Give Swirl a Whirl: A Mixed-method Analysis on Swirl Approach to Teach Applied Statistics in a Biology Student Cohort

Felicia Devina Ngan, Nanyang Technological University, Singapore
Chun Chau Sze, Nanyang Technological University, Singapore
Wilson Wen Bin Goh, Nanyang Technological University, Singapore

The Asian Conference on Education 2019
Official Conference Proceedings

Abstract
To equip biology students with data literacy skills, this study investigates the utility of retooling the electronic programming tutorial system, Swirl, for teaching applied statistics in a cohort of biology students in a 2-phased study: Phase 1 involved administration of tutorial-based course, pretest-posttest assessment and preliminary survey, while Phase 2 involved post-hoc survey and learning style analysis. The base teaching material, in both Swirl and paper-based forms received positive content evaluation amongst students and improved students' learning outcomes with significant learning gains and large effect sizes. While there is no evidence of greater learning gains in Swirl against the paper version, it does offer better palatability through its interactive and integrated programmatic components. We see Swirl's key value in early immersion of students in a formal programmatic environment while learning applied statistical theory simultaneously, and believe that this is essential for efficiently bridging theory-practice gaps for aspiring bio-data scientists.

Keywords: biological science; data literacy; education; statistics; Swirl; technology-enhanced learning

Introduction

The 21st century marks the era of technological revolution in an increasingly knowledge-driven society. In biology, Big Data is responsible for fuelling the rise in high-throughput "-omics" studies (involving the large-scale analyses of all biological molecules found within a cell or tissue). Biological research has since become data-heavy and computationally-complex. There is a pressing need for biologists to acquire strong statistical thinking and computational literacy (Carey & Papin, 2018) – two skillsets undoubtedly critical for 21st century work-readiness (Makarevitch, Frechette, & Wiatros, 2015). Biologists often could not apply their theoretical statistical knowledge to real research problems (Gore, Kadam, Chavan, & Dhumale, 2012), a situation we term as "theory-practice" gap. The superficial, touch-and-go statistical and computational training that prevailed in most biology undergraduate curriculum creates an ever-widening gulf between what is learnt by students versus the practical demands required from research activities.

1.1. Overview of The Swirl Platform

To fulfil the twin criteria of delivering statistics and programming, we selected Swirl (Statistics with Interactive R Learning) as the study platform. Swirl is a free, open-source R package developed to teach R programming (Carchedi, 2014). R is a statistical computing language that is widely adopted by non-developer users. It is a powerful platform that offers wide range of data analysis and visualisation packages for customisation and application in different fields (Becker RA, 1988). These qualities provide R with a strong advantage over commercial statistical software such as SAS and SPSS (Muenchen, 2014).

Swirl is unique from other Massive Open Online Courses (MOOCs) such as DataCamp, Khan Academy, and Coursera in that it provides students direct user interaction within the native R programmatic environment and gives them freedom for further exploration. Swirl also acts as a virtual tutor that offers guided learning at one's own pace. The interactive component comes as student responds to a series of instructional questions in Swirl and immediate feedback is given to either stimulate students' response or correct students' misconceptions.

Since the inception of Swirl in 2014 to till date, there were very few advanced statistics courses (Swirl, 2014). Moreover, despite being commended by online community as a great learning tool for data science and R, there was virtually no formal study to evaluate the effectiveness of Swirl as an educational instrument. This work served to provide empirical evidence on the use of Swirl as teaching platform for applied statistics in the context of adopting it as part of large scale, formal institutional teaching.

However, effective statistical education goes beyond just selecting a suitable instructional method. Subject matter difficulty such as the relatively abstract statistical concepts could present significant barriers in teaching and learning. For the instrument to be effective, it should also be designed with appropriate pedagogy that integrate both statistics and students' academic discipline (Feser, Vasaly, & Herrera, 2013). The lack of personalised applied statistics within the biological science field

thereof calls for technical development of the course content based on a selected applied statistics theme.

1.2. Research Objectives

The overall aim of the study was to evaluate the suitability of using Swirl as an educational platform to teach applied statistics for biology students. Given that the effectiveness of an educational platform is preconditioned on the content validity, our research objectives were framed in two parts:

(1) *Content:* to design an effective introductory course to teach a selected theme of applied statistics for biology students, incorporating pedagogical elements as recommended in the literature

(2) *Platform*: to evaluate the potential of Swirl as an instructional platform for teaching the same educational content developed in the first objective by comparing it to a Control medium (paper-based/PDF format).

2. Methodology

The materials and methods were selected based on its feasibility to deliver the study's objectives within the time span of 17 weeks.

2.1. Preparation

Course Development

Content analysis was first done to identify relevant themes in statistics for senior biology students. Out of the three themes shortlisted, the concept of False Discovery Rate (FDR) was chosen as the final teaching topic based on several considerations as exemplified in Figure 1.

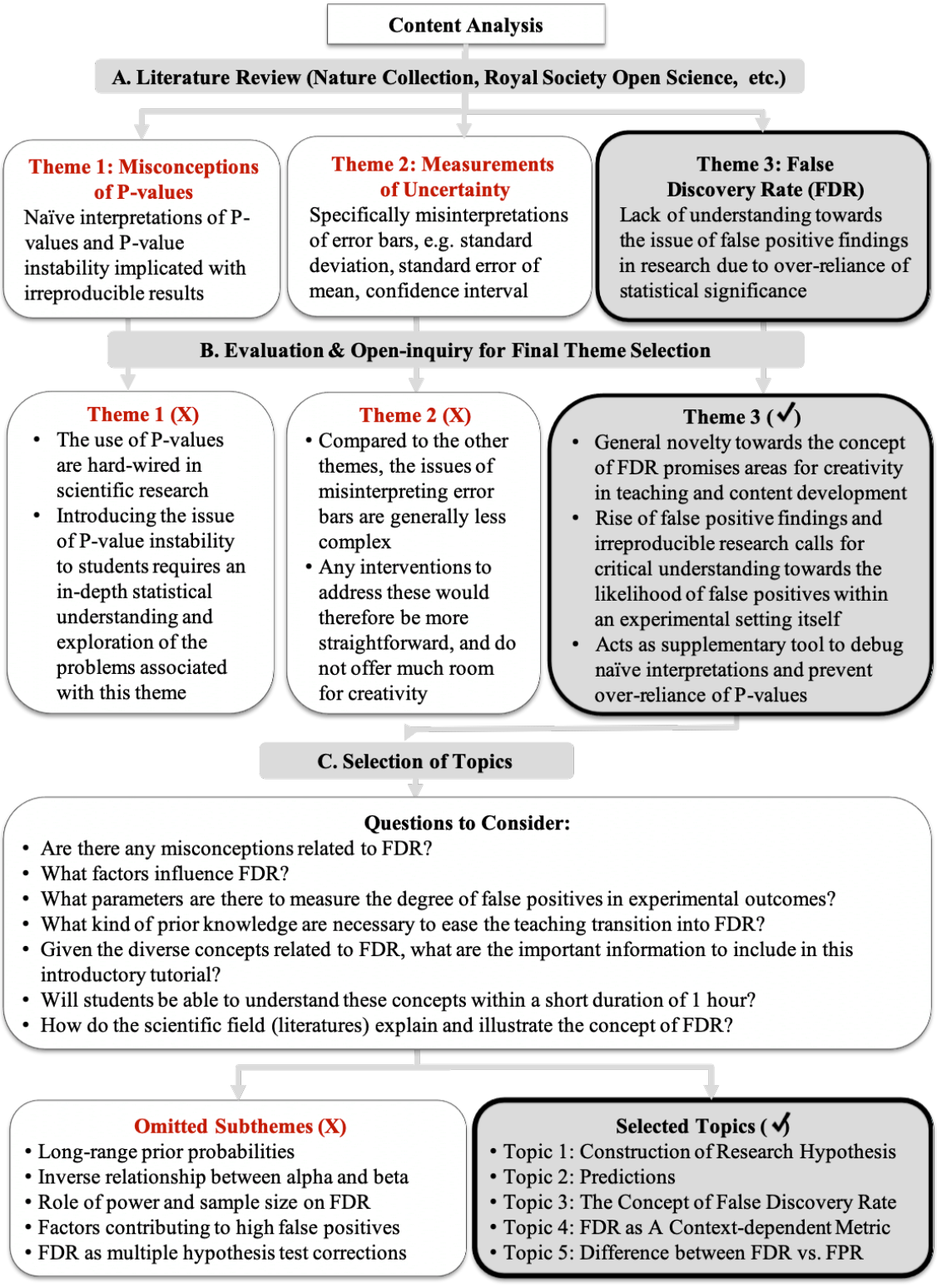


Figure 1: Content analysis performed over three major themes lasted a time span of three weeks, consisting of: (A) Literature review, (B) Evaluation of statistical themes, (C) Topics selection of the final theme

The course was developed by adopting blended mix of content-pedagogy-technology teaching strategies (D. S. Moore, 1997) and integrating statistics with biological field (Feser *et al*., 2013). The pedagogical design of the educational instrument was guided following four content evaluation criteria (Figure 2A) and quality process frameworks (Figure 2B).

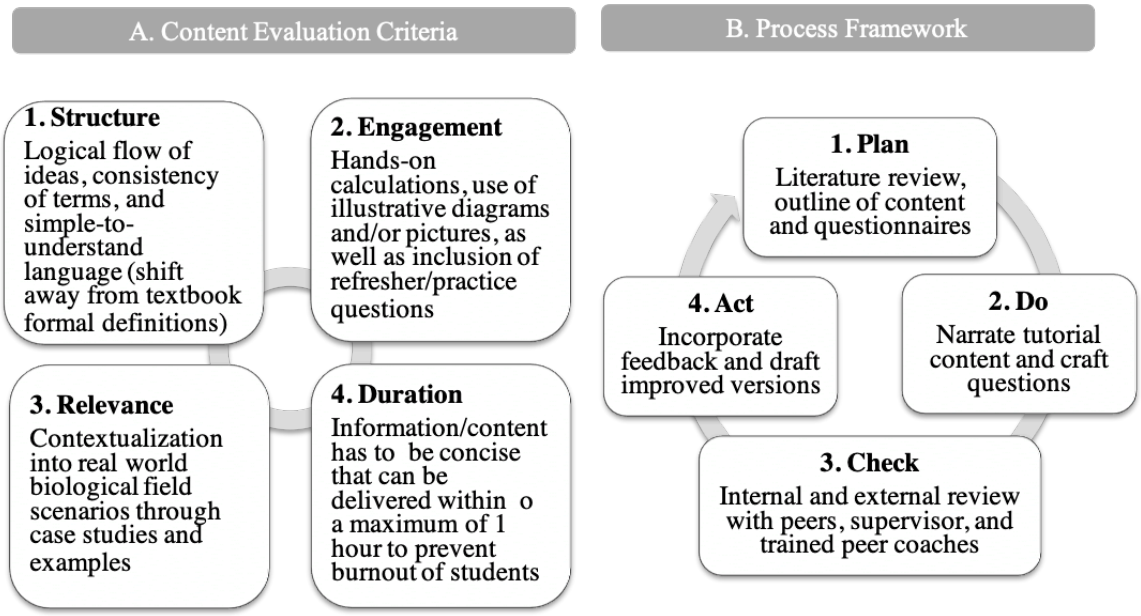


Figure 2: (A) Content evaluation criteria, (B) Process framework

The short course materialised in the form of a tutorial format, covering a total of five topics (Table 1).

