The Effects of Cue Utilization and Cognitive Load in the Detection of Phishing Emails

George Nasser[1][0000-0003-3704-5598], Ben W. Morrison[1][0000-0002-5026-4675], Piers Bayl-Smith[2][0000-0001-8014-0633], Ronnie Taib[3], Michael Gayed[2], and Mark W Wiggins[2][0000-0002-6422-9475]

[1] Charles Sturt University, Bathurst, New South Wales, Australia
[2] Macquarie University, New South Wales, Australia
[3] Data 61, CSIRO Australia

Abstract. Phishing emails represent a major threat to online information security. While the prevailing research is focused on users' susceptibility, few studies have considered the decision making strategies that account for skilled detection. One relevant facet of decision making is cue utilization, where users retrieve feature-event associations stored in long-term memory. High degrees of cue utilization help reduce the demands placed on working memory (i.e., cognitive load), and invariably improve decision performance (i.e., the information-reduction hypothesis in expert performance). The current study explored the effect of cue utilization and cognitive load when detecting phishing emails. A total of 50 undergraduate students completed: (1) a rail control task and; (2) a phishing detection task. A cue utilization assessment battery (EXPERTise 2.0) then classified participants with either higher or lower cue utilization. As expected, higher cue utilization was associated with a greater likelihood of detecting phishing emails. However, variation in cognitive load had no effect on phishing detection, nor was there an interaction between cue utilization and cognitive load. These findings have implications for our understanding of cognitive mechanisms that underpin the detection of phishing emails and the role of factors beyond the information-reduction hypothesis.

Keywords: Phishing Emails, Cue Utilization, Decision Making, Cognitive Load.

1 Introduction

1.1 The Phishing Problem

Despite the best efforts of cybersecurity companies, the average email user must still respond to approximately sixteen phishing emails a month (Symantec, 2018). In large

George Nasser, Ben Morrison, Piers Bayl-Smith, Ronnie Taib, Michael Gayed, and Mark Wiggins. The Effects of Cue Utilization and Cognitive Load in the Detection of Phishing Emails, Proceedings of AsiaUSEC'20, Financial Cryptography and Data Security 2019 (FC). February 14, 2020 Kota Kinabalu, Sabah, Malaysia Springer, 2020.

organizations, this can amount to thousands of such emails arriving in employees' inbox each year, each with the potential to seriously disrupt productivity and damage reputation (Vergelis, Shcherbakova, & Sidorina, 2019).

Over the last decade, a broad range of approaches have explored the reasons why certain users are more susceptible than others to cyberattacks (Williams, Hinds, & Joinson, 2018). However, little research has explored the strategies that users adopt when making successful decisions about an email's legitimacy, such as the skilled use of cue-based associations in memory (Johnston & Morrison, 2016; Morrison & Morrison, 2015; Morrison, Morrison, Morton, & Harris, 2013; Morrison, Wiggins, Tyler, & Bond, 2013; Wiggins, 2015; Wiggins & O'Hare, 2006). In the context of phishing detection, cue utilization is presumed to involve an individual's capacity to recognize features within an email that signal (often rapidly and unconsciously) an attempt to deceive.

When faced with a complex diagnostic task, expert decision makers automatically recognize features that cue patterns from memory, and which 'trigger' the rapid retrieval of a plausible response (e.g., a process of recognition-primed decision-making; Klein, 1993). The timely recognition of these patterns will invariably reduce the demands placed on working memory, with attentional resources being deployed selectively to task-relevant features in the environment (Haider, & Frensch, 1999). Thus, when decision-makers possess a greater capacity for cue utilization, they have additional cognitive resources to respond to incoming demands (Brouwers et al., 2017; Ericsson & Lehmann, 1996). This implies that greater levels of cue utilization may 'buffer' against the usually deleterious impacts of increased cognitive load by reducing the amount of information in the environment that needs to be processed. Such a strategy may be particularly useful in the context of phishing detection, since it is a process often engaged in tandem with other complex, resource-demanding tasks. Consistent with an information-reduction hypothesis (Haider, & Frensch, 1999), behavior associated with relatively higher cue utilization is likely to be associated with higher levels of task performance under increasing cognitive load (e.g., that arising from an increase in task complexity).

