Title

Finding the Perfect Match: Fingerprint Expertise Facilitates Statistical Learning and Visual Comparison Decision-Making

Journal

Journal of Experimental Psychology Applied, 29(2)

ISSN

1076-898X

Authors

Growns, Bethany
Mattijssen, Erwin JAT
Salerno, Jessica M
et al.

Publication Date

2023-06-01

DOI

10.1037/xap0000422

Peer reviewed

Abstract

Forensic feature-comparison examiners compare – or 'match' – evidence samples (e.g., fingerprints) to provide judgements about the source of the evidence. Research demonstrates that examiners in select disciplines possess expertise in this task by outperforming novices – yet the psychological mechanisms underpinning this expertise are unclear. This paper investigates one implicated mechanism: statistical learning, the ability to learn how often things occur in the environment. This ability is likely important in forensic decision-making as samples sharing rarer statistical information are more likely to come from the same source than those sharing more common information. We investigated the statistical learning and application of this knowledge of 46 fingerprint examiners and 52 novices. Participants completed four measures of their statistical learning about fingerprint categories (frequency discrimination judgements, bounded and unbounded frequency estimates, and same-source-likelihood judgements) before and after familiarisation to the 'ground-truth' category frequencies. Fingerprint examiners had superior domain-specific statistical learning across all measures and better applied this knowledge to identify fingerprint pairs that were more likely to have come from the same source than other fingerprints than novices. This suggests that fingerprint expertise facilitates domain-specific statistical learning. This has important theoretical and applied implications for the development of training programs and selection tools in forensic science.

Keywords: expertise, forensic science, forensics, statistical learning, distributional learning, fingerprint

Introduction

Forensic feature-comparison examiners (e.g., fingerprint examiners) perform an integral task in the justice system where they compare – or 'match' – evidence samples (e.g., fingerprints) in visual comparison tasks to provide judgements about the source of the evidence (Towler et al., 2018). Contrary to popular belief, this task is largely undertaken by human decision-makers – not computers or algorithms (Growns & Martire, 2020b; Thompson et al., 2013). Research investigating the performance of these decision-makers is therefore critical (President's Council of Advisors on Science and Technology, 2016). Evidence suggests that – although performance is not perfect – practitioners in some disciplines show *expertise* by outperforming novices in this task (Bird et al., 2010; Busey & Vanderkolk, 2005; Ericsson et al., 2018; Thompson et al., 2013; White et al., 2015). Yet much remains unknown about *how* practitioners outperform novices – or the psychological processes that underpin this expertise.

Emerging research has begun to identify the wide array of psychological processes that underpin forensic feature-comparison expertise (see Growns & Martire, 2020 for review). When completing domain-specific visual comparisons, compared to novices, face and fingerprint examiners show evidence of both analytical, deliberate and featural processing (Thompson & Tangen, 2014; Towler et al., 2017, 2017; White et al., 2015), and non-analytical, automatic and holistic processing (Busey & Vanderkolk, 2005; Searston & Tangen, 2017; Thompson & Tangen, 2014; Vogelsang et al., 2017; White et al., 2015). Compared to novices, fingerprint examiners also show evidence of greater memory retention for fingerprints (Busey & Vanderkolk, 2005; Thompson & Tangen, 2014), perform better in domain-specific visual search tasks (Robson et al., 2021; Searston & Tangen, 2017), and show different eye movement patterns when viewing fingerprints (Busey et al., 2011, 2013). In this paper, we investigate another mechanism implicated in proficient visual comparison performance by prominent mathematical theories: *statistical learning.*

Statistical learning is the mechanism by which individuals readily and automatically learn statistical information from their environment (Aslin, 2017; Frost et al., 2019). Studies show that statistical learning occurs quickly and easily – even after only brief exposure (< 10 minutes of familiarisation; Fiser & Aslin, 2001, 2002; Turk-Browne et al., 2005). People can learn a wide variety of statistical information – from conditional relationships between stimuli (e.g., Y co-occurs with Z in time or space, or *conditional statistical learning*; Fiser & Aslin, 2001; Turk-Browne et al., 2005) to the frequency and variability of distributions in the environment (e.g., W occurs more often than X; or *distributional statistical learning;* Thiessen & Erickson, 2013; Zacks & Hasher, 2002). These two processes are interrelated with substantial individual differences in both processes (Growns et al., 2020).

Critically, distributional statistical learning could be one cognitive mechanism that contributes to performance in forensic 'matching' tasks. Information theory suggests that statistical information provides important diagnostic cues in matching tasks (Bruce & Tsotsos, 2009; Busey et al., 2016; Shannon, 1948). For example, two fingerprints sharing a rare feature (e.g., a 'lake') are more likely to have a common source – or 'match' – than two fingerprints sharing a common feature (e.g., a 'bifurcation'), as the former is shared between fewer people in the general population. Forensic examiners may implicitly acquire and use domain-specific statistical knowledge via statistical learning in their casework – potentially enabling them to outperform novices presumably lacking this knowledge. This suggests that statistical learning could join other cognitive and perceptual processes in underpinning forensic science decision-making. It could also have important applied implications as forensic examiners are increasingly being asked to provide these decisions via numerical estimates or likelihood ratios (Aitken et al., 2011; Aitken & Taroni, 2004) – which can be problematic given the lack of reliable statistical databases in most forensic disciplines for examiners to rely on (even in well-established disciplines like fingerprint analysis; Mnookin, 2008).

Yet there is mixed evidence about whether examiners acquire domain-specific statistical knowledge via statistical learning in their casework. Forensic document examiners make better bounded frequency estimates (from 0-100%) of handwriting features than novices (Martire et al., 2018). Fingerprint examiners are also better able to rank the frequency of fingerprint categories than novices – however, do not make better *unbounded* frequency estimates (i.e., estimate *x* occurrence out of *y* occurrences) of those same categories (Mattijssen et al., 2020). This discrepancy may be due to the different forensic disciplines or the nature of statistical learning – a multi-faceted mechanism that encapsulates several interdependent skills including discriminating/ranking *relative* statistical frequencies and precisely estimating *exact* frequencies (Growns et al., 2020; Growns & Mattijssen, 2020; Hasher & Zacks, 1984). These abilities are interrelated and are typically correlated (Growns & Mattijssen, 2020). Forensic examiners may be able to express their statistical knowledge in some ways but not others – supported by examiners better ranking fingerprint categories but not better estimating their relative frequencies (Mattijssen et al., 2020). Yet no study has comprehensively investigated forensic examiners' acquired statistical knowledge using more than one or two statistical learning measures. Doing so is critical to better understand how examiners' statistical knowledge is best expressed to understand its role in their decision-making.

There is also limited evidence about whether forensic examiners use acquired statistical knowledge in matching tasks. Growns and Martire (2020a) investigated this by asking forensic examiners and novices to complete a matching task with novel stimuli designed to measure the use of statistical knowledge in match judgements. Importantly, examiners learned and used statistical knowledge to facilitate performance in this task. However, they were outperformed by a group of novices who were also trained in the application of statistical information in match judgements. Trained novices who were better statistical 'learners' were also better 'matchers' than poorer learners – a relationship not seen in examiners who did not receive this training. These trained novices likely outperformed examiners due to the novel stimuli used in this study. Expertise is typically domain-specific, narrow and rarely generalises beyond someone's area of experience (Bedard & Chi, 1992; Chase & Simon, 1973). It is currently unknown whether forensic examiners can apply statistical knowledge in matching tasks *within* their domain of expertise.

This study is an exploration of forensic examiners' statistical learning and its use in a visual comparison task within the domain of fingerprint analysis. We explore fingerprint examiners' and novices' domain-specific (i.e., fingerprint) statistical knowledge across three measures drawn from previous research (frequency discrimination judgements, and bounded and unbounded frequency estimates; Growns & Martire, 2020a; Growns & Mattijssen, 2020; Mattijssen et al., 2020) and their use of this knowledge in a same-source-likelihood judgement task (i.e., which of two fingerprint category pairs would provide more support for the hypothesis that both fingerprints were from the same person; see Method). We sought to examine two main research questions: 1) whether fingerprint examiners generally have better statistical learning and make more accurate same-source-likelihood judgements than a control group of novices; and 2) whether familiarisation to 'ground-truth' statistical frequencies – as in a typical statistical learning study (Fiser & Aslin, 2001, 2002) – improves either examiners' or novices' statistical learning. To do so, we measured both examiners' and novices' *a priori* statistical learning and same-source-likelihood judgements, then asked all participants to complete a 'familiarisation' task to learn the fingerprint statistical frequencies, after which we again measured statistical learning and same-source-likelihood judgements.

