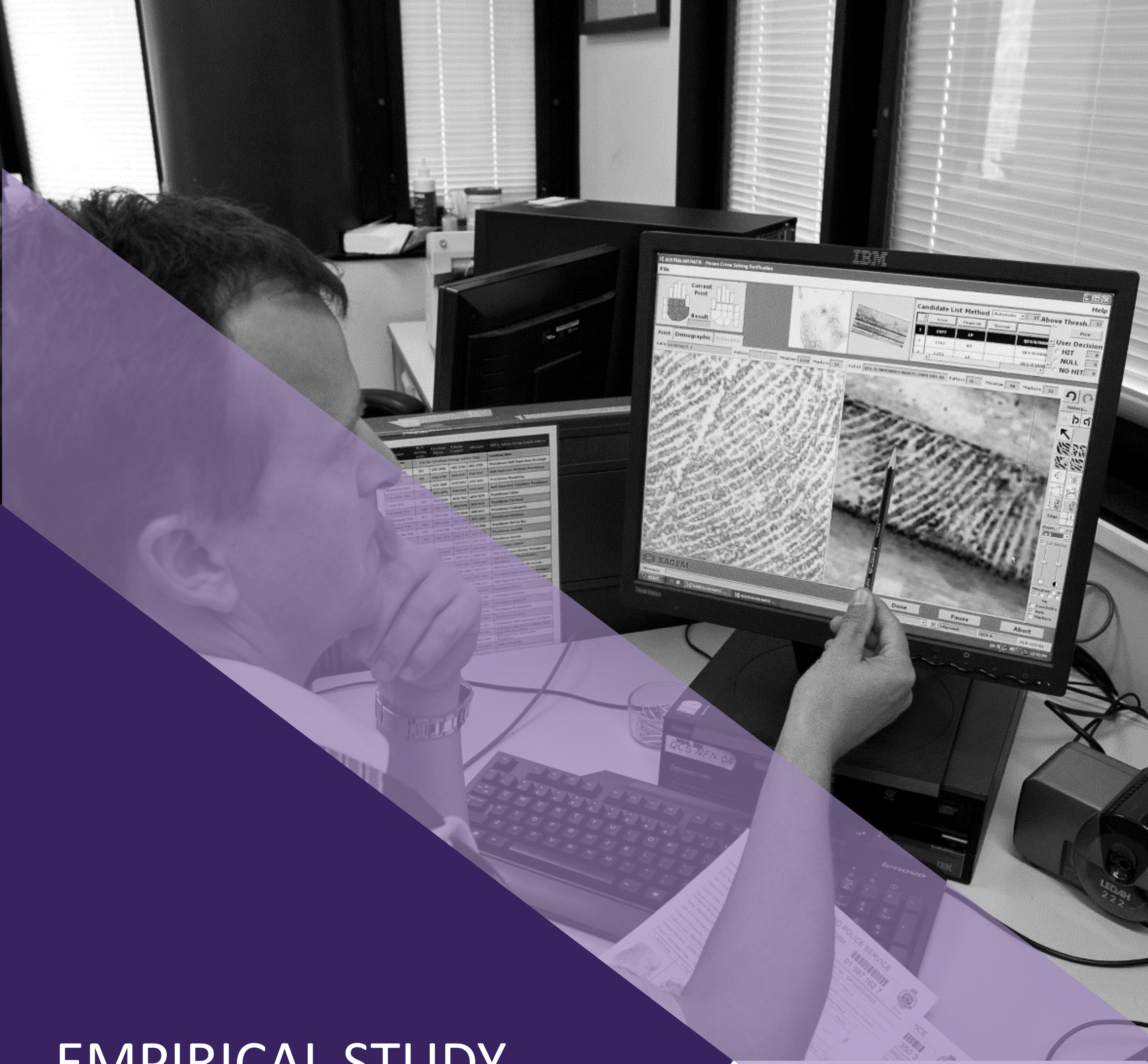


EMPIRICAL STUDY DESIGN IN FORENSIC SCIENCE

A Guideline to Forensic Fundamentals

2019

ANZPAA
Australia New Zealand
Policing Advisory Agency

NIFS
NATIONAL INSTITUTE OF FORENSIC SCIENCE AUSTRALIA NEW ZEALAND

Copyright Notice

© STATE OF VICTORIA 2019

This document is subject to copyright. Licence to reproduce this Document in unaltered form in its entirety (including with the copyright notice, disclaimer and limitation of liability notice intact) is granted to Australian and New Zealand Government bodies.