Table 1. List of topics introduced in the tutorial

Topics	Content
Topic 1	Construction of Research Hypothesis
Topic 2	Predictions 2.1 Positive and Negative Predictions 2.2 False Predictions 2.3 True Predictions 2.4 Contingency Table
Topic 3	**The Concept of False Discovery Rate** 3.1 Mini-exercises
Topic 4	**FDR as A Context-dependent Metric** 4.1 Ratio of Positive to Negative Events 4.2 Summary of FDR Calculations
Topic 5	**Difference between FDR vs. FPR** 5.1 Practice Case Calculations 5.2 Summary of Key Concepts

For the Control tutorial, the content was organised into 10 sections and totalled to 13-page editable PDF document which enabled inputs of answers by students (Appendix A). The teaching material was designed by following the four evaluation criteria of the content (Figure 2A):

(1) *Structure:* standard use of terminologies and simple-to-understand, stepwise explanations

(2) *Engagement:* adoption of visual aids (e.g. pictorial diagrams, graph simulation) to foster conceptual understanding, as well as occasional summary pointers, practice questions and a guided case study to validate students' understanding

(3) *Relevance:* the entire course was narrated based on an allergy screening scenario where students were simulated into the role of researcher investigating the screening results with the aim of identifying the false positive outcomes through the use of the false discovery rate metric

(4) *Duration:* the entire course was trialled with the help of several student volunteers from the science discipline (not the study participants) and timed for completion within 1 hour

For the Swirl tutorial, the same teaching material was adapted into Swirl using the "swirlify" R package, (Swirl, 2014b), following the workflow as shown in Figure 3. This standardisation helped to minimise confounding factors such as user attitudes and behaviour towards the use of computer-assisted learning such that any observable differences could be more directly associated with the platform differences.

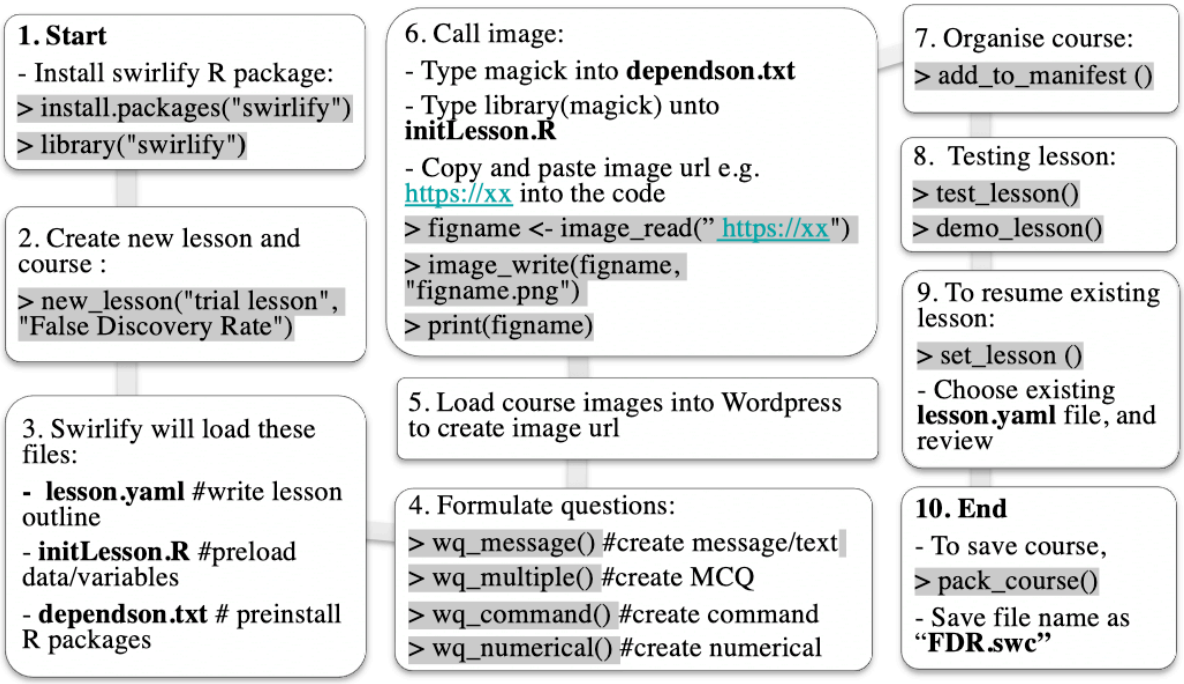


Figure 3: Process outline to develop the Swirl course on False Discovery Rate (FDR). Authoring of Swirl course was done through "swirlify" R package loaded into the R-studio (step 1). Swirlify would generate a series of files with each new course (step 2-3). The content from the control reference could be adapted into swirl course through the use of multiple question types, specifically message (for purely text description), multiple (for MCQ question), command and numerical questions were utilised for instrumentation. The course would then be organised into order (step 7) prior to trial and demo (step 8). Author of the course could exit and resume the development of course (step 9). Upon completion, the course was saved in ".swc" format which would then be ready for loading by user (step 10).

A sample of the Swirl tutorial along with its 10 characteristic elements is featured in Figures 4A and 4B below.

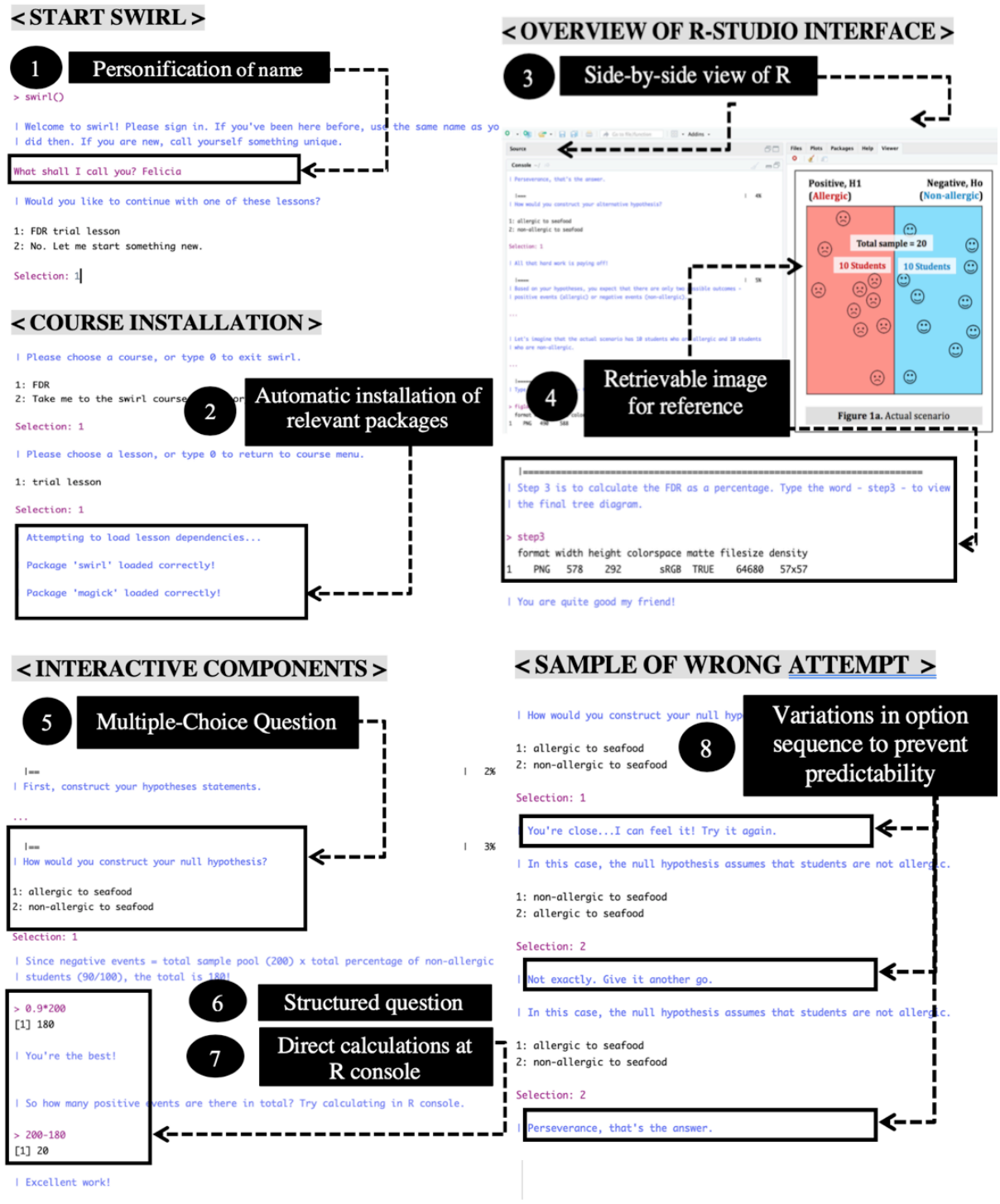


Figure 4A: Sample Swirl tutorial featuring starting interface, course installation, visual interface, interactive components and incorrect trial attempts

