Carrigan, A. J., Charlton, A., Foucar, E., Wiggins, M. W., Georgiou, A., Palmeri, T. J., & Curby, K. M. (2022). The Role of Cue-Based Strategies in Skilled Diagnosis Among Pathologists. *Human Factors*, *64*(7), 1154–1167.

Access to the published version:
https://doi.org/10.1177/0018720821990160

Topic choice: Cognition

The role of cue-based strategies in skilled diagnosis amongst pathologists

Carrigan, A.J[1,2], Charlton, A.,[3], Foucar, E.[4], Wiggins, M.W.[1,2] , Georgiou, A[5]., Palmeri, T.[6], & Curby, K.M.[1,2].

[1]Department of Psychology, Macquarie University, Sydney, Australia.

[2]Centre for Elite Performance, Expertise & Training, Macquarie University, Sydney, Australia.

[3]Department of Histopathology, Auckland City Hospital, and Department of Molecular Medicine and Pathology, University of Auckland, New Zealand.

[4]Department of Pathology, University of New Mexico, Albuquerque, USA.

[5]Centre for Health Systems and Safety Research, Macquarie University, Sydney, Australia.

[6]Department of Psychology, Vanderbilt University, Nashville, United States.

Author Contact Details:

Dr Ann Carrigan,
Centre for Elite Performance, Expertise, and Training
Department of Psychology
Macquarie University,
North Ryde, NSW 2109
Australia.
Email: ann.carrigan@mq.edu.au

Acknowledgements
For data collection, we acknowledge the support of The Royal College of Pathologists of Australasia and the Royal College of Pathologists of Australasia Quality Assurance Program. This work is supported by an ARC Discovery Project grant awarded to M.W.W (CI) K.M.C., A.G., and T.P. (DP5056000).

Word Count: 5434

Reference Word Count: 1791

Running Head: Cue based strategies in Pathology

Precis

Expertise in pathology involves accurate, reliable, and timely responses. Using a sample of histopathologists, we established differences in cue utilization that, independent of cases read per year, explained differences in their capability to accurately recognize abnormalities in static images. This has implications for the development of pathological skills in practice.

Objective: This research was designed to test whether behavioral indicators of pathology-related cue utilisation were associated with performance on a diagnostic task.

Background: Across many domains, including pathology, successful diagnosis depends on pattern recognition that is supported by associations in memory in the form of cues. Previous studies have focused on the specific information or knowledge on which medical image expertise relies. The target in this study is the more general ability to identify and interpret relevant information.

Method: Data were collected from fifty-four histopathologists in both conference and online settings. The participants completed a pathology edition of the Expert Intensive Skills Evaluation 2.0 (EXPERTise 2.0) to establish behavioral indicators of context-related cue utilization. They also completed a separate diagnostic task designed to examine related diagnostic skills.

Results: Behavioral indicators of higher or lower cue utilisation were based on the participants' performance across five tasks. Accounting for the number of cases reported per year, higher cue utilisation was associated with greater accuracy on the diagnostic task. A post hoc analysis suggested that higher cue utilisation may be associated with a greater capacity to recognise low prevalence cases.

Conclusion: This study provides support for the role of cue utilization in the development and maintenance of skilled diagnosis amongst pathologists.

Application: Pathologist training needs to be structured to ensure that learners have the opportunity to form cue-based strategies and associations in memory, especially for less commonly seen diseases.

Keywords: Medical Image Perception, Cue Utilization, Cognitive Load, Prevalence.

Histopathology is a medical pathology sub-specialty, the role of which involves the examination of tissue to correctly identify tissue features that are normal, normal variants and clinically significant pathological processes. Correct interpretation can require integration of morphological findings with clinical history. Pathologists visually inspect histopathological slides using a light microscope allowing for the interpretation and classification of diseases. Typically, tissue biopsy or resections referred to histopathologists from a variety of medical specialists have already been subjected to diagnostic tests to confirm pathology, thereby creating an environment where disease prevalence tends to be high. The diagnosis rendered by the histopathologist, is considered the gold standard and often the starting point for determining the patient's treatment and prognosis. This is especially critical for breast cancer, where early diagnosis results in a 97%, 5-year or more survival rate (Australian Institute of Health and Welfare (AIHW), 2009).

Errors within diagnostic medicine more broadly are estimated at > 10% (Goldman et al., 1983; Hoff, 2013; Kirch & Schafii, 1996; Shojania et al., 2003). In pathology, errors in cancer diagnosis are reported at up to 12% (Raab et al., 2005). The discrepancy between breast cancer staging among breast histopathologists is at a high of 40% (Elmore, et al., 2015). Elmore and colleagues (2015) conducted a large study in the US with 115 practicing pathologists and showed a 24.7% disagreement rate among pathologists interpreting breast biopsies. This rate was higher for denser breasts, among pathologists who interpreted lower weekly case volumes, worked in smaller practices or non-academic settings.

Unfortunately, errors that occur in pathology and diagnostic medicine can of course have grave consequences. For example, the under-interpretation of an atypical cancer may delay the required treatments (false negatives or misses) and conversely an overdiagnosis of normal tissue may lead to unnecessary invasive treatments (false positives or false alarms). These issues are complicated as the base rate for histopathology is non-normal. In practice, there are programs in place which record a pathologist's performance and provide the necessary feedback. For example, the Royal College of Pathologists of Australasia (RCPA) offers a Quality Assurance Program (QAP) in which pathologists must participate to maintain their registration. Given that the pathology diagnosis is final and definitive, it is important to explain the underlying processes involved in these crucial, diagnostic decisions and therefore, how errors might be prevented.

Nodine and Krupinski (1998), and more recently, Drew et al. (2013), proposed that successful diagnosis results when medical image specialists apply their fine-tuned perceptual and cognitive skills to rapidly process a layout/scene globally to reach their decision. Consistent with this account, abnormalities in a display or image can be detected rapidly, following a brief glance (Brennan et al., 2018; Carmody et al., 1981; Carrigan et al., 2018; Charness, 1996; Evans et al., 2013). Amongst cytologists, performance is above chance at detecting abnormalities in micrographs of cervical smears presented for 250ms (Evans, et al., 2013). Similarly, expert radiologists and pathologists fixate more rapidly than trainees on an abnormality and do so using fewer visual saccades (Kundel & La Follette, 1972; Kundel & Nodine, 1975; Kundel et al., 1978; Krupinski et al., 2013).

Krupinski, et al. (2013) characterised the scanning patterns of resident pathologists (pathologists in training) throughout their training and demonstrated their search patterns changed from a less efficient strategy (scanning around the entire visual field) to a more efficient strategy (targeted search) with experience. Further, visual expertise is associated with less time fixating on diagnostically irrelevant and non-diagnostic regions (Brunye et al., 2014; Krupinski et al., 2013). These capabilities suggest that experts extract global properties of an image rapidly and develop a finely tuned perceptual representation that almost instantly supports the relationship between visual stimuli and a diagnosis. This ability is also likely supported by patterns in memory gained through past experience that are triggered when presented with the stimulus (Brunye et al., 2014).

