies, and McNemar test of change. These analyses were chosen because they clearly differ on two dimensions: the nature of the IV (between subjects or within subjects) and the nature of the DV (categorical or interval/ratio). To remove any potential ambiguity, and to standardise the scenarios as much as possible, the DVs for all the chi-square and McNemar scenarios have only two levels, whereas the DVs for all the *t*-test scenarios are ratio level. Furthermore, how participants are assigned to levels of the IV (or how ordering is determined in the within subjects designs) in these scenarios is always ambiguous.

The triad judgement tasks were adapted from Rabinowitz and Hogan (2008). In each of 24 trials, participants were presented with a target research scenario and a pair of comparison scenarios. The target and comparison scenarios share deep (D) structural characteristics, surface (S) characteristics, or neither (N). Therefore, the possible comparison pairs are S-D, D-

N and S-N. Each of four targets (see section S5 of the online supplement) were presented to participants six times; twice with each possible comparison pair. The full set of 24 trials (which were randomly ordered within four randomised blocks) are presented in section S5 of the online supplement. It should be noted that the S and N comparison scenarios always differ from the target on just one structural characteristic (i.e., the nature of the IV or the nature of the DV). On each trial, participants were asked to select the comparison scenario that ‘goes best’ with the target. Their ordering was randomised. On S-D and D-N trials, selection of the D comparison scenarios was assumed to reflect structural awareness and scored as 1. Selecting either S or N (or the absence of a response) was scored as zero. Therefore, possible scores on each of these two sets of triad judgement tasks ranged from 0 to 8, with higher scores reflecting greater structural awareness. Interpretation of performance on the S-N trials is more ambiguous, as selecting S in the absence of D cannot be assumed to reflect a structural awareness deficit. Indeed, selecting S in an S-N trial would arguably be the ‘best’ choice regardless of whether or not one attends to the deep structural characteristics of the research scenarios. For these reasons, for the purposes of analysis, we scored the selection of S on these trials as 1, which produced a possible range of 0 to 8. However, it should be noted that higher or lower scores cannot be assumed to reflect either the absence or presence of structural awareness.

Explanation tasks. The explanation tasks were adapted from Yan and Lavigne’s (2014) problem categorisation task, which asked participants to categorise word problems and provide an explanation for the categories produced. For the present study, the explanation tasks were combined with the triad judgement tasks, such that participants were asked to give as many reasons as possible for why their selected comparison scenario ‘goes best’ with the target scenario. On the S-D and D-N trials these explanations give insight into whether participants identified the deep structural similarities between the target and relevant comparison scenario.

For these trials, the two essential features to be identified were always the nature of the IV/design (between or within subjects) and the nature of the DV (categorical or interval/ratio). Each correctly identified structural feature on trials where participants selected the D comparison scenario were assigned one point. Thus, scores for the S-D and D-N trial sets had a possible range of 0 to 16, with higher scores indicative of greater structural awareness. A provisional list of correct answers was included in Appendix E of our pre-registration. This was slightly expanded during coding (see section S6 of the online supplement), which was completed blind to participants' conditions. A random sample of 10% of the time 1 responses were independently coded by two researchers to assess inter-rater reliability. Weighted Kappa was 0.88, 95% CI [0.82, 0.94]. Responses on the S-N trials were not coded or analysed.

Scenario generation task. The scenario generation task was adapted from Quilici and Mayer (2002). It repurposed the target scenarios in S3 and S4 by asking participants to generate a new scenario that was “similar” to each target scenario. Presentation of these trials was randomised. For each trial, the three essential features were the number of groups or conditions (always two), the nature of the IV/design (between or within subjects) and the nature of the DV (categorical or interval/ratio). Where each of these matched the target, a point was awarded. Therefore, the maximum possible score for this task was 12, with higher scores presumed to reflect greater structural awareness. Coding was blind to participants' conditions, and a random sample of 10% of the time 1 responses were coded independently by two researchers to assess inter-rater reliability. Weighted Kappa was initially 0.62, 95% CI [0.44, 0.80], although over 90% of disagreements were small (i.e., one point on a four-point scale). Following discussion and the revision of some coding, weighted Kappa was re-calculated as 0.85, 95% CI [0.74, 0.97]. The remaining disagreements centred around ambiguous language that could reasonably be interpreted in multiple ways. The following is an example of a participant generated scenario in response to a target scenario with a between subjects IV/design and a ratio level DV that

was awarded the full three points. The raw data generated by all participants can be found at <https://osf.io/24v7c/>.

A driving instructor wants to know if flattery will boost the confidence of his students. For half of the students, he doesn't mention anything about their skills and only gives instructions. For the other half, he tells them all that they drive like Vin Diesel. He then asks them how many more sessions they think it will take for them to pass their test.

Selection skills task. Following completion of the scenario generation task, participants were asked to select an appropriate statistical analysis for each of the four target scenarios from a list of four options: independent samples *t*-test, paired samples *t*-test, chi-square test of contingencies, and McNemar test of change. Presentation of these four trials was randomised. Possible scores on this task ranged from 0 to 4, with higher scores reflecting greater selection skills.