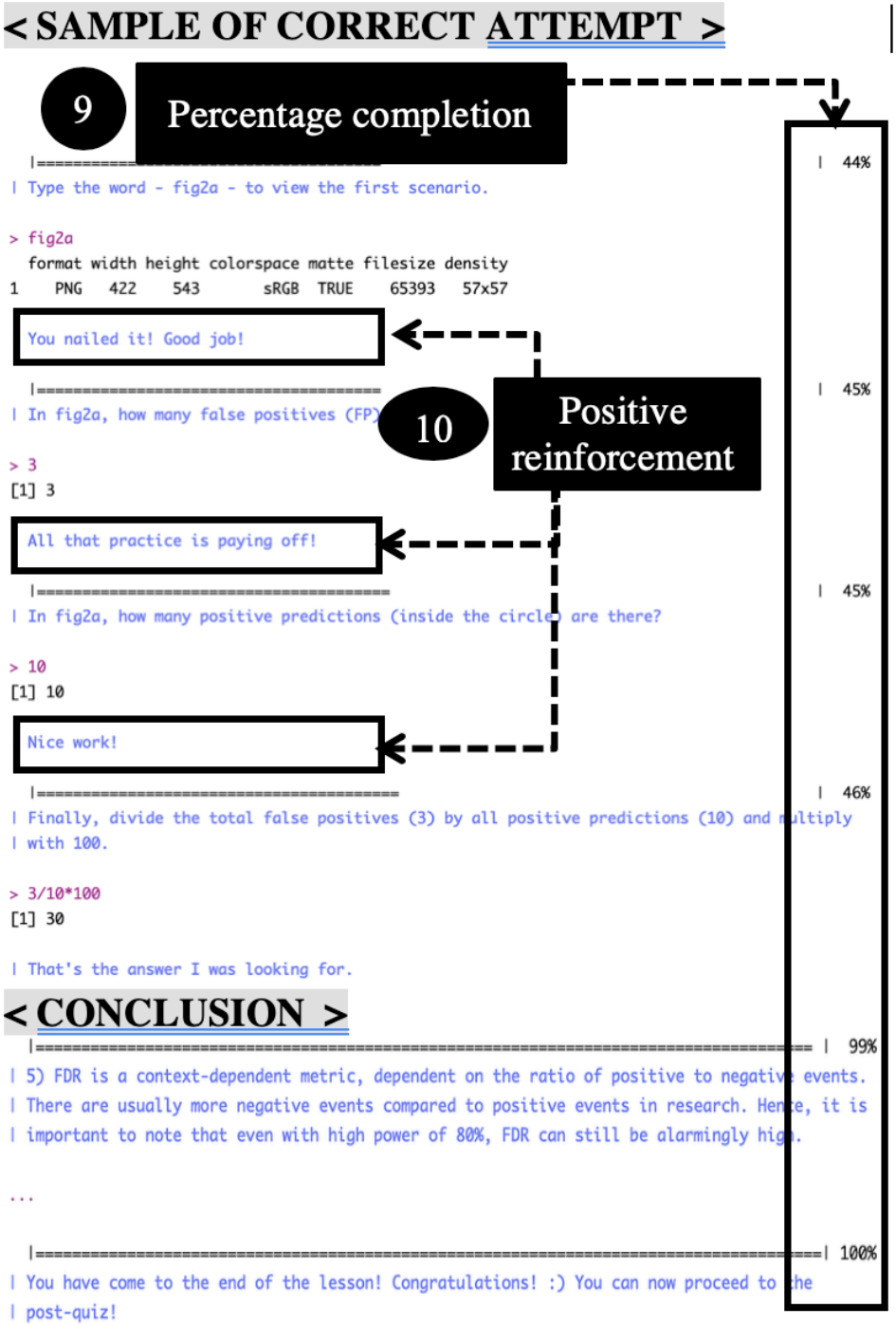


Figure 4B: Sample Swirl tutorial featuring correct trial attempts

Test Questionnaires

Each pretest and posttest (Appendix B) consisted of five closed-ended, multiple-choice questions (MCQ). The tests were administered in Google Forms with pre-specified answer keys to facilitate quick scoring and data collection. Pretest questionnaires and MCQ options were designed in parallel with the posttest for evaluation of each learning objective (Appendix D, Table 2). The validity of questions and answers were reviewed against the learning objectives and by teaching faculty. The first topic relating to the construction of hypothesis test was not tested as students were assumed to have known this introductory concept.

Survey Questionnaires

Preliminary survey items was administered immediately after the tutorial to gather students' first-hand responses towards the course quality. The post-hoc survey questionnaires were constructed to follow up on the preliminary survey responses, and consisted of 13 open-ended questions and a series of Likert-type questions which included three ranking questions and 26 short statements. The survey metrics (Appendix D, Table 3) were inspired from standardised frameworks used in evaluating technological-based instrument (Bowyer & Chambers, 2017). Both surveys (Appendix C) were were adapted into Google Forms.

Index of Learning Style (Felder & Soloman, 1997) was used to evaluate students' learning styles (Felder & Silverman, 1988). There were four categories of learning dimensions: (1) active-reflective, (2) sensing-intuitive, (3) visual-verbal and (4) sequential-global.

2.2. Implementation

12 biology students enrolled in Nanyang Technological University's "BS3033 Data Science for Biologists" were recruited and assigned non-randomly into experimental and control groups based on common timeslot. Swirl group and Control group referred to students enrolled in Swirl tutorial and paper-based (PDF) tutorials respectively. This cohort made up of senior students who were trusted to have statistical foundations needed for proper assessment of the applied statistics course and had prior familiarity with Swirl courses to provide more holistic perspectives on the use of Swirl platform.

The project was administered in two phases (Figure 5) of one hour each: Phase 1 was conducted during the mid-semester break since most students would be free from classes, whereas Phase 2 was resumed two weeks after to allow time for data analysis of first phase results and preparation of the post-hoc survey. Students' attitudes and behaviours were observed to provide supplementary qualitative data. Researcher was available to answer respondents' queries, ensured full survey completion, and obtain informed consent from participants.

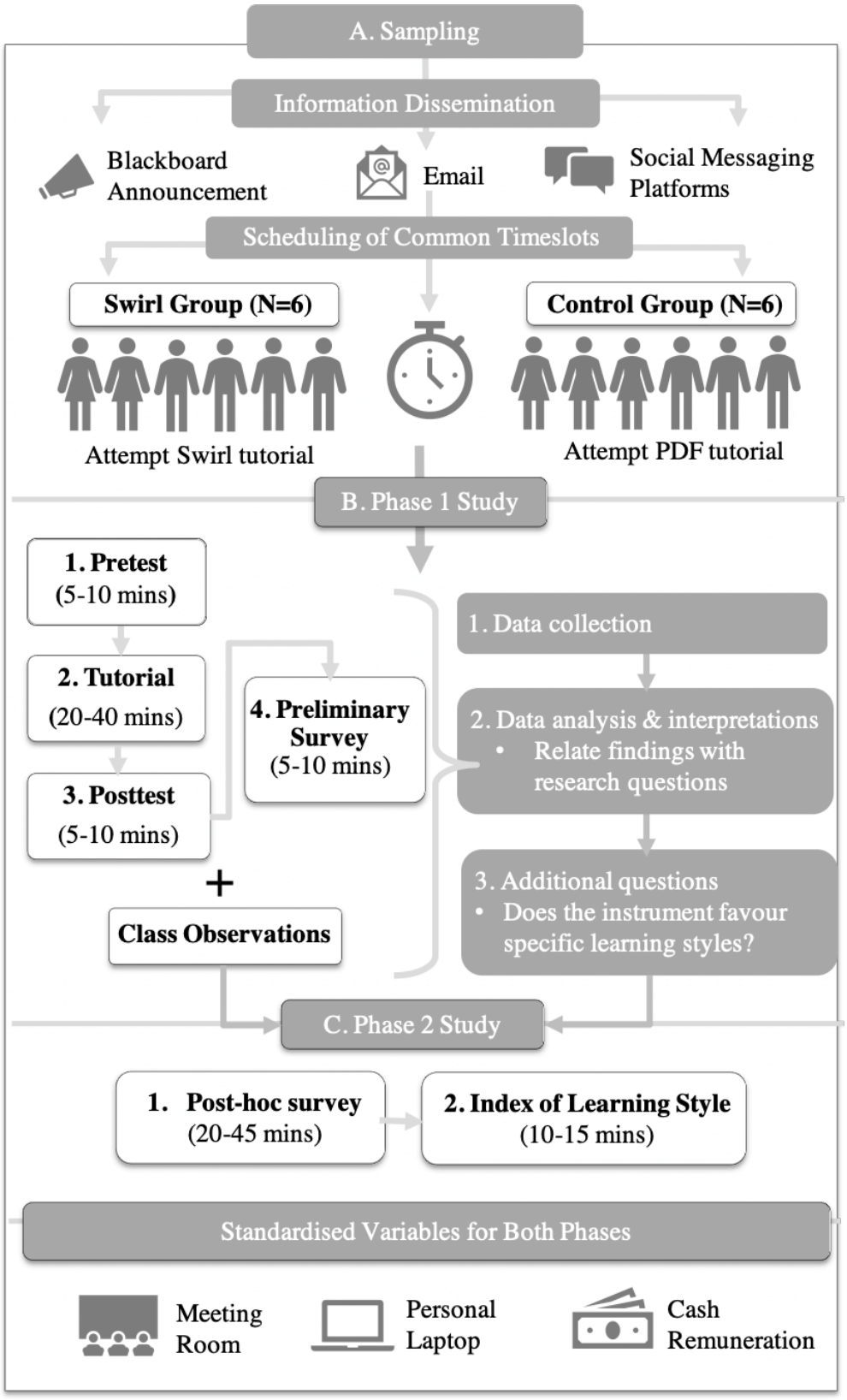


Figure 5: Process outline for the implementation stage of the study involves: (A) Sampling where information about recruitment was disseminated via three modes of communication channels, (B) Phase 1 of the study involved several tasks with approximal duration listed above; the findings were followed up in (C) Phase 2 of the study via the post-hoc and learning style surveys. External variables were controlled whenever possible

Students in Swirl group followed the instructions below to load the Swirl tutorial (Figure 6).

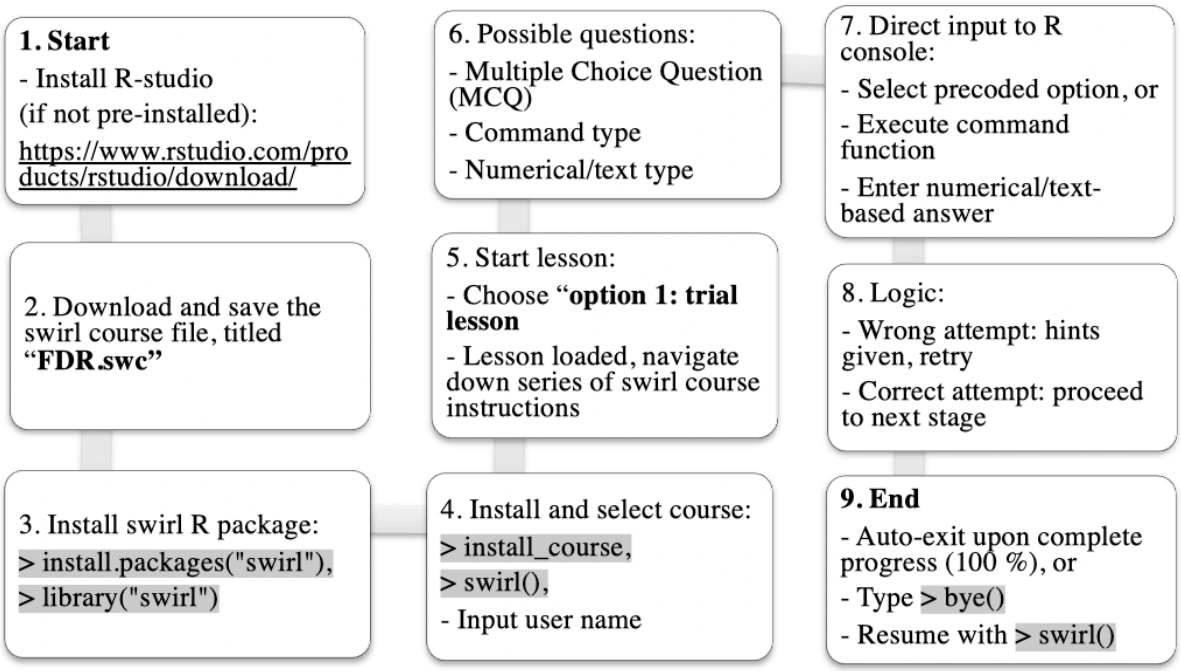


Figure 6: Process outline to install and navigate the instrument (swirl course on FDR) from a user (learner) perspective. Prerequisites for the Swirl course included installation of R-studio and swirl R package (step 1-3). The selected swirl course file (e.g. "FDR.swc") could then be loaded (step 4-5). Swirl course would direct users through a series of alternating instructional text and prompts students to answer each question. Throughout the course, users encountered different types of questions where they were required to either choose a pre-specified MCQ options, enter text command, or numerical value as answer all within the native R console (step 6-7). With each input of incorrect answer, feedback was given immediately in the form of precoded hint to prompt user to retry. With each input of correct answer, user would receive positive encouragement words which were pre-programmed in every swirl course (step 8). User could see the percentage completion with every progress made. Entire swirl course would require around 10-20 minutes, subject to individual's progress. User could choose to exit the course and resume later (step 9)

2.3. Evaluation

Mixed-method evaluation was adopted due to its holistic approach in providing insights by triangulation of both quantitative and qualitative data (Greene, Caracelli, & Graham, 1989), common in educational research.