No other reproduction, or publication, adaption, communication or modification of this Document is permitted without the prior written consent of the copyright owner, or except as permitted in accordance with the Copyright Act 1968 (Cth). All requests and inquiries concerning reproduction or use of this Document other than as permitted by this copyright notice should be directed to ANZPAA, telephone 03 9628 7211 or email Business Support at: secretariat.support@anzpaa.org.au

The State of Victoria (represented by Victoria Police) is managing the Intellectual Property of this Document on behalf of the Members of ANZPAA in accordance with the current ANZPAA Memorandum of Understanding and the Members of ANZFEC in accordance with the ANZFEC Service Level Agreement. The governance processes generally associated with ANZPAA will manage the development and review of this Document

Disclaimer

This Document has been prepared to support forensic science services in Australia and New Zealand and may not be relied upon for any other purpose.

ANZPAA has taken reasonable care to ensure that the information provided in this Document is correct and current at the time of publication. Changes in circumstances after the time of publication may impact the accuracy or completeness of the information. It is the responsibility of the user to ensure they are using the most up-to-date version of this Document.

The information contained in this Document is necessarily of a general nature only and ANZPAA makes no representation or warranty, either express or implied, concerning the suitability, reliability, completeness, currency or accuracy of this Document.

This Document is not a substitute for users obtaining independent advice specific to their needs, nor a substitute for any jurisdictionally appropriate policies, procedures, protocols or guidelines and it is not intended to take precedence over such documents. All users of this Document should assess the relevance and suitability of the information in this Document to their specific circumstances.

Third Party Resources

This Document may refer to other resources, publications or websites which are not under the control of, maintained by, associated with, or endorsed by ANZPAA ('Third Party Resources').

Links and citations to Third Party Resources are provided for convenience only.

ANZPAA is not responsible for the content, information or other material contained in or on any Third Party Resource. It is the responsibility of the user to make their own decisions about the accuracy, currency, reliability and completeness of information contained on, or services offered by, Third Party Resources.

ANZPAA cannot and does not give permission for you to use Third Party Resources. If access is sought from a Third Party Resource this is done at your own risk and on the conditions applicable to that Third Party Resource, including any applicable copyright notices.

Liability

To the maximum extent permitted by law, the State of Victoria and Members of ANZPAA do not accept responsibility or liability (including without limitation by reason of contract, tort, negligence, or strict liability) to any person for any loss, damage (including damage to property), injury, death, cost, loss of profits or expense (whether direct, indirect, consequential or special) that may arise from, or connected to, the use of, reliance on, or access to any information provided or referred to in this Document or any information provided or referred to, or service offered by any Third Party Resource.

Members of ANZPAA and ANZFEC

The National Institute of Forensic Science is a directorate within the Australia New Zealand Policing Advisory Agency (ANZPAA NIFS).

ANZPAA is established by a Memorandum of Understanding between the following members: Victoria Police; Australian Federal Police; Australian Capital Territory Policing; New South Wales Police Force; New Zealand Police; Northern Territory Police; Queensland Police Service; South Australia Police; Tasmania Police and Western Australia Police, collectively, the 'Members of ANZPAA'.

The Australia New Zealand Forensic Executive Committee (ANZFEC) is established by a Service Level Agreement between the 'Members of ANZPAA' listed above and the following forensic service providers: ACT Health Government Analytical Laboratories; ChemCentre; Forensic Science Service Tasmania; Forensic Science South Australia; Institute of Environmental Science and Research; National Measurement Institute; New South Wales Health Pathology; PathWest Laboratory Medicine WA; Queensland Health Forensic and Scientific Services and Victorian Institute of Forensic Medicine.

References in this notice to ANZPAA are references to the Members of ANZPAA and the Members of ANZFEC.

Document Control

Version Number:	2.0
Date Distributed:	September 2019
Approved by:	ANZFEC
Status and Security:	Unclassified

Cover image
© The State of Queensland – (Queensland Police Service).

