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

American Psychological Association

Scholarship of Teaching and Learning in Psychology

2332-2101/19/$12.00

2019, Vol. 5, No. 1, 23–36
http://dx.doi.org/10.1037/stl0000132

An Experimental Evaluation of StatHand: A Free Application to Guide Students' Statistical Decision Making

Peter J. Allen
University of Bristol

James Finlay, Lynne D. Roberts, and Frank D. Baughman
Curtin University

Quantitative research methods underpin psychological literacy and evidence-based practice in psychology. Despite this, many students struggle to identify appropriate statistics for different types of research questions and data types. StatHand (see https://stathand.net) is a free application that facilitates this statistical decision making process by prompting students to focus systematically on each structural characteristic of their research design. A total of 217 undergraduate psychology students were randomized to use one of four decision making aids: StatHand on an iPad, a familiar textbook, a familiar paper decision tree, or the textbook and decision tree combined. Participants were then asked to identify suitable statistics for five research scenarios. Students assigned to use StatHand demonstrated higher decision making accuracy than users of the three alternative aids (δ = .50 to .64). StatHand users also experienced lower cognitive load, higher confidence in the accuracy of their decisions and greater satisfaction with their assigned aid than one or more of the other groups. However, it took the StatHand users longer to make their decisions. Finally, there was strong evidence to support the hypothesis that StatHand is instructionally efficient, and that its use requires less effort to promote higher performance relative to the other three aids (δ = .49 to .70). StatHand can be incorporated into a variety of classroom learning activities, and educators are encouraged to consider how they can use it most effectively.

Keywords: mobile learning app, statistic selection, decision tree, instructional efficiency, Bayesian

Quantitative research methods underpin psychological literacy and evidence-based practice in psychology (American Psychological Association Presidential Task Force on Evidence-Based Practice, 2006; Cranney, Morris, & Botwood, 2015). They comprise a substantial part of most undergraduate psychology curricula (Norcross et al., 2016), and are prominently represented in the graduate attributes and course learning outcomes specified by psychology accreditation organisations around the world (e.g., American Psychological Association, 2016; Australian Psychology Accreditation Council, 2018; British Psychological Society, 2017). The quantitative research skills of psychology graduates are valued by employers (Appleby, 2018),

Peter J. Allen, School of Psychological Science, University of Bristol; James Finlay, Lynne D. Roberts, and Frank D. Baughman, School of Psychology, Curtin University.

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 authors would also like to acknowledge Casimir Ludwig, School of Psychological Science, University of Bristol, for his feedback on a draft of this article. The results reported in this article were previously presented at 10th International Conference on Teaching Statistics (ICOTS10) and the 2018 British Psychological Society Division of Academics, Researchers & Teachers in Psychology (DART-P) conference.

Correspondence concerning this article should be addressed to Peter J. Allen, School of Psychological Science, University of Bristol, 12a Priory Road, Bristol BS8 1TU, United Kingdom. E-mail: p.allen@bristol.ac.uk

and put psychology majors at an advantage relative to majors from many other disciplines when competing for graduate level positions in a wide range of industries (Halonen & Dunn, 2018).

Despite the ubiquity and value of quantitative research-methods, many undergraduate students find them to be among the most challenging components of their degrees (Waples, 2016). One skill that appears particularly lacking among students is the ability to identify appropriate statistics for different types of research questions and data types. This ability has been referred to by Ware and Chastain (1989, p. 223) as "selection skill." To illustrate, Gardner and Hudson (1999) presented upper level undergraduate and graduate-level psychology and education students with 21 research scenarios, and asked them to recall appropriate statistics for as many as possible within a 45-min period. The scenarios suggested statistics covered in typical introductory behavioral statistics textbooks (e.g., Field, 2018). Even though most students had completed at least six research-methods modules, nearly all performed poorly. On average, they read 10.9 scenarios and recalled correct statistics for just 25% of them. Follow-up interviews revealed that the students' poor performance could be attributed to a range of factors, including misinterpretation of the research scenarios, an inability to recall the names of known statistics, and confusion around measurement levels (Gardner & Hudson, 1999). More recently, Allen, Dorozenko, and Roberts (2016) asked undergraduate psychology students to reflect on the strategies they use to select statistics for scenarios like Gardner and Hudson's (1999). Although these students were, on average, in their third year of study, the strategies described by most were haphazard and inefficient. For instance, students reported searching for (potentially misleading) clues in the wording of the scenarios, scanning through textbooks and lecture notes, relying on memory or the advice of friends, and sometimes just guessing. While some students noted that a systematic decision making process could be applied to statistic selection, none could describe it clearly or completely, and most also raised issues that were irrelevant to the task at hand.

When students have been asked to recognize (as opposed to recall) suitable statistics, their performance does not appear much stronger. For example, students averaged just 45% on a multiple-choice selection skill test that Ware and Chastain (1989) administered at the end of a first year statistics module. When reflecting on this poor performance, Ware and Chastain (1989) noted that many research-methods instructors teach statistical procedures "one at a time" (p. 226), and provide relatively few opportunities for students to think about and practice selection skills. A similar point has been made by Quilici and Mayer (1996, 2002) and Yan and Lavigne (2014).

Even though not all research-methods instructors do so, it is possible to train selection skills. Ware and Chastain (1991) did this by restructuring their module to increase the emphasis placed on when to use different statistics, and observed a corresponding improvement in students' selection skills. Furthermore, selection skills appear to be built on "structural awareness," which is the ability to focus on the structural (e.g., the number and nature of variables; the hypothesized relationships between them etc.) rather than surface-level characteristics of research scenarios (Quilici & Mayer, 2002, p. 325). Like selection skills, structural awareness can be trained. Students who have undergone such training are more likely to categorize research scenarios according to how they would be analyzed (rather than, e.g., their substantive topic area) and then correctly identify the structural features defining each category. They are also better able to generate new research scenarios that are structurally similar to those seen previously, and more likely to demonstrate selection skills by applying appropriate statistics to novel research scenarios (Quilici & Mayer, 1996, 2002; Yan & Lavigne, 2014).

