



Allen, P. J., Fielding, J. L., Westermann, A., & Lafratta, A. (2021). Training Structural Awareness with StatHand: A 1 Year Follow-Up. *Teaching of Psychology*. <https://doi.org/10.1177/0098628320985080>

Publisher's PDF, also known as Version of record

License (if available):
CC BY

Link to published version (if available):
[10.1177/0098628320985080](https://doi.org/10.1177/0098628320985080)

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

This is the final published version of the article (version of record). It first appeared online via Sage Publications at <https://doi.org/10.1177/0098628320985080> . 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/>

Training Structural Awareness with StatHand: A 1 Year Follow-Up

Teaching of Psychology
1-8

© The Author(s) 2021



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0098628320985080

journals.sagepub.com/home/top



Peter J. Allen¹ , Jessica L. Fielding¹, Annabel H. Westermann¹,
and Amelia M. Lafratta¹

Abstract

Background: Allen, Fielding, East, et al. demonstrated experimentally that structural awareness, or the ability to disregard a research problem's topic and instead focus on its structural features, can be trained using StatHand (<https://stathand.net>). Most training benefits persisted for 1 week. **Objective:** The objective was to assess the longer-term effects of training. **Method:** One year after training (or control activities), 54 participants were re-administered 5 measures of structural awareness and 1 statistic selection measure. **Results:** Trained participants continued to reliably out-perform control participants on 4 measures of structural awareness, though no longer on the 5th. Over the year, decrements in trained participants' performance on the 5 structural awareness measures were mostly small. However, 1 year after training, the trained participants' statistic selection advantage had largely disappeared. **Conclusion:** Brief structural awareness training can have long-term benefits, though selecting an appropriate statistical test for common research scenarios without assistance remains a difficult task. **Teaching Implications:** Structural awareness can be trained. However, even structurally aware students cannot reliably select appropriate statistics without assistance. Training plus easy access to a decision-making aid should maximize statistic selection accuracy. Our evidence-based training methods and materials, including StatHand, can be freely used and adapted for these purposes.

Keywords

instructional technology, statistics, research methods

The ability to select appropriate statistical analyses for common research questions and designs is one of the undergraduate learning goals specified by the Society for the Teaching of Psychology Statistical Literacy Task Force (2014). Psychology students find this process difficult (Gardner & Hudson, 1999; Ware & Chastain, 1989). To make it easier, a range of decision aids have been developed. Chief amongst these are decision trees, which routinely feature in introductory statistics textbooks (e.g., Nolan & Heinzen, 2017). Paper decision trees are both popular and effective (Protsman & Carlson, 2008). However, they are also constrained by their format, which necessitates brevity and the separation of the tree from information that would aid its navigation (e.g., definitions, examples). Digital learning technologies can overcome these constraints. It was in this context that StatHand was developed (Allen, Roberts, et al., 2016). StatHand is a free iOS/iPadOS and web (see <https://stathand.net>) application that asks users a series of annotated questions about their research design and, based on the answers they provide, suggests an appropriate statistical analysis. In an experimental evaluation (Allen et al., 2019), psychology students randomized to StatHand demonstrated higher statistic selection accuracy than students randomized to three other common decision-making aids (a familiar paper decision

tree, a familiar textbook, and the decision tree and textbook combined; $\delta = 0.50-0.64$).

Despite outperforming the control groups by at least half a standard deviation, the StatHand group's performance was still underwhelming (Allen et al., 2019). On average, they identified appropriate analyses for just 35% of the research scenarios they were shown. On most university marking scales, this would be a clear "fail." This suggests that simply providing an aid like StatHand is not enough to promote accurate statistic selection. Rather, to use it effectively, students require some training.

The training described in Allen, Fielding, East, et al. (2020) targeted a mechanism proposed to underpin statistic selection competence: structural awareness. Structural awareness reflects the ability to disregard the surface/topic level features of a research design and focus instead on its deep structural characteristics (Quilici & Mayer, 2002). These structural

¹ School of Psychological Science, University of Bristol, United Kingdom

Corresponding Author:

Peter J. Allen, School of Psychological Science, University of Bristol, 12a Priory Rd, Bristol, BS8 1TU, United Kingdom.

Email: p.allen@bristol.ac.uk

characteristics include the number and nature of the design's independent and dependent variables, as well as the associations between them. Most psychology students are not naturally structurally aware (Allen, Dorozenko, & Roberts, 2016; Rabinowitz & Hogan, 2008).