According to Nodine & Mello-Thoms (2010), when an expert considers a case, ‘features’ that are extracted during the initial glance are compared against a template or ‘pattern’ in memory. This process tends to be rapid, non-conscious, and domain-specific and is dependent upon a repertoire of feature-event/object associations in memory that serve as ‘cues’ (Brunswick, 1955; Croskerry, 2009; Klein, 1989; Wiggins, 2014; 2020; Wiggins, Brouwers, Davies, et al., 2014). Carrigan et al. (2019) demonstrated that compared with non-radiologists, radiologists perceive features that are especially diagnostic to the case (subtle nodule in a chest radiograph) as more salient, lending support to the notion that feature-based cues are integral for successful performance.

Norman (2005) demonstrated that the greater and more diverse an expert’s memory for patterns, the faster their capacity to match a visual percept against a stored pattern. For example, a pathologist reviewing a breast biopsy slide might associate the destruction of normal, lobular architecture of the breast glands (feature) with the likelihood of cancer (event). As additional cases are reviewed, a repertoire of feature-event associations is accrued, enabling more precise discrimination in future encounters. As these associations become integrated to form patterns, cognitive processes become automated, reducing the demand on cognitive resources, including working memory, while maintaining accuracy and efficiency (e.g., see Curby & Gauthier, 2007; Curby et al., 2009 for evidence of domain-specific visual working memory capabilities afforded by expertise).

Indirect evidence of the role of cue utilization in supporting diagnostic performance is evident in experts’ extraction of a limited number of diagnostic features and their relatively rapid and efficient formulation of diagnoses in comparison to non-experts (e.g., Brennan et al., 2018; Carrigan et al., 2019; Carrigan et al., 2018; Evans et al., 2013; Krupinski et al., 2013; Kundel & La Follete, 1972). However, the strength of these associations in memory is likely to depend on the frequency with which they are activated. For example, patterns of features that are reviewed less frequently in practice may decay through inactivity, while a lack of exposure to specific cases may result in patterns of features that are imprecise and/or liable to inappropriate activation, resulting in diagnostic errors. Since the activation of patterns is reliant upon cues in memory, differences in the capacity to utilize cues within a given context is likely to be an important cause of differences in diagnostic accuracy.

Ideally, assessments of cue-based associations in memory would involve the identification of a universal catalogue of features that could constitute a benchmark for a 'test of existence'. However, the specific feature-event associations that experts acquire are shaped by their individual experience. Moreover, visual features may be more or less salient based on individual differences in visual detection skills, so that, despite similar levels of experience, experts are unlikely to share identical cue-based relationships in memory.

Despite differences in the specific feature-event/object relationships in memory, it remains possible for practitioners to achieve similar levels of performance, particularly in the context of diagnosis. Characterizing the nature of this behavior allows more general inferences to be drawn concerning the application or utilization of cue-based associations from memory. Relatively higher cue utilization can be inferred based on the capacity of a practitioner to: (1) rapidly identify areas of concern, (2) accurately recognize classes or categories of features, (3) greater and more rapidly differentiate associations between features, (4) discriminate relevant from less relevant features during problem-solving, and (5) demonstrate a more explicit prioritization of the acquisition of features during problem orientation (Wiggins, 2014; 2020).

Comparative behavioral assessments of cue utilization have successfully differentiated diagnostic performance across a range of non-medical contexts including rail control (Sturman et al., 2019), aviation piloting (Wiggins et al., 2014; Schriver et al., 2008), air-traffic control (Falkland & Wiggins, 2019), water safety (Wiggins et al, 2019), software engineering (Loveday et al., 2014), and electricity network power system control (Loveday et al., 2013a). In medical domains, including radiology, (Carrigan et al., 2020), paediatrics (Loveday et al., 2013b), and anaesthesiology (Crane et al., 2018), behaviour reflecting higher cue utilization is associated with greater levels of accuracy in both direct and indirect measures of diagnostic performance.

The primary goal of the present study was to test the contribution of cue utilization in the diagnostic performance of histopathologists, accounting for self-reported experience. In the case of histopathology, diagnosis requires that a clinician match features associated with a pattern of cells against a repertoire of features in memory. These features include tissue architecture, cell arrangement, cytoplasmic and nuclear shape, size, colour and density.

A greater and more discriminating repertoire of feature-event associations, reflecting a greater capacity for the utilization of cues, is likely to enable a more accurate and rapid diagnosis. Therefore, it was hypothesized that relatively higher cue utilization in the context of histopathology would be associated with relatively greater accuracy on a simulated diagnostic task, independent of experience.

METHOD

Participants

Pilot data were collected from eight histopathologists across a range of experience. Experimental data were then collected from 54 participants (32 female) who participated in a conference setting (n = 35), or online (n = 19) and who were familiar with reporting histopathological specimens. The mean age of histopathologists in the experimental group was 46 years (SD = 12 years). Forty participants were qualified pathologists and 14 were residents (pathologists in training). Participants reported an average experience as a pathologist of 14 years (*SD* = 12, range 1-52), having recently read between zero (one was retired) to 14000 cases per year (*M* = 3663, *SD* = 3554). Three participants were left-handed, all reported normal or corrected-to-normal vision, and all were naïve to the purposes of the experiment. In return for participation, participants were offered the chance to win an iPad. The study was approved by Macquarie University Human Research Ethics Committee (Medical Sciences).

Measures

The participants completed a 45-min task which included: (1) a demographic survey, (2) an assessment of cue utilization using the Expert Intensive Skills Evaluation platform (EXPERTise 2.0; Wiggins et al., 2015), and (3) a computer-based static image recognition task that provided an independent measure of diagnostic performance.

(1) Demographic Survey: Covariates

The participants were asked to indicate their age, sex, handedness, qualifications, their number of years of experience in pathology, the number of cases performed per day, the number of cases performed per year, and how frequently they played video games or a musical instrument. If they played a musical instrument, they were asked to indicate their perceived level of proficiency. These data were recorded, as there is evidence to suggest that video game players perform at a higher level on attentional tasks (e.g., Cain et al., 2014), and that playing a musical instrument may enhance visuo-spatial abilities (Tanner Boyd et al., 2008). Using five-point Likert scales, participants were also asked about their energy levels at time of experimentation, their confidence in their role, the extent to which they considered that ‘pattern recognition is an inherent skill which cannot be taught’, and their self-rated performance as a pathologist.

(2) Cue utilization

Cue utilization was assessed using the Expert Intensive Skills Evaluation platform (EXPERTise 2.0; Wiggins et al., 2015). EXPERTise 2.0. is an online assessment tool that incorporates five tasks that are designed to assess behaviour that corresponds to the utilisation of cues in a specific domain. Guided by subject-matter experts, stimuli are developed for the various tasks that represent a specific domain or context.

In the Feature Identification Task (FIT), participants are asked to identify, as quickly as possible, an abnormality or area of concern. Participants responded to one practice and 15 randomised trials, 10 of which incorporated a single area of abnormality while the remainder were normal. The anatomical location of the abnormal images comprised the breast (2), lung (2), gastrointestinal tract (4), liver (2) and salivary gland (1). The response was made under free-viewing conditions and no feedback was provided. Notably, no assumptions are made about what specific features experts should use to select a region. In the FIT, higher cue utilisation is generally associated with lower mean response latency, reflecting the rapid and targeted identification of an area, rather than an exhaustive search (Loveday et al., 2013a) (See Figure 1).