We expected that examiners would discriminate relative fingerprint frequencies and estimate these frequencies with given bounds better than novices in the pre-familiarisation phase (Growns & Martire, 2020a; Martire et al., 2018), but would not necessarily be better at estimating frequencies without bounds (Mattijssen et al., 2020). We also expected examiners to make more accurate same-source-likelihood judgements than novices at pre-familiarisation. We pre-registered that we had no *a priori* expectations about whether examiners would outperform novices on any measure post-familiarisation. Yet critically, if the direct acquisition of statistical knowledge alone improves the ability to use it when making same-source-likelihood judgements, we would expect to see novices' same-source-likelihood judgement accuracy increase from pre- to post-familiarisation. We also expected to find a relationship where better statistical 'learners' were also better able to use this information (i.e., same-source-likelihood judgements) post-familiarisation (after novices were given the opportunity to develop statistical knowledge; Growns & Martire, 2020a).

Method

Design

This study examined forensic examiners' statistical learning and same-source-likelihood judgements in a 2 (between-subjects; group: fingerprint examiners or a control group of novices) x 2 (within-subjects; familiarisation to statistical frequencies: pre-familiarisation or post-familiarisation) design (see Figure 1). The study pre-registration, de-identified data and analysis scripts are available at [links to be replaced upon acceptance: pre-registration: https://osf.io/mxe3q/?view_only=74351f651e734049aa1ea30dd9156de5; and data and analysis: https://osf.io/z3rce/?view_only=e586d680ef87480b9a5f37259ead0536].

Participants

Participants were 86 individuals (*n* = 58 fingerprint examiners, *n* = 28 other forensic practitioners) recruited through a snowball-sample method via emails sent to forensic organisations and mailing lists, and 58 novices recruited via Prolific Academia. Novices were paid $6.50 for participation in the 60-minute study, examiners were not paid for their involvement. To motivate performance, all participants were also offered the chance to win a prize for the best performance in the pre-familiarisation ($250) and post-familiarisation ($350) phases of the experiment. The second prize was larger than the first to encourage all participants to complete the entire study (Wray & Gates, 1996) – although it was not considered sufficiently large to incentivize participants to complete the study against their own best interests.

We used our pre-registered exclusion criteria to exclude participants who failed one or more (out of three) attention-check questions (*n* = 7 examiners, *n* = 3 novices). Based on our pre-registered criteria and to ensure sample homogeneity, all participants recruited via the snowball-sampling method who did not report that they were a fingerprint examiner were also excluded (*n* = 28). We also excluded participants who reported that they were aware of the paper we used to obtain the 'ground-truth' fingerprint category frequencies (*n* = 5 examiners, *n* = 2 novices).[1]

The final sample comprised 52 novices and 46 fingerprint examiners (45 from our snowball-sample method and one from Prolific; henceforth *examiners*). Novices were 32.6 years

[1] Note this exclusion criterion was not pre-registered as this specific question was added after pre-registration.

of age on average (*SD* = 11.8, *range* = 18-67), and the majority self-identified as female (53.8%) and reported English was their first language (92.3%). All novices reported living in the United States. Examiners were 37.2 years of age on average (*SD* = 7.99, *range* = 23-60), and the majority also identified as female (68.9%) and reported English was their first language (67.4%). The majority of examiners also reported currently residing in the United States (74.5%; 10.2% Australia, 6.12% Romania, 3.06% Brazil, 2.04% Canada, 2.04% India, 1.02% Switzerland, 1.02% Hungary). Examiners reported having 11.01 years of experience on average in fingerprint examination (*SD* = 7.88, *range* = 1-30), and the majority reported working in a police forensic laboratory (56.5%; 28.2% government forensic institution; 2.2% self-employed; 13.04% preferred not to specify).

We determined our sample size by the number of fingerprint examiners that were able to be recruited during our pre-registered time period for data acquisition, as well as the subsequent sample-size matched group of novices. We elected this method to determine our sample size rather than an *a priori* power analysis due to the difficulty of recruiting expert populations (e.g., Wray & Gates, 1996). However, it is important to note that the sample size recruited in this study (*n* = 46) was within the same range or exceeded that of other studies that also recruited forensic examiners (e.g., *n* = 11–26, Busey & Vanderkolk, 2005; Growns & Martire, 2020a).

Materials and Dependent Measures

Participants completed three phases during the experiment (see Figure 1): two test phases where they completed the four different tasks described below; and a stimulus familiarisation phase where they viewed fingerprint category images that appeared with their ground-truth statistical frequencies. Participants completed one test phase prior to the stimulus familiarisation phase, and one test phase after the stimulus familiarisation phase.

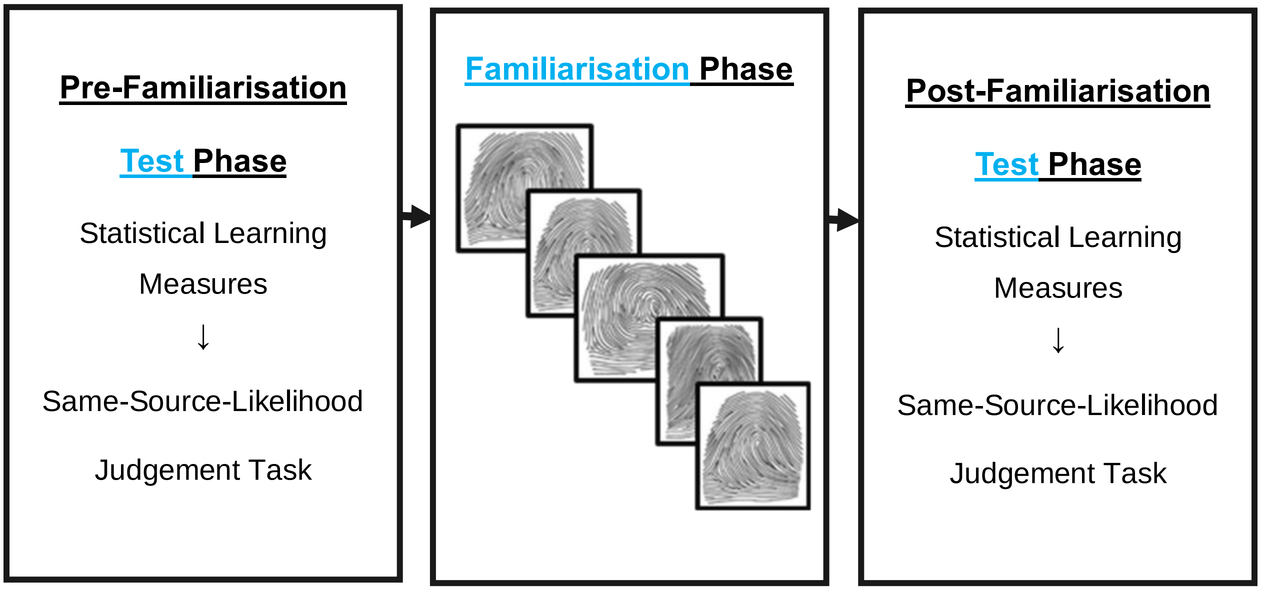

Figure 1. Procedure of experiment and examples of the fingerprint images participants viewed (middle panel; see de Jongh et al. (2019) for all images used)

Fingerprint Stimuli

All tasks were completed using fingerprint stimuli sourced and used with permission from de Jongh et al. (2019). The fingerprint stimuli were thirty-five representative exemplars from a fingerprint categorisation system developed and used by expert fingerprint examiners (de Jongh et al., 2019; Galton, 1892; Henry, 1913). Each fingerprint image was created by the same expert fingerprint examiners to be an ecologically valid example of a fingerprint category that an examiner would see in casework. The statistical frequencies for each fingerprint category (see Table 1) were calculated by two expert examiners independently classifying 10,000 fingerprint images from 2,452 randomly selected individuals from a criminal database according to the categorisation system, and a forensic trainee trained by the original experts classifying an additional 14,000 images. Each fingerprint category exemplar appears with a different statistical frequency in the general population (see Table 1).

Tasks

Participants completed four measures of their statistical learning as described below. In the pre-familiarisation test phase, participants were instructed to 'please base your answers on any previous experience that you have with fingerprints.' This was done to measure examiners' *a priori* knowledge of the fingerprint categories before any experimental familiarisation designed to induce statistical learning (with novices' responses as a baseline). In the post-familiarisation test phase, participants were instructed to 'Please base your answers on your previous experience with fingerprints and your experience with the images you just viewed' (i.e., the images viewed during stimulus familiarisation). This was done to measure both whether familiarisation could improve examiners' statistical learning and whether it could induce novices' statistical learning of the complex statistical frequencies.

Participants were instructed to base their answers in each task on any prior experience they had with fingerprints pre-familiarisation, and to base their answers post-familiarisation on any prior experience and the images they viewed during stimulus familiarisation.