Procedure

Prior to the collection of any data, this study was reviewed and approved by the research ethics committee at the University of Bristol (reference number: 75921). Training and data collection at time 1 were individual for all participants. At time 2, data collection occurred in groups of up to four participants. For 94% of participants this occurred seven days after time 1 (range = 6 to 11 days). The first four authors shared responsibility for training and data collection. To maximise standardisation, they followed the detailed protocols in the study pre-registration and practiced with each other before training or collecting data from any participants. On average, time 1 took 90 minutes for the training condition and 75 minutes for the control condition. Time 2 took, on average, 45 minutes for both conditions.

Time 1. Following standard informed consent procedures, participants were randomised to either the training or control condition in blocks of six within each year of study. This increased the likelihood of relatively equal group sizes and a similar profile of abilities

and experience across groups. Following the training/control procedures summarised below, all participants were directed to an online questionnaire containing the triad judgement, explanation, scenario generation and selection skills tasks, as well as demographic measures.

Training condition. The experimenter sat beside each participant and explained that, “there are two parts to this experiment today. In the first part I’m going to show you how to use an iPad application, and then give you some time to practice using it. Then, I’m going to take the iPad away, and ask you to complete some sorting and story writing tasks. You’ll complete these on a computer.”

The experimenter then presented a training booklet to the participant and read out the scenario at the top of the first page. Following this they asked if the participant knew the answer to the first question, “what statistical test should [the protagonist in the scenario] use?” The experimenter wrote down the participant’s response but did not provide any feedback.

The experimenter then informed the participant that “when you’re faced with a problem like this, and you don’t know which statistical test to use, there is a process that you can follow to find an answer. There are lots of books, websites and apps that will guide you through this process. Today, I want to show you an iPad app called StatHand”. At this point, the experimenter launched StatHand in landscape mode, and noted that, although there are five broad research goals on the home screen, the current experiment would be focusing exclusively on the ‘compare samples’ goal. The experimenter also noted the information icons to the right of each option, which can be used to clarify terms that are unclear.

The experimenter then used StatHand to verbally walk the participant through the decisions that must be made to identify an appropriate statistical analysis for the first training scenario. At each decision point, the experimenter read out the decision question (e.g., “what type of data is your dependent variable?”) the answer options (e.g., “nominal”, “ordinal” and “interval/ratio”) and relevant parts of the further information (e.g., “a nominal variable is

categorical...”). He or she then read out the correct answer, and related it back to the scenario (e.g., “our dependent variable is nominal because participants were asked to indicate whether they were satisfied or not satisfied with the medicine. ‘Satisfied’ and ‘not satisfied’ are mutually exclusive categories...”). When all relevant decisions had been made, the name, purpose and illustration of the analysis suggested by StatHand were read aloud, and the analysis name was written in the space provided in the training booklet.

Following this, the experimenter pointed out the ‘History’ feature in StatHand, which provides a list of all the decisions made. He or she then used it to answer the second training booklet question, “why did you choose this test?” Finally, in the mapping exercise, the experimenter illustrated the correspondence between the structural and surface characteristics of the scenario and demonstrated how the structural characteristics of the scenario were the same as those in the example on StatHand.

The experimenter then repeated this process for the second scenario in the training booklet, and then asked the participant to follow their process for the remaining two scenarios. Feedback was provided for the participant’s answers to these (e.g., “it is correct that the dependent variable is nominal, because ...”, “however, the independent variable actually has two levels, which are ...” etc.), and the participant was encouraged to correct any errors on the booklet. An example of a page from a completed training booklet can be found in Allen, Fielding, Kay, and East (in press).

Following training, the experimenter handed the participant one of the eight possible practice booklets. It was ensured that the order of implied analyses in a participant’s practice booklet differed from that in their training booklet. The participant was told that the booklet contained four more practice scenarios, which they should use StatHand to solve. No feedback was provided for responses in the practice booklets.

The full training process lasted approximately 30 minutes. It should be noted that this training was developed to be relatively authentic. That is, it is plausible that a research methods instructor would introduce students to the process of using an application like StatHand to identify appropriate statistical analyses for different types of research problems in the manner described above. However, it is recognised that they would likely do this in a small group teaching or lecture context, rather than with students individually. Randomisation and control necessitated individual training in this study.

Control condition. The experimenter sat beside each participant and explained, “there are two parts to this experiment today. In the first part I’m going to show you how to use an iPad application, and then give you some time to practice using it. Then, I’m going to take the iPad away, and ask you to complete some sorting and story writing tasks. You’ll complete these on a computer.” The experimenter opened the Paperama app on the iPad. This is a virtual origami app, in which the user is required to make a shape with a specified number of ‘folds’. There are multiple levels, with each more complex than the last. For our purposes, the app was reset to level 1 for each participant. The experimenter demonstrated how the application works by completing the first level. On a sheet of paper they recorded the number of folds made, a description of the folds, and an explanation for why those folds ‘worked’ (for example, “I folded along the diagonal between the top right and bottom left corners to make a triangle that fit within the dashed boundary lines”). After the first level, the iPad and sheet were handed to the participant, who was asked to complete levels until told to stop. They were told that they should take their time, and that writing a full description and explanation for each level was important. Participants were left to complete this task for approximately 15 minutes.