Quilici and Mayer (1996, 2002) and Yan and Lavigne (2014) used worked examples to highlight the structural features of the research scenarios in their studies. Decision trees can serve a similar pedagogic function by systematically focusing students' attention on each structural component of a research scenario, as well as the hierarchical and horizontal relationships between components (Schau & Mattern, 1997). As a tool to guide statistic selection, decision trees have a long history (e.g., Mock, 1972), and are now commonly included in statistics textbooks (e.g., Allen, Bennett, & Heritage, 2014; Field, 2018). Their inclusion in textbooks is supported by research demonstrating that deci-

sion trees can facilitate timely and accurate statistic selection, as well as research indicating their popularity among students (Carlson, Protsman, & Tomaka, 2005; Protsman & Carlson, 2008).

Despite their efficacy and popularity, traditional decision trees are usually constrained by the requirement that they fit on a single sheet of paper, or within a few pages of a textbook. Because of this, information that would assist users with navigating the tree (e.g., definitions of key terms) is either spatially separated from the tree or absent entirely (Koch & Gobell, 1999). Furthermore, the scale, complexity and nonlinearity of a decision tree can overwhelm some users, and prompt them to disengage. This phenomenon has been referred to as "map shock" (Blankenship & Dansereau, 2000).

Map shock and the space constraints associated with print media can both be overcome with hypertext. For instance, when Koch and Gobell (1999) adapted paper-based decisions trees for deployment on the Internet, they broke each into a series of decision points, and presented these to users one at a time. This approach has the advantage of prompting users to systematically engage with every salient aspect of their research design before settling on a statistical procedure. Furthermore, within the web-based tree, Koch and Gobell (1999) were able to provide definitions, examples and information on how to compute statistics, along with links to relevant external online resources. A small-scale evaluation suggested that, compared to students in a control condition, students using the web-based tree were better able to select appropriate statistics for different research designs, more confident in their selections, and found the selection process easier to complete. Koch and Gobell's (1999) web-based tree is no longer available, although a number of contemporary versions have taken its place. Some of these are freely available (e.g., Jackson, n.d.), whereas others have been published partially or completely behind a paywall (e.g., Lund Research, 2018). This new generation of web-based trees does not appear to have undergone evaluation. Furthermore, their functionality relies on a live Internet connection.

By way of contrast, mobile learning applications can be developed to maintain functionality without an Internet connection (Kretser et al., 2015). In the last decade, the use of mobile learning technologies including smart devices (e.g., smart phones, tablets) and mobile applications has increased at a rapid rate. Among higher education students in developed countries, their level of penetration is approaching 100% (Brooks & Pomerantz, 2017), and students have expressed positive attitudes toward using mobile learning technologies for a range of educational purposes (Bowen & Pistilli, 2012). This has encouraged educators to consider ways of incorporating mobile learning technologies into classroom activities (Stowell, 2015). Recent research suggests the use of these technologies promotes learning within psychology (Diliberto-Macaluso & Hughes, 2016), statistics (Ling, Harnish, & Shehab, 2014) and a range of other disciplines (Sung, Chang, & Liu, 2016).

It is within this context that StatHand was developed (see Allen, Roberts, et al., 2016; Allen et al., 2017). StatHand is a cross-platform application that aids the process of selecting appropriate statistics for a wide range of research questions and data types by prompting the user to focus systematically on each structural feature of their research problem. Native iOS applications for iPad and iPhone are available free on the Apple App Store. Users of other devices can access the mobile-compatible StatHand web application via https://stathand.net. The content and features of the iOS and web applications are identical, although the latter relies on a live Internet connection.

In a recent qualitative evaluation, 25 undergraduate psychology students participated in focus groups, and nine psychology instructors participated in semistructured interviews in which the utility, merits and limitations of StatHand were explored (Allen et al., 2017). The students liked the interactivity and accessibility of StatHand, considered it faster and more reliable than familiar alternatives, and indicated that they would recommend it to peers. The instructors saw StatHand as an aesthetically pleasing, user-friendly application likely to promote active learning and self-efficacy. Both groups suggested features that have since been incorporated into the application.

The current study extends the work of Allen et al. (2017) by subjecting StatHand to an experimental evaluation. Specifically, we random-

ized undergraduate psychology students to four decision making aids, (a) StatHand on an iPad, (b) a familiar textbook, (c) a familiar paper decision tree, or (d) the textbook and decision tree combined, and asked them to identify suitable statistics for five research scenarios. We hypothesized that participants in the StatHand condition would demonstrate a higher level of decision making accuracy (Hypothesis 1 [H1]), and self-report lower cognitive load (Hypothesis 2 [H2]), greater confidence in the accuracy of their decisions (Hypothesis 3 [H3]) and higher satisfaction with their assigned decision making aid (Hypothesis 4 [H4]) than participants in the three control conditions. As it was plausible that participants in the StatHand condition could take either more or less time to make their decisions than participants in the control conditions, a nondirectional hypothesis was made for this dependent variable. That is, we hypothesized that the decision making speed of participants in the StatHand condition would differ from that of participants in the remaining conditions (Hypothesis 5 [H5]). Finally, educational techniques or resources can be evaluated or compared in terms of their instructional efficiency. A technique or resource is considered instructionally efficient when, compared to other techniques or resources, its use requires less effort (or cognitive load) to promote higher performance (Hoffman & Schraw, 2010). We hypothesized that the StatHand decision making aid would be instructionally efficient, relative to the other three aids (Hypothesis 6 [H6]).

Method

Participants

A total of 227 second-year psychology students participated in the activities described herein as part of a class data collection exercise at the start of a psychology research-methods module at Curtin University in Australia. For most students, this was the third research-methods module in their degree. The previous two modules focused on evidence-based practice (part of a common first year health sciences curriculum) and experimental design and analysis. Of the 227 students, nine did not consent to the use of their data in this study. One further case was excluded due to excessive missing data. Therefore, the final sample size was $N = 217$ (although, due to missing data, some analyses are based on slightly fewer cases). The demographic characteristics of the final sample are reported in Table 1, where it can be seen that the four groups shared similar gender distributions, mean ages and mean marks on the experimental design and analysis module.