Frequency discrimination judgements. Participants completed 47 frequency discrimination judgement trials pre-familiarisation and 47 frequency discrimination judgements trials post-familiarisation (94 total trials; see Figure 2). Participants viewed two, three or four fingerprint exemplars per trial and were asked to discriminate between their relative frequencies by selecting the most familiar exemplar. They were asked 'which of these fingerprint types is more familiar to you?' and selected the image by clicking on it (see Figure 2).

This measure was designed to be a complex and multi-alternate measure to best measure individual differences in statistical learning (see Siegelman et al., 2017 for discussion). Correct answers were based on the true frequencies from the familiarisation phase. Accuracy was the total correct trials and group-level chance performance was 39.4% accuracy or 18.5 trials in each task (i.e., pre-familiarisation and post-familiarisation).

Bounded frequency estimates. Participants were shown 18 fingerprint exemplars pre-familiarisation (out of the total of 35 fingerprint categories), and then the remaining 17 fingerprint categories post-familiarisation to ensure they viewed different images in each phase (see Figure 2). Participants viewed individual fingerprint categories and were asked 'What percentage of the time have you seen this fingerprint type?' They provided their estimates in a textbox restricted to numbers between 0 and 100 with a limit of 3 decimal places.

Estimation accuracy was measured by computing the absolute error for each participant by calculating the absolute difference between the true frequency and the estimated frequency for each category, then averaging across the absolute error of all individual estimates. Lower absolute error values indicate better bounded estimation accuracy. Note that estimation accuracy was calculated based on the presented frequencies from the familiarisation phase (see Table 1).

Unbounded frequency estimates. Participants were shown the 17 fingerprint exemplars they did not provide bounded estimates for during pre-familiarisation (out of the total of 35 fingerprint categories), and then the remaining 18 fingerprint categories post-familiarisation (see Figure 2). Participants viewed individual fingerprints and were asked to provide two numbers: one to indicate how often 'this fingerprint occurs,' and the other indicating 'out of this many fingerprints.' Participants provided their estimates in two textboxes restricted so that the number in the second textbox had to be equal to or greater than the number in the first textbox. No other

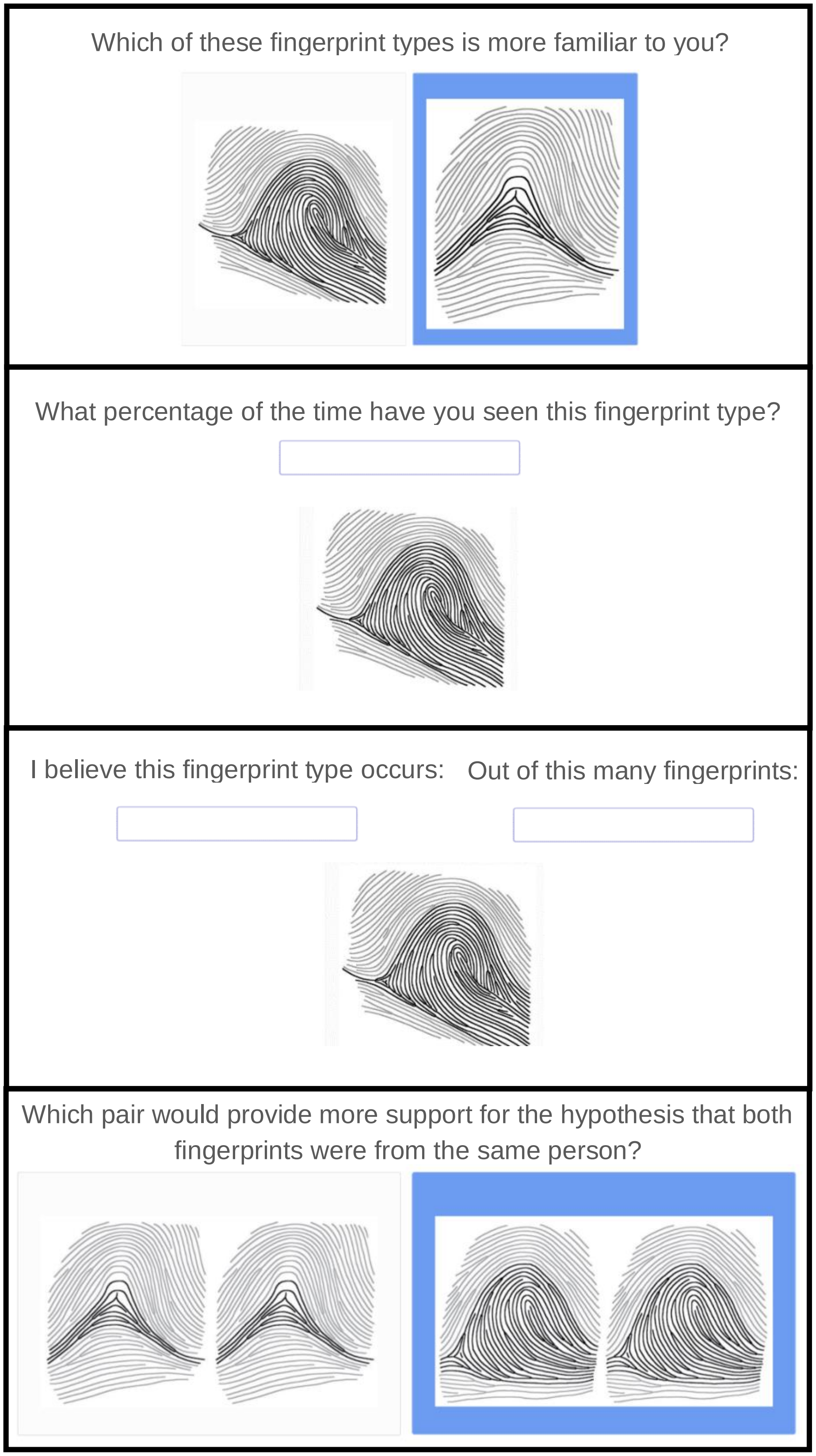

Figure 2. Example trials in all tasks: frequency discrimination judgements (top panel), bounded frequency estimates (second top panel), unbounded frequency estimates (second bottom panel), and same-source-likelihood judgements (bottom panel). Note the blue boxes in the top and bottom panels are an example of a participant highlighting their chosen option.

bounds were used to restrict participants' estimates.

Estimation accuracy was calculated by taking the absolute error for each participant by dividing the first estimate by the second estimate, multiplying that ratio by 100, subtracting the true ratio from the estimated ratio, and then averaging across the absolute error of all individual estimates. Lower absolute error values indicate better unbounded estimation accuracy.

Same-source-likelihood judgements. Participants completed 32 same-source-likelihood judgements pre-familiarisation and 32 judgements post-familiarisation (total 64; see Figure 2). Participants viewed two pairs of fingerprints and were asked 'which pair would provide more support for the hypothesis that both fingerprints were from the same person?' (see Figure 2). Each fingerprint pair always comprised two images of the exact same fingerprint category image. Participants were instructed that they would be asked to judge which pair provided stronger support for two fingerprints originating from the same source and were told to make their decisions based on the fingerprint category, rather than any other information that may usually be used in fingerprint comparisons (e.g., minutiae).

Correct answers were based on the true frequencies from the familiarisation phase where the pair of rarer fingerprint categories was the correct answer. Accuracy was the total correct trials and group-level chance performance was 50% accuracy or 16 trials both in each task (i.e., pre-familiarisation and post-familiarisation).

Procedure

Participants completed the experiment on an online survey platform (Qualtrics) and completed a CAPTCHA before beginning the experiment (Von Ahn et al., 2008). Participants first provided demographic and professional practice information. They were then given brief instructions and an opportunity to view all of the fingerprint images used in the study simultaneously to allow them to understand the categorisation system used. Participants were also instructed to adjust their browser zoom so images could be fully seen and to only take breaks at appropriate points when prompted. Participants were not restricted to a certain device to participate, and a mobile-friendly version was available via Qualtrics for those who did not complete the task on a desktop computer.

Participants then first completed the pre-familiarisation test phase where they completed all four tasks described above in a set order: frequency discrimination judgements, bounded frequency estimates, unbounded frequency estimates, and then same-source-likelihood judgements. They then completed the stimulus familiarisation phase where they viewed a stream of 430 fingerprint images in a randomised order for 12.5 minutes. These images appeared in proportion to their 'ground-truth' category frequencies ranging from 0.002 – 0.305 (as in de Jongh et al., 2019; see Table 1). Participants viewed each image for 1.5-sec with an interstimulus interval (ISI) of .25-sec (as in Growns & Mattijssen, 2020). After the stimulus familiarisation phase, participants completed the post-familiarisation test phase where they completed the same four tasks with different stimuli. To prevent statistical learning occurring in the test phases (particularly pre-familiarisation), each fingerprint exemplar appeared the exact same number of times across all tasks in each test phase. Upon completion of the experiment, participants viewed a debrief screen thanking them for their participation and informing them about the aims of the study.