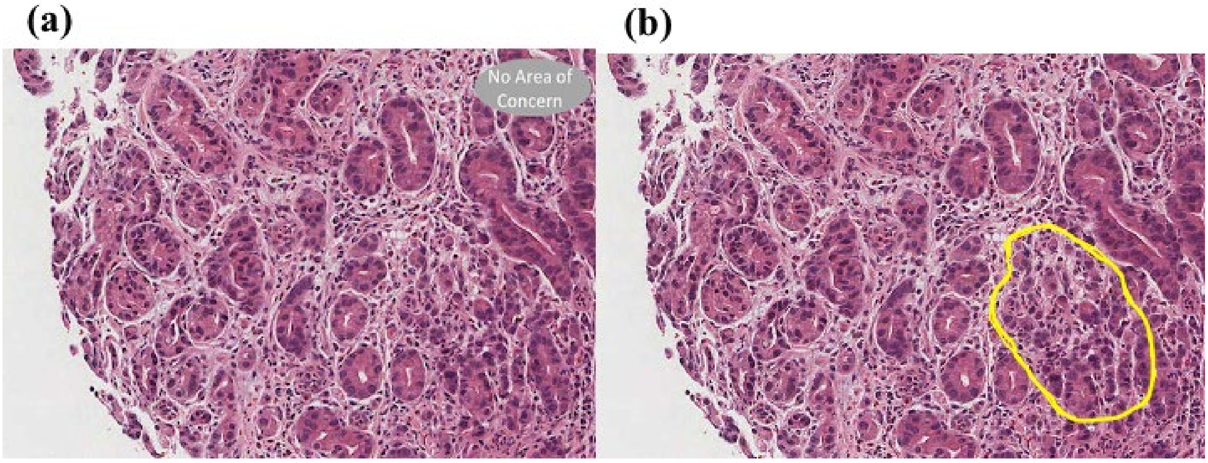

Figure 1: Exemplar image from the first of five tasks (the Feature Identification Task) from the Pathology edition of EXPERTise 2.0. Panel (a) depicts an image presented to the participants. Panel (b) depicts the image with the abnormal area outlined in yellow, representing gastric adenocarcinoma.

In the Feature Recognition Task (FRT), participants are asked to classify stimuli following a short exposure. One practice and 20 randomized, experimental scenarios were displayed for four seconds. Each scenario contained an abnormality, in response to which, participants were asked to select one from five multiple choice options to categorise the abnormality (Benign Neoplasm, Malignant Neoplasm, Reactive/Inflammatory Process, Developmental, Metabolic). No feedback was provided. In the FRT, higher cue utilization is generally associated with greater accuracy (Wiggins & O'Hare, 2003).

In the Feature Association Task (FAT), participants are asked to assess the relatedness of pairs of pathological, text-based, task-related stimuli using a six-point Likert Scale. The relatedness of the texts within each trial varied and was reviewed by a subject matter expert. Over one practice trial and 15 randomized experimental trials, the two terms were presented sequentially for 1500ms after which participants were asked to rate the relatedness of the features on a six-point Likert scale (from 1 = ‘Extremely unrelated’ to 6 = ‘Extremely related’). For example, ‘Granulation Tissue’ (feature) might be followed by ‘Tuberculosis’ (event). Seven pairs were less likely to be related, (e.g., ‘Amyloid’ (feature) followed by ‘Grass’ (event) and 8 were more likely to be related in practice (e.g., ‘Bubbly Cytoplasm’ (feature) followed by ‘Chordoma’ (event)). In the FAT, higher cue utilisation is typically associated with greater variance in the perceived relatedness of the terms as a function of response latency (Morrison et al., 2013).

In the Feature Discrimination Task (FDT), participants are presented with a short, written description of a problem-oriented scenario. On the basis of this information, they are asked to select a response from four possible options based on their typical response (i.e., ‘What would be your first response in this situation?’). Having selected a response, participants are presented with a list of features that were incorporated in the scenario and are asked to rate, with reference to the decision, the perceived importance of the features using a 10-point Likert scale (from 1 = ‘Not important at all’ to 10 = ‘Extremely important’). In the FDT, higher cue utilization is associated with greater variance across feature-relevance ratings (Weiss & Shanteau 2003; Pauley et al., 2009), where participants are more likely to select either ‘Not important at all’ (1) or ‘Extremely important’(10).

For the Feature Prioritization task (FPT), participants are required to prioritise information to solve a histopathology-related problem. The histopathologists were presented with incomplete, descriptive scenarios (‘You are doing routine reporting. As quickly as possible, access the information below that you feel is necessary to decide on your response.’), and were provided a list of 14 information screens (feature cues), varying in relevance and presented randomly in drop-down tabs from which feature-related information could be accessed (e.g., ‘Clinical Impression’, ‘Low power microscopic image’). Participants with higher cue utilisation tend to access the information in a less sequential manner, reflecting their ability to prioritize the information perceived as important to their goal. Meanwhile, participants with lower levels of cue utilisation tend to access the sub-menus in a more sequential manner, accessing features as they are listed (Wiggins & O'Hare, 1995). Higher cue utilization is associated with a lower frequency of pairs of features accessed in sequence, calculated as a proportion of the total frequency of pairs of features accessed during the scenario (Wiggins & O'Hare, 1995) (see Table 1).

Table 1: Summary of five tasks within EXPERTise 2.0.

Task	*Cognitive Process Examined*	*Task Description*	*Measure*	*Validity/Reliability*
FIT	Identification of predictive features	Identify, as quickly as possible the area of concern.	Response latency	Loveday et al. (2013) Wiggins, Brouwers, Davies, and Loveday (2014)
FRT	Identification of predictive features	Select the category of abnormality, displayed for four seconds.	Accuracy	Loveday et al. (2013)
FAT	Feature-event relationships in memory	Rate the strength of perceived associations between the feature-event	Variance divided by response latency	Morrison, Wiggins, Bond, and Tyler (2013)
FDT	Discrimination between predictive features	Rate the relative importance of features during a task-related problem-solving process.	Variance	Pauley, O'Hare, and Wiggins (2009)
FPT	Prioritization of feature-event relationships	Acquire task-related information to solve a problem-solving process.	Ratio of sequential to non-sequential menus accessed	Wiggins and O'Hare (1995) Wiggins, Stevens, Howard, Henley, and O'Hare (2002)

Note: EXPERTise = EXPERT Intensive Skills Evaluation; FIT = feature identification task; FRT = feature recognition task; FAT = feature association task; FDT = feature discrimination task; FPT = feature prioritization task.

The test-retest reliability of each of the tasks that comprise EXPERTise 2.0 has been demonstrated in the context of audiology (Watkinson et al., 2018), while test-test reliability of classifications based on performance aggregated across the five tasks has been demonstrated in the context of electrical transmission power control (Loveday et al., 2013). The construct validity of the EXPERTise 2.0 classifications has been demonstrated in radiology (Carrigan et al., 2020), transmission and distribution power control (Loveday et al., 2013a; Sturman et al., 2019), software engineering (Loveday et al., 2014), and aviation (Wiggins et al., 2014), while predictive validity has been demonstrated in audiology (Watkinson et al., 2018).