Normalised learning gains (*nlg*) adapted from Hake's normalised gain (Hake, 1988) and Cohen's *d* effect size (Cohen, 1988) were used concurrently to evaluate the extent of learning gained from the intervention. Average *nlg* was computed by taking the mean of all *nlg* scores of students.

All scores were calculated in percentages and equations used were as follow:

$$\text{Learning gains } (lg) = \text{ Posttest scores} - \text{Pretest scores} \quad (1)$$

$$\text{Normalised learning gains } (nlg) = \frac{\text{Learning gains in Equation 1}}{\text{Maximum scores} - \text{Pretest scores}} \quad (2)$$

$$\text{Effect size } (d) = \frac{(\text{Average of } X) - (\text{Average of } Y)}{\text{Pooled Standard Deviation (SD)}} \quad (3)^{\#}$$

$$\text{Pooled SD} = \frac{\sqrt{(\text{SD of } X + \text{SD of } Y)}}{2} \quad (4)^{\#}$$

[#]For equations (3) and (4), '**X**' denotes posttest scores and '**Y**' denotes pretest scores in the calculation of within-group differences in terms of pretest-posttest scores. For calculation of between-group differences in terms of *nlg* (Table 4C), '**X'** denotes mean nlg of Control group and '**Y**' denotes mean nlg of Swirl group.

For analysis of survey responses with 5-point Likert, the median of respondents' ratings in each sample was first computed, followed by taking the mean of substituent survey items to give a composite mean score for each survey metric. Negatively-expressed survey items were reverse-coded for easier mean score computations of each metric. Higher scores suggested higher attributes for the metric of interest. The preliminary and post-hoc survey were used to evaluate one and six survey metrics respectively.

Statistical analyses and graph computations were performed using GraphPad Prism (version 6.0e). For all statistical tests, a standard significance cutoff at P = 0.05 is used. Independent t-tests with assumptions of unequal variances (Welch's correction) were the primary mode of statistical analysis. Within-group *lg* differences were analysed with one-sample t-test, whereas between-groups *nlg* were analysed with two-sample t-test. One-tailed, two-sample t-tests were done to identify if one variable was significant over the other for gender and learning styles preferences. The Fisher's Exact Test was performed to evaluate any significant relationships between each pair of learning style dimensions.

3. Results

3.1. Quantitative Analysis

Pretest-posttest assessment was done to evaluate the effectiveness of each platform in promoting students' conceptual understanding. Both groups showed significant pretest to posttest scores improvement in terms of learning gains (p<0.05), with Swirl group achieving twice as much effect size (1.85) compared to the Control group (0.89) (Table 4, Figures 7A-7B).

Table 4. Summary statistics showing students' performance: (A) Swirl group, (B) Control group and (C) Normalised learning gains (*nlg*) of both groups

Variable	Mean ± Standard Deviation	Min	Max	*P*-value (Cohen's d)
(A) Swirl				
Pretest scores (%)	23.33 ± 15.06	0.00	40.00	$P = 0.042$* ($d = 1.85$)
Posttest scores (%)	63.33 ± 26.58	20.00	100.00	
Learning gains (%)	40.00 ± 28.28	0.00	80.00	
(B) Control				
Pretest scores (%)	60.00 ± 17.89	40.00	80.00	$P = 0.018$* ($d = 0.89$)
Posttest scores (%)	76.67 ± 19.66	40.00	100.00	
Learning gains (%)	16.67 ± 15.06	0.00	40.00	
(C) Average normalised learning gains (*nlg)*				
Swirl *nlg*	0.52 ± 0.14	0.00	1.00	$P = 0.36$ ($d = 0.21$)
Control *nlg*	0.44 ± 0.16	0.00	1.00	
Difference in *nlg* (%)	0.077 ± 0.21			

* denotes statistically significant with sig level of 0.05 (two-tailed, two-sample t-test).

Pretest and posttest scores between groups were compared to determine differences in terms of prior knowledge and post-intervention understanding. Table 4, Figures 7C-7B showed both groups differed significantly in terms of pretest scores (p<0.01), with Swirl group performing lower (0-40%) compared to Control group (40-80%). No significant difference was observed in the posttest scores (Figure 7D).

Because of non-comparable pretest scores, comparison based on posttest scores alone would not fairly associate students' understanding to the sole merits of the platform. Both groups' performance were further compared using *nlg*, a widely used metric in educational research to account for disparities of learning abilities and backgrounds (Hake, 1988). Comparison of both groups' *nlg* showed non-significant differences with small effect size of 0.21 (Figure 7E).

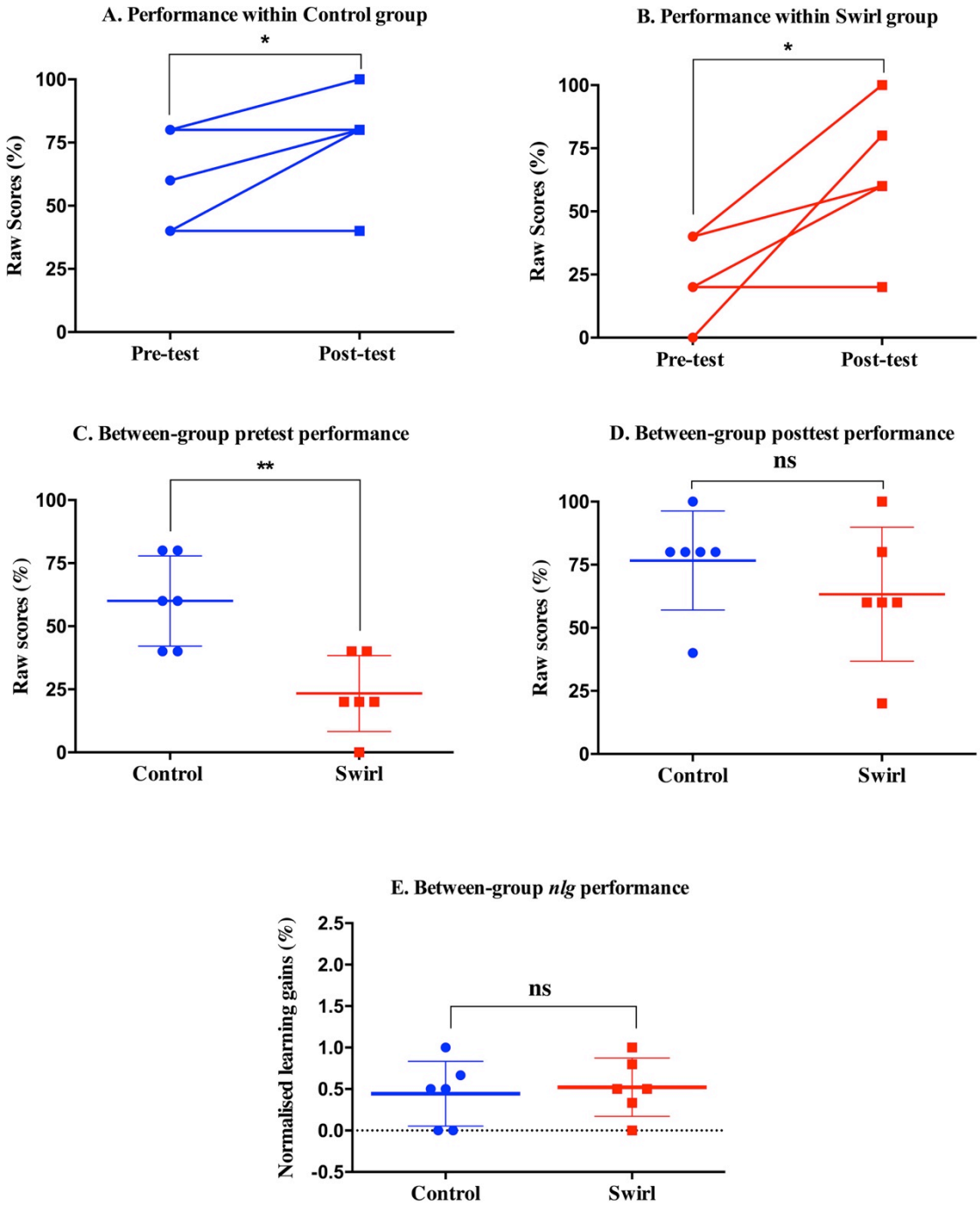


Figure 7: Comparison of students' performance with regards to:
(A) Control group, (B) Swirl group, (C) pretest scores, (D) posttest scores, (E) *nlg*.
For (A) and (B), there were two data points bearing the same pretest and posttest scores pairing of 60%- 80% and 20%-60% respectively. Statistical analyses for all figures were done using two-tailed, two-sample t-test with Welch's correction (sig level = 0.05). Legends: '*' and '**' denotes statistical significance at $p<0.05$ and $p<0.01$ respectively, and 'ns' denotes not statistically significant

Given no significant differences in *nlg* (Figure 7E), we analysed students' performance on each pair of pretest and posttest questions to determine whether the choice of platform helped to facilitate better conceptual understanding for certain

topics than the others. Both groups of students however yielded similar sequence of ranking in terms of question-based performance (Appendix D, Table 5A, 5B).

After analysing the preliminary results of students' performance, we investigated the effect of gender and learning style preferences on average *nlg*. We found no significant difference in any of the four learning style dimensions between both samples based on the Fisher's Exact Test. Interestingly, female gender and students with sensing learning style performed significantly better ($p<0.05$) than male and intuitive learners respectively (Table 6). Demographics analysis on the sensing-intuitive learning style showed that gender was similarly represented on both scales. However, there were two females in the sensing category who scored the maximum *nlg* of 1.00.

Table 6. Gender and learning style preferences of students against the average *nlg*

Variables	Swirl **(N=6)**	Control **(N=6)**	Average *nlg* **(%)**	Standard Deviation	*P*-value
Gender (Total)					
Female (5)	2	3	0.70	0.27	0.0316*
Male (7)	4	3	0.33	0.34	
Learning Profiles and Associated Learning-Teaching Dimensions					
Processing-Participation Dimension					
Active (5)	3	2	0.43	0.43	0.3598
Reflective (7)	3	4	0.52	0.32	
Perception-Content Dimension					
Sensing (6)	3	3	0.67	0.28	0.0362*
Intuitive (6)	3	3	0.30	0.35	
Input-Presentation Dimension					
Visual (11)	5	6	0.45	0.36	-
Verbal (1)	1	0	0.80	-	
Understanding-Perspective Dimension					
Sequential (3)	3	0	0.67	0.29	0.1487
Global (9)	3	6	0.42	0.37	

* denotes statistically significant with sig level of 0.05

3.2. Qualitative Analysis

Classroom observations

The student participants who showed good learner attitudes (e.g. notes-taking and careful review of content before taking the posttest) were seen to perform well. These behaviours were common in Control group, but less so for Swirl group (only one student took notes during the tutorial). In solving test questions, most students in Control session wrote their workings and used calculator, while those in Swirl session performed calculations directly on the R console. Duration for completion of entire Phase 1 was approximately 45 and 60 minutes for Swirl and Control group respectively. Investigation into time stamp records showed that Swirl group started

the post-test comparatively earlier than Control group. In Phase 2, no stark differences were observed between both groups.