We tested our hypotheses using both Bayesian and frequentist methods. A frequentist sensitivity power analysis using G*Power 3.1.4 (Faul, Erdfelder, Lang, & Buchner, 2007) indicated that we had power of .80 for detecting differences of at least $d = .64$ in our one-sided frequentist hypothesis tests (when evaluated for significance at $\alpha = .01$). For the two-sided frequentist hypothesis tests, the smallest detectable effect was $d = .68$. These effect sizes were used in fixed-*n* Bayes factor design analyses (BFDA; Schönbrodt & Wagenmakers, 2018), which indicated a .86/.83 probability of observing Bayes factors (BFs) > 3 (qualitatively defined as at least "moderate" evidence in favor of the research hypothesis; Wagenmakers, Love, et al., 2018) in our one-/two-sided Bayesian hypothesis tests. The probability of inconclusive or anecdotal evidence (BFs between 3 and .33) was estimated at around .14/.16, while the probability of false negatives (BFs $< .33$) was

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

			Age		Previous module mark	
	N	% Female	*M*	*SD*	*M*	*SD*
StatHand	50	70.00	20.96	3.63	63.69	11.93
Textbook (TB)	57	63.20	20.70	4.58	63.50	9.92
Decision Tree (DT)	55	67.30	22.27	5.97	60.48	8.19
TB + DT	55	76.40	20.63	5.02	62.93	9.61
Full sample	217	69.10	21.14	4.91	62.62	9.96

less than .01. As it could be argued that $\delta = .64$ to .68 was somewhat optimistic, we reran the BFDA using $\delta = .40$, which Hattie (2015) recently reported as the average sized effect for educational research. Assuming a population effect size of $\delta = .40$, we estimated .47/.34, .46/.50 and .07/.16 probabilities of observing BFs >3, 3–.33 and $< .33$, respectively in our one-/two-sided Bayesian hypothesis tests. Although this latter set of probabilities suggest that a larger sample size would have been desirable, our sampling was constrained by the number of students enrolled in the research-methods module. Our sample represented over 90% of the total enrolled cohort at the time of testing. Our Bayes Factor Design Analyses were run using the BFDA (Version 0.2) R package (Schönbrodt, 2017) with $n = 51$ (the smallest average cell size for our Bayesian hypothesis tests) and the number of simulations set at 10,000.

Prior to data collection, the study was reviewed and approved by the Human Research Ethics Committee at Curtin University (Reference Number: RDHS-125–15). Although participation in data collection was a requisite part of the students' research-methods module (and the collected data were used in subsequent teaching and learning activities), participation in the current research was strictly voluntary. Students indicated consent (or lack thereof) actively via the online questionnaire that was used to collect all the data (excluding prior module marks, which were extracted from university records) for this study.

Materials and Measures

Decision making aids. The decision making aids used in this study were (a) the StatHand application (Allen, Roberts, et al., 2016; Allen et al., 2017) installed on an iPad, (b) the *SPSS Statistics Version 22: A Practical Guide* (Allen et al., 2014) textbook, without the decision tree printed inside the front cover, (c) the decision tree from Allen et al. (2014) on an A4 sheet of paper, and (d) the full Allen et al. (2014) textbook, including the decision tree. The Allen et al. (2014) textbook and decision tree were selected due to their familiarity to participants. This is consistent with past studies (e.g., Carlson et al., 2005; Protsman & Carlson, 2008), where performance when using a familiar textbook has been compared to performance when using a novel decision making aid. By the time the current study was conducted, participants had been using the Allen et al. (2014) textbook as a core text for approximately one semester.

Accuracy. To measure decision making accuracy, participants were asked to specify an appropriate statistical test or procedure for each of five research scenarios (see https://osf.io/ut75r/). These scenarios were drafted by J. F. in consultation with P. A., who was the coordinator of the research-methods module from which the sample were drawn, and were then further refined following feedback from seven final-year psychology students. The scenarios spanned a range of statistical procedures taught in both the previous and current research-methods modules. As each scenario could potentially lend itself to several different analyses (e.g., the most obvious analysis for scenario five would be a one-way between subjects analysis of variance, although a Kruskal-Wallis analysis of variance could also be appropriate in some circumstances), a comprehensive list of "correct" answers for each was drawn up prior to data collection (see https://osf.io/ut75r/). Each answer was coded as either 0 (incorrect) or 1 (correct). Instances where students did not provide answers were coded as 0. Therefore, scores on this measure could range from 0 to 5, with higher scores indicating greater decision making accuracy.

Cognitive load. Following Paas and Van Merriënboer (1993), cognitive load was measured by asking participants to rate the amount of mental effort they invested in answering each scenario. They did this five times, once immediately following each scenario, on a 9-point scale ranging from 1 (*very, very low mental effort*) to 9 (*very, very high mental effort*). Scores for these five items were averaged, with higher average scores reflecting greater self-reported cognitive load.

Confidence. Participants were asked to rate their confidence in their answers for each scenario on scales ranging from 1 (*not at all confident*) to 9 (*extremely confident*). Scores for these five items were averaged, with higher average scores reflecting greater self-reported confidence.

Satisfaction. Participants were asked to indicate how useful, on a scale ranging from 1 (*not at all useful*) to 9 (*extremely useful*), their

assigned decision making aid was when identifying appropriate statistics for the research scenarios. They were then asked to indicate whether their assigned aid made identifying appropriate statistics more difficult or easier than this task would have been without access to any resources. Participants responded to this item on a 9-point scale ranging from 1 (*much more difficult*) to 9 (*much easier*). Scores for these two items were averaged, with higher scores representing a greater level of satisfaction with the assigned aid.

Speed. Speed was operationalized as the total time that the five scenarios were visible to participants, in seconds. That is, the sum of the time that elapsed between the presentation of each scenario and the participant clicking "next" to move on to the relevant cognitive load and confidence items. In the current study, those speeds ranged from 134 to 1,432 s, with a mean of 510 s (SD = 205 s). Participants spent, on average, 102 s (SD = 41 s) working on each scenario.

