[Chapter number]

Using StatHand to Improve Students' Statistic Selection Skills

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

Abstract

Psychology undergraduates find identifying appropriate analyses for common research designs difficult. Resources have been developed to aid this process, including decision trees commonly included in statistics textbooks. The use of such trees is supported by research demonstrating their efficacy and popularity. In recent years, decision trees to aid statistic selection have been adapted for digital media. One such adaptation is StatHand, a free iOS and web app that that aids statistic selection by prompting users to focus systematically on each structural feature of their research design. Previous research has suggested that simply providing students with an app like StatHand is not enough to promote accurate statistic selection. Rather, students need to be trained in its use. In this chapter we describe a brief statistic selection training activity built around the use of StatHand. The development of the activity was informed by two sets of literature. The first suggests that accurate statistic selection is a consequence of 'structural awareness'. The second pertains to the success of 'wise' psychological interventions across a range of contexts, including education. The students we have trained using our methods (N = 50) demonstrated substantially greater statistic selection proficiency than untrained students in previous research. Our training methods can be adapted for a range of contexts. The chapter appendices include our training materials and over 40 research scenarios spanning the range of analyses covered in StatHand. These can be freely adapted by instructors for both formative and summative learning activities.

Statistic Selection Skills

One of the five learning goals for an undergraduate psychology degree specified by the Society for the Teaching of Psychology's Statistical Literacy Taskforce (2014, p. 2) is the ability to "apply appropriate statistical strategies to test hypotheses". In meeting this goal, students should be able to "select … an appropriate statistical analysis for a given research design, problem, or hypothesis". For most, this is

a difficult task. Research indicates that psychology students are not good at recalling, recognizing or explaining how they would select appropriate statistical analyses for common research scenarios.

To illustrate the recall deficit, Gardner and Hudson (1999) gave students a series of typical research scenarios and asked them to specify an appropriate statistical analysis for each. Although the students were in third-year or above, they were unable to name an appropriate analysis for most scenarios. Indeed, even the highest performing student had an accuracy level of just 56%. To illustrate the recall deficit, in a multiple-choice selection task that Ware and Chastain (1989) administered at the end of a first-year psychology statistics unit, students averaged less than 45%. This was despite the researchers' (and their colleagues') initial beliefs that the task was easy enough for a typical first-year student to complete successfully. Finally, to illustrate the explanation deficit, Allen, Dorozenko, and Roberts (2016) asked undergraduate psychology students to describe how they would select appropriate statistical analyses for research scenarios similar to those developed by Gardner and Hudson (1999) and Ware and Chastain (1989). These students, who had each completed an average of three research methods courses, described haphazard and inefficient selection strategies that were unlikely to reliably lead them to appropriate analyses.

Recognition of these deficits has led educators to develop resources to facilitate the statistic selection process. Foremost amongst these resources are decision trees, which are routinely included in statistics textbooks (e.g., Allen, Bennett, & Heritage, 2019; Nolan & Heinzen, 2017). The proliferation of such trees is supported by research demonstrating their objective efficacy and subjective appeal (Carlson, Protsman, & Tomaka, 2005; Protsman & Carlson, 2008). However, despite their popularity, traditional paper-based decision trees are not without limitations. The most obvious of these limitations is brevity. They typically need to fit onto a single sheet of paper, meaning that information that would aid their navigation (definitions, examples etc.) is either spatially separated from the tree, or entirely absent.

To overcome such limitations, decision trees to aid statistic selection have been adapted for digital media (e.g., Koch & Gobell, 1999). One recent adaptation is StatHand (Allen et al., 2016, 2017). StatHand is a free iOS (available via the iOS App Store) and web app (see https://stathand.net) that aids the process of selecting appropriate statistical analyses for a wide range of circumstances. It achieves this by asking a series of questions that prompt users to focus systematically on each structural feature of their research design. The questions are annotated with relevant definitions and examples, such that relative novices can navigate the app without needing to consult additional sources. In answering these questions, the user progressively homes in on a statistical analysis

appropriate to their research design. The StatHand web app running on a mobile web-browser is illustrated in Figure 1, and the iOS version running on an iPad is illustrated in Figure 2.

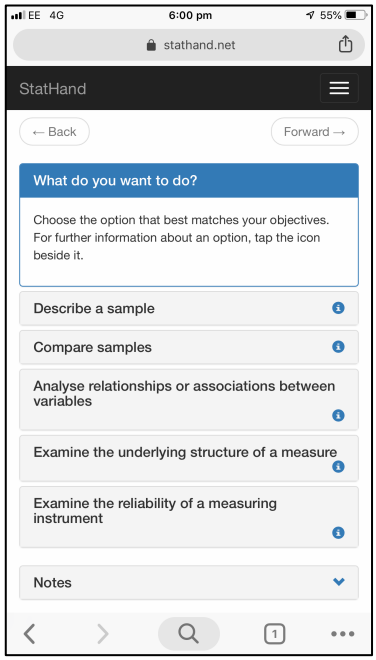


Figure 1. StatHand home screen on the Chrome mobile web-browser.

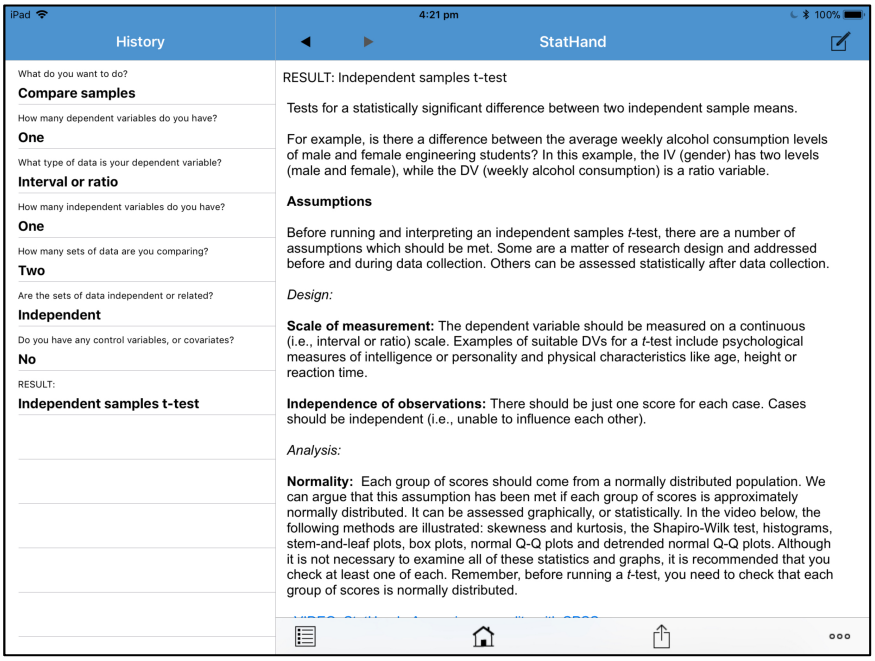


Figure 2. StatHand on an iPad. The sequence of decisions leading to an independent samples *t*-test is displayed in the History tool on the left of the screen.

In a recent experimental evaluation 217 psychology students were randomized to four different decision-making aids (StatHand, a familiar textbook, a familiar paper decision tree, or the textbook and decision tree combined) and asked to identify appropriate statistical analyses for five research scenarios (Allen, Finlay, Roberts, & Baughman, 2019). The students in the StatHand condition significantly outperformed students in the other three conditions (d = .55 to .69). However, in an absolute sense, their performance was underwhelming. On average, they identified appropriate analyses for just 1.74 (SD = 1.19) of the five scenarios. On a typical university marking scale this would be a firm 'fail'. This suggests that simply providing students with a tool like StatHand may not be enough to promote accurate statistic selection. Rather, students need to be trained in its use.