In Allen, Fielding, East, et al. (2020), 102 psychology students were randomized to a training or control group. The training group completed scaffolded activities in which they used StatHand to select appropriate statistics for four simple research designs (two-group independent and paired samples designs with ratio or dichotomous dependent variables; DVs). These activities, which took around 30 minutes to complete, highlighted the deep structural characteristics of each design, and encouraged students to reflect on how and why these characteristics were related to the statistics they selected. The control group used an origami iPad application for a similar amount of time. Following this, all participants completed five measures of structural awareness, and a statistic selection task. The first two structural awareness measures were triad judgment tasks. In these, participants chose which of two comparison research scenarios "goes best" with a target research scenario. The target and comparison scenarios shared surface (S) characteristics, deep (D) structural characteristics, or neither (N). Consistently choosing D comparison scenarios over S or N indicates structural awareness. The two triad judgment tasks were combined with two explanation tasks in which participants described why each selected comparison scenario "goes best" with the relevant target. People who consistently identify relevant structural characteristics (e.g., the nature of the design) are structurally aware. The fifth measure of structural awareness was a scenario generation task in which participants wrote new scenarios that were "similar" to the target scenarios from the triad judgment and explanation tasks. Finally, for the selection task, participants chose an appropriate statistical analysis for each target scenario. One week later, 99 participants returned for a second wave of testing. The Time 2 scenarios were structurally equivalent to those used at Time 1, although they had different surface characteristics.

We found that the trained group outperformed the control group on all five measures of structural awareness immediately following training, and again one week later ($\delta = 0.71-1.60$). At both time points, the trained group also demonstrated stronger statistic selection skills ($\delta = 0.52$ and 0.57). Finally, the only measure on which the trained group's performance declined between Times 1 and 2 was the scenario generation task. Consequently, we encouraged educators to adapt our methods and materials for classroom activities and published additional guidance and resources to facilitate this (Allen, Fielding, Kay, & East, 2020).

That the trained participants largely held on to their new abilities for one week is impressive. However, one week will not carry them through their degrees and into their professional lives. Consequently, in this study, we sought to re-visit the same sample 12 months later. We hypothesized that trained participants would continue to score higher than control participants on the five measures of structural awareness and the

Table 1. Demographic Characteristics of the Sample, Split by Condition.

	Control ($n = 32$)	Training ($n = 22$)
Age M (SD)	20.57 (1.10)	20.64 (1.40)
% Female	81.30	81.80
% White or White British	81.30	90.90
% Second/third year of study	40.60/56.30	40.90/54.50

statistic selection task. We also hypothesized that trained participants' structural awareness and statistic selection abilities would not have decreased since they were previously tested.

Method

Design

This was a pre-registered (see <https://osf.io/tv2fw>) mixed factorial experiment with one randomized between-subjects independent variable (IV; condition: training or control), one within subjects IV (time: Time 1 immediately following training, Time 2 one week later and Time 3 one year later) and six DVs. Five DVs reflect structural awareness (the S-D and D-N triad judgment 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. During testing participants also completed a series of Surface vs. Neither (S-N) triad judgment trials. As the meaning of performance on these trials is ambiguous, its analysis is confined to section S1 of the online supplement at <https://osf.io/gtxvb/>.

Participants

Based on *a priori* power considerations, we recruited 102 undergraduate psychology subject pool members to the original sample (Allen, Fielding, East, et al., 2020). Ninety-nine completed both parts of the study. We contacted each one year (\pm two weeks) after their initial training/testing and re-recruited 54 into this study. Each was compensated with their choice of subject pool credit or cash (£15/£25 for current/graduated students). Evidence indicating that these participants did not systematically differ from those in the original sample who ignored or declined our most recent invitation to participate is in S2 of the online supplement. The trained and control groups were demographically very similar (see Table 1).

Measures and Procedure

The measures and procedure in this study were identical to those used at Time 2 by Allen, Fielding, East, et al. (2020), with the exception that 11 participants (control $n = 6$, training $n = 5$) were tested online. This was due to the UK COVID-19 "lockdown," and the only deviation from our pre-registration. Testing took, on average, 50 minutes ($SD = 17$ minutes).

The average time span between Times 1 and 3 was 369 days ($SD = 14$ days).

Data Analysis

We analyzed our data using Bayesian methods, which are of particular value when researchers need to quantify the strength of evidence in favor of the null hypothesis (H_0). We had hoped to do this for our second hypothesis, which predicted the absence of change over time. A parallel set of frequentist analyses are in S3 and S4 of the online supplement. Both sets of analyses suggest the same conclusions. Our raw and processed data are available at <https://osf.io/p7b4m/> and our data dictionary is in S5 of the online supplement. Confirmatory analyses are defined as those for which a pre-registered hypothesis and analysis plan were specified. All other analyses are considered exploratory.