(3) Diagnostic Performance

Diagnostic performance was assessed using a static image recognition task. The stimuli consisted of one practice and 31 test images from various anatomical locations (e.g., breast, kidney, skin). Each image included five unique multiple-choice responses developed by two subject matter experts. Four of the images demonstrated pathologies not included in the multiple-choice options, where the correct answer was ‘none of the above’. The participants were instructed: ‘As quickly as possible, please select the most correct answer from the options below.’ The de-identified, images were provided by the subject matter experts (See Figure 2).

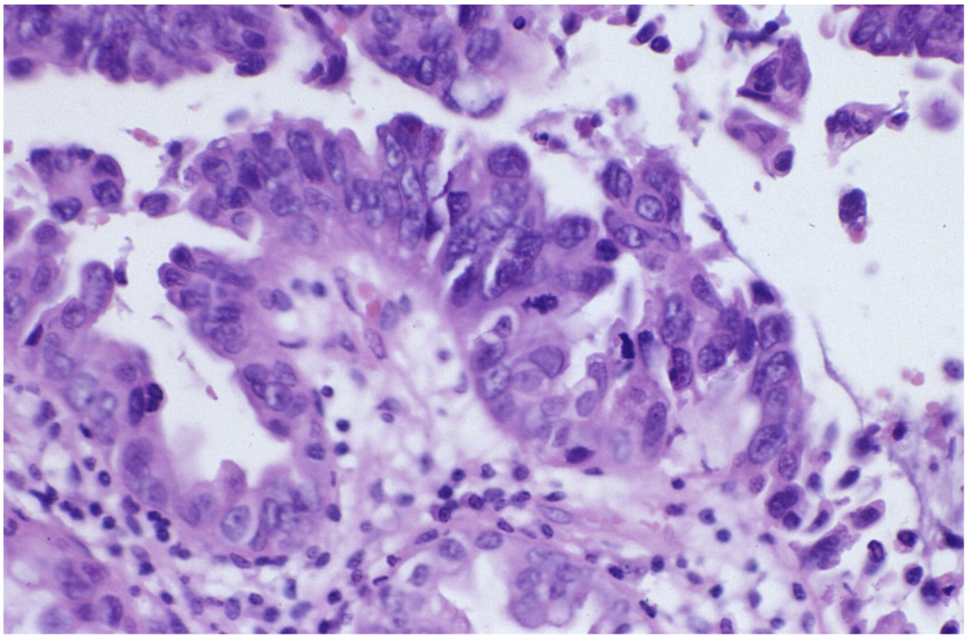

Figure 2: Exemplar of an image from the static image recognition stimuli set.
Answer = ‘c’.

Procedure

The research was conducted in accordance with the American Psychological Association Code of Ethics and was approved by Macquarie University Human Research Ethics Committee. Informed consent was obtained from all participants. EXPERTise 2.0 and the static image recognition tasks were counterbalanced across participants. All of the tasks within EXPERTise 2.0 were blocked in a set order and the scenarios were randomized within task. After completing a series of demographic questions, each task began with a practice trial followed by the EXPERTise 2.0 or static image recognition tasks.

Results

Preliminary Analysis

All statistical analyses were performed using IBM Statistical Software for the Social Sciences (SPSS Version 25). Outliers defined as +/- 2 SD were removed from the EXPERTise raw data (4 participants) leaving 50 participants for the main analysis. As the participants' raw scores for each of the five EXPERTise 2.0 tasks were on different scales, these were standardized (z scores) and then aggregated across the five tasks. The dependent variable for the image interpretation task was accuracy (% correct).

Nineteen participants completed the study online at their convenience. The study instructions included requests to perform the study on a laptop or desktop (no tablets or cell phones), and at a time when they were least likely to be interrupted. An independent samples *t* -test was performed and showed no differences in performance between those participants who completed the study in a conference setting or online; $t(52) = -0.3$, $p = .77$.

1) Demographic Survey: Covariates

A series of correlations were conducted between the demographic variables and accuracy on the static image recognition task. As expected, positive Pearson's correlations were evident between accuracy on the static image recognition task and both the number of cases performed per day; $r(52)= .39$, $p = .004$, and the number of cases performed per year; $r(52)= .37$, $p =.006$. This suggests that the diagnostic performance measure has construct validity. A positive Spearman's correlation was evident between self-rated confidence in their role as a pathologist and performance on the image interpretation task; $r_s\,(52) = 34$, $p = .01$. There was no correlation between self-reported years of experience and performance on the static image recognition task; $r(52) = .23$, $p = .097$. There were no other statistically significant correlations evident ($p > .05$). Since level of confidence is a subjective measure, it was not included in the main analysis. As the number of cases/day and per year are essentially the same measure, only the number of cases per year was included as a potential explanatory variable (covariate) in the main analysis.

(2) Cue utilization

Cue utilization was established based on the participant's performance across the five tasks included in EXPERTise 2.0. Previous research involving the five EXPERTise tasks suggests that performance can be discriminated at two levels (higher, lower) that reflect differences in behaviour during tasks that demand the utilization of cues (e.g., Brouwers et al., 2017; Carrigan et al., 2020; Falkland & Wiggins, 2019). Consequently, a *k*-means cluster analysis was employed to delineate two groups based on the standardised scores for the five EXPERTise tasks.

An inspection of the centroids that emerged following the cluster analysis was used to distinguish higher from lower performance and ensure that this pattern was uniform across the five tasks. Consistent with the hypothesis, the cluster analysis in the present study reflected uniformly, a pattern of centroids that distinguished higher from lower performance. In comparison to lower cue utilization, higher cue utilization is characterized by faster responses in the FIT, greater accuracy in the FRT, greater variance as a proportion of response latency in the FAT, greater variance in the FDT and a lower proportion of features accessed in the sequence in which they were presented in the FPT.

Eighteen participants comprised a group whose behavior was consistent with lower cue utilization, while 32 participants comprised a group whose behavior reflected higher cue utilization (see Table 2). This delineation is intended to take into account non-linear changes that occur in the rate at which cue-based associations are identified and retained in memory. There were proportionally more trainees assigned to the lower cue utilization group (see

Table 3). Four participants did not complete all of the EXPERTise 2.0 tasks, so were not assigned a cluster membership. Notably, cue utilization in this case, is a relative, not absolute, measure. Therefore, those participants assigned to Cluster 1 show patterns consistent with *relatively* lower cue utilization compared while those participants assigned to Cluster 2 showed patterns consistent with *relatively* higher cue utilization.

Table 2: Centroids for the standardized scores for each of the EXPERTise 2.0 five tasks distributed across the two groups that were delineated by a *k*-means cluster analysis.

	Cluster 1: Lower Cue Utilization Group (*n* = 18)	Cluster 2: Higher Cue Utilization Group (*n* = 32)
Task	*Centroid*	*Centroid*
Feature Identification	.92616	-.51780
Feature Recognition	-.80742	.50503
Feature Association	-.72342	.39166
Feature Discrimination	-.11779	.11648
Feature Prioritization	.23245	-.06519

Table 3: Frequency of trainee and qualified pathologists per cluster group.