Time 2

Participants completed triad judgement, explanation, scenario generation and selection skill tasks that were identical in structure (but not topic/surface content) to those used at time

1. The research scenarios used in these tasks are in S4 of the online supplement.

Data Analysis

We analysed our data using both Bayesian and frequentist methods. Both approaches lead to the same conclusions. The Bayesian results are reported in this paper, while the frequentist results can be found in section S7 of the online supplement. Our raw and processed data, which can be used to reproduce our analyses, are located at <https://osf.io/24v7c/>. Descriptions of the variables in this file, and how they were computed, are in S8 of the online supplement. All our analyses were independently reproduced and cross-checked by one or more co-authors prior to submission.

One-sided Bayesian independent samples t -tests were used to test the pre-registered hypotheses that participants in the training condition would score higher than participants in the control condition on both the immediate (time 1) and one week delayed (time 2) S-D and D-N triad judgement and explanation tasks, scenario generation tasks and selection skills tasks. We consider these confirmatory analyses.

Two-sided Bayesian independent samples t -tests were used to assess the impact of the experimental manipulation on performance on the S-N triad judgement tasks at both time 1 and time 2. Two-sided Bayesian paired samples t -tests were used to assess the impact of time (1 vs 2) on performance on the S-D and D-N triad judgement and explanation tasks, scenario generation tasks and selection skills tasks for both groups of participants. Finally, two-sided Bayesian one-sample t -tests were used to compare each group's performance to chance on the S-D, D-N and S-N triad judgement tasks (i.e., 4 out of 8) and selection skills tasks (i.e., 1 out of 4). As these analyses were not pre-registered, we consider them exploratory.

For all Bayesian t -tests, we used a default Cauchy prior with a scale parameter of $r = .707$ (Wagenmakers et al., 2018), and ran robustness analyses to determine the extent to which our conclusions would vary across a range of narrower and wider prior widths. The Bayes Factors (BFs) we report represent the probability of the observed data under the research hypothesis (H1, there is an effect, in the specified direction where applicable) versus the null hypothesis (H0, there is no effect). As such, they quantify the strength of evidence in favour of either H1 or H0. BFs between 3 and 10 are commonly considered to provide moderate evidence in favour of H1, whereas progressively larger BFs provide strong (BF = 10 to 30), very strong (BF = 30 to 100) and extreme (BF > 100) evidence in favour of H1 (see Wagenmakers et al., 2018). Conversely, BFs between .33 and .10, between .10 and .03, between .03 and .01, and < .01 provide moderate, strong, very strong and extreme evidence in favour of H0, respectively. BFs between .33 and 3 are considered non-diagnostic, as they do not provide clear evidence in favour of either H1 or H0. As a measure of effect size, we have calculated and reported δ , which is an estimate of the population standardised difference between two means (or a mean and a pre-determined value, in the case of one-sample t -tests). We can be 95% confident that the true value of δ lies within its 95% Bayesian credible interval (BCI). All BFs, δ s and associated 95% BCIs were calculated with JASP 0.9.2 (Wagenmakers et al., 2018).

Deviations from Pre-Registration

In our pre-registration we actually calculated our required sample size using $\alpha = .05$ rather than the .01 we reported using. Consequently, our analyses were somewhat underpowered for detecting the effects we planned for. Fortunately, and as we suspected on p. 3 of the pre-registration, our planned effect sizes were conservative, and our observed effects were considerably larger than these. We also reported that we had five DVs in our confirmatory analyses (there were actually six), and that scores on each set of triad judgement tasks would range from 0 to 6 (they actually ranged from 0 to 8). These were errors that should have no

impact on the interpretation of the following results. Finally, we increased compensation for paid participants from £15 to £25 to better reflect their actual time commitment. Again, it is difficult to imagine that this could have any impact on the interpretation of our results.

Results

Confirmatory Analyses

The results reported in Table 2 indicate that participants in the training condition outperformed participants in the control condition on all five measures of structural awareness immediately following training (or the control task) and again one week later. In all instances, the strength of the evidence in support of the research hypotheses could be characterised as ‘extreme’ (Wagenmakers et al., 2018), and the effect sizes could be characterised as ‘large’ (Cohen, 1988). Over these 10 tests, δ ranged from 0.71 to 1.60 ($M = 1.09$). To facilitate interpretation, condition means and their 95% BCIs are illustrated in Figure 1.

[INSERT TABLE 2 ABOUT HERE]

[INSERT FIGURE 1 ABOUT HERE]

Also illustrated in Figure 1 are the condition means and associated 95% BCIs for the selection skills DV. Immediately following training (or the control task) and again one week later there was strong evidential support for the hypothesis that our trained participants had better selection skills than our un-trained students (see Table 2). By Cohen’s (1988) conventions, these effects would be described as ‘medium’ sized ($\delta = 0.52$ and 0.57).