The next section of this chapter describes a brief statistic selection training activity built around the use of StatHand. The development of the activity was informed by two sets of literature. The first suggests that accurate statistic selection is a consequence of 'structural awareness', which can itself be trained (e.g., Yan & Lavigne, 2014). Structural awareness refers to the ability to see past the topic

area of a research problem, and focus on its structural characteristics (e.g., the number and nature of the independent and dependent variables) and the relationships between them (Quilici & Mayer, 2002). The second pertains to the success of 'wise' psychological interventions across a range of contexts, including education (Walton, 2014). Wise interventions are brief and targeted. They are designed to change specific behaviors (in both the short and longer term) by exploiting specific psychological processes. In this instance, those processes are meta-cognition and structural awareness.

The Training Activity

According to the wise framework proposed by Walton (2014), there are three elements to consider when developing a brief psychological intervention: (1) a clear and specific underlying theory/concept; (2) the recursive process that is being targeted or broken; and (3) context. Research has shown that students will enlist various cognitive and meta-cognitive strategies when learning new skills/processes (Somuncuoglu & Yildirim, 1999). However, it has become increasingly clear that when learning statistics, students are driven by a fear of failure and take a more tactical surface approach to gaining good grades (e.g., Asikainen & Gijbels, 2017; Diseth & Martinsen, 2003; Newble & Entwhistle, 1986). This contrasts with using deeper learning approaches which are considered to promote deeper semantic understanding and a lasting learning experience (e.g., Bilgin & Crowe, 2008). Meta-cognition is closely tied to deeper learning as it goes beyond simple cognition (*how we think*) and is characterized by our *awareness of how we think*. This deeper appreciation allows not only semantic understanding of concepts, but synthesis of ideas and the ability to adapt and apply this understanding in different situations (e.g. see Bayat & Tarmizi, 2010).

Our training focuses specifically on developing students' structural awareness. That is, their ability to identify and focus on the structural (e.g., IV, DV etc.) characteristics of research designs, rather than the research topic or other surface level features. We designed the intervention to break students' habits of utilizing surface level learning strategies and to encourage them to be more structurally aware, thus facilitating deep learning. The final element of a wise intervention is that it needs to be context specific – it needs to be important to the individual for it to be assimilated and have a lasting impact. Embedding this training within the curriculum automatically makes it inherently important to students. However, if not embedded within the curriculum, but perhaps used as a supplementary learning activity, the training has been designed to provide students with feedback on their progress. This ensures that they see the benefits of using a more structurally aware approach to statistic selection, which encourages this self-motivation and self-reflection (see chapter XX on reflection).

There are two phases to the training with students; a teacher-guided training phase and then a self-guided practice phase. The teacher-guided training and self-guided practice each take around 20-25 minutes to complete. Depending on how these methods are adapted, we envisage they would be suitable for a standalone lecture or class lasting no longer than one hour.

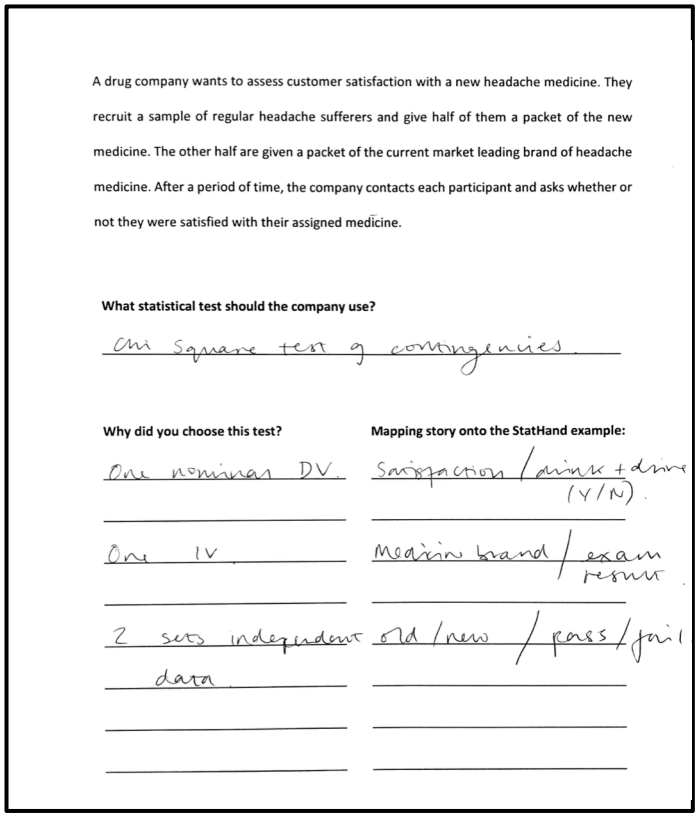
Teacher-Guided Training Phase

The teacher-guided training begins with presenting one of the four research scenarios in Appendix A to students and asking if they know how the resultant data might be analyzed. Our experience (as well as previous research; Allen, Dorozenko, & Roberts, 2016) suggests that most undergraduate students will either indicate that they do not know or will guess the statistical test that they have used most recently or most frequently. This provides an opportunity to explain that there is a systematic process that can be followed to work out how a set of data could be analyzed, and that there are many resources (e.g., flow diagrams, websites etc.) that will 'step you through' this process. (On the occasions students do identify an appropriate analysis, these resources can be pitched as tools that can be used to 'double check' or confirm their thinking.)

At this point, we introduce students to the StatHand app and demonstrate how it can be used to identify an appropriate analysis for the study in the scenario. This process involves highlighting various features of the study's design (e.g., number and nature of dependent and independent variables) and outlining how they correspond to the options and examples in StatHand. Once an appropriate statistical analysis has been identified, we use the StatHand History tool (see Figure 2) to reiterate *why* it was selected (e.g., because we had **one interval or ratio** level dependent variable and **one** independent variable with **two independent** levels). In doing so, we are highlighting the structural characteristics of the research scenario. The ability to identify such characteristics is key to the notion of structural awareness. Asking *why* also encourages students to reflect on the reasons behind particular choices, which is an important meta-cognitive skill and promotes the importance of enlisting a deeper approach to understanding as an effective decision-making process (Walton, 2014, see also chapter XX on reflection). This can also be a good time to illustrate how changing one structural feature (e.g., from **independent** to **related** samples) changes the analysis (in Figure 2, from an independent-samples to a paired-samples *t*-test). Finally, we demonstrate how our research scenario is structurally equivalent to the example in StatHand, even though their topics are quite different. We call this a 'mapping exercise'. For example, the ratio level dependent variable in the first scenario in Appendix A is *words recalled*. In the example in Figure 2, it is *drinks consumed*. Again, the ability to see

past the surface/topic characteristics of a research scenario and focus on its structural characteristics is key to structural awareness.

Students are encouraged to work more independently on the remaining three scenarios in Appendix A, receiving reduced instruction and feedback as they progress (i.e., we will talk through earlier scenarios but only confirm answers for later ones). This approach creates a safe learning environment where students are encouraged to think and work independently but are not afraid to make mistakes in the learning process. We use a handout to guide this training (see Figure 3), although PowerPoint should be similarly effective.


A drug company wants to assess customer satisfaction with a new headache medicine. They recruit a sample of regular headache sufferers and give half of them a packet of the new medicine. The other half are given a packet of the current market leading brand of headache medicine. After a period of time, the company contacts each participant and asks whether or not they were satisfied with their assigned medicine.