1.2 Study Aims

The current study was designed to test the impact of cue utilization and cognitive load on email users' ability to detect phishing emails under conditions of low, moderate, and high cognitive load. Cognitive load was manipulated using a simplified, simulated rail control task as part of a dual-task paradigm, during which participants were categorizing emails as 'trustworthy' or 'suspicious'. Behavior associated with the utilization of cues was assessed using the Expert Intensive Skills Evaluation (EXPERTise 2.0) assessment tool (Wiggins, Loveday, & Auton, 2015).

EXPERTise 2.0 comprises five tasks, each of which is designed to evaluate behavior associated with the application of cue-based associations in memory. Since cues are task-specific, an edition of the tool was developed that incorporated features associated with phishing emails. EXPERTise 2.0 has been used previously to delineate behavior associated with higher and lower cue utilization in fields as diverse a pediatric intensive care (Loveday, Wiggins, Searle, Festa, & Schell, 2013), software engineering

(Loveday, Wiggins, & Searle, 2014), and football coaching (Yee, Wiggins, Auton, Warry, & Cklamovski, 2019).

1.3 Hypotheses

Hypothesis one. Email users' performance on the phishing detection task would decline with increasing levels of cognitive load (low, moderate, and high).

Hypothesis two. Higher cue utilization, as determined by participants' performance on EXPERTise 2.0, would be associated with greater accuracy in detecting phishing emails.

Hypothesis three. An interaction would be evident between cue utilization and cognitive load where higher cue utilization would be associated with relatively smaller reductions in performance as cognitive increased.

2 Method

2.1 Participants

Fifty adult students (35 females, 15 males) were recruited as a sample of convenience from Macquarie University's SONA research recruitment system. The participants ranged in age from 18 to 45 years (M_{age} = 20.44, SD_{age} = 4.38). The mean age for males was 21.07 (*SD* = 4.21) and the mean age for females was 20.17 (*SD* = 4.48). All participants were naïve to the context of professional cybersecurity.

2.2 Materials

Expert Intensive Skills Evaluation (EXPERTise) Program Version 2.0. EXPERTise is an online platform that consists of a battery of tests, each based on empirical investigations of cue utilization. The different tasks have been individually and collectively associated with differences in performance at an operational level (Loveday, Wiggins, Harris, O'Hare, & Smith, 2014; Loveday et al., 2013). Test–retest reliability ($\kappa = .59$, $p < .05$) has been demonstrated with power control operators at six month intervals (Loveday et al., 2014) and with audiologists at 18 month intervals (Watkinson, Bristow, Auton, McMahon, & Wiggins, 2018).

Successful cue utilization is measured by individuals' ability to identify critical features quickly from an array (Feature Identification Task; FIT), categorize accurately, situations based on key features (Feature Recognition Task; FRT), quickly associate features and events in memory (Feature Association Task; FAT), discriminate between relevant features (Feature Discrimination Task; FDT), and prioritize the acquisition of information during problem resolution (Feature Prioritization Task; FPT) (Wiggins, 2014).

As cue-based associations are highly contextualized, domain-specific phishing stimuli were created for each of the EXPERTise tasks. For instance, most tasks presented users with images of emails, some of which held features that may be predictive of phishing threats (e.g., sender’s address, typographical errors, prompt for action, etc.). The stimuli were reviewed by a subject-matter expert in the field of cyber-security.

Rail Control Task. In the rail control task, participants manage the movement of trains using a simplified simulation (example screenshot seen in Figure 1; Brouwers et al., 2017). The task consisted of four green horizontal lines that represent the railway track. Various intersections occur between these lines (depicted by white portions displayed on the tracks), with the option to change the track onto a new line. Trains are depicted as red lines and assigned either an odd or even three-digit code (e.g. 555, 888). The first and third train line run from right to left, while the second and fourth train line run from left to right. The goal is to ensure that even-numbered trains terminate on even terminals and odd-numbered trains terminate at odd terminals. To correct the programmed route of the train, participants must select the ‘Change’ icon located above each train line. The direction of the track also appears under this icon. All trains progressed at the same speed with participants having seven seconds to decide whether or not to re-route the train. Participants engaged three separate conditions (each comprising 21 trains), which varied in the number of train tracks being controlled at any one time. The ordering was linear, whereby cognitive load progressively increased throughout the task, which commenced with the top two train lines (low condition), then the top three train lines (moderate condition), and finally all four train lines (high condition).