Confirmatory analyses. It was hypothesized that trained participants would score higher than control participants on each of the six DVs at Time 3. These hypotheses were tested with one-sided Bayesian independent samples t -tests. It was also hypothesized that the trained students' structural awareness and statistic selection skills would not have decreased since they were previously tested. These hypotheses were tested using Bayesian one-way repeated measures ANOVAs with planned comparisons. The planned comparisons were two-sided Bayesian paired samples t -tests.

Exploratory analyses. We used one-sided Bayesian independent samples t -tests to compare the training and control groups on each DV at Times 1 and 2, and one-way Bayesian repeated measures ANOVAs with planned comparisons to test for changes in the control group's performance on each DV since they were previously tested. Planned comparisons were also used to compare Time 1 and 2 performance for both conditions. Finally, two-sided one-sample Bayesian t -tests were used to compare each group's performance to "chance" on the S-D and D-N triad judgment and selection tasks.

The Bayesian analyses were implemented in JASP 0.13 using default prior widths ($r = .500$ and $.707$ for the ANOVAs and t -tests respectively; Wagenmakers et al., 2018), and robustness analyses (for the t -tests only) were used to determine the extent to which our conclusions would vary across a range of alternative prior widths. The Bayes Factors (BFs) we have calculated represent the probability of the observed data under the research hypothesis (H_1 , there is an effect, in the specified direction where applicable) relative to the null hypothesis (H_0 , there is no effect). As such, they quantify the strength of evidence in favor of either H_1 or H_0 . Common heuristics suggest that BFs between 3 and 10 provide moderate evidence for H_1 , whereas progressively larger BFs provide strong ($BF = 10-30$), very strong ($BF = 30-100$) and extreme ($BF > 100$) evidence for H_1 . In contrast, BFs between $.33$ and $.10$ provide moderate evidence for H_0 , whilst BFs between $.10$ and $.03$, $.03$ and $.01$, and $< .01$ provide strong, very strong and extreme evidence for

Table 2. Descriptive Statistics and Bayesian Summary Information for the Comparisons Between Conditions at Each Testing Time.

	Descriptives by Condition		Difference Between Conditions	
	Control ($n = 32$) M (SD)	Training ($n = 22$) M (SD)	BF_{+0}	δ [95% BCI]
S-D Triad Judgment				
Time 1	2.25 (2.27)	4.82 (2.44)	> 100	0.99 [0.41, 1.58]
Time 2	2.38 (2.64)	4.77 (2.56)	40.98	0.81 [0.25, 1.39]
Time 3	2.53 (2.77)	4.96 (2.57)	34.50	0.79 [0.24, 1.37]
D-N Triad Judgment				
Time 1	5.06 (1.44)	6.41 (1.14)	99.37	0.91 [0.34, 1.49]
Time 2	5.06 (1.39)	6.55 (1.44)	> 100	0.94 [0.37, 1.53]
Time 3	5.28 (1.51)	6.46 (1.44)	14.19	0.69 [0.15, 1.25]
S-D Explanation				
Time 1	1.84 (2.83)	7.77 (5.12)	> 100	1.41 [0.78, 2.04]
Time 2	1.56 (2.46)	7.09 (4.61)	> 100	1.48 [0.85, 2.12]
Time 3	2.53 (3.71)	6.59 (5.19)	44.77	0.82 [0.26, 1.40]
D-N Explanation				
Time 1	3.25 (2.59)	9.50 (3.53)	> 100	1.98 [1.30, 2.67]
Time 2	3.34 (3.40)	8.68 (4.27)	> 100	1.31 [0.69, 1.93]
Time 3	3.78 (3.09)	8.09 (4.73)	> 100	1.01 [0.43, 1.61]
Scenario Generation				
Time 1	8.53 (2.51)	10.96 (1.46)	> 100	1.02 [0.43, 1.61]
Time 2	7.81 (2.26)	9.86 (1.89)	63.14	0.86 [0.30, 1.44]
Time 3	8.47 (2.14)	9.59 (1.76)	2.84	0.47 [0.04, 1.02]
Selection Skills				
Time 1	1.63 (0.98)	2.55 (1.22)	22.59	0.74 [0.20, 1.31]
Time 2	1.13 (0.91)	1.77 (1.45)	2.82	0.47 [-0.04, 1.01]
Time 3	1.38 (0.87)	1.68 (1.17)	0.77	0.25 [-0.24, 0.77]

Note. BF_{+0} = One-sided Bayes Factor. BCI = Bayesian Credible Interval. Per van Doorn et al. (2020), all δ s and associated 95% BCIs were estimated using a two-sided default Cauchy prior with a scale parameter of $r = .707$.

H_0 , respectively (Wagenmakers et al., 2018). BFs between 3 and $.33$ are considered non-diagnostic, in the sense that they provide merely anecdotal evidence for either H_1 ($0-3$) or H_0 ($0-.33$). We have used δ , a population estimate of the standardized difference between two means, as a measure of effect size for all pairwise comparisons. Finally, we can be 95% confident that the true value of δ lies within its 95% Bayesian Credible Interval (BCI; Wagenmakers et al., 2018).