Participants completed tasks in a set order as above and trials within each task in a pseudo-randomised order where one random trial order was generated when coding the experiment and all participants completed trials in that order. This was done to minimise error variance by limiting the extent to which completing tasks/trials in different or counterbalanced orders could lead to spurious individual differences produced by different experimental experiences (Mollon et al., 2017). We pre-registered this design to ensure we could best capture individual differences in our data to investigate the relationship between statistical learning and applied use of this knowledge (i.e., same-source-likelihood judgements).

Table 1. Fingerprint frequencies and *n* adapted from de Jongh et al. (2019)

Fingerprint Category	*n*	Frequency
Tented arch	1	0.002
Tannenbaum	1	0.002
Left roofed arch	1	0.002
Right roofed arch	1	0.002
Left pseudoloop	1	0.002
Right pseudoloop	1	0.002
Composite arch	1	0.002
Left inverted loop	1	0.002
Right inverted loop	1	0.002
Plain whorl with centre dot	1	0.002
Plain whorl with a horseshoe	1	0.002
Plain whorl with a meat hook	1	0.002
Clockwise e-spiral	1	0.002
Counterclockwise e-spiral	1	0.002
Tulip	1	0.002
Double-loop with two left loops	1	0.002
Double-loop with two right loops	1	0.002
Composite whorl	1	0.002
Mushroom	1	0.002
Right pinched loop	5	0.012
Left pinched loop	6	0.014
Shared double loop with a left standing loop	6	0.014
Shared double loop with a right standing loop	6	0.014
Left central pocket loop	7	0.016
Right central pocket loop	7	0.016
Almond	8	0.019
Clockwise spiral	12	0.028
Counterclockwise spiral	12	0.028
Double-loop with a right standing loop	12	0.028
Plain arch	13	0.03
Plain whorl	13	0.03
Double-loop with a left standing loop	18	0.042
Plain whorl with a centre stick	33	0.077
Right plain loop	122	0.284
Left plain loop	131	0.305
Total	430	1

Fingerprint Category	*n*	Frequency

Note: as in Growns & Mattijssen (2020), we rounded up eight frequencies to the nearest trial (n = 1 or 0.002) to balance familiarisation length feasibility and ecological validity. As some 'ground-truth' frequencies were 1/≥1000, a familiarisation phase with these frequencies would have lasted over 25 minutes which we deemed unfeasible. We also dropped one fingerprint category from the original paper that comprised patterns that the authors were unable to assign to a specific category and thus has no exemplar image available.

Results

Descriptive Statistics and Psychometric Properties

One-sample t-tests revealed that frequency discrimination judgements were significantly better than chance in both groups pre- and post-familiarisation, but only the examiners' same-source-likelihood judgements were significantly above chance both pre- and post-familiarisation (see Table 2). Bounded and unbounded estimation accuracy was low and comparable to previous experiments in both groups pre- and post-familiarisation (e.g., Growns & Mattijssen, 2020). Psychometric properties measuring the internal reliability of each measure were close to or above recommended values for standardised psychometric measures (> .8; Streiner, 2003; see Table 3).

Table 2. One-sample t-test results between groups for frequency discrimination judgements and sensitivity

	Mean (SD)	*df*	*t*	*p*	95% CI
Frequency discrimination Judgements					
Novices: Pre-Familiarisation	22.4 (6.56)	51	4.31	< .001	[20.6, 24.3]
Novices: Post-Familiarisation	25.9 (7.65)	51	6.96	< .001	[23.8, 28.0]
Examiners: Pre-Familiarisation	32.2 (7.34)	45	12.65	< .001	[30.2, 34.4]
Examiners: Post-Familiarisation	34.7 (6.66)	45	16.54	< .001	[32.8, 36.7]
Same-source-likelihood Judgements					
Novices: Pre-Familiarisation	14.4 (6.07)	51	1.90	.063	[12.7, 16.1]
Novices: Post-Familiarisation	14.0 (7.46)	51	1.90	.064	[12.0, 16.1]
Examiners: Pre-Familiarisation	22.8 (8.89)	45	5.22	< .001	[20.2, 25.5]
Examiners: Post-Familiarisation	24.7 (6.81)	45	8.62	< .001	[22.6, 26.7]

Table 3. Psychometrics properties of all measures (Cronbach's alpha and split-half reliability *r* in brackets) between groups pre-familiarisation and post-familiarisation.

	Novices		Examiners	
	Pre-Familiarisation	Post-Familiarisation	Pre-Familiarisation	Post-Familiarisation
Discrimination Judgements	.79 (.69)	.84 (.74)	.87 (.78)	.86 (.78)
Bounded Estimates	.95 (.91)	.96 (.93)	.92 (.88)	.94 (.91)
Unbounded Estimates	.97 (.95)	.97 (.95)	.92 (.88)	.93 (.91)
Same-source-likelihood Judgements	.81 (.72)	.89 (.82)	.95 (.92)	.92 (.88)

Group Differences Between Examiners and Novices

To explore if statistical learning differed between groups or pre-familiarisation and post-familiarisation test phases, we used the *lmer* (*v. 1.1-25;* Bates et al., 2014) package in R to create mixed-effects regression models to predict each measure at the trial level from the interaction between group (fingerprint examiner or novice) and time (pre-familiarisation or post-familiarisation; see Figures 3 and 4). Confidence intervals (95%) and *p* values were calculated in each analysis using the *lmerTest* (*v 3.1-3;* Kuznetsova et al., 2017) package. Each analysis contained a random effect for participant ID which allowed intercepts to vary between participants.[2]

Frequency discrimination judgement accuracy. We conducted a mixed-effects logistic regression to predict frequency discrimination judgement accuracy at the trial level from the

[2] We also conducted exploratory analyses of all pre-registered results with examiners' years of experience taken into account (see Supplementary Analyses on OSF for details). The pattern of results was largely consistent, except that the main effect of group was significant in the unbounded estimation accuracy group analysis such that examiners were more accurate than novices (and the interaction was no longer significant).

interaction between group and time. Examiners (*M* = 33.5*, SD* = 7.0) were 2.54 times more likely to make a correct frequency discrimination judgement than novices (*M* = 24.2, *SD* = 7.1; *b* = .93, 95% CI [.68, 1.19], *z* = 7.23, *p* < .001; see Figure 3), and all participants were also 1.37 times more likely to make a correct judgement in the post-familiarisation test phase (*M* = 30.0, *SD* = 8.41) than the pre-familiarisation phase (*M* = 27.0, *SD* = 8.47; *b* = .31, 95% CI [.20, .43], *z* = 5.30, *p* < .001). The interaction between group and phase was not significant (*b* = -.03, 95% CI [-.21, .15], *z* = .32, *p* = .751).

Bounded estimation accuracy. We conducted a mixed-effects linear regression to predict bounded estimation accuracy at the trial level from the interaction between group and time. Examiners' bounded estimation accuracy (*M* = 15.1, *SD* = 16.8) was 7.67% better (i.e., lower absolute error) than novices' (*M* = 22.4, *SD* = 23.2; *b* = -7.67, 95% CI [-13.22, -2.11], $t_{(104.85)}$ = 2.73, *p* = .008; see top panel of Figure 4), but all participants' accuracy did not significantly differ between the pre-familiarisation and post-familiarisation test phases (*b* = 1.26, 95% CI [-2.67, .16], $t_{(3332)}$ = 1.74, *p* = .082). The interaction between group and phase was not significant (*b* = .75, 95% CI [-1.32, 2.81], $t_{(3332)}$ = .71, *p* = .480). Note that lower absolute error indicates better estimation accuracy, so a decrease in score indicates better performance.