What statistical test should the company use?

Chi Square test of contingencies

Why did you choose this test?

One nominal DV

One IV

2 sets independent data

Mapping story onto the StatHand example:

Satisfaction / drink + drive (Y/N)

Making brand / exam result

old / new / pass / fail

Figure 3. Handout page completed by a student.

Self-Guided Practice Phase

After the teacher-guided training phase of our activity, we give students another four scenarios (see Appendix B) formatted in a handout like the one illustrated in Figure 3. Students are instructed to work through this handout independently and told they won't receive feedback from the teacher on whether the statistics they identify are correct. This part of the wise intervention process encourages students to assimilate their learning from the training phase when there is no external motivation to find the correct answer. It is clear in wise interventions that the behavior change has to be salient for the individual to have a lasting and effective impact.

Note that we have deliberately restricted our activity to just four tests, which differ on just two structural characteristics (see Figure 4). This was because our intent was to highlight to students the importance of attending to the structural features of research designs, not to teach an extensive range of different tests and procedures. However, teachers may wish to swap the scenarios in Appendices A and B with any of the 41 scenarios in Appendix C, which cover the full range of analyses described in

StatHand. The scenarios in Appendix C can also be freely used or adapted by educators for a range of additional formative and summative learning activities.

		Dependent variable	
		Nominal	**Ratio**
Independent variable	**Independent**	Chi-square test of contingencies	Independent samples *t*-test
	Paired	McNemar test of change	Paired samples *t*-test

Figure 4. The tests used in our training activity.

Our Findings

We ran our training with *N* = 50 first-, second- and third-year psychology students, and coded their responses to each element of our self-guided practice handout. Answers were coded as correct, incorrect or absent, although incorrect and absent were merged for the inferential analyses.

Our students were able to correctly identify appropriate statistical tests for 81% (*SD* = 25%) of the scenarios in the practice handout. This was significantly and substantially higher than the 34.8% accuracy level achieved by the untrained students in the Allen et al. (2019) sample, $t(49) = 12.72$, $p < .001$, $d = 1.85$. Further analyses indicated that our trained students were significantly better able to identify appropriate statistics for some scenarios than they were for others, Cochran's Q (3, $N = 50$) = 15.48, $p = .001$. Specifically, they were significantly less likely to identify an appropriate statistic for the paired-samples *t*-test scenario (64%) than they were for the independent samples *t*-test (88%) and chi-square (88%) scenarios (Bonferroni corrected *p*s = .004). Identification accuracy for the McNemar test scenario (82%) was also higher than that for the paired-samples *t*-test scenario, though not significantly so (Bonferroni corrected $p = .065$).

When asked "why did you choose this test?", our students correctly identified an average of 85.4% of the relevant structural characteristics for each scenario. However, as illustrated in Table 1, performance levels for some scenarios and characteristics were higher than for others. In particular, students seemed less able to correctly identify the structural characteristics of the paired samples *t*-test scenario, with several of them confusing the number of levels of the IV (aka. the number of data sets) with the number of IVs, which led them to a factorial, rather than a one-way design. This may be an artefact of the nature of the scenario we used, which was perhaps less 'typical' than the scenarios used for the other three tests. However, this possibility requires further investigation.

Table 1. *Correctness of Students' Responses to the Question, "Why Did You Choose This Test", Split by Scenario and Structural Characteristic*

	Percentage of Correct Responses					*M*	Q	*p*
	No. DVs	DV Data Type	No. IVs	No. Data Sets	Design Type			
Independent samples *t*-test	$96^{a}{}_{a}$	$82^{b}{}_{ab}$	$94^{a}{}_{a}$	$90^{ab}{}_{a}$	$90^{ab}{}_{a}$	90.4	13.09	.011
Paired samples *t*-test	$88^{a}{}_{a}$	$68^{b}{}_{a}$	$76^{ab}{}_{b}$	$72^{b}{}_{b}$	$72^{b}{}_{b}$	75.2	14.80	.005
Chi-square test of contingencies	$94^{a}{}_{a}$	$86^{a}{}_{b}$	$94^{a}{}_{a}$	$92^{a}{}_{a}$	$86^{a}{}_{ab}$	90.4	10.50	.033
McNemar test of change	$92^{a}{}_{a}$	$86^{ab}{}_{b}$	$88^{ab}{}_{ab}$	$82^{ab}{}_{ab}$	$80^{b}{}_{ab}$	85.6	10.86	.028
M	92.5	80.5	88.0	84.0	82.0			
Q	3.18	10.77	14.09	14.88	10.22			
p	.364	.013	.003	.002	.017			

Note. Percentages on the same row with the same superscript letters do not differ significantly at a Bonferroni corrected α of .05. Percentages in the same column with the same subscript letters do not differ significantly at a Bonferroni corrected α of .05. These pairwise comparisons are McNemar tests of change. *M* = Mean percentage of correct responses for the relevant scenario or structural characteristic. Q = Cochran's Q for differences between percentages of correct responses across scenarios or structural characteristics. *p* = *p*-values associated with reported Cochran's Q values.

Finally, in the mapping exercise, students correctly matched 64.5% of the Appendix B scenarios' structural characteristics with their corresponding surface/topic characteristics and with the corresponding surface/topic characteristics of the StatHand examples. As illustrated in Figure 5, there was relatively little variation in average correctness across the four scenarios, $F(3, 147) = 1.90$, $p = .131$, partial $\eta^2 = .04$. The larger number of 'absent' responses for the paired-samples *t*-test scenario can be attributed to our coding. We

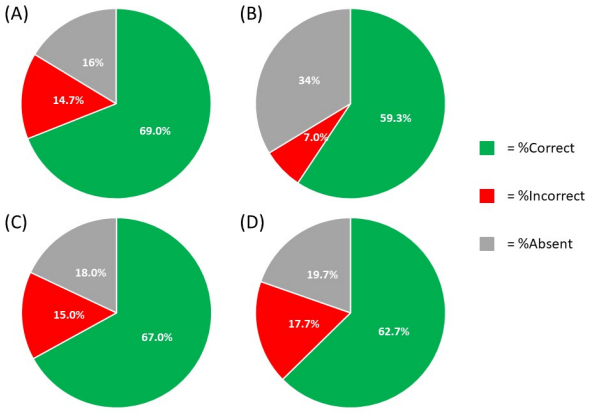


Figure 5. Percent of correct, incorrect and absent answers in the mapping exercise for the (A) independent samples *t*-test, (B) paired samples *t*-test, (C) chi-square test of contingencies and (D) McNemar test of change scenarios.

automatically converted relevant responses to absent in instances where an incorrect test had been identified.

Final Remarks

This chapter describes an activity we developed to help students use StatHand to select appropriate statistical analyses for different research designs. The data we collected from *N* = 50 undergraduate psychology students suggested that this activity is effective. Specifically, compared to the untrained students in the Allen et al. (2019) sample, our students were substantially better able to identify appropriate statistics for four common research designs. They were also proficient at declaring the structural characteristics that led them to the selection of particular analyses. Finally, the majority were capable of correctly matching the structural characteristics of different research scenarios with their corresponding surface/topic characteristics, as well as the surface/topic characteristics of the examples in StatHand. However, there was more room for improvement on this task.