Similar Survey Ratings on Perceived Difficulty Level (M2) and Perceived Ability (M3)

Table 7A showed that both groups demonstrated similar perceptions where they were moderately neutral (Mean = 2.50) and positive (Mean = 3.58) towards perceived difficulty level of course and perceived ability towards performance respectively. Results of survey items M2a-b on ranking of topics based on difficulty and abstraction level were omitted from the study due to poor construct validity. Both groups found the tutorial explanations and exercises comparatively easy compared to the assessments (pretest and posttest). For item M3c, majority of Swirl group responded with "Agree" while Control group gave "Neutral" stance. Swirl group appeared less confident for posttest performance and expressed the need for more time to review the tutorial and attempt the quizzes.

Table 7A. Breakdown of survey metrics (M2, M3) with similar ratings[#] from both groups

ID	Components	Mean ± S.D. (Control)	Mean ± S.D. (Swirl)	Comparison
M2	**Perceived difficulty of course content**	**2.50±0.31** "Agree"	**2.50±0.50** "Agree"	**0.00±0.19** (Control ≈ Swirl)
c	The explanations in the course content were easy for me to understand **(reverse-coded)**	2.00	2.00	NA
d	The exercises were too easy for me with or without the hints and answer keys **(reverse-coded)**	3.00	3.00	NA
e	The questions asked in the quizzes were relatively difficult to do	2.50	2.50	NA
M3	**Perceived ability**	**3.58±0.15** "Agree"	**3.58±1.02** "Agree"	**0.00±0.87** (Control ≈ Swirl)
a	I would think my performance in the pre-test quiz was generally good (Scored at least 3 out of 5 points)	3.00	3.00	NA
b	I would think my performance in the post-test quiz was generally good (Scored at least 3 out of 5 points)	4.00 "Agree"	3.50 "Agree"	Control > Swirl
c	I would need more time to review the course and remember the concepts **(reverse-coded)**	3.00 "Neutral"	2.00 "Agree"	Control >> Swirl
d	I would need more time to complete the quizzes **(reverse-**	3.50 "Disagree"	4.00 "Disagree"	Control < Swirl

	coded)			
e	My general performance for the course was not affected by how fast other participants complete the study	4.00	4.00	NA
f	My general performance for the course was not affected by the amount of monetary rewards I receive from participating in the study	4.00 "Agree"	5.00 "Strongly Agree"	Control < Swirl

[#] Items which did not involve mean ratings or show any group differences were indicated as "NA". The approximately equal sign (≈), single comparison sign ("<" or " >"), and double signs ("<<" or ">>") denotes negligible difference, difference of less than 1-Likert point, and difference of at least 1-Likert point respectively. Reference code for mean Likert scores: "1.00-1.49" (Strongly Disagree), "1.50-2.49" (Disagree), "2.50-3.49" (Neutral), "3.50-4.49" (Agree), "4.50-5.00" (Strongly Agree). Reverse-coded: "1.00-1.49" (Strongly Agree), "1.50-2.49" (Agree), "2.50-3.49" (Neutral), "3.50-4.49" (Disagree), "4.50-5.00" (Strongly Disagree).

Open-ended questions further investigated students' opinions on tutorial design, course and quiz duration, and impact of distraction on performance.

Table 7B. Open-ended post-hoc survey responses following survey metrics (M2 and M3)

Key Points	Key Evidence
Most Control group students were supportive of the use of guided hints	"The guided approach does help me in my learning"
Most students enrolled in Swirl tutorial were less receptive to guided hints	"...Direct hints is not good for my learning, as I will tend to think less and answer straight away..."
Both groups were not confident of their posttest performance	"...discomfort with the subjects", "tend to feel uneasiness when faced with a mathematically-related concept"
Some students were pressurised to complete posttest earlier due to peer pressure	"...started to guess my answers when others finished earlier", "I was pressured to complete the questions a bit faster...even though not sure"
The Control students were generally focused during the study	"...tend to 'get into the zone' and...block off external distractions"

Although most students expressed they were slow learners, they recorded similar range of duration as the given trial duration to review new concept and attempt assessments.

Higher Survey Ratings by Swirl Group on Perceived Course Quality (M1), Ease of Use of Self-guided Tutorials (M5) and Learner Engagement (M8)

Table 7C showed that although both groups were positive on the course quality, Swirl students had better impression than their Control counterparts specifically on course layout and structure (item M1d). While both groups were positive towards provision of hints in self-guided tutorials, they were neutral towards learning a new concept using this format alone. Swirl group showed better impression on effort to understand new concept in self-guided tutorials (item M5b) compared to Control group. While both groups preferred the bite-sized delivery of information, higher level of learner engagement (M8) was found amongst Swirl group (Mean = 4.00) than Control group (Mean = 3.13).

Both groups agreed that course duration was "just right" in the preliminary survey (Table 7C, M1d). Interestingly, this contradicted the single respondent in the Swirl group indicating preference of needing more time to review the course during post-hoc survey (Table 7A, M3c).

Table 7C. Breakdown of survey metrics (M1, M5, M8) with higher ratings[#] from Swirl group

ID	Components	Mean ± S.D. (Control)	Mean ± S.D. (Swirl)	Comparison
M1	**Perceived course quality (preliminary survey)**	**4.13±0.25** "Agree"	**4.25±0.29** "Agree"	**-0.13±0.04** (Control < Swirl)
a	The course was well-organised and structured	4.00 "Agree"	4.50 "Strongly Agree"	Control < Swirl
b	The course was easy to follow and engaging	4.00	4.00	NA
c	The course was relevant and beneficial to my learning	4.50	4.50	NA
d	Time given to complete the course was just right	4.00	4.00	NA
M5	**Perceived ease of learning from self-guided tutorials**	**3.13±0.52** "Neutral"	**3.25±0.96** "Neutral"	**-0.13±0.44** (Control < Swirl)
a	Self-guided tutorials are an easy way for students to learn independently since hints and answer keys would be provided	4.00	4.00	NA
b	It takes a lot of effort to fully understand a new concept from a self-guided tutorial **(reverse-coded)**	2.50 "Neutral"	3.00 "Neutral"	Control < Swirl
c	I would find it easier to learn when the information is delivered in a small, bite-sized manner	4.00	4.00	NA

d	I would find it easier to learn when the information is delivered in a detailed and comprehensive manner **(reverse-coded)**	2.00	2.00	NA
M8	**Learner engagement**	**3.13±0.25** "Neutral"	**4.00±0.00** "Agree"	**-0.88±0.25** **(**Control < Swirl)
a	I was fully engaged throughout the study	3.50 "Agree"	4.00 "Agree"	Control < Swirl
b	I found it difficult to stay focused during the study (e.g. my mind sometimes wander off) **(reverse-coded)**	2.75 "Neutral"	4.00 "Disagree"	Control << Swirl

[#] Items which did not involve mean ratings or show any group differences were indicated as "NA". The approximately equal sign (≈), single comparison sign ("<" or " >"), and double signs ("<<" or ">>") denotes negligible difference, difference of less than 1-Likert point, and difference of at least 1-Likert point respectively. Reference code for mean Likert scores: "1.00-1.49" (Strongly Disagree), "1.50-2.49" (Disagree), "2.50-3.49" (Neutral), "3.50-4.49" (Agree), "4.50-5.00" (Strongly Agree). Reverse-coded: "1.00-1.49" (Strongly Agree), "1.50-2.49" (Agree), "2.50-3.49" (Neutral), "3.50-4.49" (Disagree), "4.50-5.00" (Strongly Disagree).

Table 7D. Open-ended post-hoc survey responses following survey metrics (M1, M5 and M8)

Key Points	Key Evidence
Both groups agreed visual aids, logical structure, practice questions and summary sections were most useful	"diagrams were informative, intuitive, and aesthetically pleasing", " logic and structure is really good and helpful for understanding", "simple exercises to confirm understanding also help...summary sections are very helpful in cementing concepts"
Most students favoured bite-sized delivery of information in learning new concept	"detailed and comprehensive in terms of overall content quality but delivery to be small and well-paced would be the ideal combination"
Students were concerned on the depth of teaching and learning within Swirl platform as self-guided tutorial	"elementary", "self-limiting", "structured for success", "may have limited avenues to clarify doubts beyond what was already programmed within the course"
Both groups generally agreed that Swirl is an engaging learning platform	"cultivate personal interest", "real-life example", "pictorial illustrations", "hands-on calculations"
Swirl group showed that learning hands-on with Swirl helps to familiarise with coding	"my prior familiarity with swirl tutorials made me comfortable", "interaction with R console gives the illusion that I can code...this build confidence in students that coding may be manageable"

Some factors that led to students' disengagement were mainly attributed to the content rather than the platform: "*questions were repetitive and a bit predictable.*" Other reasons raised by students for losing their focus included fatigue from trying to consolidate the newly-learnt concepts in one sitting.

Higher Survey Ratings by Control Group on Perceived Usefulness of Integrating statistics and R (M4), Learner Attitude and Motivation towards Instrument (M6) and towards Swirl (M7)

Table 7E showed that Control group preferred integrated learning of statistics and R (item M4a), while Swirl group preferred separate courses for these subjects (item M4b). Despite their preference, both groups strongly agreed on the usefulness of integrating statistics and R (Control Mean = 4.50, Swirl Mean = 5.00) to prepare them for biological science (item M4c), and to improve their performance and competency in both subjects (item M4d). Interestingly, students who took the Control tutorial felt more motivated learning with guided hints than those in Swirl tutorial. Generally, strong support were given towards formal development of Swirl tutorials (item M7b).

Table 7E. Breakdown of survey metrics (M4, M6, M7) with higher ratings[#] from Control

ID	Components	Mean ± S.D. (Control)	Mean ± S.D. (Swirl)	Comparison
M4	**Perceived usefulness on integrating statistics and R as single course**	**3.75±0.28** "Agree"	**3.50±1.29** "Agree"	**0.25±1.01** (Control > Swirl)
a	I would find it more effective to learn both statistics and R as part of a single, integrated course	3.50 "Agree"	3.00 "Neutral"	Control > Swirl
b	I would find it more effective to learn statistics and R as two separate, independent courses **(reverse-coded)**	3.00 "Neutral"	2.00 "Agree"	Control >> Swirl
c	I would find enrolling in an integrated course of statistics and R useful to prepare me for Biological Science field	4.50 "Strongly Agree"	5.00 "Strongly Agree"	Control < Swirl
d	I would think learning an integrated course of statistics and R can help to improve my competency and performance in both fields	4.00	4.00	NA
M6	**Learner attitude and motivation towards instrument**	**3.75±0.50** "Agree"	**3.25±0.96** "Neutral"	**0.50±0.46** (Control > Swirl)
a	Overall, I am satisfied with the learning outcomes that I gained from the short course on False Discovery Rate	4.00	4.00	NA
b	Overall, I consider participating in the study a well-spent investment of my free time	4.00	4.00	NA

c	I found myself becoming complacent in attempting the exercises since there were guided hints and answers provided **(reverse-coded)**	3.00 "Neutral"	2.00 "Agree"	Control >> Swirl
d	I found myself more motivated in attempting the exercises since there were helpful hints and answers to correct my understanding of the concepts	4.00 "Agree"	3.00 "Neutral"	Control >> Swirl
e	If you could turn back time and do the study all over again, would you choose to do it differently to improve your understanding towards the course?	2 "Yes" 2 "No"	3 "Yes" 3 "No"	NA
M7	**Learner attitude and motivation towards Swirl in general**	**3.75±0.35** "Agree"	**3.67±0.76** "Agree"	**0.08±0.41** (Control > wirl)
a	I would be more motivated to complete a tutorial via swirl compared to conventional tutorial in the form of worksheet	4.00 "Agree"	3.00 "Neutral"	Control >> Swirl
b	I would support the school to build more of swirl tutorials to teach modules such as statistics and R programming	4.00 "Agree"	4.50 "Strongly Agree"	Control < Swirl
c	I would be interested in getting committed myself in developing Swirl tutorials to help others develop stronger foundations in applied statistics and R programming	3.50	3.50	NA

[#] Items which did not involve mean ratings or show any group differences were indicated as "NA". The approximately equal sign (≈), single comparison sign ("<" or " >"), and double signs ("<<" or ">>") denotes negligible difference, difference of less than 1-Likert point, and difference of at least 1-Likert point respectively. Reference code for mean Likert scores: "1.00-1.49" (Strongly Disagree), "1.50-2.49" (Disagree), "2.50-3.49" (Neutral), "3.50-4.49" (Agree), "4.50-5.00" (Strongly Agree). Reverse-coded: "1.00-1.49" (Strongly Agree), "1.50-2.49" (Agree), "2.50-3.49" (Neutral), "3.50-4.49" (Disagree), "4.50-5.00" (Strongly Disagree).