Robustness analyses for our pre-registered hypothesis tests are reported in S9 of the online supplement. These analyses indicate that our conclusions are not dependent on the particular prior width we used. Indeed, our conclusions remain unchanged across the full range of sensible prior widths.

Exploratory Analyses

On the S-N triad judgement tasks at time 1, the control participants ($M = 5.92$, $SD = 2.13$) showed a clear bias in favour of the S scenarios, $BF_{10} = 4.06E+05$, $\delta = 0.87$, 95% BCI [0.55, 1.20]. However, there was insufficient evidence to determine whether the trained participants ($M = 4.62$, $SD = 2.18$) were biased in favour of either type of scenario, $BF_{10} = 0.97$, $\delta = 0.27$, 95% BCI [-0.01, 0.54]. This is reflected in the greater preference for the S scenarios demonstrated by the control participants relative to the trained participants at time 1, $BF_{10} = 11.78$, $\delta = 0.55$, 95% BCI [0.17, 0.94]. By time 2 both the control ($M = 6.08$, $SD = 2.16$) and trained ($M = 5.17$, $SD = 1.87$) groups showed a clear bias in favour of the S scenarios (control group $BF_{10} = 1.29E+06$, $\delta = 0.93$, 95% BCI [0.60, 1.25]; training group $BF_{10} = 2.82E+02$, $\delta = 0.59$, 95% BCI [0.27, 0.90]), and there was insufficient evidence to claim any difference in the relative strength of these time 2 preferences, $BF_{10} = 1.90$, $\delta = 0.40$, 95% BCI [0.02, 0.80].

As can be seen in the left-hand section of Table 3, 10 of the 12 tests for differences between time 1 and time 2 performance on the S-D and D-N triad judgement and explanation tasks, scenario generation tasks and selection skills tasks were either non-diagnostic or supportive of the H_0 . In most of these instances, performance did not change over time. On the two instances where a change was observed, it was in a negative direction. The trained group's performance on the scenario generation tasks decreased moderately between times 1 and 2; as did the control group's selection skills.

[INSERT TABLE 3 ABOUT HERE]

Finally, the right-hand section of Table 3 illustrates the extent to which each group's time 1 and time 2 performance on the S-D and D-N triad judgement tasks and selection skills tasks differed from 'chance' levels. When given the choice between S or D comparison scenarios at both time 1 and time 2, the control participants demonstrated a clear bias in favour of S, whereas the trained participants demonstrated a somewhat weaker bias in favour of D.

However, when given the choice between D or N, both groups demonstrated a clear bias in favour of D at both time points. The strength of the trained participants' bias was over twice that of the control participants'. Finally, at both time points the trained participants demonstrated selection skills that were clearly above 'chance' levels, although less so at time 2 than time 1. The control participants performed above 'chance' levels at time 1. However, at time 2 our data were non-diagnostic, and unable to provide support for either H1 (performance differs from chance) or H0 (performance is consistent with chance).

Robustness analyses for our exploratory analyses are reported in S10 of the online supplement. These robustness analyses indicate that the conclusions we made in support of H1 were not dependent on the prior width we used. The conclusions we made in support of H0 would have strengthened somewhat if very wide priors were used. Conversely, the use of very narrow priors would have generally resulted in the inability to draw clear conclusions. Finally, our non-diagnostic BFs would have remained so across all reasonable prior widths.

Discussion

In this pre-registered experiment, we designed and evaluated a novel method of training structural awareness. Structural awareness reflects the ability to see past the surface or topic characteristics of a research design and focus instead on its deep structural characteristics and the associations between them (Quilici & Mayer, 2002). Most psychology students are not naturally structurally aware. During training, students were engaged with scaffolded activities in which they used StatHand to select appropriate statistics for four simple research designs. These activities highlighted the deep structural characteristics of each design, and repeatedly prompted students to reflect on how and why these structural characteristics were related to their choice of statistic. Immediately following training, and again one week later, we found that the trained students out-performed a control group of un-trained students on five measures of structural awareness. These 10 effects were 'large' (Cohen, 1988), ranging from $\delta = 0.71$ to

1.60. To help put these effects in perspective, Hattie (2015) recently estimated that the average effect size for interventions to improve achievement in higher education is around $d = 0.40$. Less than one-fifth of such effects are above $d = 0.71$.