	Cluster 1: Lower Cue Utilization Group (*n*=18)	Cluster 2: Higher Cue Utilization Group (*n*=32)
Trainees	7	4
Qualified	11	28

(3) Diagnostic performance

Diagnostic performance was measured as percentage accuracy on the static image recognition task; Mean accuracy = 66.55% (*SD* = 12.4). Single-sample *t*-tests on mean accuracy relative to chance, where chance was 20% (1 in 5 possible responses), showed that pathologists performed above chance on the image interpretation task; $t(53) = 27.59$, $p < .0001$.

To address our first aim, a between-groups analysis of covariance (ANCOVA) was performed, incorporating two levels of cue utilization (higher/lower) as the independent variable, and the number of cases per year included as a covariate, to test its contribution to diagnostic performance; $F(1,47) = 8.4$, $p = .006$, $\eta_p^2 = .14$. [1]An inspection of the means indicated that, controlling for cases per year, those histopathologists with higher cue utilization ($M = 71.57$, $SD = 9.21$) performed more accurately on the static image recognition task compared to pathologists with lower cue utilization ($M = 60.93$, $SD = 13.32$) (see Figure 3).

[1] As the sample sizes in the cue utilisation groups were unequal, a Levene's test of equality was performed, comparing the two categories of cue utilisation on diagnostic performance, $F(48) = 3.36$, $p = .073$. This suggests that equal variances can be assumed.

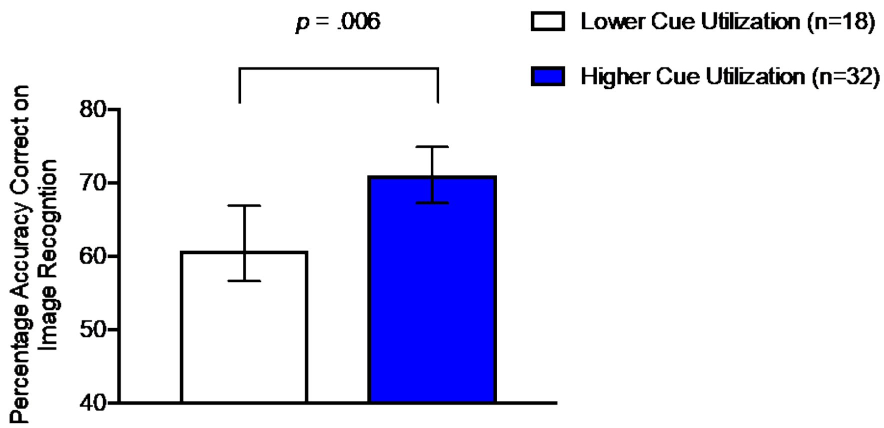

Figure 3: Percentage accuracy mean scores on the static image recognition task, distributed across cue utilization, controlling for cases read per year, for fifty histopathologists. Error bars represent 95% confidence intervals. Note: Chance is 20%.

An independent samples *t*-test, incorporating cue utilization as the grouping variable, failed to reveal a statistically significant difference between the groups in the number of cases read per year, $t(48) = -1.7$, $p = .1$. This suggests that cue utilization and cases per year read by pathologists contribute independently to the difference in performance on the diagnostic task.

Although the histopathologists performed above chance on the static image recognition task, their performance as a group was far from ceiling ($M = 66.55\%$). It is possible that these results reflect that the sample included pathologists in training (14/40) who are yet to become familiar with the pathologies included in the test set. Alternatively, while the stimuli were carefully selected by two subject matter experts, it may be the case that the performance of participants reflects the inherent variability in real-world images, and possibly within the image set. These factors include the difference in prevalence of specific types of cases in different practice settings.

In a post hoc, exploratory analysis, we first tested whether disease frequency and the rate at which histopathologists were likely to have encountered similar cases previously in their practice, may be related to their performance on the static image recognition task. Of the cases in the set, 12/31 cases were classified by a subject matter expert, and cross checked with available epidemiology data, as ‘rare’, while 19 of the 31 cases were considered ‘common’. The case with the lowest accuracy across the cohort was a rarely encountered disease occurring in a skin biopsy specimen, where six of the 50 pathologists (5/6 with higher cue utilization and 1/6 with lower cue utilization) diagnosed the case accurately. For this condition, there is a possibility that the low accuracy resulted from the relative unfamiliarity of the case.

A mixed-repeated ANCOVA testing the accuracy data from the static image recognition task, with disease frequency (common/rare) as a within-subjects variable, cue utilization (higher/lower) comprising a between-subjects variable, and the number of cases per year as a covariate, revealed a main effect for disease frequency, $F(1,47) = 66.30$, $p < .0001$, $\eta_p^2 = .59$. An inspection of the means indicated that, consistent with expectations, accuracy was greater for cases that were classified as common ($M = 74.60$, $SD = 11.29$) rather than rare ($M = 53.50$, $SD = 15.9$).

A main effect was also evident for cue utilization, $F(1,47) = 8.1$, $p = .007$, $\eta_p^2 = .15$, and an inspection of the means revealed that those participants with higher cue utilization recorded greater accuracy for both the more common cases ($M = 78.84$, SD =13.65) and the rarer cases ($M = 58.85$, $SD = 19.23$) compared to histopathologists with lower cue utilization (common: $M = 70.37$, $SD = 18.3$; rare: $M = 48.16$, $SD = 25.81$), controlling for the number of cases per year (see Figure 4).

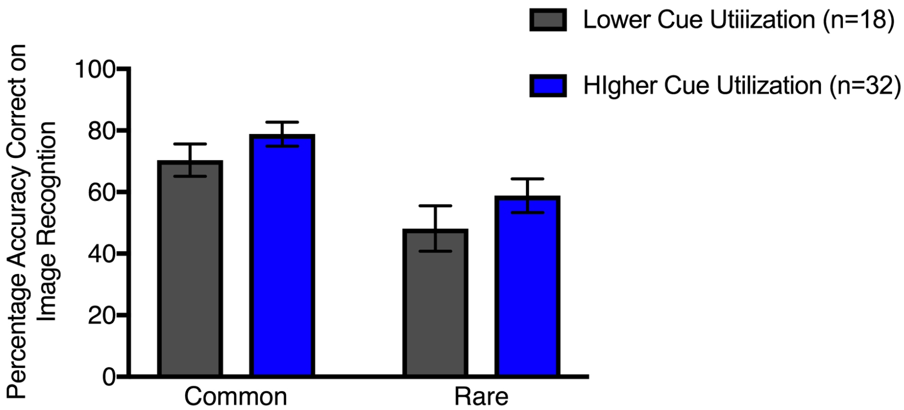

Figure 4: Percentage accuracy mean scores on the static image recognition task distributed across cue utilization, for the common and rare cases, controlling for cases read per year, for fifty histopathologists. Error bars represent 95% confidence intervals.