Results

Confirmatory Analyses

The Time 3 rows in Table 2 (and rightmost third of each graph in Figure 1) indicate that the trained participants continued to reliably out-perform the control participants on four measures of structural awareness (the S-D and D-N triad judgment and explanation tasks; median $\delta = 0.80$), though no longer on the fifth (the scenario generation task; $\delta = 0.47$). Furthermore, by the third testing session, the trained group no longer reliably outperformed the control group on the selection skills task ($\delta = 0.25$). To aid interpretation, the δ s and associated 95% BCIs for these comparisons are illustrated by the darkest bars in Figure 2.

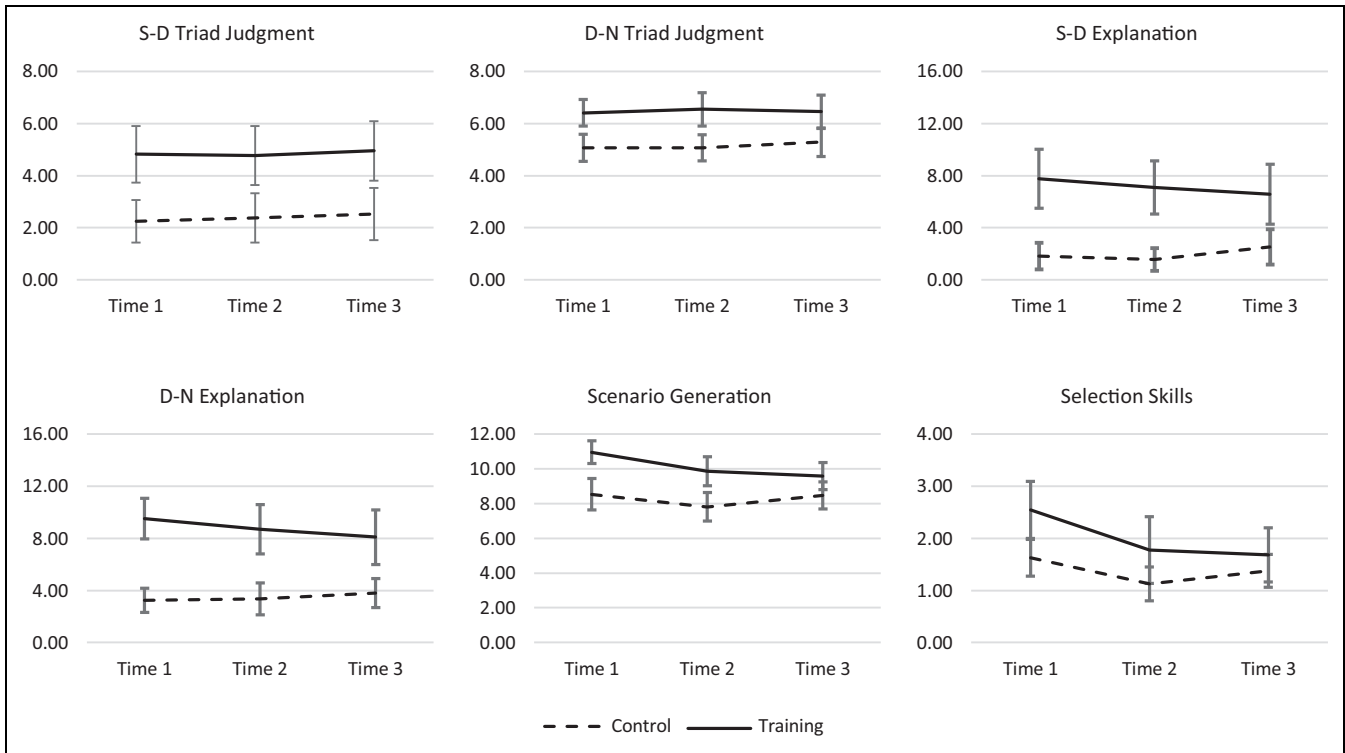


Figure 1. Means and 95% Bayesian credible intervals for each condition at each testing time. Note. To aid interpretation, the Y-axis on each graph spans the full possible range of values for each outcome variable.