Both the direction and size of our effects are consistent with the effects of previous experimental efforts to train structural awareness. For example, in Experiment 1 of Quilici and Mayer (1996), students who studied research scenarios grouped by structure were more likely to sort subsequent scenarios on the basis of structure and less likely to sort on the basis of surface characteristics than control students. These effects were large ($d \approx 1.30$ to 2.23)¹. In their second experiment, Quilici and Mayer (1996) found that students who studied structure-emphasising scenario sets were more likely to sort subsequent scenarios on the basis of structure, and less likely to sort on the basis of surface characteristics than students who studied surface-emphasising scenario sets, or no scenario sets ($d \approx 1.10$ to 1.96). In Experiment 3, Quilici and Mayer (1996) found that students who studied structure-emphasising worked examples of statistical tests were more likely to apply appropriate statistical tests to subsequently presented research scenarios than students who had studied surface-emphasising examples ($d \approx 0.38$). In later research, Quilici and Mayer (2002) reported that students taught to sort research scenarios on the basis of structural characteristics were more likely to subsequently sort on the basis of structure than control students ($d \approx 0.55$). A similar effect was observed for a subsequent scenario generation task ($d \approx 0.31$ to 0.59). Most recently, Yan and Lavigne (2014) reported that students presented with schema-emphasising worked examples were better able to categorise research scenarios based on structural features ($d \approx 1.70$ to 1.80) than students in a control condition. The trained students also correctly identified a greater number of structural features for each category of scenarios ($d \approx 1.40$ to 1.70). Taken in

¹ The effect sizes reported in this paragraph are estimates based on information available in the cited articles.

combination with ours, these results clearly suggest that structural awareness can be trained with relatively little effort, and that the effects of such training can be substantial, at least in the immediate term.

Although Quilici and Mayer (1996, 2002) and Yan and Lavigne (2014) did not investigate the longer-term effects of training, we did. It is pleasing to note that, on four of five measures, our trained students were no less structurally aware one-week following training than they were in its immediate aftermath. The extent to which they can ‘hold on’ to this skill for longer periods of time is currently unknown, and worthy of further research. On the fifth measure, scenario generation, there was a modest decrease in the trained students’ performance between time 1 and time 2 ($\delta = -0.48$). The reasons for this are unclear and, considering the absence of time related effects on the other four structural awareness measures, it should be considered with a degree of caution until replicated. It should also be noted that, even after experiencing a 0.48 standard deviation decrease over time, the trained group still substantially out-performed the control group on the time 2 scenario generation task ($\delta = 0.81$).

As our triad judgement tasks involved a forced choice, we were able to assess the extent to which participants demonstrated a preference (relative to ‘chance’) for D scenarios in the presence of S and N scenarios, as well as the extent to which they showed a preference for S scenarios in the presence of N scenarios. D scenarios shared the deep structural (but not surface) characteristics of the target scenarios. S scenarios shared the surface (but not deep structural characteristics) of the target scenarios. N scenarios did not share the targets’ deep structural or surface characteristics. Selecting D over S or N reflects structural awareness. Selecting S over N is more ambiguous and cannot be assumed to reflect a lack of structural awareness. Indeed, it is arguably the ‘best’ choice regardless of whether or not one typically attends to structural characteristics of research designs. Our trained students demonstrated a clear preference for D in the presence of both S and N. This preference was very strong on the

D-N trials, but more modest on the S-D trials. Our un-trained students also showed a preference for D in the presence of N, although this was less than half the strength of the preference shown by the trained students. However, on the S-D trials, they clearly preferred S. Finally, on the S-N trials, the control participants showed a clear preference for S at both time points. The trained students showed no clear preference at time 1, and a preference for S at time 2. We interpret these findings as broadly consistent with the hypothesis that our training increased students' structural awareness. They are also broadly consistent with the findings of Rabinowitz and Hogan (2008), who administered a similar set of triad judgement tasks to un-trained postgraduate psychology and education students. Like our un-trained students, the students who had completed at least one statistics module in Rabinowitz and Hogan's (2008) sample only showed a preference for D when it was paired with N. (Students with no statistics experience showed no clear preference for either D or N.) When D was paired with S, Rabinowitz and Hogan's (2008) less experienced students (0 to 1 modules), as well as our un-trained students, showed a preference for S. The more experienced students (2 to 8 modules) in Rabinowitz and Hogan's (2008) sample showed no clear preference for either S or D. Combined, these findings indicate that most psychology students naturally display limited structural awareness. The current study demonstrates how some simple training can remediate this, at least in the short term.

Arguably, structural awareness is only useful to the extent that it drives other adaptive behaviours. One such behaviour is statistic selection. Compared to our un-trained students, the students we trained to be structurally aware also demonstrated greater statistic selection skills at both time 1 and time 2. Furthermore, at both time points they were performing at levels well above chance. However, their performance was still far from exemplary. At time 1 they averaged 58.5%. By time 2 this had decreased (though not in an inferential sense) to 47.5%. By comparison, the un-trained first-year students in Ware and Chastain's (1989) statistics

module averaged less than 45% on a set of multiple-choice statistic selection tasks. This prompted Ware and Chastain (1991) to revise their module such that it explicitly trained selection skills. The students in the revised module significantly out-performed those in the original, but still only averaged 56.8%. More recently, Allen and colleagues' (2019) un-trained students had an average statistic selection accuracy level below 25%. Allen and colleagues' (2019) selection items were open-ended and spanned a wide range of analyses, which made them more challenging than the multiple-choice selection items in the current study. However, the students in the Allen et al. (2019) sample had access to one of four decision making aids, whereas the students in the current study performed in exam-like conditions. When considered in combination, these findings indicate that students find statistic selection difficult, but that training can work. Furthermore, they suggest that training targeting the psychological mechanism believed to underlie selection competence (i.e., structural awareness) may be as effective as training targeting the behaviour itself. However, a direct comparison between these different training methods should be conducted before firm conclusions about their relative efficacy are drawn. Furthermore, considering that there are relatively few real-life circumstances in which one would need to select a statistic in exam-like conditions, it would be useful to understand how trained students would perform when given access to an evidence-based decision-making aid like StatHand, or the graphic organiser developed by Ware and Chastain (1991).