For a small number of images (4), the correct response was 'none of the above', since the pathology depicted was not listed as one of the other alternatives. Therefore, we explored whether the relationship between cue utilization and accuracy differed, depending upon the presence or absence of a pathology listed as an alternative. A mixed-repeated ANCOVA, incorporating two levels of cue utilization as a between-subjects factor (higher/ lower) and the presence or absence of a listed pathology as a within-groups variable, tested whether there were differences in the accuracy of responses. The estimated number of cases read per year was employed as a covariate in the analysis.

The results revealed a statistically significant interaction between cue utilization and the presence or absence of a listed pathology, $F(1,47) = 7.42$, $p = .009$, $\eta_p^2 = .14$. Post hoc contrasts (Bonferroni corrected) indicated that those participants with lower cue utilization ($M = 52.54$, $SD = 38.53$) were more likely to record a false positive, rather than correctly respond with 'none of the above', compared to those participants with higher cue utilization

($M = 79.04$, $SD = 28.7$) (Note: this analysis is exploratory, as there was a disproportionate number of not-listed cases (4) compared with the listed cases (27) (see Figure 5).

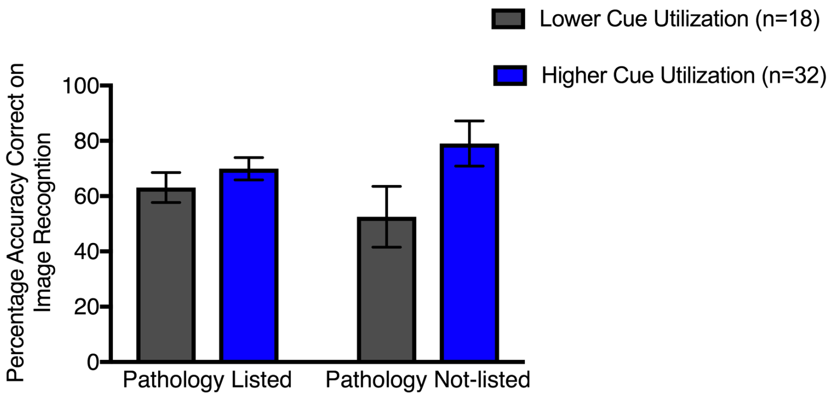

Figure 5: Percentage mean scores on the static image recognition task, distributed across cue utilization, for the pathology listed and not-listed cases, controlling for cases read per year, for fifty histopathologists. Error bars represent 95% confidence intervals.

In combination, the data suggest that differences in cue utilization contribute to performance on a diagnostic task, over and above the estimated number of cases read per year, and that these results may in part, be driven by disease prevalence both in our image set, and epidemiologically, along with the ability to recognize when a pathology is present or not present in an image.

Discussion

The overall aim of this study was to investigate whether diagnostic assessments of cue utilization differentiate performance in the context of histopathology, accounting for self-reported experience (number of pathology cases interpreted per year). Consistent with the hypothesis, higher cue utilization was associated with greater accuracy on a simulated diagnostic interpretation task requiring evaluation of tissue with significant abnormalities. A positive relationship was also evident between the estimated number of cases read per year and diagnostic accuracy. However, no relationship was evident between the estimated number of cases read per year and cue utilization. Consistent with previous research, this suggests that cue utilization and experience, in the form of the self-reported estimates of the number of cases read per year, contribute independently to diagnostic accuracy (Carrigan et al., 2020; Crane et al., 2018; Loveday et al., 2013b, Sturman et al., 2019).

This study provides evidence for differences on diagnostic performance based on cue utilization, independent of the number of cases read per year. However, there are other factors to consider that also contribute to expertise, including personality, genetic and developmental factors (Hambrick et al., 2016), that were not evaluated in the current study. Further, there is evidence to suggest that some pathologists have an inherent domain-general ability for visual recognition, independent of experience (e.g., Trueblood et al., 2018). It is also likely that in practice, if the disease frequency was inconsistent, 'experience' would begin to affect diagnostic outcomes and accuracy, independent of the number of cases read per year. Future research examining these issues is warranted, particularly comparing extreme values of cue utilization, where differences and variability in performance across the tasks can be maximized, and equal sample sizes can be assured. Importantly, the findings in the present study are consistent with Carrigan et al. (2020) who showed that among radiologists, cue utilization was positively associated with performance on a diagnostic task, independent of experience.

Although performance on the diagnostic task was above chance, it should be noted that overall accuracy was 66.55%. This is unsurprising as none of the participants reported being accustomed to diagnostic reporting from histopathology images on screens. The simulated diagnostic task presented a target tissue abnormality in a setting that was unfamiliar, rather than the normal pathological examination using an optical to view multiple fields at multiple magnifications. Although many laboratories are moving from optical microscopy to digital whole slide images, this was not the case for the current sample. Furthermore, the experimental setting lacked the opportunity for corroboration from colleagues or other resources such as text books or online images that would normally be available in routine practice when diagnostic uncertainty is encountered. Although presenting its own challenges, our diagnostic setting reproduces the essence of the routine practice setting. Unfortunately, the routine practice setting is not yet amenable to the study of activation of highly-refined cue-based associations in memory and how these associations are able to better discriminate differences in capability.

The post hoc analysis suggested that the types of cases and the prevalence of cases to which pathologists are exposed may be associated with differences in accuracy. Specifically, in comparison to participants with lower cue utilization, participants with higher cue utilization demonstrated relatively greater accuracy for less common pathologies, and where the pathology was not listed as a specific option for selection (i.e., the correct option was 'none of the above'). This has important training implications given that discordance errors are higher for histopathologists who interpret lower weekly case volumes, worked in smaller practices or non-academic settings.

Since improvements in cue utilization are thought to occur through the internalization of feature-event associations in memory, learning opportunities are necessary that enable, as quickly as possible, the extraction of features and their association with specific pathologies in memory are necessary (Wiggins, 2014). In the absence of supervised operational exposure, structured simulation such as reviewing and testing using image libraries with feedback, offers an alternative with the potential to facilitate the acquisition of cue-based associations at a rate faster and more efficiently than might be afforded through work-based exposure. With further research, evaluations based on normative assessments of performance could be employed to ensure successive improvements in cue utilisation.

Opportunities for future research also lie in examining whether cue utilization moderates performance differently depending upon the prevalence of cases. In the current study, our exploratory findings suggested that disease frequency or prevalence was related to diagnostic performance. Low target prevalence (i.e., rare or uncommon targets, few cases are truly abnormal) generally results in elevated miss rates (Evans et al., 2011; Mitroff & Biggs, 2014; Wolfe et al., 2005). Evans et al. (2011) demonstrated this 'low prevalence effect' in a cytology screening context where in the United States, the epidemiological prevalence of cervical cancer is around 1% (Benard et al., 2004). By contrast, disease prevalence in the context of histopathology tends to be higher.

Typically, in a breast histopathology setting, prior to the evaluation by a histopathologist, the patient and tissue samples have already been subjected to diagnostic testing and procedures such as mammograms, Prostate Specific Antigen (PSA) testing for prostate cancer, polyps biopsied during routine colonoscopies, or surgical excision. Further, specimens are accompanied by clinical information and surgical reports. This is likely to create an environment that potentially primes the confirmation of an abnormality. Testing the existence of 'abnormality priming' is necessary to consider the effects of prevalence on pathologists' cue utilization on a longer-term basis.