We trained our students individually, as they were participants in a larger randomized controlled trial (RCT). However, we see no reason why this activity can't be adapted for use in small group teaching and/or lectures. In such contexts, students could initially work individually, and then pair up to compare and discuss their answers prior to a class-level discussion. Furthermore, there are no compelling reasons why educators need to print handouts (PowerPoint should be equally effective, and far more environmentally conscious), restrict their lessons to just four statistical analyses (there are scenarios reflecting 10 times this number in Appendix C, including all the analyses in the Passion-Driven Statistics curriculum, see chapter XX) or even deliver these lessons face-to-face (guided online tutorials, see chapter XX, or homework activities with pre-programmed feedback for use in a flipped classroom context, see chapter XX, could be developed relatively easily). The important point is that students are prompted to focus systematically on the structural characteristics of the research designs they are exposed to, and the implications of these characteristics for selecting appropriate statistics.

Beyond their use in undergraduate research methods and statistics classes, StatHand and the accompanying training materials may be useful in a range of contexts where people need to consider (or double-check) how quantitative data could or should be analyzed. For example, given research demonstrating that even higher-level students have limited statistic selection skills and confidence (e.g., Allen, Dorozenko, & Roberts, 2016; Gardner & Hudson, 1999), students undertaking capstone or dissertation projects may find the resources we have developed to be useful. So might students undertaking research internships, UREs (undergraduate research experiences) or service-learning

projects (see chapter XX), or students who have moved into psychology as mature learners and/or from non-science disciplines (e.g., students on MSc conversion courses in the UK). Finally, our experience suggests that some colleagues, particularly those with limited statistical training, may also benefit from StatHand and the opportunity to practice using it.

Although the findings reported herein are pleasing, they raise a number of new questions requiring investigation. First, how would students perform on such tasks without the aid of the application? Though there are relatively few circumstances where one would need to rely purely on memory to complete a statistic selection task, such circumstances do exist (e.g., when 'put on the spot' by a supervisor). How much training would be required to bring students to the level of 'expert', where they could reliably (a) identify the all the relevant structural features of a research design, (b) construct a conceptual model in which the relationships between structural features are represented, and (c) integrate that model with existing knowledge to select an appropriate statistic? Second, the individual training we have run and evaluated represents a 'proof of concept'. However, as an actual method of teaching students it is very inefficient. We have described how it could be adapted for a classroom or online context and have no reason to believe that such adaptations would be less effective than individual training. However, this claim should be tested empirically. Furthermore, efforts should be made to determine a minimally effective training dose. Ideally, these efforts will be experimental, such that causal links between training and student learning/performance can be established. Finally, research is needed to investigate the extent to which training of the nature described herein actually impacts on performance on tasks believed to be reflective of structural awareness. Several such tasks have been proposed in previous research, including triad judgement tasks (Rabinowitz & Hogan, 2008), explanation tasks (Yan & Lavigne, 2014) and scenario generation tasks (Quilici & Mayer, 2002). Indeed, this is the primary focus of the RCT that we alluded to earlier.

References

Allen, P. J., Baughman, F. D., Roberts, L. D., van Rooy, D., Rock, A. J., & Loxton, N. J. (2017). *StatHand: An interactive decision tree mobile application to guide students' statistical decision making*. Canberra, Australia: Australian Government Department of Education and Training.

Allen, P., Bennett, K., & Heritage, B. (2019). *SPSS Statistics: A practical guide* (4th ed.). Melbourne, Australia: Cengage.

Allen, P. J., Dorozenko, K. P. & Roberts, L. D. (2016). Difficult decisions: A qualitative exploration of the statistical decision making process from the perspectives of psychology students and academics. *Frontiers in Psychology, 7,* Article 188. doi:10.3389/fpsyg.2016.00188

Allen, P. J., Finlay, J., Roberts, L. D., & Baughman, F. D. (2019). An experimental evaluation of StatHand: A free application to guide students' statistical decision making. *Scholarship of Teaching and Learning in Psychology, 5,* 23-36. doi:10.1037/stl0000132

Allen, P. J., Roberts, L. D., Baughman, F. D., Loxton, N. J., van Rooy, D., Rock, A. J., & Finlay, J. (2016). Introducing StatHand: A cross-platform mobile application to support students' statistical decision making. *Frontiers in Psychology, 7,* Article 288. doi:10.3389/fpsyg.2016.00288

Asikainen, H., & Gijbels, D. (2017). Do students develop towards more deep approaches to learning during studies? A systematic review on the development of students' deep and surface approaches to learning in higher education. *Educational Psychology Review, 29,* 205. doi:10.1007/s10648-017-9406-6

Bayat, S., & Tarmizi, R. A. (2010). Assessing cognitive and metacognitive strategies during algebra problem solving among university students. *Procedia – Social and Behavioral Sciences, 8,* 403-410. doi:10.1016/j.sbspro.2010.12.056

Bilgin, A., & Crowe, S. (2008). Approaches to learning in statistics. *Asian Social Science, 4,* 36-42. doi:10.5539/ass.v4n3p36

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

Diseth, A., & Martinsen, O. (2003). Approaches to learning, cognitive style, and motives as predictors of academic achievement. *Journal of Educational Psychology*, 23, 195-207. doi:10.1080/01443410303225

Gardner, P. L., & Hudson, I. (1999). University students' ability to apply statistical procedures. *Journal of Statistics Education*, 7(1). Retrieved from https://ww2.amstat.org/publications/jse/secure/v7n1/gardner.cfm

Koch, C., & Gobell, J. (1999). A hypertext-based tutorial with links to the web for teaching statistics and research methods. *Behavior Research Methods, Instruments, & Computers*, 31, 7-13. doi:10.3758/bf03207686

Newble, D. I., & Entwistle, N. J. (1986). Learning styles and approaches: Implications for medical education. *Medical Education*, 20, 162-175. doi:10.1111/j.1365-2923.1986.tb01163.x

Nolan, S. A., & Heinzen, T. E. (2017). *Statistics for the behavioral sciences* (4th ed.). New York, NY: Worth.

Protsman, L., & Carlson, M. (2008). Graphic organizers can facilitate selection of statistical tests: Part 2 - Correlation and regression analysis. *Journal of Physical Therapy Education*, 22, 36-41. doi:10.1097/00001416-200807000-00006

Quilici, J. L., & Mayer, R. E. (2002). Teaching students to recognize structural similarities between statistics word problems. *Applied Cognitive Psychology*, 16, 325-342. doi:10.1002/acp.796

Rabinowitz, M. & Hogan, T. M. (2008). Experience and problem representation in statistics. *American Journal of Psychology*, 121, 395–407. doi:10.2307/20445474

Society for the Teaching of Psychology Statistical Literacy Task Force. (2014). Statistical literacy in the undergraduate psychology curriculum. Retrieved from https://teachpsych.org/Resources/Documents/otrp/resources/statistics/STP_Statistical%20Literacy_Psychology%20Major%20Learning%20Goals_4-2014.pdf

Somuncuoglu, Y., & Yildirim, A. (1999). The relationship between achievement goal orientations and the use of learning strategies. *The Journal of Educational Research*, 92, 267-277. doi:10.1080/00220679909597606

Walton, G. M. (2014). The new science of wise psychological interventions. *Current Directions in Psychological Science*, 23, 73-82. doi:10.1177/0963721413512856

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

Yan, J., & Lavigne, N. C. (2014). Promoting college students' problem understanding using schema-emphasizing worked examples. *Journal of Experimental Education, 82*, 74-102. doi:10.1080/00220973.2012.745466

Appendix A: Scenarios Used in Teacher-Guided Training

A lecturer wants to know if listening to classical music when studying improves memory. She recruits a sample of first year students and asks half to memorize a word list whilst listening to classical music. She asks the other half to memorize the same word list in silence. She then records how many words from the list each student is able to recall. *Answer: Independent samples t-test.*