Instructional efficiency. We adopted the deviation model of instructional efficiency proposed by Paas and van Merriënboer (1993), which views relative efficiency in terms of the standardized difference between performance (or accuracy) and effort (or cognitive load). Positive scores indicate greater accuracy than would be predicted based on self-reported cognitive effort expended. Negative scores indicate the reverse.

Additional exploratory items. Participants were asked to indicate whether they would have preferred a different decision making aid, either in addition to, or instead of their assigned aid. Those who answered in the affirmative were then asked to specify what exactly they would have preferred. It should be noted that although participants completed the experimental tasks individually, they did so in laboratory classes in the presence of participants assigned to all four experimental conditions. They were therefore primed to consider at least two other potential decision making aids (and perhaps three, if they were able to distinguish between the versions of the textbook that did and did not include the decision tree). Finally, participants in the combined (textbook with decision tree) condition were asked to indicate which they made most use of. The three available response options were (a) the decision tree, (b) the rest of the textbook, or (c) both equally. No hypotheses relating to these items were specified.

Context. The measures described above were embedded in a larger online questionnaire also used to collect data for subsequent class activities and assessments. Other measures in the questionnaire included two Big Five personality factor scales, a satisfaction with life scale, a measure of music preferences, an attitude toward statistics scale, and several single item measures. The data collected using these measures were not used in any of the analyses that follow.

Procedure

Data were collected during the first set of laboratory classes for a second-year psychology research-methods module at Curtin University. There were approximately 20 students in each class. Each class began with a brief introductory presentation, which included an overview of the current study and a reminder that participation in research should always be voluntary and that consent should always be informed. Students were then block randomized into the four experimental conditions: (a) StatHand, (b) textbook, (c) decision tree, or (d) textbook and decision tree combined. This involved students self-organizing into groups of four, and using the random sequence feature on https://random.org to assign each group member to one of the four conditions. During this process students were blind to the nature of each condition. The relevant decision aids were then distributed to students, who were subsequently directed to the URL for the study's information sheet. After reading the information sheet, students were asked whether or not they consented to the use of their data in the current study. It was made clear that, regardless of whether or not they consented to participate in the current study, their data would be made available (in an anonymized form) for students to analyze in class and assessment activities throughout the semester. Following this, students commenced the online questionnaire in which the experimental tasks were embedded.

On reaching the section of the questionnaire containing the experimental tasks, students were informed that they would be presented with six research scenarios, and should indepen-

dently use their assigned decision making aid to identify an appropriate statistical test for each. These first of these was a practice scenario, for which feedback was provided. Following each scenario students completed the cognitive load and confidence measures. Following all six scenarios, students completed the satisfaction and additional exploratory items. Finally, at the end of the questionnaire, students provided their age, gender and student ID number. The latter was requested to enable matching of questionnaire responses with previous module marks. It took students an average of 24.8 min (SD = 4.3 min) to complete the entire questionnaire, although it is not possible to determine exactly what proportion of that was spent on the experimental tasks.

Results

We analyzed our data using both Bayesian and frequentists methods. Both approaches suggest the same conclusions. The Bayesian results are reported here, while the frequentist results are located at https://osf.io/ut75r/.

One-sided Bayesian t tests (using a default Cauchy prior with a scale parameter of r = .707; Wagenmakers, Love, et al., 2018) were used to test H1 to H4 and H6. Two-sided Bayesian t tests were used for H5. In all instances, StatHand was compared against each individual control aid, as well as the three control aids combined. The results of these tests, along with relevant descriptive statistics and effect sizes are reported in Table 2. To facilitate interpretation, condition means and their 95% credible intervals are illustrated in Figure 1.

The BFs reported in Table 2 represent the probability of the observed data under the research hypothesis (H_1, there is an effect, in the specified direction where applicable) versus the null hypothesis (H_0, there is no effect). As such, they can be used to quantify the strength of evidence in favor of either H_1 or H_0. According to the commonly used heuristic classification scheme (see Wagenmakers, Love, et al., 2018), BFs between 3 and 10 reflect moderate evidence in favor of H_1, whereas progressively larger BFs represent strong (BF = 10 to 30), very strong (BF = 30–100) and extreme (BF > 100) evidence in favor of H_1. Conversely, BFs between .33 and .10, between .10 and .03, between .03 and .01, and smaller than .01 reflect moderate, strong, very strong and extreme evidence in favor of H_0, respectively. Finally, BFs in the range of .33 to 3 are considered nondiagnostic in the sense that they do not provide clear evidence in favor of either H_1 or H_0. Effect size δ is an estimate of the population standardized difference between two independent means, and we can be 95% confident that the true value of δ lies within its 95% credible interval. All BFs, δs and their associated 95% credible intervals were calculated in JASP 0.8.6 (Wagenmakers, Love, et al., 2018).

The results presented in Table 2 and in Figure 1 indicate that the StatHand users demonstrated higher performance accuracy than users of the three control aids. These effects were relatively strong (δ = .50 to .64). The results also suggest that StatHand users experienced lower cognitive load, higher confidence and greater satisfaction than at least one other user group. However, StatHand users tended to take longer to identify and specify a statistic for each scenario. Finally, there is strong evidence to indicate that StatHand can be considered instructionally efficient relative to the other three aids (δ = .49 to .70).