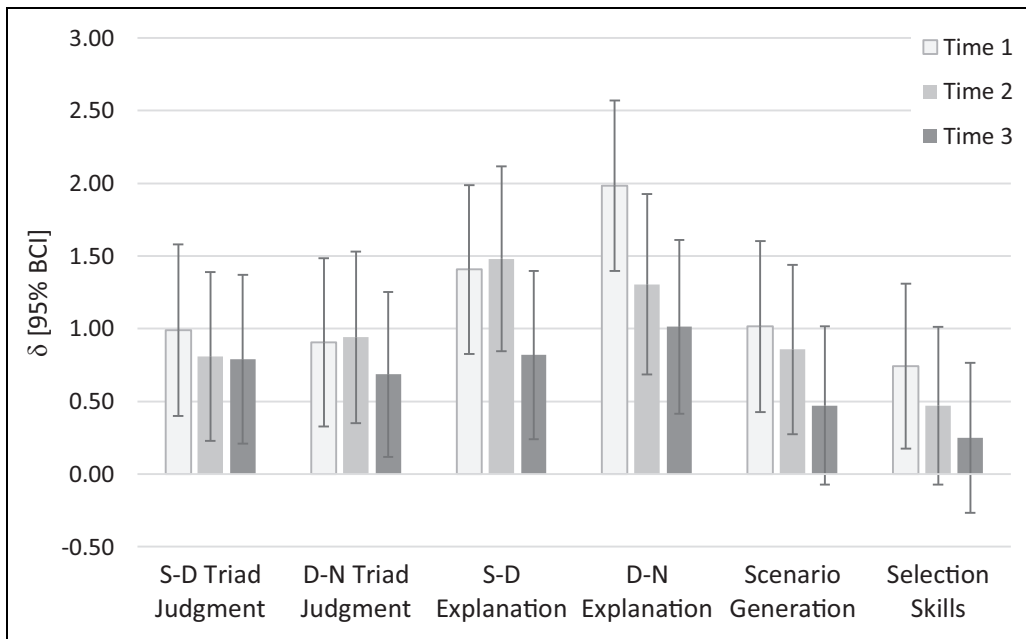


Figure 2. Standardized differences (δ) between conditions at each testing time. Note. BCI = Bayesian Credible Interval. Per van Doorn et al. (2020), all δ s and BCIs were estimated using a two-sided default Cauchy prior with a scale parameter of $r = .707$. A positive δ indicates that the trained group outperformed the control group on the relevant outcome variable.

The right-hand side of Table 3 (and solid lines in Figure 1) indicates that decrements in the trained group’s performance on the five structural awareness measures from Times 1 to 3

(median $\delta = -0.24$) and 2 to 3 (median $\delta = -0.10$) were mostly small. Of the 10 BFs for these comparisons, seven indicated moderate and two suggested anecdotal evidence in favor of H_0 .

Table 3. Bayesian Summary Information for the Differences Between Testing Times, by Condition.

		Control (n = 32)			Training (n = 22)		
		MD	BF ₁₀	δ [95% BCI]	MD	BF ₁₀	δ [95% BCI]
S-D Triad Judgment		ANOVA BF ₁₀ = 0.13			ANOVA BF ₁₀ = 0.14		
Time 1	Time 2	0.13	0.22	0.09 [-0.24, 0.42]	-0.05	0.22	-0.02 [-0.41, 0.37]
	Time 3	0.28	0.27	0.14 [-0.19, 0.48]	0.14	0.23	0.06 [-0.33, 0.45]
Time 2	Time 3	0.16	0.21	0.07 [-0.26, 0.40]	0.18	0.25	0.10 [-0.30, 0.49]
D-N Triad Judgment		ANOVA BF ₁₀ = 0.13			ANOVA BF ₁₀ = 0.13		
Time 1	Time 2	0.00	0.19	0.00 [-0.33, 0.33]	0.14	0.24	0.08 [-0.31, 0.48]
	Time 3	0.22	0.24	0.11 [-0.22, 0.44]	0.05	0.23	0.03 [-0.36, 0.42]
Time 2	Time 3	0.22	0.25	0.12 [-0.21, 0.46]	-0.09	0.23	-0.05 [-0.44, 0.34]
S-D Explanation		ANOVA BF ₁₀ = 0.70			ANOVA BF ₁₀ = 0.25		
Time 1	Time 2	-0.28	0.50	-0.24 [-0.58, 0.10]	-0.68	0.34	-0.18 [-0.58, 0.21]
	Time 3	0.69	0.39	0.20 [-0.13, 0.54]	-1.18	0.45	-0.24 [-0.65, 0.16]
Time 2	Time 3	0.97	1.03	0.32 [-0.02, 0.67]	-0.50	0.25	-0.10 [-0.49, 0.29]
D-N Explanation		ANOVA BF ₁₀ = 0.16			ANOVA BF ₁₀ = 0.31		
Time 1	Time 2	0.09	0.19	0.03 [-0.30, 0.36]	-0.82	0.37	-0.20 [-0.60, 0.20]
	Time 3	0.53	0.31	0.17 [-0.16, 0.50]	-1.41	0.73	-0.31 [-0.73, 0.09]
Time 2	Time 3	0.44	0.27	0.14 [-0.19, 0.48]	-0.59	0.26	-0.11 [-0.50, 0.28]
Scenario Generation		ANOVA BF ₁₀ = 0.34			ANOVA BF ₁₀ = 27.17		
Time 1	Time 2	-0.72	0.63	-0.26 [-0.61, 0.07]	-1.09	17.24	-0.67 [-1.14, -0.21]
	Time 3	-0.06	0.19	-0.02 [-0.35, 0.31]	-1.36	35.33	-0.74 [-1.23, -0.27]
Time 2	Time 3	0.66	0.69	0.28 [-0.06, 0.62]	-0.27	0.27	-0.12 [-0.52, 0.27]
Selection Skills		ANOVA BF ₁₀ = 1.14			ANOVA BF ₁₀ = 3.47		
Time 1	Time 2	-0.50	4.03	-0.44 [-0.80, -0.09]	-0.77	2.28	-0.46 [-0.90, -0.04]
	Time 3	-0.25	0.38	-0.20 [-0.54, 0.13]	-0.86	7.84	-0.59 [-1.05, -0.15]
Time 2	Time 3	0.25	0.33	0.18 [-0.15, 0.52]	-0.09	0.23	-0.05 [-0.44, 0.34]