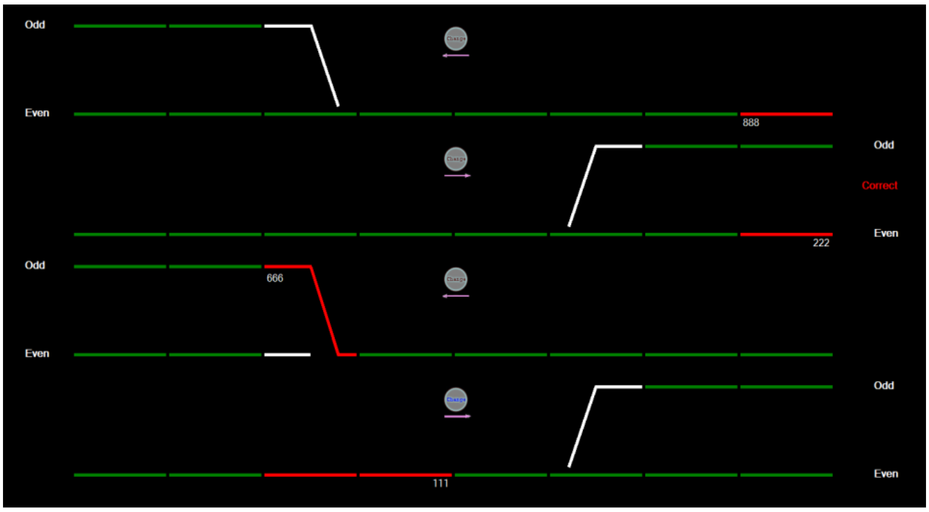

Fig. 1. The simulated rail control task display for the high load condition.

Phishing Detection Task. Phishing emails were taken from Berkeley PhishTank and modified to an Australian context. The emails included 45 phishing emails and 45 legitimate emails. Participants responded to the emails at their own pace, and the task finished when all three conditions of the rail control task had been completed. The participants were required to respond to the emails, which varied in legitimacy as either: Trustworthy or Suspicious. After participants made a decision, they selected the Next button at the bottom of the screen, which opened a new email. This task was administered through a web-based email client simulator that was programmed to randomize the presentation of emails for each participant.

2.3 Apparatus

Two LG® IPSTM EA53s Desktop Monitors (24’’ display size; LG Display, Yeongdeungpo District, Seoul, South Korea) were used in this experiment. The monitors connected to two Lenovo® IdeacentreTM 310S-07F (Lenovo, Quarry Bay, Hong Kong) workstations each equipped with 8GB of RAM and running a Windows 10 operating system. Each computer connected to a Microsoft® Optical wired mouse (Redmond, Washington, USA) that enabled participants to complete the tasks. The screen on the left of the participant operated the rail control task and the computer on the right of the participant operated the phishing detection task. EXPERTise operated through the same computer as the phishing detection task.

2.4 Procedure

The participants completed the study in individual sessions of one hour. The monitor positioned on the left of participants operated the rail control task. Participants were taken through a practice simulation of the low load condition. Participants were then informed that the task would progressively increase in complexity, starting with two active train lines, then increasing to three active train lines and finishing with all four train lines active.

The computer screen positioned to the right of the participant operated the phishing email detection task. Participants were instructed that they were to correctly identify the incoming emails as either ‘Trustworthy’ or ‘Suspicious’. Once they had indicated a response, a ‘Next’ button would appear at the bottom of the screen. Participants were instructed not to attend to the rail control task at the expense of the phishing detection task, and that equal attention levels should be directed to both tasks. After completing this task, participants were instructed to complete EXPERTise on the same computer. Each of the five tasks (FIT, FAT, FDT, FPT and FAT) were accompanied by a detailed description of the task requirements on the initial screen.

3 Results

3.1 Data Reduction

Consistent with the process outlined by Wiggins, Griffin, and Brouwers (2019), EXPERTise raw scores were standardized to *z*-scores and aggregated together to create a total EXPERTise score for each participant. In preparation for a comparison of performance, a median split categorized participants as demonstrating either relatively higher or lower levels of cue utilization (Wiggins et al., 2019).