Analysis of the exploratory items completed by participants following the main experimental tasks indicated a similar profile of preferences for the StatHand and combined (textbook and decision tree) groups (see Table 3). Specifically, over half of the members of each of these two groups reported that they would not have preferred to use an alternative decision making aid to identify appropriate statistics for the five scenarios. Under 10% of each group indicated that they would have preferred an alternative instead of their assigned aid, and the remainder indicated that they would have preferred an alternative as well as their assigned aid. Predictably, the most common alternative aids named by the StatHand group were the decision tree and textbook (in combination, nearly 80% of responses), while the most common alternative aid named by the combined (textbook and decision tree) group was StatHand (or "the iPad app"; 73% of responses). The profiles for the remaining two groups indicated greater dissatisfaction with their assigned aids (most notably for the textbook only group), and a corresponding increase in preferences for alternative aids, either instead of, or as well as those they were assigned. Around 70% of preferred alternatives

Table 2
Descriptive Statistics for Each Condition, and Bayesian Summary Information About the Differences Between StatHand and the Three Control Conditions (Individually and in Combination)

	Descriptive statistics by condition			Differences between StatHand and control conditions		
	N	*M*	*SD*	Mean difference	BF	δ [95% credible interval]
Accuracy						
StatHand	50	1.74	1.19			
Textbook (TB)	57	1.00	1.02	.74	**73.25**	.62 [.23, 1.01]
Decision Tree (DT)	55	1.02	.89	.72	**89.13**	.64 [.25, 1.04]
TB + DT	55	1.13	1.06	.61	**12.46**	.50 [.13, .89]
All controls	167	1.05	.99	.69	**>100**	.63 [.32, .95]
Cognitive load						
StatHand	48	5.32	1.62			
TB	56	5.51	1.70	−.19	.35	.10 [−.26, .47]
DT	55	6.03	1.55	−.71	**3.84**	.40 [.03, .78]
TB + DT	54	5.72	1.64	−.40	.73	.22 [−.14, .59]
All controls	165	5.75	1.63	−.43	1.09	.24 [−.06, .55]
Confidence						
StatHand	50	3.40	1.60			
TB	56	3.18	1.53	.22	.39	.12 [−.24, .48]
DT	53	2.64	1.56	.76	**5.66**	.44 [.07, .82]
TB + DT	53	2.84	1.33	.57	2.19	.34 [−.02, .73]
All controls	162	2.89	1.49	.51	2.50	.31 [.00, .63]
Satisfaction						
StatHand	49	6.21	1.91			
TB	56	4.27	1.91	1.95	**>100**	.97 [.55, 1.38]
DT	54	6.03	1.84	.19	.32	.09 [−.27, .46]
TB + DT	54	5.93	1.55	.29	.45	.15 [−.22, .52]
All controls	164	5.39	1.94	.82	**7.67**	.40 [.10, .71]
Speed						
StatHand	50	603.22	217.67			
TB	57	511.60	218.67	91.62	1.63	.38 [.00, .75]
DT	55	449.03	139.64	154.20	**>100**	.80 [.38, 1.21]
TB + DT	55	485.26	210.43	117.97	**6.70**	.51 [.10, .89]
All controls	167	482.31	193.76	120.91	**>100**	.57 [.25, .89]
Instructional Efficiency						
StatHand	48	.50	1.18			
TB	56	−.08	.97	.58	**10.60**	.49 [.11, .88]
DT	55	−.28	.89	.79	**>100**	.70 [.29, 1.11]
TB + DT	54	−.09	.96	.59	**12.43**	.50 [.13, .91]
All controls	165	−.15	.94	.65	**>100**	.62 [.29, .95]

Note. BF = Bayes factor. BF_{+0}s are reported where the StatHand condition was hypothesized to have the higher mean. BF_{-0}s are reported where the StatHand condition was hypothesized to have the lower mean. For speed, two sided BF_{10}s are reported. BFs >3 are in bold text. All δs and associated 95% credible intervals were estimated using a two sided default Cauchy prior with a scale parameter of $r = .707$.

for the decision tree and textbook only groups were the textbook and decision tree respectively. Another 20% of responses for each group related to StatHand. Finally, the students in the combined (textbook and decision tree) group were asked to indicate which part of their decision making aid they made most use of. For nearly 80%, this was the decision tree, which goes some way toward explaining the relative dissatisfaction expressed by students in the textbook only condition.

Discussion

This article reports the results of an experimental evaluation of StatHand, an application designed to facilitate the selection of appropriate statistics for a wide range of research de-

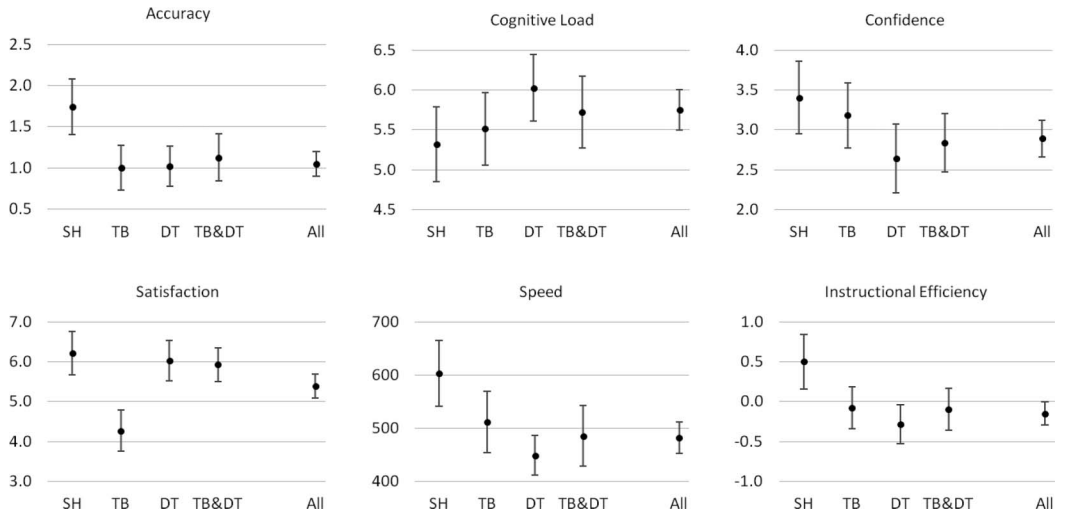

Figure 1. Means and Bayesian 95% credible intervals for each condition (and the three control conditions in combination) on each dependent variable. SH = StatHand; TB = Textbook; DT = Decision Tree; TB&DT = Textbook and Decision Tree combined; All = all three control conditions combined.

signs (Allen, Roberts, et al., 2016; Allen et al., 2017). We found that students assigned to StatHand were able to identify appropriate statistics for more research scenarios than students assigned to the control aids. These effects were relatively large, and consistent with our first hypothesis (H1). They were also consistent with the work of Carlson and colleagues (2005; Protsman & Carlson, 2008) and Koch and Gobell (1999) who observed that students assigned to paper and hypertext decision trees outperformed students in control conditions on tasks similar to ours. Ostensibly, these effects occur because tools like StatHand promote the development of structural awareness (Quilici & Mayer, 1996, 2002). However, this hypothesis requires testing in future research.