Conclusion

This study was designed to test the role of cue utilization in the diagnostic performance of histopathologists. Accounting for the self-reported number of cases reported per year, higher cue utilization was associated with higher accuracy on an independent, simulated diagnostic task. This effect was greatest where the pathology depicted on the image was not listed as a potential classification option and ‘none of the above’ was the accurate response.

Despite the limitations associated with the use of a simulated task that potentially lacks ecological validity, the results raise important issues concerning the relationship between motivation and cue utilization, particularly in response to challenging tasks. The outcomes also highlight the value of an approach that does not assume that experts use the same cues or features during problem-resolution. Our exploratory image analyses showed that factors such as disease prevalence are also likely to influence accuracy. Crucially, the outcomes of the study have important implications for the development and maintenance of skills amongst pathologists: Exposure sufficient to acquire and utilize cue-based strategies appears to be associated with the highest level of performance on the diagnostic task.

Key Points

- In pathology, accurate diagnosis involves accurate, reliable and timely responses.
- Pathological cue utilization is positively associated with performance on a diagnostic task, independent of the number of cases reported per year.
- A post hoc analysis suggested that higher cue utilisation may be associated with a greater capacity to recognise rarer and low prevalence cases.
- Targeted training that encourages the acquisition and utilization of cue-based strategies is recommended.

ORCID iD

Ann J. Carrigan https://orcid.org/0000-0002-2525-9241

References

Benard, V. B., Eheman, C. R., Lawson, H. W., Blackman, D. K., Anderson, C., Helsel, W., ... & Lee, N. C. (2004). Cervical screening in the national breast and cervical cancer early detection program, 1995–2001. *Obstetrics & Gynecology*, *103*(3), 564-571.

Brennan, P. C., Gandomkar, Z., Ekpo, E. U., Tapia, K., Trieu, P. D., Lewis, S. J., Wolfe, J.M., & Evans, K. K. (2018). Radiologists can detect the 'gist' of breast cancer before any overt signs of cancer appear. *Nature Scientific Reports*, *8*, 8717.

Brouwers, S., Wiggins, M. W., Griffin, B., Helton, W. S., & O'Hare, D. (2017). The role of cue utilization in reducing the workload in a train control task. *Ergonomics, 60*, 1–55.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62*, 193–217. doi: 10.1037/h0047470

Brunye, T. T., Carney, P. A., Allison, K. H., Shapiro, L. G., Weaver, D. L., & Elmore, J. G. (2014). Eye movements as an index of pathologist visual expertise: a pilot study. *PloS one*, *9*(8), e103447.

Cain, M. S., Prinzmetal, W., Shimamura, A. P., & Landau, A. N. (2014). Improved control of exogenous attention in action video game players. *Frontiers in Psychology*, *5*, 69.

Carmody, D. P., Nodine, C. F., & Kundel, H. L. (1981). Finding lung nodules with and without comparative visual scanning. *Perception & Psychophysics*, *29*, 594-598.

Carrigan, A. J., Curby, K., Morerel, D., & Rich, A.N. (2019). Exploring the effect of context and expertise on attention: is attention shifted by information in medical images? *Attention, Perception & Psychophysics*, *81*, 1283-1296. doi.org/10.3758/s13414-019-01695-7

Carrigan, A. J., Magnussen, J., Georgiou, A., Curby, K. M., Palmeri, T. J., & Wiggins, M. W. (2020). Differentiating Experience from Cue Utilization in Radiological Assessments. *Human Factors*, 0018720820902576.

Carrigan, A. J., Wardle, S. G., & Rich, A. N. (2018). Finding cancer in mammograms: if you know it's there, do you know where? *Cognitive Research: Principles and Implications*, *3*, 10. https://doi.org/10.1186/s41235-018-0096-5.

Charness, N., Krampe, R., & Mayr, U. 1996. The role of practice and coaching in entrepreneurial skill domains: An international comparison of life-span chess skill acquisition. *The road to excellence: The acquisition of expert performance in the arts and sciences, sports, and games*, pp. 51-80.

Crane, M. F., Brouwers, S., Wiggins, M. W., Loveday, T., Forrest, K., Tan, S. G. M., & Cyna, A. M. (2018). "Experience Isn't Everything": How Emotion Affects the Relationship Between Experience and Cue Utilization. *Human Factors*, 60, 685-698.

Croskerry, P., A. (2009). A universal model of diagnostic reasoning. *Academic Medicine, 84*, 1022 - 1028.

Curby, K. M., & Gauthier. I. (2007). A visual short-term memory advantage for faces. *Psychonomic Bulletin and Review*, 14(4), 620-628.

Curby, K. M., Glazek, K., & Gauthier, I. (2009). A visual short-term memory advantage for objects of expertise. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 94 – 107.

Drew, T., Evans, K., Võ, M. L. H., Jacobson, F. L., & Wolfe, J. M. (2013). Informatics in radiology: what can you see in a single glance and how might this guide visual search in medical images*? Radiographics, 33*(1), 263-274.

Elmore, J. G., Longton, G. M., Carney, P. A., Geller, B. M., Onega, T., Tosteson, A., N. Nelson, H. D., Pepe, M. S., Allison, K. H., and Schnitt, S. J. (2015). Diagnostic concordance among pathologists interpreting breast biopsy specimens, *JAMA,* 313, 1122-1132.

Evans, K.K., Georgian-Smith, D., Tambouret, R., Birdwell, R.L., & Wolfe, J.M. (2013). The gist of the abnormal: Above-chance medical decision making in the blink of an eye. *Psychonomic Bulletin Review, 20,* 1170 – 1175.

Evans, K. K., Tambouret, R. H., Evered, A., Wilbur, D. C., & Wolfe, J. M. (2011). Prevalence of abnormalities influences cytologists' error rates in screening for cervical cancer. *Archives of Pathology & Laboratory Medicine*, *135*(12), 1557-1560.

Falkland, E. C., & Wiggins, M. W. (2019). Cross-task cue utilisation and situational awareness in simulated air traffic control. *Applied Ergonomics*, *74*, 24-30.

Goldman, L., Sayson, R., Robbins, S., Cohn, L. H., Bettmann, M., & Weisberg, M. (1983). The value of the autopsy in three medical eras. *New England Journal of Medicine*, *308*, 1000-1005.

Hambrick, D. Z., Macnamara, B. N., Campitelli, G., Ullén, F., & Mosing, M. A. (2016). Beyond born versus made: A new look at expertise. In *Psychology of learning and motivation* (Vol. 64, pp. 1-55). Academic Press.

Hoff, S. R. (2013). Breast cancer: missed interval and screening-detected cancer at full-field digital mammography and screen-film mammography—results from a retrospective review. *Radiology, 264*, 378–386

Kirch, W., & Schafii, C. (1996). Misdiagnosis at a university hospital in 4 medical eras - report on 400 cases. *Medicine, 75*, 29–40.

Klein, G. (1989). Recognition-primed decisions (RPD). *Advances in Man-Machine Systems*, 5, 47 – 92.