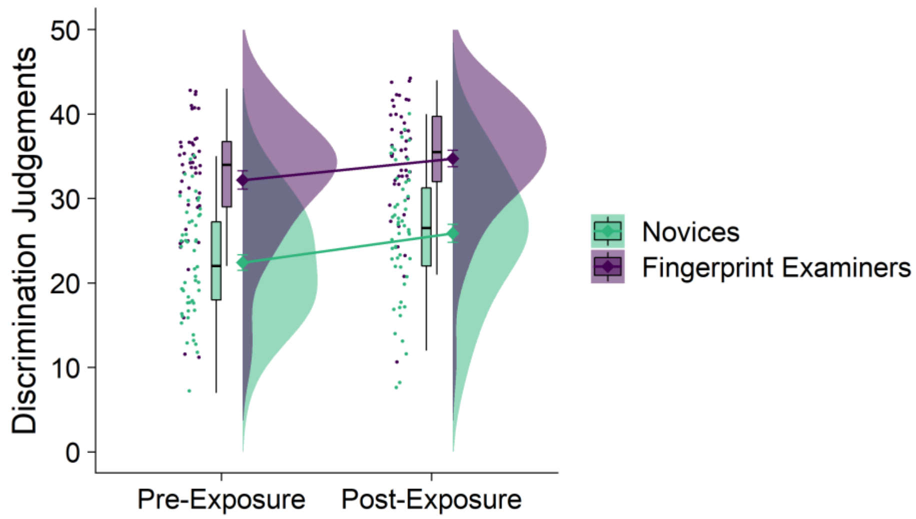

Figure 3. Average discrimination judgement accuracy by experience and familiarisation phase. Raincloud plots depict (left-to-right) the jittered participants' averaged data points, box-and-whisker plots, means (represented by diamonds) with error bars representing ± 1 *SE*, and frequency distributions.

Unbounded estimation accuracy. We conducted a mixed-effects linear regression to predict unbounded estimation accuracy at the trial level from the interaction between group and time. Unbounded estimation accuracy did not significantly differ between groups (b = -5.70, 95% CI [-11.67, .26], $t_{(103.39)}$ = 1.89, p = .061; see bottom panel of Figure 4) or between the pre-familiarisation and post-familiarisation test phases (b = .41, 95% CI [-.88, 1.69], $t_{(3332)}$ = .62, p = .536). However, there was a significant interaction between group and phase (b = -.2.10, 95% CI [-3.97, -.23], $t_{(3332)}$ = 2.20, p = .028). Follow-up comparisons revealed that examiners' unbounded estimation accuracy was 1.70% better post- (M = 11.1, SD = 14.9) than pre-familiarisation (M = 12.8, SD = 15.4; b = 1.70, 95% CI [.33, 3.06]), whereas novices' accuracy didn't significantly differ pre- (M = 18.5, SD = 22.8) to post-familiarisation (M = 18.9, SD = 22.4; b = -.41, 95% CI [-1.69, .88]). Note that lower absolute error indicates better estimation accuracy, so a decrease in score indicates better performance.

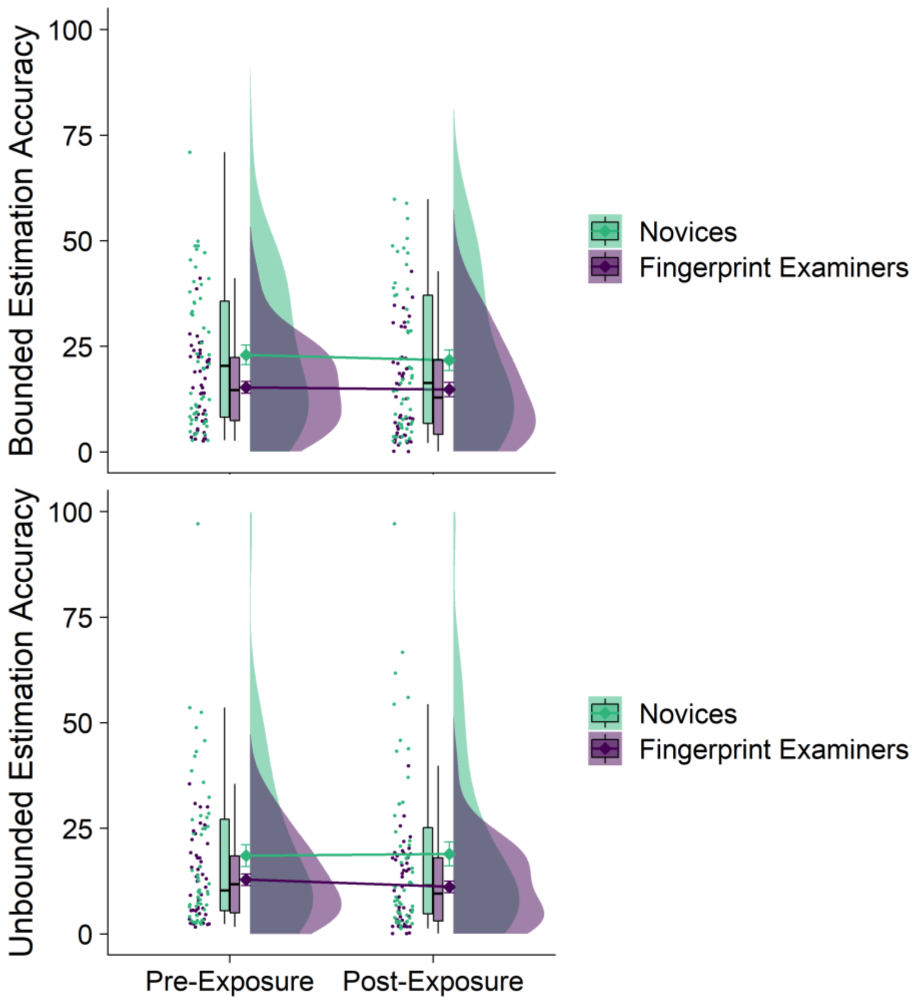

Figure 4. Average bounded (top panel) and unbounded (bottom panel) estimation accuracy by expertise and familiarisation phase. Raincloud plots depict (left-to-right) the jittered participants' averaged data points, box-and-whisker plots, means (represented by diamonds) with error bars representing ± 1 *SE*, and frequency distributions. Note that lower absolute error indicates better estimation accuracy, so a decrease in score indicates better performance.

Same-source-likelihood judgement accuracy. We conducted a mixed-effects logistic regression to predict same-source-likelihood judgement accuracy at the trial level from the interaction between group and time (see Figure 5). Examiners (M = 23.8, SD = 12.3) were 3.79 times more likely to make accurate same-source-likelihood judgements than novices (M = 14.2, SD = 6.8; b = 1.33, 95% CI [.89, 1.78], z = 5.94, p < .001; see Figure 5), but all participants' accuracy did not significantly differ between the pre-familiarisation and post-familiarisation test phases (b = -.05, 95% CI [-.20, .09], z = .72, p = .473). However, there was a significant interaction between group and phase (b = .44, 95% CI [.20, .68], z = 3.59, p < .001). Follow-up comparisons revealed that examiners were 1.47 times more likely to make an accurate same-source-likelihood judgement pre- (M = 22.8, SD = 8.89) than post-familiarisation (M = 24.7, SD = 6.81; b = -.40, 95% CI [-.66, -.14]), compared to the differences between novices' accuracy pre- (M = 14.4, SD = 6.07) and post-familiarisation (M = 14.0, SD = 7.46; b = .06, 95% CI [-1.8, .30]).

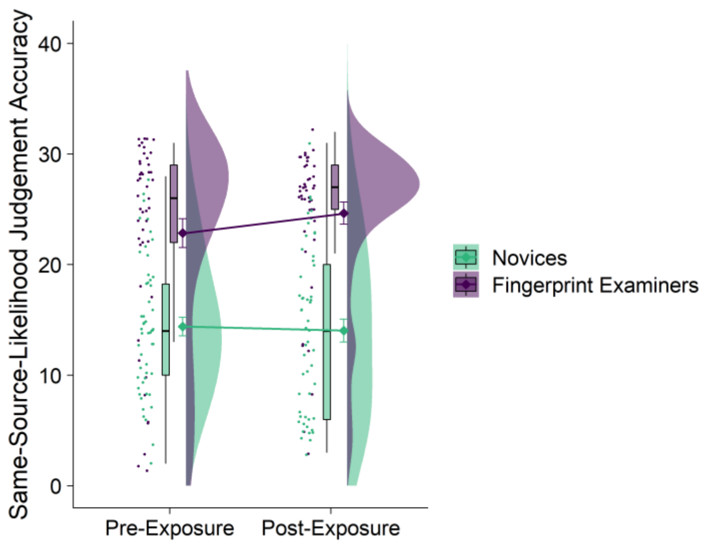

Figure 5. Average same-source-likelihood judgement accuracy by expertise and familiarisation phase. Raincloud plots depict (left-to-right) the jittered participants' averaged data points, box-and-whisker plots, means (represented by diamonds) with error bars representing ± 1 *SE*, and frequency distributions.

Individual Differences in Statistical Learning and Same-Source-Likelihood Judgement Accuracy

To explore the relationship between individual differences in statistical learning and same-source-likelihood judgements, we conducted linear regression models to predict same-source-likelihood judgement accuracy from the interaction between group and each statistical learning measure at the trial level. We investigated these relationships post-familiarisation only after novices had been given the opportunity to develop statistical knowledge.