Students assigned to StatHand not only outperformed students in the textbook condition, but also students in both the paper decision tree and combined conditions. The performance benefits of StatHand relative to the decision tree may be linked to at least two factors. First, StatHand requires that users engage with all necessary decision points before a statistic is suggested. Unlike a paper decision tree, it is not possible to 'gloss over' difficult decisions in StatHand. Second, StatHand provides the guidance necessary for making informed decisions. If, for example, a student requires the definition of a "nominal variable" to progress their decision making, this definition is readily available within the application. The performance benefits of StatHand relative to the decision tree/

Table 3
Participants' Preferences for Assigned Decision Making Aids Versus Alternative Aids, Split by Experimental Condition

	No (%)	Yes, instead of assigned aid (%)	Yes, in addition to assigned aid (%)
StatHand	52.0	8.0	40.0
Textbook (TB)	10.5	43.9	45.6
Decision Tree (DT)	25.5	20.0	54.5
TB + DT	57.4	9.3	33.3

textbook combination may be linked to the spatial relationship between decision points and guiding content. In the combined condition, students had to temporarily leave the decision tree to locate definitions and other explanatory information in the textbook. In the StatHand condition, such guidance could be accessed with a single tap, and without needing to navigate away from one's current place within the decision making sequence. Of course, the precise reasons why students using StatHand outperformed those using other aids require more serious consideration in future research.

In contrast to Carlson and colleagues (2005; Protsman & Carlson, 2008), students in our textbook only condition did not appreciably underperform relative to students in our decision tree and combined conditions. However, they were less satisfied with their assigned aid, and more likely to have indicated a preference for an alternative. As has been observed in similar contexts, there is not always a clear association between student satisfaction and student performance (Allen & Baughman, 2016; Sizemore & Lewandowski, 2009).

Even though students using StatHand outperformed those using the control aids, their performance was still rather underwhelming. On average, they identified appropriate statistics for just 1.74 of the five scenarios. This suggests that simply providing students with a tool like StatHand is not enough to promote accurate statistic selection. Rather, to be maximally effective, StatHand needs to be integrated into the research-methods curriculum. Allen, Roberts, et al. (2016) provide some suggestions for achieving this based on the unified theory of acceptance and use of technology (Venkatesh, Morris, Davis, & Davis, 2003). Specifically, they recommend demonstrating StatHand at the outset and throughout the course, linking it to existing teaching resources, minimizing competition from other sources of interaction, and encouraging students to use it consistently and repeatedly in both methods and non-methods modules. Intuitively these recommendations make sense, although their efficacy has not yet been tested.

Our second, third and fourth hypotheses were partially supported. There was evidence to suggest that students in the StatHand condition experienced lower cognitive load (H2), higher confidence in the accuracy of their decisions (H3) and greater satisfaction with their assigned decision making aid (H4) than at least one of the control groups. Specifically, StatHand users reported lower cognitive load and higher confidence than the decision tree users, and higher satisfaction than the textbook users. The remaining comparisons between StatHand and the control aids on cognitive load, confidence and satisfaction were largely non-diagnostic, and the estimated effect sizes were generally small.

That the use of StatHand appears no more cognitively taxing than the use of other common decision making aids can be seen as a positive outcome when one considers its relative complexity (e.g., there are over 65 unique pathways through StatHand, but only 37 through the decision tree in Allen et al., 2014) and novelty to the students in the sample. That using StatHand was somewhat less cognitively taxing than using the decision tree is not surprising considering the sparseness of the latter. For similar reasons, it makes sense that students using StatHand would be more confident in their decisions than students using the paper decision tree, as they were able to compare the experimental research scenarios with the examples provided within the application. Similar reassurance was possible for students in the textbook and combined conditions, but not for those in the decision tree only condition. That StatHand was no less satisfactory to students than the decision tree and decision tree/textbook combination, and was clearly more satisfactory than the textbook in isolation, should give some confidence to teachers looking to introduce students to new evidence-based instructional technologies like StatHand while remaining mindful of how student satisfaction is associated with student engagement and success (Strahan & Credé, 2015), as well as their own career progression (Hornstein, 2017).

In partial support of H5, it took students in the StatHand condition longer to make their decisions than students assigned to the decision tree and combined conditions. Given the relative novelty and complexity of StatHand, this result makes intuitive sense. These effects were medium-to-large by Cohen's (1988) conventions. However, it should be noted that a medium-to-large effect in this context translates to just 25–30 s per scenario. This suggests that educators need not be concerned that integrating StatHand into the research-methods curric-

ulum will take time away from other important activities.

Finally, our data provided strong evidential support for the hypothesis that StatHand is instructionally efficient relative to the other three aids (H6). These effects were fairly large, and indicate that using StatHand requires less effort to promote stronger performance, relative to the three control aids. In an era when educators experience regular pressure to 'achieve more with less' (Mitchell, Leachman, & Masterson, 2017), this is arguably the strongest reason to advocate for the regular integration of StatHand into quantitative research-methods teaching and learning activities. The adoption of techniques and materials that are instructionally efficient gives students opportunities to increase the rate, amount and quality of their learning, which can free up time and cognitive resources for more advanced study or non-academic pursuits (Hoffman & Schraw, 2010).