Note. MD = Mean Difference. BF₁₀ = Two-sided Bayes Factor. BCI = Bayesian Credible Interval. ANOVA BF₁₀ = BF₁₀ for the Bayesian Repeated Measures ANOVA for the differences between testing times for the relevant condition and outcome variable. A positive MD/δ indicates that performance on the relevant dependent variable increased from the earlier to the later testing time.

The remaining comparison, between scenario generation scores at Time 1 and Time 3, indicated a clear reduction in performance (BF₁₀ = 35.33, δ = -0.74). However, most of this reduction occurred between the first two waves of testing (δ = -0.67), rather than the second and third. There was also a clear reduction in selection skills between Times 1 and 3 (BF₁₀ = 7.84, δ = -0.59), though very little of this occurred since the second testing session (δ = -0.05). Robustness analyses for all pairwise confirmatory analyses are in S6 of the online supplement.

Exploratory Analyses

The remaining rows of Table 2 indicate that there were large, reliable differences between the trained and control participants on all five measures of structural awareness at Times 1 (median δ = 1.02) and 2 (median δ = 0.94). At Time 1 there was also a large difference between the selection skills of the two groups (δ = 0.74). However, by Time 2, this difference had reduced to the extent that it could only suggest anecdotal support for H1 (BF₁₀ = 2.82, δ = 0.47). These effects are illustrated as mean differences in Figure 1 and standardized differences in Figure 2.

The remaining rows in Table 3 (and dashed lines in Figure 1) indicate that the control participants' performance on the structural awareness measures was stable across the three waves of testing (median δ = 0.11). However, their selection skills were

somewhat more variable. Disregarding the trained participants' early drops on the scenario generation and selection skills DVs, their performance across the first two waves of testing was stable (median δ = -0.10).

Table 4 illustrates how each group's performance on the S-D and D-N triad judgment and selection skills tasks differed from "chance" levels. When given the choice between S and D comparison scenarios the control participants consistently selected S. The trained participants showed a weak preference for the D scenarios, although the corresponding BFs were non-diagnostic. When given the choice between D or N, both groups showed a clear preference for D at all three time points. The strength of the trained participants' preference more than doubled the control participants' preference. Finally, both groups achieved performance at levels clearly above "chance" on the Time 1 selection skills task. At Times 2 and 3 the picture was less clear. The effect sizes suggest that the performance of both groups was always above chance levels, however the corresponding BFs indicate that such conclusions should be made tentatively in the absence of more data. Robustness analyses for all pairwise exploratory analyses are in S7 of the online supplement.

Discussion

This study extends findings reported in Allen, Fielding, East, et al. (2020) by demonstrating that many of the benefits of brief

Table 4. Bayesian Summary Information for the Differences from “Chance,” by Testing Time and Condition.