Frequency discrimination judgement accuracy. Same-source-likelihood judgement accuracy was significantly predicted by group (b = 9.44, 95% CI [8.72, 10.16], $t_{(4602)} = 25.75$, $p < .001$; see top panel of Figure 6), but was not significantly predicted by frequency discrimination judgement accuracy (b = -.03, 95% CI [-.87, .26], $t_{(4602)} = 1.05$, $p = .294$). The interaction between group and frequency discrimination judgement accuracy was significant (b = 1.67, 95% CI [.78, 2.55]), $t_{(4602)} = 3.70$, $p < .001$). Examiners' same-source-likelihood judgement accuracy increased with their frequency discrimination judgement accuracy (b = 1.36, 95% CI [.68, 2.04]), whereas no reliable relationship between frequency discrimination and same-source-likelihood judgement accuracy was seen in novices (b = -.30, 95% CI [-.87, .26]).

Bounded estimation accuracy. Same-source-likelihood judgement accuracy was significantly predicted by group (b = 10.83, 95% CI [9.93, 11.74], $t_{(1662)} = 23.45$, $p < .001$; see middle panel of Figure 6), and was also significantly predicted by bounded estimation accuracy such that same-source-likelihood judgement accuracy increased as absolute error decreased (b = -.04, 95% CI [-.06, -.02], $t_{(1662)} = 4.01$, $p < .001$). This suggests that same-source-likelihood judgement accuracy increases with bounded estimation accuracy, as lower absolute error indicates higher estimation accuracy. The interaction between group and bounded estimation accuracy was not significant (b = -.03, 95% CI [-.07, <.01], $t_{(1662)} = 1.94$, $p = .052$). This suggests the relationship between the two measures did not significantly differ between groups.

Unbounded estimation accuracy. Same-source-likelihood judgement accuracy was significantly predicted by group (b = 11.68, 95% CI [10.86, 12.50], $t_{(1760)} = 28.02$, $p < .001$; see bottom panel of Figure 6), and was also significantly predicted by unbounded estimation

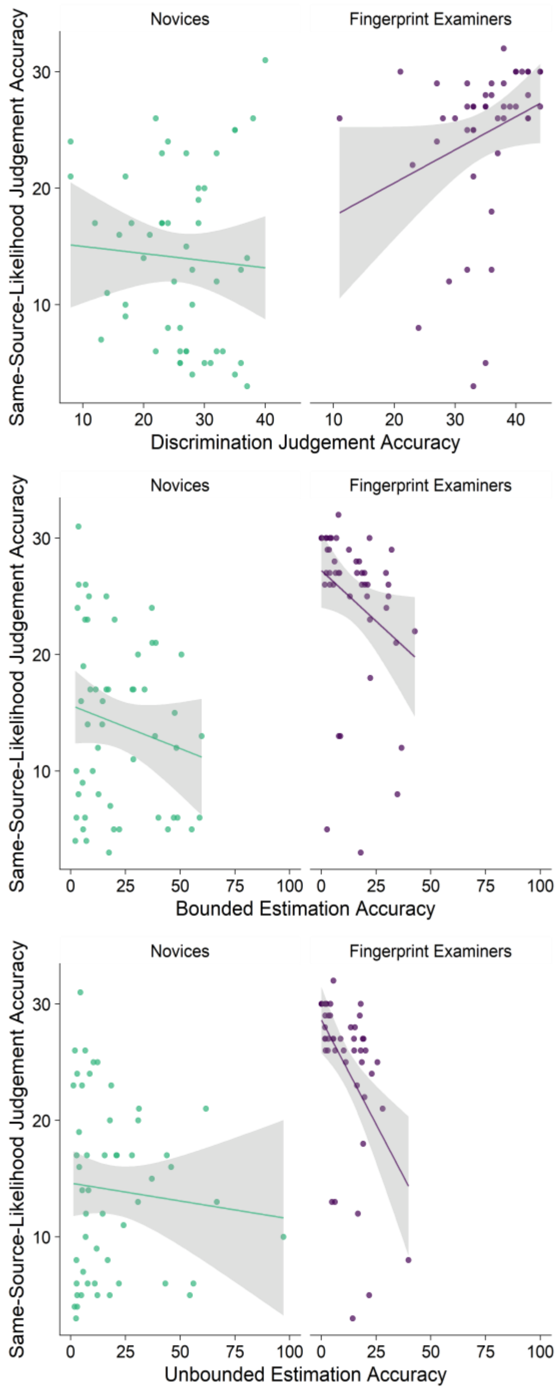

Figure 6. Average same-source-likelihood judgement accuracy and average discrimination judgement accuracy (top panel), average bounded estimation accuracy (middle panel) and average unbounded estimation accuracy (bottom panel) between groups. Regression lines are displayed with 95% confidence interval bands.

accuracy such that same-source-likelihood judgement accuracy increased as absolute error decreased (b = -.02, 95% CI [-.04, <-.01], $t_{(1760)}$ = 2.24, p = .026). This suggests that same-source-likelihood judgement accuracy increases with unbounded estimation accuracy, as lower absolute error indicates higher estimation accuracy. The interaction between group and bounded estimation accuracy was also significant (b = -.11, 95% CI [-.15, -.07], $t_{(1760)}$ = 5.91, p < .001). Examiners' same-source-likelihood judgement accuracy increased more strongly as absolute error decreased (b = -.13, 95% CI [-.16, -.10]), whereas the same relationship in novices increased less strongly (b = -.02, 95% CI [-.04, <-.01]).

Discussion

This study was an empirical investigation of fingerprint examiners' acquired domain-specific statistical learning and use of statistical knowledge in a same-source-likelihood task. Fingerprint examiners had better acquired statistical knowledge than novices, even pre-familiarisation: they better discriminated between rare and common fingerprint categories, and better estimated category frequencies with given bounds (i.e., 0-100%). Just as individuals can extract and encode statistical information from their environment via statistical learning (Fiser & Aslin, 2001; Siegelman et al., 2017a), fingerprint examiners also appear to learn the frequency of fingerprint categories in their work. This suggests that statistical learning joins other cognitive mechanisms as a process that underlies their expertise – something that is also consistent with research linking statistical learning and other forms of expertise (Daikoku & Yumoto, 2020; Francois & Schön, 2011; Mandikal Vasuki et al., 2016, 2017).

Yet consistent with previous research (Mattijssen et al., 2020), examiners weren't better able to estimate the same frequencies without given bounds than novices (i.e., estimate x occurrence out of y occurrences). They did however provide better estimates without bounds after familiarisation – an effect not seen for the novices. This discrepancy between bounded and unbounded estimates may be due to the difficulty of estimating explicit frequencies without boundaries – specific bounds may make it easier to conceptualise frequencies. Yet it is important to note that their unbounded estimation accuracy did improve post-familiarisation. These results further emphasise the multi-faceted nature of statistical learning (Growns et al., 2020; Hasher & Zacks, 1984), but also have important implications for how examiners should express their statistical knowledge. Nevertheless, fingerprint examiners appear to have domain-specific statistical knowledge above and beyond novices' – likely acquired via statistical learning during their increased familiarisation to fingerprints in their casework.

Importantly, this study demonstrated that fingerprint examiners can also *use* their acquired statistical knowledge better in a same-source-likelihood task than novices. Fingerprint examiners more accurately judged which fingerprint categories provided stronger evidence of fingerprints originating from the same source than novices – accuracy even improved from pre-to-post familiarisation suggesting that frequency familiarisation better calibrated their judgements. However, novices' same-source-likelihood accuracy did not improve post-familiarisation even though their ability to discriminate rare and common fingerprints did improve. This suggests that the development of statistical knowledge alone isn't sufficient to *use* this knowledge to determine the likelihood of two fingerprints coming from the same source – possessing and utilising statistical knowledge appear to be separate skills. Fingerprint examiners' advantage may arise via their experience or training – similar to how novices can be trained to use statistical knowledge to improve matching accuracy (e.g., Growns & Martire, 2020a). Importantly, their accuracy improved post-familiarisation and future work could thus focus on targeted training to improve examiners' utilization of acquired statistical knowledge in decision-making.

We also observed a relationship between statistical learning and same-source-likelihood judgements where better statistical 'learners' were also better 'matchers' than poorer learners.[3] Participants with higher estimation accuracy (bounded or unbounded) were more accurate in the same-source-likelihood task than those with lower accuracy – this relationship was even stronger for examiners' unbounded estimates. Examiners who better discriminated rare and common fingerprints categories were also better able to apply their knowledge than poorer discriminators (although this relationship was not seen in novices). This suggests that there appears to be an underlying relationship between statistical learning and its explicit use – examiners that are better statistical 'learners' may be more accurate *using* their statistical knowledge in their work than poorer 'learners.' However, this relationship may be differentially impacted by experience across different statistical learning measures.