A polling was also done to investigate the most important reasons for participating in the study (Figure 8).

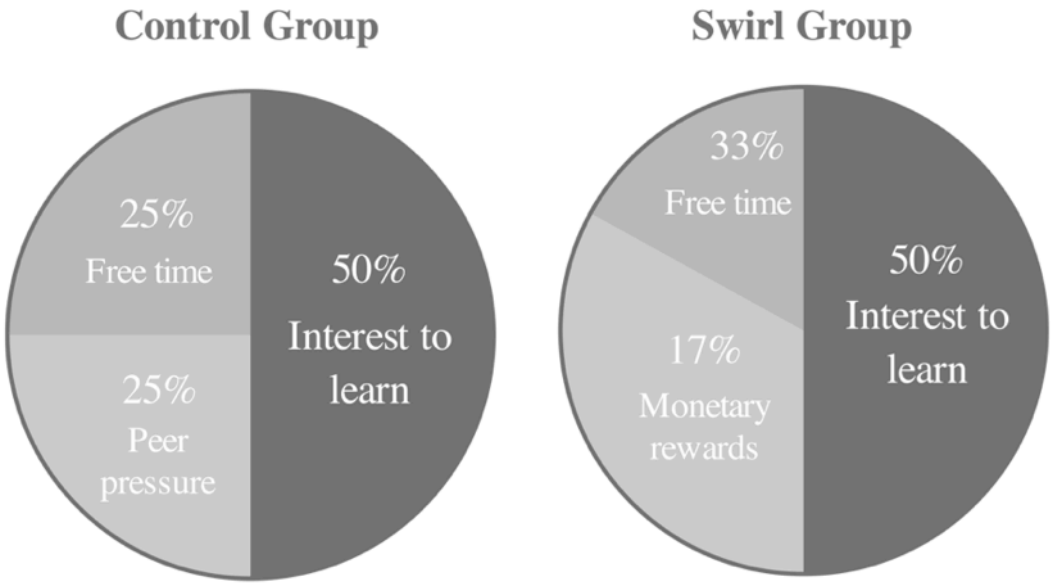


Figure 8: Both student groups shared the same top and second motivating factors (interest to learn and free time), differing only in the third factor (Peer pressure/monetary awards).

Table 7F. Open-ended post-hoc survey responses following survey metrics (M4, M6, M7)

Key Points	Key Evidence
Students were generally supportive on integration of statistics and R, placing a special emphasis on importance of learning R	"can enhance the practical application of statistical knowledge and R functions and facilitate better learning", "R is an amazing platform for statistical computation...being able to run simulations (and freely changing variables)...can visualise certain concepts better than just hearing from lecturer..."
Students felt the suitability of the course delivery was dependent on the purpose of learning, motivation and confidence in the subject matter of statistics and R, e.g. students who were not confident of either statistics and R knowledge preferred learning them separately	"Integrated course would be more suitable for those who have a strong background in either statistics or R programming", "learning statistics on its own allows more in-depth understanding of statistics", "learning 2 unfamiliar things at the same time can be daunting at the start."

4. Discussion

To answer the two-part objectives of this study, we first evaluated the course quality in delivering applied statistics concepts of FDR, followed by comprehensive analysis on Swirl platform.

4.1. Content Evaluation

The consistent positive survey impressions and improved performance of students showed that the educational content was effective in promoting learning outcomes regardless of the mode of platforms. It was likely that the incorporation of pedagogical elements (conceptual and contextual illustrations, practice-and-drill questions, guided case study) into the more abstract concepts of the course helped to enhance the conceptual understanding of students. However, inclusion of practice questions with similar level difficulty and long texts contributed to some degree of repetitions and fatigue which may compromise students' learning. More varied questions featuring different scenarios of biology-related research problems and increasing the difficulty of questions could be done to minimise predictability and promote higher cognitive engagement. Such findings further reinforced the need for more pedagogical efforts in statistics education.

4.2. Evaluation of Swirl Platform

Having attained high satisfaction in content quality as a precondition to evaluating the effectiveness of the instructional method (Kirkpatrick, 1998), we move on to investigate any differential performance and qualitative benefits attributed by the specific platforms.

Course Outcomes

We found that there was no evidence of greater learning conferred by the interactive delivery of Swirl tutorial as opposed to the conventional, passive paper-based platform. This was similar to observations wherein automated tutoring did little effect in improving students' achievement (Palocsay & Stevens, 2008). In many ways, Swirl can be likened to computer-assisted instruction (CAI). The sole use of CAI in teaching statistics (to diverse subject disciplines, including biology) had shown mixed results on several reports (Capper, 1985; Cotton, 1991).

While this showed that promoting interactive learning may not guarantee better performance, it is pertinent to recognise limitations in the metrics used. Even with normalisation to offset pretest scores, interpretations based on *nlg* (Figures 7A, 7B) and average *nlg* may not accurately portray students' achievement as it disregards the issue of losses (correct attempt in pretest, but incorrect attempt in the corresponding posttest) (Miller et al., 2010), which was clearly evident from the question-based performance (Appendix D, Tables 5A-5B). It was not possible to determine whether losses were attributed to students randomly guessing during the pretest or simply a reflection of the theory-practice gap. For more reliable comparison between disparate groups, other studies recommended the use of regression-based ANCOVA (Analysis of Covariance) which utilised pretest score as covariate to the corresponding posttest or learning gain (Weber, 2009), but this was not explored in this study due to inability

to satisfy the assumptions needed. In the case of a future study with larger sample size, further stratification based on students' pretest scores for comparison to their posttest achievement may yield interesting insights such as differential performance of low and high achievers with CAI (Owusu, Monney, Y. Appiah, & Wilmot, 2010).

Learning Effectiveness and Engagement Level

The multimedia modality of Swirl granted students the convenience of visualising the course concepts and tutorial questions while working at the solutions first-hand at the R console. This helps to enhance students' learning experiences and also promotes constructive thinking process (Nickerson, 1995), critical to fostering effective statistical learning. Swirl's bite-sized delivery of information and fast calculations enabled by R is also useful in offsetting students' negative experiences related to performing manual complex, long calculations. This reportedly help to reduce learning barriers towards statistics (D. S. Moore, 1997). By shifting the focus of the course to teaching (and learning) the principles behind why certain methodology is adopted in tackling real world biology research problems, it makes statistics learning more palatable to biology students.

The immediate, virtual feedback that students received with every incorrect attempt promoted stress-free learning upon performing trial-and-error. This continuous positive reinforcement system (Tsai, 1992) helped to facilitate positive emotional engagement amongst students. When designed effectively, Swirl's guided but stepwise session has potentially greater value over conventional tutorials that condone complacent behaviours such as decision to not attempt the questions at all and "peek" into the provided answer straightaway. Future adoption of Swirl as part of formal teaching may include graded assessment to encourage higher participation of independent learning.

Notably, we found strong concern towards missing functionality to document previous attempts in current version of Swirl course. The design of Swirl course could be further optimised by grouping into shorter courses of 10-15 minutes to prevent cognitive overload (Swirl Development Team, 2014a/b) and allow students to backtrack and review previous information. While exact matching of user input to Swirl's built-in answer may serve as stringent checks towards instilling good coding habits from the start, many felt that it could impede progress for them who were untrained in programming. Such issues could have been addressed by explicitly introducing students to the Swirl's in-built "play()" command that enable users to pause course progress such that they could perform trial-and-error calculations or even retrieve images which were embedded in the earlier parts of the course for review (N. A. Carchedi, 2014).

Swirl as a form of CAI was found to enhance learning experience of both low and high achievers group (Owusu et al., 2010): Self-paced learning allowed for drilling of fundamental knowledge at the convenience of time, privacy and feedback for the former group (Cotton, 1991) while enabling faster learning for the latter group (Capper, 1985). Although in our trials, the Swirl group completed their studies about 10-15 minutes earlier than Control group, we do not think this accurately reflect students' learning pace: Both supposedly fast and slow learners were required to complete all session tasks within the allotted 1-hour period. Awareness of a timed

study may induce artefactual effects that influenced students' attitudes to progress more quickly than proceed at their own comfortable pace.

Attitude and Motivation

Beyond the functional benefits, Swirl's potential in becoming an effective self-teaching tool was ultimately (and unsurprisingly) dependent on the attitude and motivation of each individual. A student who did not find value in the educational platform may not actually invest their cognitive efforts to learn as much as another student who believed in the effectiveness of the method (J. Moore, 2018). The high positive perception on the use of guided hints amongst Control students actually led to higher motivation in attempting the self-guided tutorial exercises, which in turn translated to higher motivation in doing interactive Swirl tutorial compared to the conventional format, and the exact opposite patterns applied for Swirl group. The Control tutorial incorporated guided hints in similar fashion but most students did not respond negatively against it, suggesting poorer learner attitudes or learning incompatibility amongst Swirl participants.

Categorical variables

Studies on the role of gender on performance in statistics or in general academic performance (Zogheib, Zogheib, & Saheli, 2015) suggest that the observation of females outperforming males in this study could simply be attributed to females exhibiting positive learner attitudes (i.e., taking notes and attentive to details). Likewise, sensing learners could have outperformed intuitive learners due to comfort with details, repetition and fact memorisation (Felder & Silverman, 1988), all of which were critical factors to perform in the timed study of the given content.

Research Limitations

While the case study approach allowed in-depth exploration of both qualitative and quantitative data on individual level, the data obtained here may not be representative to extended biology student populations. However, there may be (uncontrollable) confounding factors such as guessing of answers, misreading of questions or possibly not putting in efforts to recall previously-learnt concepts which contributed to measurement errors of the results. The 1-week time lapse between Phase 1 and 2 may also compromise the qualitative analysis of the metrics since insights gathered from post-hoc survey typically involved retrospective thinking rather than immediate reactions of students. Mismatch between self-perception and reality sometimes occur as well.

4.3. Research Implications

This study reaffirms the commonly-held notion that biology students generally lack statistical skills. Although negative perceptions such as fear of learning applied statistics should be alleviated by selecting students who have also read data science-related courses in this study, such aversion still present nonetheless. The lack of confidence in statistics coupled with unfamiliarity with the computational language R may lead to the false impression of steep learning curve due to learning of both statistics and R simultaneously. Hence, not all respondents were favourable towards

learning an integrated course combining both of these subjects despite acknowledging its usefulness. The inconsistent positioning of preference versus usefulness for integrating both subjects stemmed from the fact that students viewed appropriateness of the delivery format was subjective on the purpose of teaching. Most importantly, since there is a demand for Swirl tutorials and that supplementation of R was perceived useful for teaching applied statistics, it reinforces the need for schools to incorporate integrated learning platform of statistics and R such as Swirl.