Krupinski, E. A., Graham, A. R., & Weinstein, R. S. (2013). Characterizing the development of visual search expertise in pathology residents viewing whole slide images. *Human Pathology*, *44*, 357-364.

Kundel, H. L., & La Follette Jr, P. S. (1972). Visual search patterns and experience with radiological images. *Radiology, 103*, 523-528.

Kundel, H. L., & Nodine, C. F. (1975). Interpreting chest radiographs without visual search. *Radiology*, *116*, 527-532.

Kundel, H. L., Nodine, C. F., & Carmody, D. (1978). Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Investigative Radiology, 13*, 175-181.

Loveday, T., Wiggins, M. W., Harris, J. M., O'Hare, D., & Smith, N. (2013a). An objective approach to identifying diagnostic expertise among power system controllers. *Human Factors*, *55*, 90-107.

Loveday, T., Wiggins, M. W., & Searle, B. J. (2014). Cue utilization and broad indicators of workplace expertise. *Journal of Cognitive Engineering and Decision Making*, *8*, 98-113.

Loveday, T., Wiggins, M. W., Searle, B. J., Festa, M., & Schell, D. (2013b). The capability of static and dynamic features to distinguish competent from genuinely expert practitioners in pediatric diagnosis. *Human Factors*, *55*, 125-137.

Mitroff, S. R., & Biggs, A. T. (2014). The ultra-rare-item effect: Visual search for exceedingly rare items is highly susceptible to error. *Psychological Science*, *25*, 284-289.

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*, 211–226.

Nodine, C. F., & Krupinski, E. A. (1998). Perceptual skill, radiology expertise, and visual test performance with NINA and WALDO. *Academic Radiology*, *5*, 603-612.

Nodine, C.F., & Mello-Thoms, C. (2010). The role of expertise in radiologic image interpretation (p 139-156). In Samei, & E. A. Krupinski (Eds.). *The handbook of medical image perception and techniques*. Cambridge: Cambridge University Press.

Norman, G (2005). Research in clinical reasoning: past history and current trends. *Med. Educ. 39*, 418-427.

Pauley, K., O'Hare, D., & Wiggins, M. W. (2009). Measuring expertise in weather-related aeronautical risk perception: The validity of the Cochran–Weiss–Shanteau (CWS) Index. *The International Journal of Aviation Psychology,19*, 201–216.

Raab, S. S., Grzybicki, D. M., Janosky, J. E., Zarbo, R. J., Meier, F. A., Jensen, C., & Geyer, S. J. (2005). Clinical impact and frequency of anatomic pathology errors in cancer diagnoses. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, *104*, 2205-2213.

Schriver, A. T., Morrow, D. G., Wickens, C. D., & Talleur, D. A. (2008). Expertise differences in attentional strategies related to pilot decision making. *Human Factors*, *50*(6), 864-878.

Shojania, K. G., Burton, E. C., McDonald, K. M., & Goldman, L. (2003). Changes in rates of autopsy-detected diagnostic errors over time: a systematic review. *Journal of the American Medical Association, 289,* 2849–2856.

Sturman, D., Wiggins, M. W., Auton, J. C., & Helton, W. S. (2019). Cue utilisation predicts control room operators' performance in a sustained visual search task. *Ergonomics*, 1–13. https://doi.org/10.1080/00140139.2019.1680873

Tanner Boyd, I. J., Van Sickle, K., Schwesinger, W., Michalek, J., & Bingener, J. (2008). Music experience influences laparoscopic skills performance. *JSLS: Journal of the Society of Laparoendoscopic Surgeons*, *12*(3), 292.

Trueblood, J. S., Holmes, W. R., Seegmiller, A. C., Douds, J., Compton, M., Szentirmai, E., Woodruff, M., Huang, W., Stratton, C., & Eichbaum, Q. (2018). The impact of speed and bias on the cognitive processes of experts and novices in medical image decision-making. *Cognitive Research: Principles and Implications*, *3*, 28.

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*, 681-687.

Weiss, D. J., & Shanteau, J. (2003). Empirical assessment of expertise. *Human Factors, 45,* 104–116.

Wiggins, M. W. (2014). The role of cue utilization and adaptive interface design in the management of skilled performance in operations control. *Theoretical Issues in Ergonomics Science, 15*, 283–292.

Wiggins, M.W. (2020). A behaviour-based approach to the assessment of cue utilisation: implications for situation assessment and performance, *Theoretical Issues in Ergonomics Science*, DOI: 10.1080/1463922X.2020.1758828

Wiggins, M. W., Azar, D., Hawken, J., Loveday, T., & Newman, D. (2014). Cue-utilisation typologies and pilots' pre-flight and in-flight weather decision-making. *Safety Science*, *65*, 118-124.

Wiggins, M. W., Brouwers, S., Davies, J., & Loveday, T. (2014). Trait-based cue utilization and initial skill acquisition: implications for models of the progression to expertise. *Frontiers in Psychology*, *5*, 541.

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*, 0018720818814299.

Wiggins, M. W., Loveday, T., & Auton, J. (2015). *EXPERT Intensive Skills Evaluation* (EXPERTise 2.0) Test: Macquarie University, Sydney, AUS.

Wiggins, M. W., & O'Hare, D. (1995). Expertise in aeronautical weather-related decision making: A cross-sectional analysis of general aviation pilots. *Journal of Experimental Psychology, 1*, 305–320.

Wiggins, M., & O'Hare, D. (2003). Weatherwise: Evaluation of a cue-based training approach for the recognition of deteriorating weather conditions during flight. *Human Factors*, *45*, 337-345.

Wiggins, M. W., Stevens, C., Howard, A., Henley, I., & O'Hare, D. (2002). Expert, intermediate and novice performance during simulated pre-flight decision-making. *Australian Journal of Psychology, 54,* 162–167.

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Cognitive psychology: rare items often missed in visual searches. *Nature*, *435*, 439.

Ann J. Carrigan is a postdoctoral researcher at Macquarie University and works in radiology as a medical sonographer. Her research focuses on medical image perception and expertise in diagnostic medicine.

Amanda Charlton is an Anatomical Pathologist at Auckland City Hospital, her clinical work involves making diagnoses on tissue biopsies. Her interests include image perception, medical education, artificial intelligence and bias in diagnostic pathology.

Elliott Foucar, MD is an Adjunct Professor in the Department of Pathology, University of New Mexico School of Medicine. His major areas of interest are general diagnostic surgical pathology, dermatopathology, cytopathology, and autopsy pathology.

Mark W. Wiggins is a professor of organizational psychology at Macquarie University. His research focuses on skill acquisition and expertise in advanced technology environments.

Andrew Georgiou is a professor of Diagnostic Informatics at Macquarie University. His research focuses on diagnostic and organizational communication within the healthcare system.

Thomas J. Palmeri is a professor of psychology, ophthalmology, and visual sciences at Vanderbilt University. His research focuses on perceptual categorization, category learning, visual learning, visual memory, perceptual expertise, object and face recognition, automaticity, perceptual decision making, mathematical, computational, and neural modelling.

Kim M. Curby is an associate professor of cognitive psychology at Macquarie University. Her research focuses on the cognitive mechanisms and neural substrates that underlie expert performance in visually based domains.