An environmental scientist is interested in factors that influence public acceptance of recycled water. He recruits a sample of home owners and shows each two pictures of sinks full of water. In the first picture, the water is clear. In the second picture the water is a light brown color. Following each picture, each resident is asked whether or not they would support building a new water recycling plant that would produce water of the color illustrated. *Answer: McNemar test of change.*

A psychology student is studying the effects of auditory interference on reaction time. On each experimental trial participants quickly press a keyboard spacebar in response to a flash of light. For half of the trials, spoken word poetry is played in the background. The remaining trials are completed in silence. Each participants' average reaction time for both the auditory interference and silent trials is recorded. *Answer: Paired samples t-test.*

A drug company wants to assess customer satisfaction with a new headache medicine. They recruit a sample of regular headache sufferers and give half of them a packet of the new medicine. The other half are given a packet of the current market leading brand of headache medicine. After a period of time, the company contacts each participant and asks whether or not they were satisfied with their assigned medicine. *Answer: Chi-square test of contingencies.*

Appendix B: Scenarios Used in Self-Guided Practice

A researcher wants to know if imagining being in the presence of others influences charitable behavior. She asks half of her participants to imagine that they are alone and the other half to imagine that they are in a busy café. She then presents each participant with information about an animal welfare charity and asks them how much they would be willing to donate. *Answer: Independent samples t-test.*

An introductory psychology lecturer is interested in understanding whether his students' perceptions of the scientific status of psychology are influenced by the topic they have most recently studied. Following a lecture on Freud's theories he asks everyone in the class to indicate whether or not they consider psychology to be a scientific discipline. Following a lecture on psychopharmacology he asks them all the same question. *Answer: McNemar test of change.*

A head of department wants to know if Statistics 100 grades differ from Psychology 100 grades. All students in the department take both of these units at the same time. The head of department records each student's grades for each of the units. *Answer: Paired samples t-test.*

An occupational psychologist wants to understand the effects of including the word 'please' in an email request. She composes two versions of an email asking for some brief information about daily work habits. The word 'please' is included in the first version, but not in the second. She sends out the first version to half of the people in her contact list, and the second version everyone else. She records whether or not each contact provides the information she requested within one week. *Answer: Chi-square test of contingencies.*

Appendix C: Additional Scenarios

This Appendix contains 41 research scenarios which map onto the statistics, tests and procedures covered in StatHand (see Table C1). They are organized according to the five broad data analysis objectives that are presented to users when first opening the application (or visiting https://stathand.net). For many of the scenarios, there is no definitive 'correct' technique for analyzing the data they would likely generate. Having said that, most tend to suggest one obvious technique, whilst opening up the possibility for alternatives on further consideration. Consequently, they might best be used as discussion starters rather than, for example, multiple choice items for which there can only be a single 'correct' answer.

Table C1. *The Statistics, Tests and Procedures Described in StatHand, Grouped by Data Analysis Objective. The Objectives Listed Correspond with the Five Options Presented to Users on the StatHand Home Screen*

Objective	**Statistics, Tests and Procedures Described in StatHand**
Describe a sample	Bar graph; category count; histogram; interquartile range; Mean; median; mode; pie chart; range; standard deviation; stem-and-leaf plot.
Compare samples	ANCOVA (independent samples and mixed; one way and factorial); ANOVA (independent samples, repeated measures and mixed; one way and factorial); chi-square (goodness of fit and contingencies); Cochran's Q test; Friedman two-way ANOVA; Kruskal-Wallis one-way ANOVA; Mann-Whitney U test; McNemar test of change; *t*-test (one sample, independent samples and paired samples); Wilcoxon signed-rank test (one sample and paired samples).
Analyze relationships or associations between variables	Chi-square test of contingencies (with Phi or Cramer's V); correlation coefficients (point-biserial, rank-biserial, Spearman's and Pearson's); eta; linear regression (bivariate and multiple; standard and hierarchical); logistic regression (binary and multinomial; standard and hierarchical); ordinal regression (standard and hierarchical).
Examine the underlying structure of a measure	Confirmatory factor analysis; exploratory factor analysis; principal components analysis.
Examine the reliability of a measuring instrument	Cohen's kappa; Cronbach's alpha; intraclass correlation; Kuder-Richardson 20; Weighted kappa.

Note. Adapted from "Introducing StatHand: A cross-platform mobile application to support students' statistical decision making," by P. J. Allen, L. D. Roberts, F. D. Baughman, N. J. Loxton, D. Van Rooy, A. J. Rock, and J. Finlay, 2016, *Frontiers in Psychology, 7,* Article 288, p. 6.

Describe a sample

Scenario 1: Anwar and Sally have just finished collecting data for a large study examining the methods that people use to find reliable information on the internet. Prior to reporting their results, they need to describe their sample. From each participant, they collected the following demographic information: age, marital status, highest level of education and annual income. Which measures of central tendency and dispersion should they report for each of these variables, and how should they be graphed?

*Marital status is a nominal variable. Consequently, its **mode** should be reported, along with the **number of categories** of marital status represented in the sample. The number of people endorsing each marital status category can be captured in **bar graph**. (Or, alternatively, the proportion of the sample endorsing each marital status category can be captured in a **pie chart**.)*

*Highest level of education is most likely an ordinal variable. Anwar and Sally can report the **median** highest level of education, along with either a **range** or **interquartile range**. A **bar graph** can be used to visualize the distribution of education in the sample.*

*Age and annual income are both continuous (ratio level) variables. Assuming they are normally distributed, a **mean** and **standard deviation** can be reported for each. If they are skewed, which seems particularly likely for income, a **median** can be reported as well. Graphically, the distributions of both age and income can be captured using **histograms** or **stem-and-leaf plots**.*

Compare samples

Scenario 2: The manager of a voluntary extra tuition program wants to know whether or not 'regular' attendees (i.e., students who go to 5 or more sessions per semester) achieve higher end-of-semester grades than non-attendees. However, he suspects that his study may be confounded by the fact that regular attendees also tend to be smarter students! Consequently, he wants to use IQ as a control variable in his analyses. What statistical analysis would you advise him to conduct?

*The manager has a between subjects (or independent groups) design with one ratio level dependent variable (end-of-semester grades) and one independent variable with two levels (regular attendee vs. non-attendee). He also has one covariate. Assuming he can meet the relevant assumptions, a **one-way analysis of covariance (ANCOVA)** can be used to analyze his data. Alternatively, hierarchical multiple regression could be used. Both should lead to the same conclusions.*

Scenario 3: A researcher is interested in whether female or male students who play Tetris, Call of Duty, or no computer games over 10 days have significantly different mental rotation speeds (measured in milliseconds). He randomly divides 60 participants into three groups: 10 female and 10 male participants were asked to play Tetris for 20 minutes a day, 10 female and 10 male participants were asked to play Call of Duty for 20 minutes a day, and 10 female and 10 male participants were instructed to play no video games over the 10 days. The researcher is interested in whether playing a video game influences participants' mental rotation speeds, and if so, which game is most effective/deleterious. The researcher further wants to know if participants' mental rotation speeds differ according to gender. Finally, the researcher believes that the participants average daily 'screen time' may also influence their mental rotation speeds and wants to control for this variable in his analyses. Which statistical analysis would you recommend to this researcher?

This researcher appears to have a between subjects (independent groups) design, with a ratio level dependent variable (mental rotation speed) and two independent variables (game and gender). He also has a ratio level covariate (screen time). Consequently, a ***factorial between groups ANCOVA*** *can be used to analyse his data.*

Scenario 4: A friend of yours is running an experiment which involves measuring the self-esteem of 40 children, and then randomizing them into two conditions. The children in the experimental condition are then praised after displaying good behavior, whereas the children in the control condition are not. After a period of time, the self-esteem of each child is measured for a second time. Your friend wants to know if any changes in self-esteem observed between the pre- and post-tests are influenced by the experimental manipulation and would like to include the children's ages (which she has also recorded) as a control variable in her study. What statistical analysis would you recommend?