	Control (<i>n</i> = 32)			Training (<i>n</i> = 22)		
	MD	BF ₁₀	δ [95% BCI]	MD	BF ₁₀	δ [95% BCI]
S-D Triad Judgment						
Time 1	-1.75	> 100	-0.72 [-1.12, -0.33]	0.82	0.65	0.30 [-0.11, 0.71]
Time 2	-1.63	22.96	-0.57 [-0.95, -0.20]	0.77	0.53	0.27 [-0.13, 0.68]
Time 3	-1.47	7.58	-0.49 [-0.86, -0.13]	0.96	0.81	0.33 [-0.08, 0.75]
D-N Triad Judgment						
Time 1	1.06	> 100	0.69 [0.31, 1.09]	2.41	> 100	1.99 [1.25, 2.79]
Time 2	1.06	> 100	0.72 [0.33, 1.11]	2.55	> 100	1.67 [1.00, 2.37]
Time 3	1.28	> 100	0.80 [0.40, 1.21]	2.46	> 100	1.61 [0.95, 2.29]
Selection Skills						
Time 1	0.63	31.85	0.60 [0.22, 0.98]	1.55	> 100	1.17 [0.62, 1.75]
Time 2	0.13	0.25	0.13 [-0.21, 0.46]	0.77	2.75	0.48 [0.05, 0.92]
Time 3	0.38	2.39	0.40 [0.05, 0.75]	0.68	4.14	0.52 [0.09, 0.97]

Note. MD = Mean Difference. BF₁₀ = Two-sided Bayes Factor. BCI = Bayesian Credible Interval. Chance was defined as 4/8 on the triad judgment tasks and 1/4 on the selection skills tasks. A positive mean difference/δ indicates performance at a level above chance.

structural awareness training persisted for 12 months. For instance, in our third wave of testing, trained participants continued to reliably out-perform control participants on four measures of structural awareness. These effects (median $\delta = 0.80$) were not as big as those observed for the same measures in waves one and two (median $\delta = 1.20$ and 1.12 respectively). However, they were still “large” (Cohen, 1988), and around twice the typical size of effects for interventions aimed at boosting achievement in higher education (Hattie, 2015). They were also consistent with the immediate effects of other experimental attempts to train structural awareness (Quilici & Mayer, 1996, 2002; Yan & Lavigne, 2014). To our knowledge, this is the first time that longer-term effects have also been studied. This is not surprising, given the rarity of longer-term follow-ups of experimental interventions in education (Watts et al., 2019).

On the fifth measure of structural awareness, scenario generation, the trained participants no longer reliably outperformed the control participants, despite an effect size of $\delta = 0.47$. This suggests insufficient statistical power. We only managed to re-recruit around half of the original Allen, Fielding, East, et al. (2020) participants. However, it is worth noting that those we did re-recruit did not obviously differ from those we did not (see S2 of the online supplement), suggesting that our effects are not merely an artefact of non-random attrition.

For the trained participants, performance on the structural awareness measures was generally stable over time (median $\delta = -0.11$), with the one large drop, on the scenario generation task, occurring mostly in the week following training. On the triad judgment tasks they showed a weak and unreliable preference for D on the S-D trials (median $\delta = 0.30$) and a very clear preference for D on D-N trials (median $\delta = 1.67$). The performance of the control group was similarly stable over time (median $\delta = 0.11$). In the triad judgment tasks at all three time points they preferred S over D (median $\delta = -0.57$), but D over N (median $\delta = 0.72$). However, the

strength of their preference for D over N was only half the size of the trained group’s. When considered alongside previous research (Allen, Dorozenko, & Roberts, 2016; Rabinowitz & Hogan, 2008), these findings indicate that psychology students are not inherently structurally aware and are unlikely to become so during their undergraduate years without training. Our data suggest that such training can be brief, and its effects can be lasting.

Despite continuing to out-perform the control group on most measures of structural awareness, by Time 3 the trained group’s statistic selection advantage had largely disappeared ($\delta = 0.25$). However, their performance was still modestly above chance levels, and just over the UK undergraduate “pass” threshold of 40%. The control group’s Time 3 performance was just under that threshold. These findings suggest that the relationship between structural awareness and statistic selection may be more complex than previously thought (Allen, Fielding, East, et al., 2020). They also indicate that selecting an appropriate statistical test for common research scenarios without assistance is stubbornly difficult for most students (see also Gardner & Hudson, 1999; Ware & Chastain, 1989, 1991). Finally, they suggest that any conceptual replication of this study should include an additional IV: with vs. without an aid during statistic selection. As there are few situations beyond exams where students would need to select a statistic “blind,” this would provide a more authentic assessment of the impact structural awareness training has on statistic selection skills.

In conclusion, students’ structural awareness should not be assumed, but can be trained. However, even structurally aware students cannot reliably select appropriate statistics without assistance. Training combined with easy access to a decision-making aid should maximize statistic selection accuracy. Instructors are encouraged to freely use and adapt our evidence-based training methods and materials (see particularly Allen, Fielding, Kay, & East, 2020; Allen, Fielding, East, et al., 2020), including StatHand, for these purposes.


Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: 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.