We find it instructive to evaluate our research against the eight psychology SoTL (Scholarship of Teaching and Learning) 'gold standards' articulated by Wilson-Doenges and colleagues (e.g., Wilson-Doenges, Troisi, & Bartsch, 2016). The current study clearly meets five of these standards, and partially meets a sixth. First, it was theoretically and empirically based. Specifically, it was informed by research demonstrating a lack of statistic selection skills and structural awareness amongst psychology students (Allen et al., 2019; Rabinowitz &

Hogan, 2008), as well as research demonstrating the efficacy of wise interventions (Walton, 2014) and various methods of statistic selection and structural awareness training (Quilici & Mayer, 2002; Ware & Chastain, 1989). Second, it was an experimental study in which individual students were randomised to the two levels of our IV. Third, our sample size was informed by (conservative) considerations about likely and minimally interesting effect sizes. Although not ‘large’, it was sufficient for our purposes and respectable by the standards of our discipline (Bakker, van Dijk, & Wicherts, 2012). Fourth, our primary statistical analyses were pre-registered and Bayesian. The former should offer readers some assurances about the likely replicability of our findings. The latter enabled us to quantify the extent to which our data support claims about both the presence and absence of effects. The ability to quantify the strength of evidence in favour a null hypothesis is a particular strength of Bayesian statistics, and allowed us to conclude that, by most measures, our trained students’ structural awareness did not decrease between time 1 and time 2. Fifth, when designing and running this study we closely followed best-practice ethical recommendations for SoTL research in psychology (e.g., Swenson & McCarthy, 2012). The sixth standard, which we partially met, is the use of longitudinal designs. We collected data immediately following training and again one week later. This allowed us to conclude that our trained students ‘held on’ to their newly developed structural awareness for a modest period of time. Whether or not the benefits of training are longer lasting is a question we cannot currently answer. The final two gold standards advocated by Wilson-Doenges et al. (2016) are the use of diverse samples and mixed methods. Our participants were all psychology undergraduates at a highly selective British university. Although we have no obvious reasons to believe that students elsewhere, and in other disciplines, would not respond similarly to our training, these matters require further investigation. Finally, our methods were exclusively quantitative. Future researchers should

seek to triangulate the benefits (or otherwise) of structural awareness training via multiple methods.

In conclusion, this paper describes a pre-registered experiment in which we designed and evaluated a novel method of training structural awareness. Our data indicated that it clearly worked. However, as an actual method of instruction, it is not very efficient. We encourage educators to consider how they can optimise and adapt our methods and materials for deployment in a classroom context or online activity. Indeed, these would make sensible directions for future research. Of particular interest may be the extent to which the ‘dose’ of training can be reduced without significantly compromising its efficacy. Educators may also want to substitute the four statistical analyses we trained with for any of the other analyses covered in StatHand. There are 41 research scenarios that can be used for this purpose in Allen et al. (in press). Finally, we encourage educators to evaluate their efforts and let us (and the broader psychology SoTL community) know how things went.

References

- Allen, P. J., Baughman, F. D., Roberts, L. D., van Rooy, D., Rock, A. J., & Loxton, N. J. (2017). *StatHand: An interactive decision tree mobile application to guide students' statistical decision making*. Canberra, Australia: Australian Government Department of Education and Training.
- Allen, P., Bennett, K., & Heritage, B. (2019). *SPSS Statistics: A practical guide* (4th ed.). Melbourne, Australia: Cengage.
- Allen, P. J., Dorozenko, K. P. & Roberts, L. D. (2016). Difficult decisions: A qualitative exploration of the statistical decision making process from the perspectives of psychology students and academics. *Frontiers in Psychology*, 7, Article 188. doi:10.3389/fpsyg.2016.00188
- Allen, P. J., Fielding, J. L., Kay, R. H. S., & East, E. C. (in press). Using StatHand to improve students' statistic selection skills. In A. Beyer & J. Peters (Eds.), *For the love of (teaching) undergraduate statistics*. Washington, DC: Society for the Teaching of Psychology. Retrieved from <http://teachpsych.org/ebooks/>
- Allen, P. J., Finlay, J., Roberts, L. D., & Baughman, F. D. (2019). An experimental evaluation of StatHand: A free application to guide students' statistical decision making. *Scholarship of Teaching and Learning in Psychology*, 5, 23-36. doi:10.1037/stl0000132
- Allen, P. J., Roberts, L. D., Baughman, F. D., Loxton, N. J., van Rooy, D., Rock, A. J., & Finlay, J. (2016). Introducing StatHand: A cross-platform mobile application to support students' statistical decision making. *Frontiers in Psychology*, 7, Article 288. doi:10.3389/fpsyg.2016.00288

- Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, 7, 543-554. doi:10.1177/1745691612459060
- Carlson, M., Protsman, L., & Tomaka, J. (2005). Graphic organizers can facilitate selection of statistical tests: Part 1 - Analysis of group differences. *Journal of Physical Therapy Education*, 19, 57-65. doi:10.1097/00001416-200507000-00008
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Faul, F., Erdfelder, E., Lang, A. -G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191. doi:10.3758/BF03193146
- Gardner, P. L., & Hudson, I. (1999). University students' ability to apply statistical procedures. *Journal of Statistics Education*, 7(1). Retrieved from <https://ww2.amstat.org/publications/jse/secure/v7n1/gardner.cfm>
- Hattie, J. (2015). The applicability of visible learning to higher education. *Scholarship of Teaching and Learning in Psychology*, 1, 79-91. doi:10.1037/stl0000021
- Koch, C., & Gobell, J. (1999). A hypertext-based tutorial with links to the web for teaching statistics and research methods. *Behavior Research Methods, Instruments, & Computers*, 31, 7-13. doi:10.3758/bf03207686
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, 4, Article 863. doi:10.3389/fpsyg.2013.00863
- Nolan, S. A., & Heinzen, T. E. (2017). *Statistics for the behavioral sciences* (4th ed.). New York, NY: Worth.

- Protsman, L., & Carlson, M. (2008). Graphic organizers can facilitate selection of statistical tests: Part 2 - Correlation and regression analysis. *Journal of Physical Therapy Education*, 22, 36-41. doi:10.1097/00001416-200807000-00006
- Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology*, 88, 144–161. doi:10.1037/0022-0663.88.1.144
- Quilici, J. L., & Mayer, R. E. (2002). Teaching students to recognize structural similarities between statistics word problems. *Applied Cognitive Psychology*, 16, 325-342. doi:10.1002/acp.796
- Rabinowitz, M. & Hogan, T. M. (2008). Experience and problem representation in statistics. *American Journal of Psychology*, 121, 395–407. doi:10.2307/20445474
- Schönbrodt, F. D., & Wagenmakers, E. -J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, 25, 128-142. doi:10.3758/s13423-017-1230-y
- Society for the Teaching of Psychology Statistical Literacy Task Force. (2014). Statistical literacy in the undergraduate psychology curriculum. Retrieved from https://teachpsych.org/Resources/Documents/otrp/resources/statistics/STP_Statistical%20Literacy_Psychology%20Major%20Learning%20Goals_4-2014.pdf
- Swenson, E. V., & McCarthy, M. A. (2012). Ethically conducting the scholarship of teaching and learning research. In R. E. Landrum & M. A. McCarthy (Eds.), *Teaching ethically: Challenges and opportunities* (pp. 21-29). Washington, DC: American Psychological Association.
- van Doorn, J., van den Bergh, D., Bohm, U., Dablander, F., Derks, K., Draws, T.,... Wagenmakers, E. -J. (2019). The JASP guidelines for conducting and reporting a Bayesian analysis. *PsyArXiv*. doi:10.31234/osf.io/yqxfr

- Wagenmakers, E. -J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J.,... Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, 25, 58–76. doi:10.3758/s13423-017-1323-7
- Walton, G. M. (2014). The new science of wise psychological interventions. *Current Directions in Psychological Science*, 23, 73-82. doi:10.1177/0963721413512856
- Ware, M. E., & Chastain, J. D. (1989). Computer-assisted statistical-analysis: A teaching innovation? *Teaching of Psychology*, 16, 222-227. doi:10.1207/s15328023top1604_16
- Ware, M. E., & Chastain, J. D. (1991). Developing selection skills in introductory statistics. *Teaching of Psychology*, 18, 219-222. doi:10.1207/s15328023top1804_4
- Wilson-Doenges, G., Troisi, J. D., & Bartsch, R. A. (2016). Exemplars of the gold standard in SoTL for psychology. *Scholarship of Teaching and Learning in Psychology*, 2, 1-12. doi:10.1037/stl0000050
- Yan, J., & Lavigne, N. C. (2014). Promoting college students' problem understanding using schema-emphasizing worked examples. *Journal of Experimental Education*, 82, 74-102. doi:10.1080/00220973.2012.745466

Table 1

Demographic Characteristics of the Full Sample, Split by Experimental Condition

	Control	Training
<i>N</i>	52	50
Age <i>M</i> (<i>SD</i>)	19.78 (1.16)	19.92 (1.59)
% Female	84.62	80.00
% White or White British	80.77	86.00
% First/Second Year of Study	38.46/50.00	38.00/52.00

Table 2

Descriptive Statistics and Bayesian Summary Information for the Pre-Registered Hypothesis Tests