3.2 Cue Utilization, Cognitive Load, and Phishing detection

A 2x3 mixed-repeated ANOVA, incorporating two categories of cue utilization (high and low) as a between-groups variable, and three levels of cognitive load (low, moderate, and high) as a within-groups variable examined whether any significant difference existed in performance on the phishing detection task. The decision performance values on the phishing detection task were taken from the efficiency scores, which considered the number of correctly identified phishing emails as a proportion of the total number of emails to which participants responded.

The ANOVA results revealed no main effect for cognitive load on the phishing detection task, $F(2, 48) = 2.84$, $p = .06$ (two-tailed), $\eta p^2 = .06$. As the result was in the opposite direction to our hypothesis, a decision was made not to correct the *p*-value for one-tail. This means that increases in cognitive load had no adverse impact on participants' performance during the phishing detection task and hypothesis one was not supported.

The results revealed a statistically significant main effect for cue utilization, $F(1, 48) = 4.15$, $p = .02$ (one-tailed), $\eta p^2 = .08$ (medium effect), with higher cue utilization (M= .54, $SE = .03$) associated with greater accuracy on the phishing detection task in comparison to participants with lower cue utilization ($M = .46$, $SE = .03$) (see Figure 2). This result supported hypothesis two.

As participant could respond to the emails at their own pace (and therefore, potentially manage their cognitive load via their rate of response on the phishing email task), an independent *t*-test was used to test for a difference in the number of emails reviewed between the higher and lower cue utilization groups. The results did not reveal a statistically significant difference, $t(48) = -.31$, $p = .761$. The higher cue utilization group responded to a mean of 40.80 *(SD* = 14.60) emails and the low cue group responded to a mean of 39.50 *(SD* = 15.87) emails.

Hypothesis three explored whether an interaction existed between cue utilization and cognitive load, and performance on the phishing detection task. However, the results failed to reveal any statistically significant interaction between cue utilization and cognitive load, $F(2, 48) = 0.25$, $p = .391$, $\eta_p{}^2 = .005$. Therefore, there were no differences in accuracy based on cue utilization and accounting for differences in cognitive load (see Figure 2).

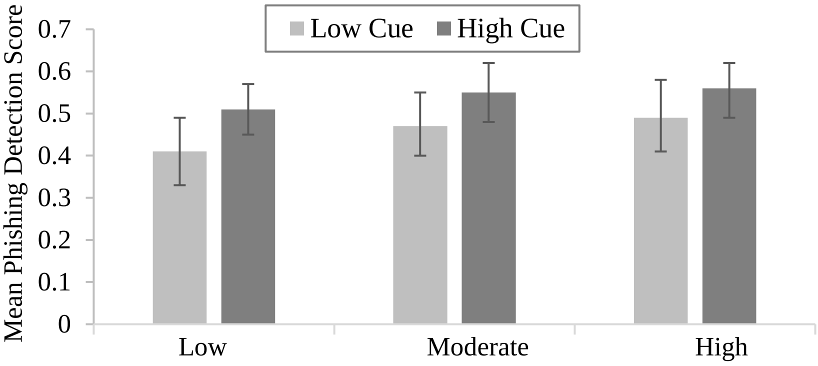

Fig. 2. The mean performance on the phishing detection task for high and low cue utilization groups across the three levels of cognitive load (error bars are 95% CI).

4 Discussion

The current study tested the effects of cue utilization and cognitive load on the detection of phishing emails. The purpose was to investigate the decision making strategies of skilled email users when formulating accurate assessments as to the legitimacy of an email.

4.1 Cognitive Load

Contrary to the hypothesis, email users' performance on the phishing detection task was not adversely impacted by increasing levels of cognitive load (low, moderate, and high). Instead, the results indicated a trend whereby performance on the phishing task increased with each additional level of cognitive load. The observed trend may be due to a practice effect on the rail control task (Falleti, Maruff, Collie, & Darby, 2006). All participants began the task with the low load condition and progressively increased to the high condition. The initial exposure to the low load condition is likely to have familiarized participants with the task and naturally improved their performance on the subsequent conditions, despite increases in task demands. Furthermore, the improved performance suggests that the cognitive load task might not have been sufficiently challenging to disrupt participants' cognitive resources. Instead, the task may have increased participants arousal to a level that improved decision performance (Jackson, Kleitman, & Aidman, 2014).