CONTENTS

PREAMBLE

This document was developed by a working group comprising forensic science practitioners, researchers and cognitive psychologists, with the aim to provide a resource that supports the continuous improvement of forensic science and its application in the criminal justice system. It is anticipated that this document will be used to both assess the ability of an existing empirical study to demonstrate the validity of a forensic science method or opinion, and as a point of reference when designing a new empirical study with the same purpose. Feedback on the contents of this document is welcomed and encouraged, to ensure that it continues to evolve to meet the needs of the forensic science community. All feedback and enquiries should be directed to: secretariat.nifs@anzpaa.org.au secretariat.nifs@anzpaa.org.au

INTRODUCTION

Empirical evidence is based on experimentation, systematic observation, or measurement, as opposed to theory or pure logic. In recent years there have been numerous questions raised about the empirical evidence that supports forensic science,[1] as well as an increasing effort to address these questions. However, there still exists a wide range of views and awareness in relation to what constitutes a good quality empirical study.

To this end, this document has been developed as a guide for evaluating or undertaking an empirical study in forensic science. It is not an exhaustive resource; rather, it consolidates numerous principles and practices and should form a helpful starting point. References that provide more detailed guidance on some of the aspects discussed have also been included. It is important to consider where expertise from other individuals may be required to ensure a robust empirical study is performed and reported. This will likely include individuals from different organisations, disciplines and possibly fields of expertise not previously considered.

It is intended that this document may be applied to all forensic science disciplines, whether they involve analytical instrumentation, or are more feature comparison based where human interpretation informs the result. It is important to note that analytical disciplines will generally still have some element of human interpretation, which should be tested (e.g. comparing chemical spectra or interpreting mixed DNA profiles etc). Where possible, the applicability to analytical and feature comparison disciplines has been noted throughout this document to acknowledge the different requirements for each.

[1] National Research Council, "Strengthening Forensic Science in the United States: A Path Forward," 2009; PCAST, "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods," 2016.

CHECKLIST

The following checklist has been developed to provide a broad overview of the requirements of a good quality empirical study in forensic science. For each question on the checklist, further guidance is provided in the subsequent pages of this document to assist in identifying whether the requirements have been met.

While a 'no' or 'partly' response for one question may not completely undermine the quality of a study, caution should be used when interpreting the results obtained and consideration should be given to the limitations that may arise as a result.

QUESTION	YES	PARTLY	NO
Is the claim under test well defined and appropriate? – see 'Claims' section	□	□	□
Has the empirical study been well designed? – see 'Experimental Design' section	□	□	□
Has a sufficient sample size been used? – see 'Sample Size' section	□	□	□
Have ground truth known materials been used that adequately reflect the materials encountered in forensic casework? – see 'Test Materials' section	□	□	□
Have the results been appropriately described and reported? – see 'Results and Reporting' section	□	□	□
Are the limitations of the empirical study clearly outlined? – see 'Limitations' section	□	□	□
Do the conclusions appropriately reflect the inferences that can be made from the results? – see 'Conclusions and Implications' section	□	□	□
Has the empirical study been critically reviewed and published? – see 'Review and Publication' section	□	□	□

CONSIDERATIONS

CLAIMS

The first consideration when designing or assessing an empirical study is to check whether the claim that is under test is well-defined. For the purposes of this document, the term 'claim' is used to refer broadly to the method, opinion or ability that is to be validated. This term has been selected as its use in forensic science has increased in recent years and it is intended to be interchangeable with other terms such as element,[2] assertion and hypothesis.

There is potential for the term claim to be interpreted as something that cannot be substantiated or is assumed beyond one's capabilities and even belief. However, in the context of this document it is not intended to have any negative connotations; rather, the term is routinely used in science to describe something that can be tested and therefore validated.

TESTABLE AND SPECIFIC

A claim should be testable and specific. A testable claim is one that sets an expectation that can be met or not met. The claim should generally involve a 'can' or 'does' statement and also provide information about how performance will be measured compared to one's expectation.

Examples include:

- Gas Chromatography-Mass Spectrometry can separate chemicals.
- Fingerprint experts can match fingerprints.