This design most obviously lends itself to a ***mixed model ANCOVA****, as there is one repeated measures independent variable (time), one between subjects independent variable (praise vs control), a continuous covariate (age), and a (presumably) interval level dependent variable (self-esteem).*

Scenario 5: Jake believes that the type of music you listen to while studying may have an impact on test scores. Jake randomly assigns 60 students to listen to either, rock, country or classical music while they study a passage of text. After an hour of study, the students are given a test on the contents of the passage. What statistical analysis should Jake use to test his hypothesis?

In this scenario, there is one continuous (interval or ratio) level dependent variable (test scores), and one independent variable (music type) with three levels. It's a between subjects (independent groups) design,

as different students have been assigned to each type of music. Consequently, a ***one-way between subjects/groups ANOVA*** *can be used to analyze Jake's data. If significant, it can be followed by either planned comparisons or post-hoc tests. If Jake's dependent variable is considered ordinal and/or he can't meet the assumptions of the parametric ANOVA,* ***a Kruskal-Wallis ANOVA*** *can be used instead.*

Scenario 6: You work in an animal laboratory and have been asked to investigate whether rats can be 'bred' to perform well on a T-maze task. (This is a commonly used task requiring that a rat learn which features of the maze identify where food is located.) As a secondary consideration, you've been asked to look at whether performance is also influenced by the nature of the environment in which the rats were raised. You have access to a group of rats that have been selectively bred to perform exceptionally well on this task (the 'bright' rats), a group that have been selectively bred to perform poorly on the task (the 'dull' rats) and a group who were bred without regard for their maze performance (the 'control' rats). Furthermore, half of each group has been raised in an 'enriched' environment, whilst the other halves have been raised in an 'impoverished' environment. Performance on the T-maze is measured as the time that it takes for the rat to find the food, averaged over five trials. There is one trial every 48 hours, and testing begins when each rat is exactly 60-days old. Your objective is to find out if and how breeding and environment influence T-maze performance.

This appears to be a between subjects (independent groups) design, as there are different subjects in each of the six breeding/environment groups. There are two independent variables (breeding and environment), and the dependent variable (time) is measured on a ratio scale. Consequently, a ***factorial between groups ANOVA*** *can be used to analyse this data.*

Scenario 7: You are interested in whether a short (2-week) course of mindfulness therapy is effective in reducing parental stress levels for young mothers with postnatal depressive symptoms, and whether any improvements are evident at two-month, six-month and one-year follow-ups. At each of the four testing sessions, stress is measured using the 15 item "Parental Stress Scale". Participants complete this scale by responding to each item on a scale ranging from 1 (strongly disagree) to 5 (strongly agree). These responses are then summed to give a total stress score between 15 and 75 for each participant. Which statistical analysis would you use here?

If we assume that the dependent variable (stress scores) is at least interval level, a ***one-way repeated measures ANOVA*** *can be used to analyse this data. If it is decided that stress is an ordinal level variable and/or the assumptions of the parametric ANOVA are not met, a* ***Friedman two-way ANOVA*** *can be used instead.*

Scenario 8: The National Cycling Association encourages members to wear fluorescent vests at all times, as it believes that doing so makes them more visible on the roads and thus safer. However, they have not yet tested this belief. To begin to do so, they hire a scientist who programs a driving simulator to drop a virtual cyclist into a driving simulation at random intervals. When a cyclist appears, the 'driver' must respond, as quickly as possible, by pressing a button located on the simulator steering wheel. The simulator automatically records the time (in milliseconds) that it takes the driver to react to the presence of the cyclist. In a complete testing session, a driver is required to respond to 30 cyclists: 15 wearing fluorescent vests, and 15 in normal, non-fluorescent clothing (the 'control' cyclists). Furthermore, five fluorescent cyclists and five control cyclists are dropped into the simulation during 'daylight hours'; five of each type of cyclist are dropped into the simulation at 'dusk'; and the remaining five of each type of cyclist are dropped into the simulation at 'night'. The 30 trials are fully randomized, and reaction times are averaged across each type of trial (e.g., fluorescent at dusk; control at dusk etc.) for each participant. How might the data collected in this experiment be analyzed?

In this experiment, there is a ratio level dependent variable (reaction time) and two repeated measures independent variables (cyclist type and time of day). Consequently, it can be analyzed using a ***factorial repeated measures ANOVA****.*

Scenario 9: A friend of yours is running an experiment which involves measuring the self-esteem of 40 children, and then randomizing them into two conditions. The children in the experimental condition are then praised after displaying good behavior, whereas the children in the control condition are not. After a period of time, the self-esteem of each child is measured for a second time. Your friend wants to know if any changes in self-esteem observed between the pre- and post-tests are influenced by the experimental manipulation. What statistical analysis would you recommend?

This design most obviously lends itself to a ***mixed model ANOVA****, as there is one repeated measures independent variable (time), one between subjects independent variable (praise vs control) and a (presumably) interval level dependent variable (self-esteem).*

Scenario 10: A vending machine manufacturer wants to know which brands of cola students prefer. He sets up a machine in a busy university hallway, and stocks it with an equal amount of three cola brands: Coca-Cola, Pepsi and Royal Crown. He then returns 48 hours later and counts up the number of cans of each cola that have been sold. Which statistical test should the manufacturer use to determine whether or not the students prefer some brands more than others?

In this scenario, the manufacturer is seeking to compare observed category membership frequencies (i.e., the number of cans of each type of cola sold) with a set of expected category membership frequencies (i.e., the number of cans of each type of cola that would be sold if students made their selections randomly, and expressed no true preferences). This can be achieved with a ***chi-square test for goodness of fit****.*

Scenario 11: You are interested in whether Star Trek fans or Star Wars fans are more likely to be married or not married. Which statistical analysis would you use?

Here, we wish to compare samples. There is one nominal 'dependent variable' (marital status), and one nominal 'independent variable' (series preference) with two levels. The obvious analytic technique here is the ***chi-square test of contingencies****, which can be used to determine whether or not marital status is contingent on (or related to) the preference for Star Trek vs. Star Wars.*

Scenario 12: The Dean of Psychology at City University suspects that the 'fail rate' for Statistics 100 is higher than that for the other two compulsory first semester psychology classes (Behavioral Science 100 and Human Biology 100). There were 200 students enrolled in these three classes last semester. The Dean has a record of whether each student passed (coded as '1') or failed (coded as '0') each class. How should she test her hunch?

This is a repeated measures design, with a dichotomous dependent variable (passed vs. failed). There are three levels of the independent variable (Statistics 100, Behavioral Science 100 and Human Biology 100). A ***Cochran's Q test*** *could be used by the Dean to determine whether the proportion of fail grades differs significantly across the three classes. If the Cochran's Q test was statistically significant, it should be followed up with a series of McNemar tests of change.*

Scenario 13: The City Council are facing financial difficulties and need to make some cuts to the services they provide to residents. They have come up with four possibilities, each of which will result in approximately the same reduction in expenditure: (a) halve the amount of money spent on maintaining public recreation spaces; (b) close the Emergency Room at one of the City's three public hospitals; (c) sell 40% of City Park to private developers; or (d) reduce the number of police patrolling the streets by 20%. Because they're facing an election in 12-months, the Council are keen to pursue the option likely to cause the least amount of outrage amongst their constituency. Consequently, they have asked a sample of residents to rank the four possibilities outlined above from 1 = 'most preferred option' to 4 = 'least preferred option'. They now need to analyze this data. What would you recommend?