4.2 Cue Utilization

Consistent with the hypothesis, higher cue utilization was associated with greater accuracy in discriminating phishing from non-phishing emails. This suggests that behavior associated with the utilization of cue-based associations in memory is associated with an increased likelihood in detecting phishing emails while undertaking a concurrent task.

These results are broadly consistent with previous research where the detection of phishing emails is presumed to be dependent upon the capacity to identify key features, such as spelling and email addresses that signify the possibility that an email is untrustworthy (Williams et al., 2018).

4.3 Cue Utilization, Cognitive Load, and Phishing Detection

Hypothesis three was not supported insofar as no interaction was evident between cue utilization and cognitive load. The result suggests that performance on the phishing email task was not due to differences in the capacity of participants with higher cue utilization to better manage the cognitive load associated with the rail control task, but was due possibly to an inherent capability to either recognize or maintain an awareness that enabled the discrimination of phishing from non-phishing emails (Brouwers et al., 2017; Loveday et al., 2014).

These results, in particular, have implications for an explanation of phishing email detection based on an information-reduction hypothesis (Haider & Frensch, 1999). Indeed, it suggests that alternative theoretical perspectives may be involved, including the possibility that respondents are making judgements based on a template or prototype of trustworthy emails, and/or the detection of phishing emails is dependent upon a heightened level of awareness for features that characterize emails that are untrustworthy.

4.4 Limitations

A notable limitation of the current work was the use of an equal number of phishing and legitimate emails in the Phishing Detection Task. In reality, most users will receive far fewer phishing emails than legitimate ones. As such, the ratio adopted may be problematic when considering a truth-default theory in human communication (Levine, 2014). However, achieving realistic base-rates in an experimental design is challenging, as it would require participants to assess a significantly greater number of emails overall. Future studies may wish to address this limitation, as well as other experimental artefacts that may impact the generalizability of the findings to real-world environments.

4.5 Conclusion

The current study provides an exploration of the cognitive processes associated with decision making in cybersecurity. We found an improvement in discrimination based on participants' utilization of cues associated with the detection of phishing emails. These results provide support for the proposition that the detection of phishing emails is based on the recognition of specific features that reflect untrustworthy emails. The use of cue-based training interventions has proven effective in other domains (e.g., Morrison, Wiggins, & Morrison, 2018), and these findings imply potential value in their adoption in the cyber-security domain.

References

1. Brouwers, S., Wiggins, M. W., Griffin, B., Helton, W. S., & O'Hare, D. (2017). The role of cue utilisation in reducing the workload in a train control task. *Ergonomics, 60*(11), 1500-1515.
2. Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology, 47*(1), 273-305. doi:10.1146/annurev.psych.47.1.273
3. Falleti, M. G., Maruff, P., Collie, A., & Darby, D. G. (2006). Practice effects associated with the repeated assessment of cognitive function using the CogState battery at 10-minute, one week and one month test-retest intervals. *Journal of Clinical and Experimental Neuropsychology, 28*(7), 1095-1112. doi:10.1080/13803390500205718
4. Haider, H., & Frensch, P. A. (1999). Information reduction during skill acquisition: The influence of task instruction. *Journal of Experimental Psychology: Applied, 5*(2), 129-151. doi: 10.1037/1076-898X.5.2.129
5. Jackson, S. A., Kleitman, S., & Aidman, E. (2014). Low cognitive load and reduced arousal impede practice effects on executive functioning, metacognitive confidence and decision making. *Public Library of Science One, 9*(12), e115689-e115689. doi:10.1371/journal.pone.0115689
6. Johnston, D., & Morrison, B. W. (2016). The application of naturalistic decision-making techniques to explore cue use in rugby league playmakers. *Journal of Cognitive Engineering and Decision Making, 10*(4), 391-410. doi:10.1177/1555343416662181
7. Klein, G. (1993). *A recognition-primed decision (RPD) model of rapid decision making Decision making in action: Models and methods.* (pp. 138-147). Westport: Ablex.
8. Levine, T. R. (2014). Truth-default theory: A theory of human deception and deception detection. *Journal of Language and Social Psychology, 33*(4), 378-392. doi:10.1177/0261927X14535916
9. Loveday, T., Wiggins, M. W., Harris, J. M., O'Hare, D., & Smith, N. (2013). An objective approach to identifying diagnostic expertise among power system controllers. *Human Factors: The Journal of Human Factors and Ergonomics Society, 55*(1), 90-107. doi:10.1177/0018720812450911
10. Loveday, T., Wiggins, M. W., & Searle, B. (2014). Cue utilization and broad indicators of workplace expertise. *Journal of Cognitive Engineering and Decision Making, 8*(1), 98-113. doi: 10.1177/1555343413497019
11. Loveday, T., Wiggins, M. W., Searle, B. J., Festa, M., & Schell, D. (2013). The capability of static and dynamic features to distinguish competent from genuinely expert practitioners in pediatric diagnosis. *Human Factors, 55*(1), 125-137. doi:10.1177/0018720812448475
12. Morrison, B. W., & Morrison, N. M. V. (2015). Diagnostic cues in major crime investigation. In M. W. Wiggins & T. Loveday (Eds.), *Diagnostic Expertise in Organizational Environments* (pp. 91-98). Surrey, England: Ashgate Publishing.
13. Morrison, B. W., Morrison, N. M. V., Morton, J., & Harris, J. (2013). Using critical-cue inventories to advance virtual patient technologies in psychological assessment. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration* (OzCHI '13), Haifeng Shen, Ross Smith, Jeni Paay,