A specific claim is one for which an empirical study can be designed, without the need for too many different tests and measures. For example, the claim that 'forensic science works' is too broad as there are many disciplines and sub-disciplines of forensic science that would need to be examined. It would also be difficult to determine what evidence would indicate that forensic science is working or not.

'Trained bloodstain pattern analysts can determine whether blood was deposited via an air-borne or transfer mechanism'

This is a well-defined claim:

This claim can be tested in a robust, empirical manner. It specifies who should be tested, the type of test materials and what is being discriminated, namely, the ability for analysts to correctly determine whether a pattern was the result of air-borne blood or transferred blood.

'Bloodstain pattern analysis is a valid science'

This is a poorly-defined claim:

This claim is too broad to be tested in a controlled manner. The claim doesn't address how the analysis should be performed, the type of patterns to be differentiated, or the numerous possibilities that can be encountered in casework situations.

CLAIMS AND SUB-CLAIMS

There will be numerous claims for any one discipline. The level at which the claim is pitched is also important to ensure that is can be validated. In most cases it will be necessary to break a claim into multiple sub-claims.

An example is the comparison of footwear impressions, where the claim '*Experts can match impressions to the shoe that created the impression*' can be broken down to include the following sub-claims:

- Different shoes have class characteristics that are persistent and reproducible.
- Shoe outsole characteristics are transferred to an impression during deposition.
- Experts can differentiate between class characteristics and pattern types.

[2] ANZPAA NIFS, "A Guideline to Forensic Fundamentals," 2016, www.nifs.org.au.

- Experts can detect and compare randomly acquired characteristics.
- Experts can interpret the significance of matching and non-matching characteristics.

It may be easier to test individual sub-claims in separate empirical studies, rather than addressing the wider claim in a single empirical study. The same can apply to analytical disciplines, where the use of analytical instrumentation is often accompanied by the requirement for human interpretation.

An example is the analysis of explosives, where the claim *'Explosives can be detected and analysed'* can be broken down to include the following sub-claims:

- Analytical instrumentation can detect the presence of explosives.
- Experts can interpret the output of analytical instrumentation to confirm the presence of explosives.
- Experts can infer explosive potential from the quantity of explosives available.

Another useful reference when developing claims and sub-claims is the hierarchy of propositions, which refers to the scale of propositions that are formulated in forensic science for case assessment and interpretation. The scale is made up of three main levels; offence, activity and source.[3] In some disciplines, references are also made to the sub-source level.[4] These levels essentially relate to the guilt of the person (offence), the events that could explain the forensic evidence (activity) and the results of the comparison of questioned and reference materials (source). This is important because if the analysis in forensic casework is addressing activity level propositions, the claim should address the how well the method, expert or instrument performs at the activity level.

The development of claims and empirical studies for activity level assessments or scene interpretation/reconstruction can be challenging, due to the wide array of scenarios, actions and variables that must be considered. While a claim such as "Firearms experts can accurately reconstruct events at a shooting scene" may appear testable, it is composed of a large number of possible sub-claims such as "experts can determine muzzle to target range based on physical characteristics"; "experts can determine the trajectory of fired bullets"; "experts can differentiate between entry and exit wounds on a human body" etc. In each of these claims, there are a wide number of variables which must be operationalised (either manipulated or controlled), a large number of potential confounding variables, and a lack of potential knowledge about the ability to extrapolate between conditions. While it is still possible to empirically evaluate activity level claims, studies may be more complex to design, and require greater attention to study design factors described below.

Whenever a claim is being considered, it is important to understand what empirically derived information may already be available. A comprehensive literature review will provide guidance on what similar claims may have already been examined, as well as any work that may have already been performed in relation to the claim in question.

[3] R Cook, IW Evett, G Jackson, PJ Jones, and JA Lambert. 1998. A hierarchy of propositions: deciding which level to address in casework, *Science and Justice* 38: 231-239.

[4] D Taylor, J Bright, and J Buckleton. 2014. The factor of two issue in mixed DNA profiles. *Journal of Theoretical Biology* 363: 300-306; D Taylor, D Abarno, T Hicks, and C Champod. 2016. Evaluating forensic biology results given source level propositions, *FSI Genetics* 21: 54-67; D Taylor, A Biedermann, T Hicks, and C Champod. 2018. A template for constructing Bayesian networks in forensic biology cases when considering activity level propositions, *FSI Genetics* 33: 136-146.