Swirl makes an appealing learning platform for students who wish to acquire both statistical and computational proficiency for those who recognise the benefits that Swirl can deliver for their learning. The sceptical reactions pertaining to Swirl interface could be attributed more to unfamiliarity rather than total aversion to such technology. With enough exposure and practice to Swirl tutorials, students expressed that they gained comfort and confidence working with Swirl. It could also be because students had preconceived notions of what teaching resources "should" be like in biology: Many courses were delivered using traditional lecture slides and that students were accustomed to comprehensive study of textbooks for understanding concepts not taught in details during classroom (J. Moore, 2018).

This work reviews the adoption of digital technology in facilitating higher education learning in Singapore, with an eye towards contributing to the local educational research efforts (Luke, Freebody, Shun, & Gopinathan, 2005). Adoption of Swirl is hoped to offer the value of allowing the faculty to make more strategic use of class hours to address challenging aspects of the course, beyond what was programmed within the Swirl tutorial. This would in turn enhance the teaching effectiveness of applied statistics to biology students. Collectively, Swirl seems to have greater potential in serving as a practice-and-drill instrument rather than primary mode of instruction for teaching applied statistics for students. This was supported by effectiveness of CAI as supplementary teaching tool to complement statistics teaching in tertiary education (Basturk, 2005). A longitudinal cohort study could be explored to evaluate the effectiveness of Swirl in improving statistical and computational R proficiency amongst biology students.

5. Conclusion

Proper education and training to develop statistical and computational literacy amongst biology students is crucial to prepare them for the data-heavy research settings in the near future. This study serves as empirical work for evaluating the suitability of Swirl as an instructional platform for teaching standardised applied statistics content customized for biologists. Triangulating quantitative performance of students with qualitative opinions interfacing with Swirl, the study showed that learning in Swirl did not yield superior results compared to conventional medium, which suppressed its potential as stand-alone teaching instrument and raised the question of whether adaptation of courses into Swirl was a worthwhile effort. This study provides preliminary evidence on the use of Swirl platform as particularly useful for teaching applied form of statistics while immersing learners in the native R programmatic environment, by tapping on data analysis functions of R and by facilitating an interactive, multimedia learning environment that encouraged active engagement amongst students. The study also identified some limitations and suggestions that may guide future development of Swirl courses. The low cost of

development involved due to the open-source nature of Swirl makes it an attractive alternative instrument for delivering lessons over commercial statistical platforms in the context of university teaching. The long-term adoption of Swirl as a supplementary teaching tool could possibly help to develop independent learning, enhance statistical thinking and improve data processing skills in biology students living in the 21st century. Last but not least, successful adoption of Swirl in statistics education is preconditioned on student buy-in to Swirl approach in helping them learn both applied statistics and R.

References

Basturk, R. (2005). The Effectiveness of Computer-Assisted Instruction in Teaching Introductory Statistics. Educational Technology & Society, 8(170-178).

Becker RA, C. J., Wilks AR. (1988). The New S Language: a programming environment for data analysis and graphics. . *Chapman and Hall*.

Bowyer, J., & Chambers, L. (2017). Evaluating blended learning: Bringing the elements together. *Research Matters: A Cambridge Assessment Publication, 23*(17-26).

Capper, J., & Copple, C. . (1985). *Computer use in education: Research review and instructional implications.* Washington, DC: Center for Research into Practice.

Carchedi, N. A. (2014). *TURNING THE R CONSOLE INTO AN INTERACTIVE LEARNING ENVIRONMENT WITH SWIRL.* (Master of Science), Johns Hopkins University. Retrieved from https://jscholarship.library.jhu.edu/bitstream/handle/1774.2/37310/CARCHEDI-THESIS-2014.pdf?sequence=1&isAllowed=y

Carey, M. A., & Papin, J. A. (2018). Ten simple rules for biologists learning to program. *PLoS Computational Biology, 14*(1), e1005871. doi:10.1371/journal.pcbi.1005871

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

Cotton, K. (1991). Computer-Assisted Instruction. *Northwest Regional Educational Library: School Improvement Research Series, 10*.

Felder, R. M., & Silverman, L. K. (1988). Learning and Teaching Styles In Engineering Education. *Engineering Education, 78*(7), 674-681.

Felder, R. M., & Soloman, B. A. (1997). Index of Learning Styles Questionnaire. Retrieved from https://www.webtools.ncsu.edu/learningstyles/

Feser, J., Vasaly, H., & Herrera, J. (2013). On the Edge of Mathematics and Biology Integration: Improving Quantitative Skills in Undergraduate Biology Education. *CBE Life Sciences Education, 12*(2), 124-128. doi:10.1187/cbe.13-03-0057

Gore, A. D., Kadam, Y. R., Chavan, P. V., & Dhumale, G. B. (2012). Application of biostatistics in research by teaching faculty and final-year postgraduate students in colleges of modern medicine: A cross-sectional study. *International Journal of Applied and Basic Medical Research, 2*(1), 11-16. doi:10.4103/2229-516X.96792

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a Conceptual Framework for Mixed-Method Evaluation Designs. *Educational Evaluation and Policy Analysis, 11*(3), 255-274. doi:10.3102/01623737011003255

Hake, R. (1988). Interactive-engagement versus traditional methods: A six- thousand-student survey of mechanics test data for introductory physics courses. *Am. J. Phys, 66*, 64-74.

Kirkpatrick, D. (1998). *Evaluating training programs: the four levels.* (2 ed.). San Francisco, CA:: Berrett-Koehler.

Luke, A., Freebody, P., Shun, L., & Gopinathan, S. (2005). Towards Research-based Innovation and Reform: Singapore schooling in transition. *Asia Pacific Journal of Education, 25*(1), 5-28. doi:10.1080/02188790500032467

Makarevitch, I., Frechette, C., & Wiatros, N. (2015). Authentic Research Experience and "Big Data" Analysis in the Classroom: Maize Response to Abiotic Stress. *CBE Life Sciences Education, 14*(3), ar27. doi:10.1187/cbe.15-04-0081

Miller, K., Lasry, N., Reshef, O., Dowd, J. E., Araujo, I., & Mazur, E. (2010). *Losing it: The Influence of Losses on Individuals'* Normalized *Gains.* . Paper presented at the Physics Education Research Conference Portland, Oregon.

Moore, D. S. (1997). New Pedagogy and New Content: The Case of Statistics. International Statistical Review. *65*(123-137). doi:10.1111/j.1751-5823.1997.tb00390.x

Moore, J. (2018). Efficacy of Multimedia Learning Modules as Preparation for Lecture-Based Tutorials in Electromagnetism. *Education Sciences, 8*(1), 23.

Muenchen, R. A. (2014). The popularity of data analysis software. Retrieved from http://r4stats.com/articles/popularity/

Nickerson, R. S. (1995). Can Technology Help Teach for Understanding? In Software Goes to School: Teaching for Understanding with New Technologies.

Owusu, K., Monney, K., Y. Appiah, J., & Wilmot, E. (2010). *Effects of computer-assisted instruction on performance of senior high school biology students in Ghana* (Vol. 55).

Palocsay, S. W., & Stevens, S. P. (2008). A Study of the Effectiveness of Web-Based Homework in Teaching Undergraduate Business Statistics. *Decision Sciences Journal of Innovative Education, 6*(2), 213-232. doi:doi:10.1111/j.1540-4609.2008.00167.x
Swirl. (2014). Swirl Course Repository. Retrieved from https://github.com/swirldev/swirl_courses

Tsai, Y. M.-H. (1992). The effects of different systems of positive reinforcement on computer-based learning *Retrospective Theses and Dissertations*.

Weber, E. (2009). Quantifying Student Learning: How to Analyze Assessment Data. *The Bulletin of the Ecological Society of America, 90*(4), 501-511. doi:doi:10.1890/0012-9623-90.4.501

Zogheib, S., Zogheib, B., & Saheli, A. E. (2015). University Students' Achievement in Mathematics: The Role of Student's Gender, Instructor's Gender, Educational Level, and Experience*The Mathematics Educator, Vol. 16, No.1, 77- 92*.

Contact email: wilsongoh@ntu.edu.sg

Appendix A

Control Tutorial

Appendix B
Pretest Questionnaires
Correct answers highlighted in grey

Question 1

Which of the following statement is true?

A. beta is the probability of getting a true positive, that is rejecting the null hypothesis w not true
B. power is the probability of identifying a true negative, that is not rejecting the null hy when it is true
C. (1-alpha) is the probability of getting a false negative, that is not rejecting the null hyp when it is not true
D. alpha is the probability of getting a false positive, that is rejecting the null hypothes is true
E. I do not know this concept

Question 2

Which of the following formula represents false discovery rate?

A. $\frac{False\ Positives}{All\ Predictions} x\ 100\%$

B. $\frac{False\ Positives}{All\ Positive\ Predictions} x\ 100\%$

C. $\frac{False\ Positives}{All\ Negative\ Predictions} x\ 100\%$

D. $\frac{False\ Positives}{All\ Positive\ Events} x\ 100\%$

E. $\frac{False\ Positives}{All\ Negative\ Events} x\ 100\%$

F. I do not know this concept

Question 3

You are the Game Master for a gene discovery game in SBS Camp.
For Round 1, you allocate an equal number of positive and negative gene signals.
For Round 2, you decide to increase the number of positive signals available.

So how will you compare the false discovery rate (FDR) in Round 1 and Round 2?

A. Both rounds have the same FDR
B. Round 1 has a lower FDR than in Round 2.
C. Round 1 has a higher FDR than in Round 2.
D. Not enough information to decide
E. I do not know this concept

Question 5

Out of the 10
false negative
positive outc

A. 22%
B. 26%
C. 77%
D. 83%
E. I do no

Question 4

The results of a court trial are as such:

- **There is an equal chance that the suspects present are likely to be guilty as they are innocent**
- **For suspects who are really guilty, Judge declares "Guilty" 80% of the time ("True Positive")**
- **For suspects who are actually innocent, Judge declares "Guilty" 5% of the time ("False Positive")**

So what is the chance of wrongly accusing a suspect given all the judgements declared as "Guilty"? Round off your answer to nearest whole number.

A. 2.5% (FP only)
B. 5.0% (FPR)
C. 6.0% (FDR)
D. 20.0% (FP/False Pred (FP+FN))
E. I do not know this concept

Posttest Questionnaires

Correct answers highlighted in grey

Question 1

A recent personalised genomics test has an alpha of 5% and power of 70%. How do you interpret these specifications?

A. The chance of getting a false positive, a false negative and true positive is 5%, 30%, and 70%
B. The chance of getting a false positive, a false negative and true positive is 5%, 70%, and 30%
C. The chance of getting a false positive, a false negative and true positive is 30%, 5%, and 70%
D. The chance of getting a false positive, a false negative and true positive is 30%, 70%, and 5%
E. I do not know this concept

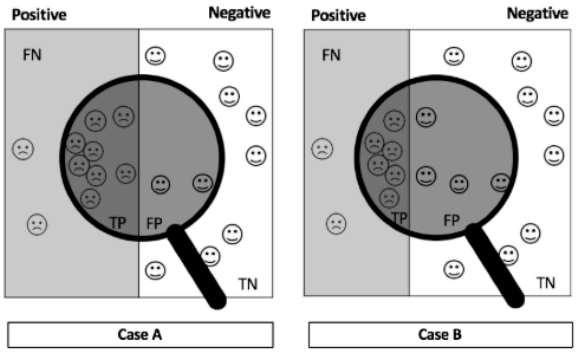
Question 2

Which of the following components are NOT required to determine the false discovery rate of a study?