Despite the potential impact of this study, it is not without limitations of the type common to psychology scholarship of teaching and learning (SoTL) research (Wilson-Doenges, Troisi, & Bartsch, 2016). First, this study was cross-sectional rather than longitudinal, and thus has little to say about whether the gains achieved by the StatHand group will generalize beyond the immediate study context, or be sustained over any period of time. Second, even though our sample size was respectable by the standards of psychology (Bakker, van Dijk, & Wicherts, 2012), it was still less than ideal. The consequences of this are most obviously reflected in the width of the credible intervals around the effect sizes reported in Table 2. For example, the population effect size δ was estimated as .62 for the first comparison reported in Table 2, but could plausibly range between .23 (small by Cohen's, 1988, conventions) and 1.01 (or larger than large). The only way of increasing the precision of these estimates is to increase the sample sizes on which they are based. Third, our sample was fairly homogenous, having been sourced from a single psychology course in a single Australian university. This carries implications for external validity that can only be addressed with replication. Fourth, our methods were exclusively quantitative, and thus we were unable to triangulate the benefits (or otherwise) of using StatHand via multiple methods. This limitation is, however, somewhat mitigated by the qualitative evaluation reported in Allen et al. (2017). The use of longitudinal designs, large and diverse samples, and mixed methods are four of the eight 'gold standard benchmarks' to which Wilson-Doenges et al. (2016) propose all psychology SoTL researchers should aspire.

The remaining gold standards proposed by Wilson-Doenges et al. (2016) include (a) situating psychology SoTL in a theoretical and/or empirical context, (b) employing true experimental designs, (c) using advanced statistical methods, and (d) maintaining the highest ethical standards. Our study reflects all of these standards. First, it was informed by research demonstrating the "selection skill" deficit among students (Allen, Dorozenko, & Roberts, 2016; Gardner & Hudson, 1999; Ware & Chastain, 1989, 1991), the theoretical construct of "structural awareness" (Quilici & Mayer, 1996, 2002) and previous work on paper and hypertext decision trees that aid statistic selection (Carlson et al., 2005; Koch & Gobell, 1999; Protsman & Carlson, 2008). Second, this was an experimental study in which we randomized individual students to the four levels of our independent variable. Third, we used Bayesian analytic techniques in recognition of the benefits they afford pragmatic researchers (Wagenmakers, Marsman, et al., 2018). However, we have also reported the equivalent null hypothesis significance tests for readers not yet familiar with our approach (see https://osf.io/ut75r/). It is important to note that both sets of analyses suggest very similar conclusions. Finally, when designing and running this study, we closely followed the recommendations provided by Roberts and Allen (2015) for navigating the complex ethical issues that can arise when conducting SoTL research.

In conclusion, this article reports an experimental evaluation of StatHand, which is a cross-platform application designed to help students identify suitable statistics for a wide range of research designs. Second-year psychology undergraduates were randomized to four different statistical decision making aids (StatHand, a familiar textbook, a familiar paper decision tree, or the textbook and decision tree combined) and asked to select appropriate statistics for five typical research scenarios. We found that the students assigned StatHand completed these tasks with higher accuracy than students assigned the other three aids. There was also some

evidence to suggest that the StatHand users experienced lower cognitive load, higher confidence in the appropriateness of the statistics they selected, and greater satisfaction with their assigned aid than at least one other user group. However, it took them longer to make their selections. Finally, our data provided strong evidential support for the hypothesis that StatHand is instructionally efficient relative to the other three aids. We believe that StatHand 'works' because its use promotes structural awareness. However, research to validate this claim is ongoing. In the interim, we hope that educators and students find StatHand to be a useful guide through what can sometimes seem like an overwhelming maze of statistical tests and techniques.

References

Allen, P. J., & Baughman, F. D. (2016). Active learning in research methods classes is associated with higher knowledge and confidence, though not evaluations or satisfaction. *Frontiers in Psychology, 7,* 279. http://dx.doi.org/10.3389/fpsyg.2016.00279

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: Australian Government Department of Education and Training.

Allen, P., Bennett, K., & Heritage, B. (2014). *SPSS Statistics version 22: A practical guide*. Melbourne, Australia: Cengage Learning.

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,* 188. http://dx.doi.org/10.3389/fpsyg.2016.00188

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,* 288. http://dx.doi.org/10.3389/fpsyg.2016.00288

American Psychological Association. (2016). Guidelines for the undergraduate psychology major: Version 2.0. *American Psychologist, 71,* 102–111. http://dx.doi.org/10.1037/a0037562

American Psychological Association Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist, 61,* 271–285. http://dx.doi.org/10.1037/0003-066X.61.4.271

Appleby, D. C. (2018). Preparing psychology majors to enter the workforce: Then, now, with whom, and how. *Teaching of Psychology, 45,* 14–23. http://dx.doi.org/10.1177/0098628317744944

Australian Psychology Accreditation Council. (2018). *Accreditation standards for psychology programs*. Melbourne, Australia: Author. Retrieved from https://www.psychologycouncil.org.au/sites/default/files/public/APAC_Accreditation_Standards_2018_Jan_Version_for_Online_Publishing_Single.pdf

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7,* 543–554. http://dx.doi.org/10.1177/1745691612459060

Blankenship, J., & Dansereau, D. F. (2000). The effect of animated node-link displays on information recall. *Journal of Experimental Education, 68,* 293–308. http://dx.doi.org/10.1080/00220970009600640

Bowen, K., & Pistilli, M. D. (2012). *Student preferences for mobile app usage*. Louisville, CO: Educause Center for Analysis and Research. Retrieved from https://library.educause.edu/resources/2012/9/student-preferences-for-mobile-app-usage

British Psychological Society. (2017). *Standards for the accreditation of undergraduate, conversion and integrated Masters programmes in psychology*. Leicester, United Kingdom: Author. Retrieved from https://www.bps.org.uk/sites/bps.org.uk/files/Accreditation/Undergraduate%20Accreditation%20Handbook%20(2017).pdf

Brooks, D. C., & Pomerantz, J. (2017). *ECAR study of undergraduate students and information technology, 2017*. Louisville, CO: Educause Center for Analysis and Research. Retrieved from https://library.educause.edu/resources/2017/10/ecar-study-of-undergraduate-students-and-information-technology-2017

Carlson, M., Protsman, L., & Tomaka, J. (2005). Graphic organizers can facilitate selection of statistical tests: Part 1–Analysis of group differences. *Journal, Physical Therapy Education, 19,* 57–65. http://dx.doi.org/10.1097/00001416-200507000-00008

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

Cranney, J., Morris, S., & Botwood, L. (2015). Psychological literacy in undergraduate psychology education. In D. S. Dunn (Ed.), *Oxford handbook of undergraduate psychology education* (pp. 863–872). New York, NY: Oxford University Press.