This is a repeated measures (within subjects) design with one ordinal dependent variable (ranked preference) and one independent variable with four levels (the four cost reduction options). It lends itself most obviously to a ***Friedman two-way ANOVA****. If statistically significant, it should be followed with a series of Wilcoxon signed rank tests to identify which pairs of options differ significantly.*

Scenario 14: Your friend works in an animal laboratory and has been asked to find out which of three nutritional supplements produce optimal cognitive performance in rats. She sets up an experiment in which 40 rats are randomized to four groups. Members of one group are given a placebo, whilst the remaining groups are given supplements A, B and C. She then places all 40 rats in a maze, and gives each a score between 1(st) and 40(th) based on the order in which they successfully complete it (i.e., the first rat across the maze 'finish line' is given a score of 1, the second across gets a score of 2, and so on). What statistical analysis would you advise to your friend?

This is a between subjects (or independent groups) design with one ordinal dependent variable (completion rank), and one independent variable with four levels (placebo, supplement A, supplement B and supplement C). It lends itself most obviously to a ***Kruskal-Wallis one-way ANOVA****. If the ANOVA is statistically significant, the friend should be advised to follow it with a series of Mann-Whitney U tests to identify which pairs of supplements differ significantly.*

Scenario 15: You have been hired to investigate whether a new vaccination impacts on subjective lethargy (tiredness). A sample of university students are randomized into two groups. The first group are given the vaccination, whilst the second group are given a placebo. Each student is then asked to rate their lethargy on 3-point scale, where 1 = not at all tired, 2 = somewhat tired, 3 = very tired. What statistical test would be appropriate for comparing the average levels of lethargy reported by each group?

The ***Mann-Whitney U test*** *can be used to compare two independent samples of ordinal data.*

Scenario 16: Before implementing a new anti-bullying intervention program at the 12 primary schools under their authority, the local school district asks each school's counsellor whether they currently believe their school has a 'bullying problem'. The counsellors responded to this question by answering either 'yes' or 'no'. At the conclusion of the program, each counsellor was again asked the same question. What statistical analysis should be used to determine whether or not the counsellors' perceptions of bullying at their schools changed between the two points in time?

The ***McNemar test of change*** *can be used to determine whether or not category membership on a binary dependent variable (counsellors' perceptions of whether or not a bully problem exists) changes between*

two points in time (before and after the intervention). Note that a McNemar test of change requires that both variables are dichotomous, and a repeated measures design.

Scenario 17: Lena believes that the Matching Figures Test (MFT), a visual identification test, is too difficult for children who are younger than five years old. The MFT consists of 24 items. For each item, a child is shown a picture for two seconds then, after a five second pause, is asked to select the same picture from a three-picture line-up. If a child is simply selecting at random (or guessing), we would expect him/her to select the correct picture on approximately 8 of the 24 trials (24/3 = 8). Lena has tested 75 five-year-olds and would like to know if they are performing on the MFT at a level that is any better than chance. What statistical test should she use here?

Lena wants to compare the mean of a sample of ratio level data against a predetermined value, 8, which represents a level of performance equivalent to chance. Assuming the MFT data are reasonably normally distributed, she should do this using a ***one-sample t-test****. If normality cannot be assumed, a* ***one-sample Wilcoxon signed-rank test*** *may be more appropriate.*

Scenario 18: The AFL (Australian Football League) Commission is interested to know whether West Coast Eagles supporters and Dockers supporters differ in the average amount of money they spend per season on AFL merchandise. Which statistical analysis would you recommend?

The commission want to compare two independent samples of ratio level data. Assuming these samples are reasonably normally distributed, an ***independent samples t-test*** *should be used. If normality cannot be assumed, the Commission should consider using a* ***Mann-Whitney U test*** *instead.*

Scenario 19: You are interested to know if a new fitness program is effective. You measure the weight of 30 participants prior to starting the program and again after completing the program. Which statistical test would you use to compare the two sets of measurements?

This is a repeated measures (within subjects) design, with a ratio level dependent variable and one independent variable with two levels (before vs. after completing the program). The resultant data could be analysed with ***a paired samples t-test****. If the assumptions of this test cannot be met, a* ***Wilcoxon signed rank test*** *could be considered instead.*

Scenario 20: In 2010, the residents of Capitol City were asked to indicate whether they were 'very dissatisfied', 'dissatisfied', 'neither satisfied nor dissatisfied', 'satisfied' or 'very satisfied' with the performance of the city council. The median response was 'neither satisfied nor dissatisfied'. Earlier this year, residents were asked the same question. The mayor would like to know whether satisfaction with the council has improved since the previous survey.

The mayor wants to compare the median of a sample of ordinal data collected earlier this year against a specified value (i.e., the 2010 median). The appropriate statistic for doing this would be a ***one-sample Wilcoxon signed-rank test****.*

Scenario 21: The National Cricket Association is keen to test the effects of their new 'subliminal advertising' campaign. They recruit a sample of 30 community members and ask them to rank 10 sports from 1 = 'most favorite' to 10 = 'least favorite'. The participants are then taken to a movie theatre and asked to watch an action film. Unbeknownst to the participants, positive subliminal references to cricket have been inserted throughout the film. At the conclusion of the film the participants are again asked to rank the 10 sports. The Association would like to know if the typical rank that participants assign to cricket improves between the two points in time. What statistical analysis could they use to determine this?

A ***Wilcoxon signed rank test*** *can be used to compare two related samples or ordinal (or ranked) data.*

Analyze relationships or associations between variables

Scenario 22: Is whether or not someone has a mortgage related to whether they prefer to watch commercial or non-commercial television news programs? What statistical test could be used to determine whether or not these two variables are related?

A ***chi-square test of contingencies****, along with a* ***phi coefficient*** *could be used to examine the nature of the relationship between these two dichotomous variables. if either or both of the variables had more than two levels, Cramer's V should be used in place of phi.*

Scenario 23: A teacher would like to know whether there is an association between physics grades and gender amongst her 10th grade students. What statistical test would you recommend.

A ***point biserial correlation coefficient*** *can be used to quantify the relationship between a dichotomous variable (gender) and a continuous (i.e., interval or ratio level) variable (physics grades). Alternatively, if the grades can be considered the 'dependent variable' in this study, an* ***independent samples t-test*** *can be used instead. Both will lead the teacher to the same basic conclusion.*

Scenario 24: A doctor would like to know whether there is an association between BMI (Body Mass Index) classification (underweight, normal, overweight, or obese) and dog ownership (yes or no) amongst his patients. What statistical test would you recommend?

The doctor is interested in the relationship between an ordinal variable (BMI classification) and a dichotomous variable (dog ownership). Such a relationship can be quantified with a ***rank biserial***

***correlation coefficient**. Alternatively, if BMI classification can be considered the 'dependent variable' in this study, a **Mann-Whitney U test** can be used instead. Both will lead the doctor to the same basic conclusion.*

Scenario 25: The Banana Growers Association (BGA) want to know if there is a relationship between retailer size (categorized as 'small', 'medium' or 'large') and the banana wholesale prices the retailers are able to negotiate. What statistic should they use to address this research question?

*Assuming that the relationship between size and price is monotonic, and that the BGA has price and size data for a range of different retailers, **Spearman's rho** can be used to quantify the strength and direction of the relationship between these two variables.*

Scenario 26: A doctor is interested in finding out whether his patients' weight is related to how much TV they watched in the previous week. What statistical analysis would you recommend to him?