Paul Calder, and Theodor Wyeld (Eds.). ACM, New York, NY, USA, 531-534. DOI: http://dx.doi.org/10.1145/2541016.2541085

14. Morrison, B. W., Wiggins, M. W., Bond, N. W., & Tyler, M. D. (2013). Measuring relative cue strength as a means of validating an inventory of expert offender profiling cues. *Journal of Cognitive Engineering and Decision Making, 7*(2), 211-226. doi:10.1177/1555343412459192
15. Morrison, B. W., Wiggins, M. W., & Morrison, N. (2018). Utility of expert cue exposure as a mechanism to improve decision-making performance among novice criminal investigators. *Journal of Cognitive Engineering and Decision Making, 12*(2), 99-111. https://doi.org/10.1177/1555343417746570
16. Symantec. (2018). Internet Security Threat Report. Retrieved from https://www.symantec.com/content/dam/symantec/docs/reports/istr-23-2018-en.pdf
17. Vergelis, M., Shcherbakova, T., & Sidorina, T. (2019). Spam and phishing in 2018. Retrieved from https://securelist.com/spam-and-phishing-in-2018/89701/
18. Watkinson, J., Bristow, G., Auton, J., McMahon, C. M., & Wiggins, M. W. (2018). Postgraduate training in audiology improves clinicians’ audiology-related cue utilisation. *International Journal of Audiology, 57*(9), 681-687.
19. Wiggins, M. (2014). The role of cue utilisation and adaptive interface design in the management of skilled performance in operations control. *Theoretical Issues in Ergonomics Science, 15*(3), 282-292. doi:10.1080/1463922X.2012.724725
20. Wiggins, M. W. (2015). Cues in diagnostic reasoning. In M. W. Wiggins, & T. Loveday (Eds.), *Diagnostic Expertise in Organizational Environments* (pp. 1-11). Surrey, England: Ashgate Publishing.
21. Wiggins, M. W., Griffin, B., & Brouwers, S. (2019). The potential role of context-related exposure in explaining differences in water safety cue utilization. *Human Factors, 61*(5), 825-838. https://doi.org/10.1177/0018720818814299
22. Wiggins, M. W., Loveday, T., & Auton, J. (2015). *EXPERT Intensive Skills Evaluation* (EXPERTise 2.0) Test: Macquarie University, Sydney, AUS.
23. Wiggins, M., & O'Hare, D. (2006). Applications of micro-simulation in cognitive skills development. In W. Karwowski (Ed.), *International encyclopedia of ergonomics and human factors* (2nd ed., pp. 3262-3267). United Kingdom: Taylor & Francis.
24. Williams, E. J., Hinds, J., & Joinson, A. N. (2018). Exploring susceptibility to phishing in the workplace. *International Journal of Human-Computer Studies, 120*, 1-13. doi:10.1016/j.ijhcs.2018.06.004
25. Yee, D. J., Wiggins, M. W., Auton, J. C., Warry, G., & Cklamovski, P. (2019). Technical and social cue utilization in expert football coaches. *Sport, Exercise, and Performance Psychology*. Advance online publication. https://doi.org/10.1037/spy0000170