	Time 1 (Immediately Following Training/Control Task)				Time 2 (One Week Later)			
	Descriptives by Condition		Difference Between Conditions		Descriptives by Condition		Difference Between Conditions	
	<i>N</i>	<i>M</i> (<i>SD</i>)	BF	δ [95% BCI]	<i>N</i>	<i>M</i> (<i>SD</i>)	BF	δ [95% BCI]
S-D Triad Judgement								
Control	52	2.02 (2.25)			51	2.33 (2.57)		
Training	50	5.12 (2.60)	4.53E+06	1.22 [0.78, 1.65]	48	5.10 (2.67)	3.20E+04	0.99 [0.59, 1.43]
D-N Triad Judgement								
Control	52	4.92 (1.66)			51	5.14 (1.60)		
Training	50	6.30 (1.52)	1.30E+03	0.81 [0.39, 1.22]	48	6.40 (1.66)	2.22E+02	0.71 [0.31, 1.13]
S-D Explanation								
Control	52	1.52 (2.42)			51	1.69 (2.58)		
Training	50	7.46 (5.22)	3.66E+08	1.42 [0.98, 1.87]	48	7.67 (5.42)	5.01E+07	1.35 [0.91, 1.80]
D-N Explanation								
Control	52	2.89 (2.52)			51	3.45 (3.70)		
Training	50	8.94 (4.56)	2.88E+10	1.60 [1.17, 2.05]	48	8.96 (5.00)	1.84E+06	1.20 [0.76, 1.64]
Scenario Generation								
Control	52	8.58 (2.64)			51	8.12 (2.19)		
Training	50	10.54 (2.22)	4.54E+02	0.75 [0.36, 1.16]	48	9.88 (1.84)	1.02E+03	0.81 [0.39, 1.23]
Selection Skills								
Control	52	1.67 (0.96)			51	1.18 (0.89)		
Training	50	2.34 (1.35)	15.37	0.52 [0.14, 0.91]	48	1.90 (1.37)	27.85	0.57 [0.19, 0.97]

Note. BF = Bayes Factor. BCI = Bayesian Credible Interval. One-sided BF_{+0S} are reported throughout. Per van Doorn et al. (2019), all δ s and associated 95% BCIs were estimated using a two-sided default Cauchy prior with a scale parameter of $r = .707$.

Table 3

Bayesian Summary Information for the Exploratory Analyses

	Difference Between Times 1 and 2			Difference from Chance ^a					
	MD ^b	BF	δ [95% BCI]	Time 1			Time 2		
				MD ^c	BF	δ [95% BCI]	MD ^c	BF	δ [95% BCI]
S-D Triad Judgement									
Control	0.28	0.28	0.15 [-0.12, 0.41]	-1.98	2.28E+05	-0.85 [-1.17, -0.53]	-1.67	7.63E+02	-0.62 [-0.93, -0.31]
Training	-0.02	0.16	-0.01 [-0.28, 0.27]	1.12	8.86	0.41 [0.13, 0.69]	1.10	5.79	0.39 [0.12, 0.68]
D-N Triad Judgement									
Control	0.20	0.24	0.13 [-0.13, 0.40]	0.92	1.26E+02	0.53 [0.24, 0.83]	1.14	3.20E+03	0.68 [0.39, 0.99]
Training	0.08	0.17	0.06 [-0.22, 0.33]	2.30	3.94E+11	1.49 [1.12, 2.15]	2.40	2.50E+10	1.40 [1.03, 1.78]
S-D Explanation									
Control	0.14	0.17	0.06 [-0.21, 0.34]						
Training	0.08	0.16	0.02 [-0.25, 0.29]						
D-N Explanation									
Control	0.51	0.30	0.16 [-0.11, 0.43]						
Training	-0.06	0.16	-0.02 [-0.29, 0.25]						
Scenario Generation									
Control	-0.45	0.35	-0.17 [-0.44, 0.10]						
Training	-0.81	30.69	-0.48 [-0.78, -0.18]						
Selection Skills									
Control	-0.51	75.20	-0.51 [-0.81, -0.23]	0.67	2.88E+03	0.67 [0.36, 0.97]	0.18	0.39	0.17 [-0.08, 0.46]
Training	-0.44	0.98	-0.27 [-0.55, 0.01]	1.34	2.00E+06	0.96 [0.64, 1.31]	0.90	5.22E+02	0.62 [0.32, 0.94]

Note. MD = Mean Difference. BF = Bayes Factor. BCI = Bayesian Credible Interval. Two-sided BF_{10s} are reported throughout.

^a Chance was defined as 4/8 on the triad judgement tasks and 1/4 on the selection skills tasks. ^b A positive mean difference indicates that performance on the relevant variable increased from time 1 to time 2. ^c A positive mean difference indicates performance at a level above chance.

Figure 1. Means and 95% BCIs for each condition at each time point on the variables in the pre-registered hypothesis tests. To aid interpretation, the Y-axis on each graph spans the full possible range of values for each DV.