Diliberto-Macaluso, K., & Hughes, A. (2016). The use of mobile apps to enhance student learning in introduction to psychology. *Teaching of Psychology, 43,* 48–52. http://dx.doi.org/10.1177/0098628315620880

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power anal-

ysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191. http://dx.doi.org/10.3758/BF03193146

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics* (5th ed.). London, United Kingdom: SAGE.

Gardner, P. L., & Hudson, I. (1999). University students' ability to apply statistical procedures. *Journal of Statistics Education*. Advance online publication. Retrieved from https://ww2.amstat.org/publications/jse/secure/v7n1/gardner.cfm

Halonen, J. S., & Dunn, D. S. (2018). Embedding career issues in advanced psychology major courses. *Teaching of Psychology, 45,* 41–49. http://dx.doi.org/10.1177/0098628317744967

Hattie, J. (2015). The applicability of visible learning to higher education. *Scholarship of Teaching and Learning in Psychology, 1,* 79–91. http://dx.doi.org/10.1037/stl0000021

Hoffman, B., & Schraw, G. (2010). Conceptions of efficiency: Applications in learning and problem solving. *Educational Psychologist, 45,* 1–14. http://dx.doi.org/10.1080/00461520903213618

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education, 4,* 1304016. http://dx.doi.org/10.1080/2331186X.2017.1304016

Jackson, M. (n.d.). *Statistical test flowchart*. Retrieved from http://www.statsflowchart.co.uk/

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. http://dx.doi.org/10.3758/BF03207686

Kretser, H. E., Wong, R., Roberton, S., Pershyn, C., Huang, J. M., Sun, F. P., . . . Zahler, P. (2015). Mobile decision-tree tool technology as a means to detect wildlife crimes and build enforcement networks. *Biological Conservation, 189,* 33–38. http://dx.doi.org/10.1016/j.biocon.2014.08.018

Ling, C., Harnish, D., & Shehab, R. (2014). Educational apps: Using mobile applications to enhance student learning of statistical concepts. *Human Factors and Ergonomics in Manufacturing & Service Industries, 24,* 532–543. http://dx.doi.org/10.1002/hfm.20550

Lund Research. (2018). *Statistical test selector*. Retrieved from https://statistics.laerd.com/premium/sts/index.php

Mitchell, M., Leachman, M., & Masterson, K. (2017). *A lost decade in higher education funding: State cuts have driven up tuition and reduced quality*. Washington, DC: Center on Budget and Policy Priorities. Retrieved from https://www.cbpp.org/research/state-budget-and-tax/a-lost-decade-in-higher-education-funding

Mock, T. J. (1972). A decision tree approach to the methodological decision process. *The Accounting Review, 47,* 826–829.

Norcross, J. C., Hailstorks, R., Aiken, L. S., Pfund, R. A., Stamm, K. E., & Christidis, P. (2016). Undergraduate study in psychology: Curriculum and assessment. *American Psychologist, 71,* 89–101. http://dx.doi.org/10.1037/a0040095

Paas, F. G. W. C., & Van Merriënboer, J. J. G. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance-measures. *Human Factors, 35,* 737–743. http://dx.doi.org/10.1177/001872089303500412

Protsman, L., & Carlson, M. (2008). Graphic organizers can facilitate selection of statistical tests: Pt. 2–Correlation and regression analysis. *Journal, Physical Therapy Education, 22,* 36–41. http://dx.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,* 144–161. http://dx.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,* 325–342. http://dx.doi.org/10.1002/acp.796

Roberts, L. D., & Allen, P. J. (2015). Exploring ethical issues associated with using online surveys in educational research. *Educational Research and Evaluation, 21,* 95–108. http://dx.doi.org/10.1080/13803611.2015.1024421

Schau, C., & Mattern, N. (1997). Use of map techniques in teaching applied statistics courses. *The American Statistician, 51,* 171–175.

Schönbrodt, F. D. (2017). *BFDA: An R package for Bayes factor design analysis* (version 0.2). Retrieved from https://github.com/nicebread/BFDA

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review, 25,* 128–142. http://dx.doi.org/10.3758/s13423-017-1230-y

Sizemore, O. J., & Lewandowski, G. W., Jr. (2009). Learning might not equal liking: Research methods course changes knowledge but not attitudes. *Teaching of Psychology, 36,* 90–95. http://dx.doi.org/10.1080/00986280902739727

Stowell, J. R. (2015). Using technology effectively in the psychology classroom. In D. S. Dunn (Ed.), *Oxford handbook of undergraduate psychology education* (pp. 265–274). New York, NY: Oxford University Press.

Strahan, S., & Credé, M. (2015). Satisfaction with college: Re-examining its structure and its relationships with the intent to remain in college and academic performance. *Journal of College Student*

Retention, 16, 537–561. http://dx.doi.org/10.2190/CS.16.4.d

Sung, Y.-T., Chang, K.-E., & Liu, T.-C. (2016). The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. *Computers & Education, 94,* 252–275. http://dx.doi.org/10.1016/j.compedu.2015.11.008

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *Management Information Systems Quarterly, 27,* 425–478. http://dx.doi.org/10.2307/30036540

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. http://dx.doi.org/10.3758/s13423-017-1323-7

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., . . . Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25,* 35–57. http://dx.doi.org/10.3758/s13423-017-1343-3

Waples, J. A. (2016). Building emotional rapport with students in statistics courses. *Scholarship of Teaching and Learning in Psychology, 2,* 285–293. http://dx.doi.org/10.1037/stl0000071

Ware, M. E., & Chastain, J. D. (1989). Computer-assisted statistical-analysis: A teaching innovation? *Teaching of Psychology, 16,* 222–227. http://dx.doi.org/10.1207/s15328023top1604_16

Ware, M. E., & Chastain, J. D. (1991). Developing selection skills in introductory statistics. *Teaching of Psychology, 18,* 219–222. http://dx.doi.org/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. http://dx.doi.org/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. http://dx.doi.org/10.1080/00220973.2012.745466

Received August 7, 2018
Revision received November 7, 2018
Accepted November 9, 2018 ■