*Assuming the relationship between these two variables is linear, it will be best captured by **Pearson's product moment correlation coefficient**.*

Scenario 27: The Computer Retailers Association (CRA) would like to quantify the strength of the relationship (if any) between annual income and preferred operating system (Windows, MacOS, Linux or Other) amongst their members. What statistic would you recommend?

***Eta** is a symmetric measure of association between one nominal variable (operating system) and one continuous (i.e., interval or ratio level) variable (income).*

Scenario 28: A teacher wants to know if she can predict end-of-semester exam scores using mid-semester exam scores. What would you recommend?

*The criterion variable here is continuous (end-of-semester exam scores), as is the predictor variable (mid-semester exam scores). **Bivariate regression** can be recommended to the teacher.*

Scenario 29: You work at a university library and have been tasked with finding out which students accrue the largest 'overdue fines'. The head librarian has provided you with a data file that gives you the total amount of fines (in dollars) accrued by each borrower during the previous 12 months, along with a range of additional information (e.g., each borrower's course of study, age, gender, number of items borrowed etc.). What statistical analysis would you use?

*The intent here is to predict the size of fines (a continuous variable) using a range of continuous and categorical predictor variables. This scenario suggests **standard multiple regression**, following the dummy coding of any categorical predictors with more than two levels (e.g., course of study).*

Scenario 30: A teacher wants to know if, after controlling for students' mid-semester exam scores, she can predict their end-of-semester exam scores with their written assignment marks. The students complete two written assignments in addition to the two exams during the semester. What statistical analysis should she use?

In this scenario, there is a continuous criterion variable (end-of-semester exam scores), two predictor variables (the two sets of written assignment marks) and one control variable (mid-semester-exam scores). ***Hierarchical multiple regression*** *appears appropriate.*

Scenario 31: You are interested in whether the following factors in combination can predict whether or not participants who have been sexually harassed in their workplaces reported the incident to their supervisors: (a) relationship status (coded as 0 = single and 1 = in a relationship); (b) feminist ideology (measured using a self-report scale); (c) frequency of the harassment (measured on a 5-point scale); and (d) offensiveness of the behavior (measured on a 10-point scale). What statistical analysis would be appropriate here?

In this scenario, we have a dichotomous criterion variable (whether or not the incident was reported) and several predictor variables (relationship status etc.). There do not appear to be any covariates in the study, and thus an appropriate analysis would be a ***standard binary logistic regression.***

Scenario 32: You're working at a travel agency and have already established that wealthier customers are more likely to travel internationally during their holidays. Your manager wants to know if she can improve her ability to predict whether or not customers travel internationally by incorporating a few more demographic variables (age, gender, number of flights booked in the previous 12 months etc.) into her statistical model. The manager hopes that she will be able to use this model to refine her advertising materials. What statistical analysis would you consider using here?

This scenario appears to be suggesting a ***hierarchical binary logistic regression*** *to predict whether or not customers travel internationally (a dichotomous criterion variable). Income would be added to this model on the first block, followed by the remaining demographic variables on the second block. The manager seems to be hoping that the full model will have significantly greater predictive utility than the model containing only income as a predictor.*

Scenario 33: A large electronics retailer wants to use the information in their customer database to develop a model that can be used to predict which type of tablet device customers are most likely to purchase. Over the preceding 12 months, they have sold 10,000 tablets, which have been categorized by operating system: iOS, Windows, Android, and Other. In addition to the type of tablet purchased,

the retailer also knows the following about each tablet customer: number of other mobile devices owned, annual income, age, gender and postcode. What statistical technique will you suggest to the retailer?

In this scenario, there is a nominal criterion variable that has four levels (iOS, Windows, Android and Other). There are several predictor variables (income, age etc.), though postcode will need to be transformed to be useful. (Perhaps postcodes can be grouped by socioeconomic level?) Once the issue with postcodes has been addressed, multinomial logistic regression seems to be an appropriate recommendation. If the retailer has a rationale for the order in which predictors should be entered into the model, ***hierarchical multinomial logistic regression*** *can be used. If not, all predictors can be entered simultaneously into a* ***standard multinomial logistic regression****.*

Scenario 34: The editor of an academic journal is interested the degree to which he can predict his reviewers' recommendations ('reject', 'accept with major revisions', 'accept with minor revisions' or 'accept') based on the characteristics of the papers they are reviewing. These characteristics include (a) length; (b) the number of *p*-values reported; and (c) number of previous papers published by the author. What statistical technique would you recommend.

The editor has an ordinal criterion variable (recommendation), and multiple predictor variables (which are all continuous in this case), suggesting ordinal regression. If the editor has a rationale for the order in which predictors should be entered into the model, ***hierarchical ordinal regression*** *can be used. If not, all predictors can be entered simultaneously into a* ***standard ordinal regression****.*

Examine the underlying structure of a measure

Scenario 35: A researcher is developing a new eighty-item measure to assess mental ability and wants to know whether there is a small set of constructs underlying the questionnaire. Which statistical analysis would you recommend?

Either ***exploratory factor analysis*** *or* ***principal components analysis*** *could be used here. The researcher will need to decide which is more appropriate for his or her purposes.*

Examine the reliability of a measuring instrument

Scenario 36: Kate and Phil have video-recorded 100 interactions in a busy shopping center car park. Each has independently classified each interaction as either 'aggressive', 'friendly' or 'neutral'. Now, they want to compute a statistic that will indicate the extent to which their classifications are consistent. What would you recommend?

Cohen's kappa *can be used to as an index of inter-rater reliability here.*

Scenario 37: David presented 200 participants with a 10-item questionnaire that measured perceived social support. Participants responded to the items on a 5-point scale ranging from 1 (strongly disagree) to 5 (strongly agree). David wants to make sure that the measure is internally consistent. Which statistical analysis should David use?

Cronbach's alpha *can be calculated to assess the internal consistency of a unidimensional Likert scale.*

Scenario 38: Researchers are interested estimating the test-retest reliability of a recently developed measure of the big five personality factors (openness, agreeableness, neuroticism, extraversion and conscientiousness). They have administered the measure to the same group of participants on two separate occasions, and now need some statistical advice.

An ***intraclass correlation coefficient*** *can be used here. Note that the researchers will need to compute a separate coefficient for EACH sub-scale/factor in the measure.*

Scenario 39: Two psychiatrists independently watched 50 video recordings of interviews with patients diagnosed with schizophrenia and counted the number of symptoms displayed by each patient. They now want to compute a statistic that will indicate the degree with which their symptom counts agree. What would you suggest?

An ***intraclass correlation coefficient*** *can be used to assess inter-rater reliability, when the ratings have been made on a continuous scale.*

Scenario 40: A team of researchers are examining social anxiety levels in first year university students. They administer a measure of social anxiety to a sample of 150 first-year university students and want to assess the reliability of the measure in this sample. They have come to you to ask how to do this. The measure is unidimensional, and participants responded to each item with either 'true' or 'false'. Which statistical analysis would you recommend?

KR20 (Kuder-Richardson 20) *can be calculated to assess the internal consistency of a unidimensional scale with a dichotomous response format.*

Scenario 41: Kate and Phil have video-recorded 100 unambiguously aggressive interactions in a busy shopping center. They have then independently rated each interaction as 'very aggressive', 'aggressive' or 'mildly aggressive'. Now, they want to compute a statistic that will indicate the extent to which their classifications are consistent. What would you recommend?

*As the interactions have been classified into ordered categories, **weighted kappa** should be used an index of inter-rater reliability.*