A. The number of true positives
B. The total number of positive predictions
C. The ratio between positive and negative events
D. The ratio between positive and negative predictions
E. I do not know this concept

Question 3

Case A and Case B represent the outcomes for two different experimental settings.



Which of the following statement is most likely to be true?

A. Case A has a higher false discovery rate than case B (should be lower)
B. Case A has a lower false discovery rate than case B
C. Case A has a higher false positive rate than case B (should be lower)
D. Case A has a lower ratio of positive to negative events than case B (should be higher)
E. I do not know this concept

Question 4

Typically, a drug candidate will have 50% chance of failure and 50% chance of success during the intermediate stage of clinical trial. Given this clinical trial has an alpha of 10% and power of 70%, what is the false discovery rate? Round off your answer to the nearest whole number.

A. 5%
B. 10%
C. **13%**
D. 25%
E. I do not know this concept

Question 5

The following table summarises the epidemiology results for 200 residents on HIV prevalence:

Possible Test Outcomes	State of truth	
	HIV-free Residents	HIV-positive Residents
Outcomes declared "Positive"	5	40
Outcomes declared "Negative"	95	60

What is the false positive rate for this study? Round off to nearest whole number.

A. 3%
B. **5%**
C. 8%
D. 11%
E. I do not know this concept

Appendix C
Preliminary Survey

1) Course content *

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
The course was well-organised and structured	○	○	○	○	○
The course was easy to follow and engaging	○	○	○	○	○
The course was relevant and beneficial to my learning	○	○	○	○	○
Time given to complete the course was just right	○	○	○	○	○

2) What aspects of this course were most useful or valuable or interesting for you? E.g. specific theme/exercises/concepts *

Your answer

3) How would you improve the course? Feel free to feedback about anything *

Your answer

4) Why did you choose to sign up for this study? *

- ☐ Free time
- ☐ Interest to learn
- ☐ Peer pressure
- ☐ Monetary rewards
- ☐ Other:

Post-hoc Survey

School of Biological Sciences
Post Hoc Survey (Phase 2)

M2: Perceived difficulty of the course (***omitted from results due to confusion)***

- **Rank the concepts based on:**
  - **Increasing level of difficulty**: *1 (being least difficult) to 5 (most difficult)*
  - **Increasing level of abstractness**: *1 (being least abstract) to 5 (most abstract)*

- 5 Main Topics Covered:
  - **Construction of Research Hypotheses**
  - **Predictions**
    - Positive and Negative Predictions
    - False Predictions
    - True Predictions
    - Summary of Test Outcomes
  - **The Concept of False Discovery Rate**
    - Mini-exercises
  - **FDR as a context-dependent metric**
    - Ratio of positive to negative events
    - Summary of FDR Calculations
  - **DIfference between FDR vs. FPR**
    - Practice case calculations
    - Summary of all important points

Open-ended Questions for Posthoc Survey

(Q1) To further improve your learning, how do you think the exercises should be designed in terms of the format or delivery, the level of difficulty and the quantity of questions? Do you think having a guided approach (e.g. direct hints) help you in your learning or do you feel otherwise? Please outline your thought processes and justify your preference.

(Q2) With reference to your answers in (2a) and (2b), what are the factors that go through your mind when you have to estimate your expected performance for the trial quizzes? Elaborate these in details. You may want to consider factors such as but not limited to the motivations to participate in the study, discomfort with numbers/calculation cases, optimal study time, level of difficulty of the quiz, the degree of understanding towards the content, etc.

(Q3) How long do you think is a comfortable duration for you to review the content in order to fully understand the concepts pertaining to the "False Discovery Rate" (as covered in the trial)? Please specify the time duration in minutes/hours and justify why. In deciding what is a comfortable duration, you may consider your usual study habits and the amount of time you spent to learn a new concept related to statistics/mathematics.

(Q4) How long do you think is a comfortable duration for you to complete pre-test and post-test quizzes respectively? Please specify the time duration in minutes/hours and justify why. There were a total of 5 questions in pre-test and post-test quiz respectively.

(Q5) Were you able to complete the session earlier ahead of other participants? If not, did you get distracted or affected when others actually completed the study earlier? How did they affect you? Please specify your reactions and elaborate on the things which went through your mind during the incident. Could you have skipped reading certain texts, misread the questions in the quizzes or probably guessed the answers? Or, did you stay focused and were not bothered at all since you really wanted to understand the concepts and questions

(Q6) Elaborate on your thought processes on why a particular format of course delivery (integrated vs independent) would help you gain a better understanding of the two fields of interest (Statistics and R). You may may want to consider factors such as but not limited to the breadth and width of the content, the presentation, the speed of learning, the ease or difficulty of learning, the relevance to real life applications, your personal study habits, etc.

(Q7) Explain the reasons that motivate your opinions on the relative ease of learning from self-guided tutorial. You may want to consider factors such as but not limited to the scale of use, the quality of explanations and facilitation, the flexibility of learning beyond classroom, the general reaction of SBS students to this idea, the interactivity, and etc.

(In Swirl Group)

(Q8) Were there any difficulties, challenges or discomfort that you face (technical or otherwise) while loading any of the study materials such as the swirl course and questionnaires? Please specify a few examples or instances (if any) and explain how it has affected your learning.

(In Control grp)

(Q8) Were there any difficulties, challenges or discomfort that you face (technical or otherwise) while reviewing/learning the study materials such as the course in PDF format and questionnaires? Please specify a few examples or instances (if any) and explain how it has affected your learning.

(Q9) What aspects of swirl made you more motivated or less motivated to complete a tutorial using swirl compared to a conventional worksheet? Please justify your answer.

Motivation for Reattempt

If you could turn back time and do the study all over again, would you choose to do it differently to improve your understanding towards the course?

a. Yes
b. No

(Q10-Yes) If you have chosen "Yes", explain why you may decide to attempt it differently and also what would you have done differently? You may want to consider certain strategies or practices that you usually adopt to help you study for your lessons in school, such as but not limited to reviewing the content again, redo the exercises, write out/make notes, talking about the learning points, etc.

(Q10-No) If you have chosen "No", explain why you may decide not to. Is it because you have prepared or performed sufficiently well for the lesson? You may want to consider factors such as but not limited to the lack of motivation to perform, disinterest to learn the concept, eager desire to quickly rush through the content and leave, discomfort with statistics, etc.

(Q11) Besides having pictorial illustrations and summary pointers in the course, were there any other factors which kept you engaged on the course throughout? You can want to consider factors such as but not limited to your typical attention span, real case examples, hands-on calculation practices, personal interest in the lesson itself, motivation to do well, and etc.

(Q12) If you found yourself losing focus, when did you actually start to switch off or feel bored along the way? Please indicate specific period of time when you started to get distracted, and also the sections of the lesson which you find boring/confusing. If this does not apply to you, simply state "NA"

(Q13) Could you think of any factors that made you switch off mentally along the way? Does it concern your typical attention span, the format of the delivery, the length of information, the timing of the study, the level of difficulty of the content, and etc.

Appendix D: Supplementary Data

Table 2. Rationales for learning objectives (LO) associated with each pair of pretest and posttest questions

ID	Learning objectives	Rationale for each question pair
LO1	**Evaluates understanding of Topic 2:** Able to interpret the meanings of different probability outcomes such as alpha, beta, and power	**Question 1** The aim of the pretest was to assess students' theoretical knowledge on the definitions of parameters associated with hypothesis testing. To refresh students' memory, explanations on the four parameters (alpha, beta, power, and 1-alpha) were included under Topic 2 of the tutorial. The posttest version served as a slight variation of pre-test, where an understanding towards all three definitions (alpha, beta, power) were tested by matching the right numerical value onto each parameter.
LO2	**Evaluates understanding of Topic 2, 3, 4:** Recall the parameters used in calculating False Discovery Rate (FDR)	**Question 2** The pretest was designed to identify students with prior exposure to the mathematical formula associated with FDR. The posttest served as an indirect version to assess students' understanding towards the three key parameters (alpha, power, ratio of positive to negative events) introduced in the tutorial. The notion of "events" and "predictions" were only introduced explicitly in the tutorial. Being able to differentiate these two terminologies showed that students paid attention to the details in the tutorial. Getting correct answers for both pretest and posttest implied good understanding towards the calculations involved to solve FDR.
LO3	**Evaluates understanding of Topic 3, 4:** Understand that FDR is a context-dependent metric, dependent on the ratio of positive to negative events	**Question 3** In the pretest version, students were evaluated on the ability to recognise that varying the amount of positive and negative signals or data had an impact on increasing FDR of an experiment. The same objective was tested on posttest, but framed using an illustrative diagram which was introduced in the tutorial. Being able to interpret the diagram demonstrated students' understanding of the context-dependency nature of FDR from the tutorial itself.
LO4	**Evaluates understanding of Topic 3, 4, 5:** Know how to apply and solve for False Discovery Rate and False Positive Rate given a problem statement/case study	**Question 4** In order not to penalise students who did not have prior knowledge of FDR and yet investigate those who have prior understanding, the pretest was expressed in terms of simple probability question. The posttest meanwhile was framed in a manner similar to tutorial's case study to assess students' ability to solve for FDR. This required students to understand the concept of FDR and apply the formula.
		Question 5 The main objective was to test students' ability to differentiate the concept of FDR from False Positive Rate (FPR).

Table 3. List of qualitative survey metrics (M1-M8) analysed through a series of survey items introduced as part of preliminary and posthoc study

ID	Qualitative Survey Metrics
Preliminary survey	
M1	**Perceived course quality:**
a.	Structure (relates to M2, M5)
b.	Engagement (relates to M8)
c.	Relevance (relates to M4, M6, M7)
d.	Duration (relates to M3)
Post-hoc survey	
M2	**Perceived difficulty of course content:**
a.	Topics introduced
b.	Explanations
c.	Exercises/practice questions
d.	Assessment (pretest and posttest)
M3	**Perceived ability:**
a.	Pretest
b.	Posttest
c.	Course duration
d.	Quiz duration
e.	Distraction
f.	Monetary rewards
M4	**Perceived usefulness of integrating statistics and R as a single course:**
a-b.	Learning preference for integrated course vs independent course
c.	Usefulness of integrated course for biological field
d.	Usefulness of integrated course for improvement in statistics and R
M5	**Perceived ease of use of self-guided tutorials**:
a.	Provision of hints and solutions
b.	Effort to learn new concept
c-d.	Learning preference for bite-sized vs comprehensive release
M6	**Learner attitude and motivation towards specific study of the**
a.	**intervention:**
b.	Learning satisfaction
c-d.	Participation satisfaction
e	Provision of hints and solutions
	Reattempt of study
M7	**Learner attitude and motivation towards Swirl in general:**
a.	Completion of Control vs Swirl tutorial
b.	Support for formal development of Swirl
c.	Support for informal development of Swirl
M8	**Learner engagement:**
a-b.	Engagement vs disengagement