Practical Implications

Forensic examiners assert that they can explicitly estimate relevant statistical frequencies in their work (Biedermann et al., 2013; Daeid et al., 2020; Nordgaard et al., 2012; Willis et al., 2015) and are increasingly being asked to provide their decisions via numerical estimates (Aitken et al., 2011; Aitken & Taroni, 2004). Our results demonstrated that whilst fingerprint examiners can provide better bounded estimates than novices, their performance was not perfect – bounded estimates were off by 15.1% on average. This is consistent with previous research in other forensic disciplines – frequency estimates made by forensic document examiners are off by approximately 20% on average (Martire et al., 2018). It is also consistent with broader cognitive science research demonstrating that individuals are typically poor at providing exact frequency

[3] Note we only measured this relationship post-familiarisation when novices had been given the opportunity to acquire statistical knowledge.

estimates and judging the base-rate of events (Bar-Hillel, 1980; Brenner et al., 1996; Lee & Danileiko, 2014). There are many decision-making biases that can impact how people encode and recall frequency information (e.g., the *base-rate* fallacies where individuals tend to underweight prior information; Kahneman & Tversky, 1973; Welsh & Navarro, 2012). Nevertheless – whilst our results suggest that these explicit estimates would not rely on 'numbers from nowhere' (Bali et al., 2019; Risinger, 2013), they do indicate they would include some error if based solely on examiners' personal experience. This has the potential to mislead fact-finders in court.

These results highlight the importance of developing reliable statistical databases in all forensic disciplines – such as those that exist in nuclear DNA analysis – to ensure that only reliable and valid forensic evidence is presented in court. Such databases do not exist in many forensic disciplines, even in fingerprint comparison (Mnookin, 2008) – emerging databases are often only restricted to specific geographical regions (e.g., Gutierrez-Redomero et al., 2011; Gutiérrez-Redomero et al., 2012; Johnson et al., 2017). Examiners thus likely rely on their own implicit statistical knowledge developed over the course of their work in decision-making and when making explicit evidential strength judgements. Whilst our results suggest they likely do this better than novices, their performance will be imperfect. Without the existence of reliable and generalisable statistical databases, examiners can only rely on their implicit knowledge. This emphasises the importance of funding the development of such statistical databases in many forensic feature-comparison disciplines.

Limitations and Future Directions

It is important to note that we were unable to measure perceptual sensitivity or response bias in the same-source-likelihood task in this study – as matching performance is typically assessed in a signal detection framework (Phillips et al., 2001). We were unable to measure sensitivity or bias as we could not create the 'match' and 'non-match' trials necessary to calculate these measures due to stimuli restrictions (i.e., there was only one image of each category). This means that we can only draw conclusions about how examiners use rare or common statistical information in making applied statistical decisions – not whether they can use this knowledge to facilitate better sensitivity to the presence of 'target' stimuli (i.e., diagnostic cues signifying samples 'match') amongst noise (Phillips et al., 2001). It will be important for future research to assess the use of statistical knowledge in matching tasks using signal detection measures. Nevertheless, our results do highlight that fingerprint examiners *can* use rare and common information better when making same-source-likelihood decisions than novices.

It is also important to note that statistical learning is not the only cognitive mechanism that underpins forensic science expertise, nor is distributional learning of fingerprint categories – as investigated in this paper – the only statistical information that examiners may rely on. Fingerprint examiners likely also rely on more complex statistical knowledge in their work – such as fingerprint minutiae that can co-occur in unique spatial relationships (Phillips et al., 2001) or in different areas of a fingerprint (Busey & Parada, 2010). Individuals are also able to extract and encode spatial co-occurrences in the visual environment – a process that is inter-related to learning distributions and variations in the environment (Growns et al., 2020; Thiessen et al., 2013; Thiessen & Erickson, 2013). It will therefore be important for future research to further investigate the many facets of statistical learning that may be implicated in fingerprint expertise.

It is also important to note that fingerprint examiners in all jurisdictions do not necessarily use the categorisation system used in this study (de Jongh et al., 2019; Mattijssen et al., 2020). Whilst most fingerprint examiners are aware of the broader 'arch', 'loop' and 'whorl' categorisation system (Galton, 1892; Henry, 1913; Mattijssen et al., 2020), this study used a different categorisation system that was developed to better assess the diagnostic value of each category (de Jongh et al., 2019). Although we allowed all participants to view the categorisation system used upon commencing the study, examiners' performance could have been impaired if they were not familiar with the system. It is also important to note that examiners often do not use higher-level categorical information to make decisions in their casework (de Jongh et al., 2019; Mattijssen et al., 2020) – instead relying on lower-level featural information like fingerprint minutiae (Busey & Parada, 2010; Phillips et al., 2001) or handwriting features (Johnson et al., 2017; Martire et al., 2018). Future research should further explore forensic examiners' statistical learning of lower-level featural information.

Conclusion

This study examined a psychological process theorised to be important in forensic feature-comparison expertise (statistical learning) in the domain of fingerprint examination. Fingerprint examiners demonstrated better statistical knowledge of fingerprint categories and used this knowledge in a same-source-likelihood task better than novices – better 'learners' also better applied this knowledge. Examiners' outperformance of novices demonstrates that statistical learning is another characteristic of fingerprint expertise. However, examiner performance was not perfect and these results emphasise the importance of developing reliable statistical databases in many forensic feature-comparison disciplines.

References

Aitken, C., Berger, C. E., Buckleton, J. S., Champod, C., Curran, J., Dawid, A., Evett, I. W., Gill, P., Gonzalez-Rodriguez, J., & Jackson, G. (2011). Expressing evaluative opinions: A position statement. *Science and Justice*, *51*(1), 1–2.

Aitken, C., & Taroni, F. (2004). *Statistics and the evaluation of evidence for forensic scientists*. John Wiley & Sons.

Aslin, R. N. (2017). Statistical learning: A powerful mechanism that operates by mere exposure. *Wiley Interdisciplinary Reviews: Cognitive Science*, *8*(1–2).

Bali, A. S., Edmond, G., Ballantyne, K. N., Kemp, R. I., & Martire, K. A. (2019). Communicating forensic science opinion: An examination of expert reporting practices. *Science & Justice*, 1–9.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*(3), 211–233.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *ArXiv Preprints*.

Bedard, J., & Chi, M. T. (1992). Expertise. *Current Directions In Psychological Science*, *1*(4), 135–139.

Biedermann, A., Garbolino, P., & Taroni, F. (2013). The subjectivist interpretation of probability and the problem of individualisation in forensic science. *Science & Justice*, *53*(2), 192–200. https://doi.org/10.1016/j.scijus.2013.01.003

Bird, C., Found, B., & Rogers, D. (2010). Forensic document examiners' skill in distinguishing between natural and disguised handwriting behaviors. *Journal of Forensic Sciences*, *55*(5), 1291–1295.

Brenner, L. A., Koehler, D. J., Liberman, V., & Tversky, A. (1996). Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes*, *65*(3), 212–219.

Bruce, N. D., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, *9*(3), 5–5.

Busey, T. A., & Parada, F. J. (2010). The nature of expertise in fingerprint examiners. *Psychonomic Bulletin & Review*, *17*(2), 155–160.

Busey, T. A., & Vanderkolk, J. R. (2005). Behavioral and electrophysiological evidence for configural processing in fingerprint experts. *Vision Research*, *45*(4), 431–448.

Busey, T., Nikolov, D., Yu, C., Emerick, B., & Vanderkolk, J. (2016). Characterizing human expertise using computational metrics of feature diagnosticity in a pattern matching task. *Cognitive Science*, *41*, 1717–1759.

Busey, T., Yu, C., Wyatte, D., & Vanderkolk, J. (2013). Temporal sequences quantify the contributions of individual fixations in complex perceptual matching tasks. *Cognitive Science*, *37*(4), 731–756.

Busey, T., Yu, C., Wyatte, D., Vanderkolk, J., Parada, F., & Akavipat, R. (2011). Consistency and variability among latent print examiners as revealed by eye tracking methodologies. *Journal of Forensic Identification*, *61*(1), 60–91.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81.

Daeid, N. N., Biedermann, A., Champod, C., Hutton, J., Jackson, G., Neocleous, T., Spiegelhalter, D., Willis, S., Kitchin, D., & Wilson, A. (2020). *The use of Statistics in legal proceedings: A primer for the courts*.

Daikoku, T., & Yumoto, M. (2020). Musical expertise facilitates statistical learning of rhythm and the perceptive uncertainty: A cross-cultural study. *Neuropsychologia*, *146*, 107553.

de Jongh, A., Lubach, A. R., Lie Kwie, S. L., & Alberink, I. (2019). Measuring the rarity of fingerprints patterns in the Dutch population using an extended classification set. *Journal of Forensic Sciences*, *64*, 108–119.