ORCID iD

Peter J. Allen  <https://orcid.org/0000-0002-9690-1545>

Open Practices Disclosure



The raw and processed data for this study are openly available for download at <https://osf.io/p7b4m/>. Code is not available as the analyses were performed in JASP. However, the analyses can be fully reproduced with reference to the data dictionary in section S5 of the online supplement: <https://osf.io/gtxvb/>. The materials used in this study are identical to the Time 2 materials published as an online supplement to Allen, Fielding, East, et al. (2020). They are openly available for download at <https://osf.io/2hkat/>. The data collection methods for this study were pre-registered at <https://osf.io/tv2fw/>, along with the hypotheses and data analysis plan. There was one deviation from the pre-registration: We had to test a small number of participants online due to the 2020 COVID-19 “lockdown” in the UK. This deviation is described in the paper. Prior to any data collection, this study was approved by the School of Psychological Science Research Ethics Committee at the University of Bristol (reference number: 96025). The complete Open Practices Disclosure for this article can be found at <http://journals.sagepub.com/doi/suppl/10.1177/0098628320985080>. This article has received badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at <http://www.psychologicalscience.org/publications/badges>

References

- 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. <https://doi.org/10.3389/fpsyg.2016.00188>
- Allen, P. J., Fielding, J. L., East, E. C., Kay, R. H. S., Steele, C. S., & Breen, L. J. (2020). Using StatHand to train structural awareness and promote the development of statistic selection skills. *Scholarship of Teaching and Learning in Psychology*. Advance online publication. <http://doi.org/10.1037/stl0000177>
- Allen, P. J., Fielding, J. L., Kay, R. H. S., & East, E. C. (2020). Using StatHand to improve students’ statistic selection skills. In A. Beyer & J. Peters (Eds.), *For the love of teaching undergraduate statistics* (pp. 178–203). Society for the Teaching of Psychology. <http://teachpsych.org/ebooks/lovestats>
- 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*(1), 23–36. <https://doi.org/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. <https://doi.org/10.3389/fpsyg.2016.00288>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.
- Gardner, P. L., & Hudson, I. (1999). University students’ ability to apply statistical procedures. *Journal of Statistics Education, 7*(1). <http://jse.amstat.org/secure/v7n1/gardner.cfm>
- Hattie, J. (2015). The applicability of visible learning to higher education. *Scholarship of Teaching and Learning in Psychology, 1*(1), 79–91. <https://doi.org/10.1037/stl0000021>
- Nolan, S. A., & Heinzen, T. E. (2017). *Statistics for the behavioral sciences* (4th ed.). 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*(2), 36–41. <https://doi.org/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*(1), 144–161. <https://doi.org/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*(3), 325–342. <https://doi.org/10.1002/acp.796>
- Rabinowitz, M., & Hogan, T. M. (2008). Experience and problem representation in statistics. *American Journal of Psychology, 121*(3), 395–407. <https://doi.org/10.2307/20445474>
- Society for the Teaching of Psychology Statistical Literacy Task Force. (2014). *Statistical literacy in the undergraduate psychology curriculum*. Society for the Teaching of Psychology. https://teachpsych.org/Resources/Documents/otrp/resources/statistics/STP_Statistical%20Literacy_Psychology%20Major%20Learning%20Goals_4-2014.pdf
- van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., Evans, N. J., Gronau, Q. F., Haaf, J. M., Hinne, M., Kucharský, Š., Ly, A., Marsman, M., Matzke, D., Gupta, A. R. K. N., Sarafoglou, A., Stefan, A., Voelkel, J. G., & Wagenmakers, E.-J. (2020). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*. Advance online publication. <https://doi.org/10.3758/s13423-020-01798-5>
- Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E.-J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., de Jong, T., van den Bergh, D., Sarafoglou, A., Steingrover, H., Derks, K., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP.

- Psychonomic Bulletin & Review*, 25(1), 58–76. <https://doi.org/10.3758/s13423-017-1323-7>
- Ware, M. E., & Chastain, J. D. (1989). Computer-assisted statistical-analysis: A teaching innovation? *Teaching of Psychology*, 16(4), 222–227. https://doi.org/10.1207/s15328023top1604_16
- Ware, M. E., & Chastain, J. D. (1991). Developing selection skills in introductory statistics. *Teaching of Psychology*, 18(4), 219–222. https://doi.org/10.1207/s15328023top1804_4
- Watts, T. W., Bailey, D. H., & Li, C. (2019). Aiming further: Addressing the need for high-quality longitudinal research in education. *Journal of Research on Educational Effectiveness*, 12(4), 648–658. <https://doi.org/10.1080/19345747.2019.1644692>
- Yan, J., & Lavigne, N. C. (2014). Promoting college students' problem understanding using schema-emphasizing worked examples. *Journal of Experimental Education*, 82(1), 74–102. <https://doi.org/10.1080/00220973.2012.745466>