Dyer, A. G., Found, B., & Rogers, D. (2006). Visual attention and expertise for forensic signature analysis. *Journal of Forensic Sciences*, *51*(6), 1397–1404.

Ericsson, K. A., Hoffman, R. R., Kozbelt, A., & Williams, A. M. (2018). *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge University Press.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, *12*(6), 499–504.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *28*(3), 458–467.

Francois, C., & Schön, D. (2011). Musical expertise and statistical learning of musical and linguistic structures. *Frontiers in Psychology*, *2*, 167.

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin Journal*.

Galton, F. (1892). *Finger prints* (Issues 57490–57492). Macmillan and Company.

Growns, B., & Martire, K. A. (2020a). Forensic feature-comparison expertise: Statistical learning facilitates visual comparison performance. *Journal of Experimental Psychology: Applied*, 1–18. https://doi.org/10.31234/osf.io/pzfjb

Growns, B., & Martire, K. A. (2020b). Human factors in forensic science: The cognitive mechanisms that underlie forensic feature-comparison expertise. *Forensic Science International: Synergy*, *2*, 148–153. https://doi.org/10.1016/j.fsisyn.2020.05.001

Growns, B., & Mattijssen, E. J. A. T. (2020). Distributional Statistical Learning: How and How Well Can It Be Measured? *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*.

Growns, B., Siegelman, N., & Martire, K. A. (2020). The multi-faceted nature of visual statistical learning: Individual differences in learning conditional and distributional regularities across time and space. *Psychological Bulletin & Review*, *27*, 1291–1299.

Gutierrez-Redomero, E., Alonso-Rodríguez, C., Hernández-Hurtado, L. E., & Rodríguez-Villalba, J. L. (2011). Distribution of the minutiae in the fingerprints of a sample of the Spanish population. *Forensic Science International*, *208*(1–3), 79–90.

Gutiérrez-Redomero, E., Rivaldería, N., Alonso-Rodríguez, C., Martín, L. M., Dipierri, J. E., Fernández-Peire, M. A., & Morillo, R. (2012). Are there population differences in minutiae frequencies? A comparative study of two Argentinian population samples and one Spanish sample. *Forensic Science International*, *222*(1), 266–276. https://doi.org/10.1016/j.forsciint.2012.07.003

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist*, *39*(12), 1372–1388.

Henry, E. R. (1913). *Classification and uses of finger prints*. HM Stationery Office, printed by Darling and son, Limited.

Johnson, M. E., Vastrick, T. W., Boulanger, M., & Schuetzner, E. (2017). Measuring the frequency occurrence of handwriting and handprinting characteristics. *Journal of Forensic Sciences*, *62*(1), 142–163.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13).

Lee, M. D., & Danileiko, I. (2014). Using cognitive models to combine probability estimates. *Judgment and Decision Making*, *9*(3), 259–273.

Mandikal Vasuki, P. R., Sharma, M., Demuth, K., & Arciuli, J. (2016). Musicians' edge: A comparison of auditory processing, cognitive abilities and statistical learning. *Hearing Research*, *342*, 112–123.

Mandikal Vasuki, P. R., Sharma, M., Ibrahim, R. K., & Arciuli, J. (2017). Musicians' online performance during auditory and visual statistical learning tasks. *Frontiers in Human Neuroscience*, *11*, 1–12.

Martire, K. A., Growns, B., & Navarro, D. J. (2018). What do the experts know? Calibration, precision, and the wisdom of crowds among forensic handwriting experts. *Psychonomic Bulletin & Review*, *25*(6), 2346–2355. https://doi.org/10.3758/s13423-018-1448-3

Mattijssen, E. J. A. T., Witteman, C. L. M., Berger, C. E. H., & Stoel, R. D. (2020). Assessing the frequency of general fingerprint patterns by fingerprint examiners and novices. *Forensic Science International*, *313*, 110347. https://doi.org/10.1016/j.forsciint.2020.110347

Mnookin, J. L. (2008). The validity of latent fingerprint identification: Confessions of a fingerprinting moderate. *Law, Probability and Risk*, *7*, 127.

Mollon, J. D., Bosten, J. M., Peterzell, D. H., & Webster, M. A. (2017). Individual differences in visual science: What can be learned and what is good experimental practice? *Vision Research*, *141*, 4–15. https://doi.org/10.1016/j.visres.2017.11.001

Nordgaard, A., Ansell, R., Drotz, W., & Jaeger, L. (2012). Scale of conclusions for the value of evidence. *Law, Probability and Risk*, *11*(1), 1–24. https://doi.org/10.1093/lpr/mgr020

Phillips, V. L., Saks, M. J., & Peterson, J. L. (2001). The application of signal detection theory to decision-making in forensic science. *Journal of Forensic Sciences*, *46*(2), 294–308.

President's Council of Advisors on Science and Technology. (2016). *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*.

Risinger, D. M. (2013). Reservations about likelihood ratios (and some other aspects of forensic 'Bayesianism'). *Law, Probability and Risk*, *12*(1), 63–73.

Robson, S. G., Tangen, J. M., & Searston, R. A. (2021). The effect of expertise, target usefulness and image structure on visual search. *Cognitive Research: Principles and Implications*, *6*(1), 1–19.

Searston, R. A., & Tangen, J. M. (2017). Expertise with unfamiliar objects is flexible to changes in task but not changes in class. *PLoS ONE*, *12*(6), 1–14.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/doi:10.1002/j.1538-7305.1948.tb01338.x

Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2017a). Towards a theory of individual differences in statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*, 1–10.

Siegelman, N., Bogaerts, L., & Frost, R. (2017b). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods*, *49*(2), 418–432.

Streiner, D. L. (2003). Being inconsistent about consistency: When coefficient alpha does and doesn't matter. *Journal of Personality Assessment*, *80*(3), 217–222.

Thiessen, E. D., & Erickson, L. C. (2013). Beyond word segmentation: A two-process account of statistical learning. *Journal of Current Directions in Psychological Science*, *22*(3), 239–243.

Thiessen, E. D., Kronstein, A. T., & Hufnagle, D. G. (2013). The extraction and integration framework: A two-process account of statistical learning. *Psychological Bulletin*, *139*(4), 792–814.

Thompson, M. B., & Tangen, J. M. (2014). The nature of expertise in fingerprint matching: Experts can do a lot with a little. *PLoS ONE*, *9*(12), 1–23.

Thompson, M. B., Tangen, J. M., & McCarthy, D. (2013). Expertise in fingerprint identification. *Journal of Forensic Sciences*, *58*(6), 1519–1530.

Towler, A., White, D., Ballantyne, K., Searston, R. A., Martire, K. A., & Kemp, R. I. (2018). Are forensic scientists experts? *Journal of Applied Research in Memory and Cognition*, *7*(2), 199–208.

Towler, A., White, D., & Kemp, R. I. (2017). Evaluating the feature comparison strategy for forensic face identification. *Journal of Experimental Psychology: Applied*, *23*(1), 47–58.

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General*, *134*(4), 552–564. https://doi.org/10.1037/0096-3445.134.4.552

Vogelsang, M. D., Palmeri, T. J., & Busey, T. A. (2017). Holistic processing of fingerprints by expert forensic examiners. *Cognitive Research: Principles and Implications*, *2*(1), 15–15. https://doi.org/10.1186/s41235-017-0051-x

Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. %J S. (2008). *recaptcha: Human-based character recognition via web security measures*. *321*(5895), 1465–1468.

Welsh, M. B., & Navarro, D. J. (2012). Seeing is believing: Priors, trust, and base rate neglect. *Organizational Behavior and Human Decision Processes*, *119*(1), 1–14.

White, D., Jonathon Phillips, P., Hahn, C. A., Hill, M., & O'Toole, A. J. (2015). Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society B: Biological Sciences*, *282*, 1–8. https://doi.org/10.1098/rspb.2015.1292

Willis, S., McKenna, L., McDermott, S., O'Donell, G., Barrett, A., Rasmusson, B., Nordgaard, A., Berger, C., Sjerps, M., & Lucena-Molina, J. (2015). ENFSI guideline for evaluative reporting in forensic science. *European Network of Forensic Science Institutes*.

Wray, J., & Gates, B. (1996). Problems of recruiting participants for nursing research: A case study. *University of Alberta Health Sciences Journal*, *1*(5), 366–373.

Zacks, R. T., & Hasher, L. (2002). Frequency processing: A twenty-five year perspective. In P. Sedlmeier & T. Betsch (Eds.), *Etc. Frequency Processing and Cognition* (pp. 21–36). Oxford University Press.