EXPERIMENTAL DESIGN

Experiments should be constructed to ensure that results will be fit for purpose and robust enough to withstand critical review. There are numerous references that provide relevant experimental design guidance in the fields of science, mathematics and cognitive psychology. The purpose of this section is to highlight specific considerations that are often important for empirical studies in forensic science. These include:

- Test Materials
- Sample Size
- Operationalising Variables
- Confounding Variables
- White Box and Black Box Design
- Open Set and Closed Set
- Blinding of Participants and Assessors
- Use of Controls
- Inconclusive Responses
- Replication.

TEST MATERIALS

Wherever possible empirical studies should be conducted with materials for which the answer is known, commonly referred to as ground truth known materials. Testing methods, participants or theories on samples with unknown origin does not provide information about accuracy as it is impossible to know for certain whether or not the answer is correct. Ideally, study materials should be created for assessment purposes using known methods and items. For some disciplines, this is a relatively simple although potentially time consuming exercise (e.g. obtaining latent fingerprints, DNA samples, fired bullets or shed fibres). For other disciplines, ground truth known materials may be more difficult to create or obtain and often have significant ethical or practical implications (e.g. materials fire investigation, pathology and bloodstain pattern analysis).

INFERRING GROUND TRUTH

Where ground truth known materials cannot be created, it may be possible to use casework material if the true answer can be reliably inferred from another source. For example, material for cause of death investigations in pathology cannot be created, but a selection of cases may be able to be sourced where the death was captured on CCTV or can be verified through other reliable means. Secondary sources of this kind can provide ground truth knowledge, and allow the use of the case materials for subsequent testing. This inference of truth must be from a highly reliable and accurate source, and should not involve expert opinion on the claim being tested. Using expert opinion to infer ground truth on casework materials for a test of the accuracy of the expert opinion suffers from circular reasoning and potential error. If it is not possible to create or obtain ground truth known materials in any way, then the appropriateness of the claim should be evaluated.

REPLICATING CASEWORK

It is also important to ensure that the test materials used reflect the range of materials and difficulty encountered in casework and that conditions are consistent with those in an operational setting. For example, a ‘difficult’ pair of non-matching fingerprints would be selected to appear as though they came from the same source (e.g. a close non-match from a large database search), and a difficult pair of matching fingerprints would be selected to appear as though they came from different sources (e.g. though distortion or reversal). If examinations are always conducted using original material, the test material should be original, unless it has been demonstrated that non-original materials do not impact accuracy. Similarly, testing only complete, high quality samples will not explore the accuracy of the method on partial, distorted and degraded material. Studies that do not use degraded, partial stimuli may overestimate the true rate of performance. Likewise, tests for claims which involve multiple hypotheses, such as bloodstain pattern analysis, fire investigation or pathology should include as many potential scenarios within the testing paradigm as possible, to avoid examiners choosing between only two to three possibilities. A validation is not a test for 100% performance; it is a tool to determine when a method works and when it does not. In fact, evidence of 100% correct responses is an indication that the test materials were not sufficiently complex. The experimental design should have considered the range of outcomes possible to ensure that the outer bounds of the claim are assessed.

SAMPLE SIZE

The nature of forensic science, as well as practical and ethical considerations often mean that obtaining sufficient sample sizes is difficult. It is important to include enough data points or observations to be able to generalise to the required population; therefore, the expected variability in your population will influence the required sample size. Low variability may mean fewer data points or observations are required, while high variability will generally mean that a greater number of data points or observations are required to fully estimate the range of possible answers. There is no hard and fast rule as to what constitutes a sufficient sample size, and each study should be assessed individually. But it is important to specify before conducting any analyses or even before collecting any data, how large the sample size will be for the experiment, so it is clear exactly when data collection will end. It is often necessary to consult with someone who is experienced with experimental methodology and statistics for advice; however, some guidance is provided below.

STATISTICAL POWER

In many studies, the focus is on establishing that there is a difference between two or more populations. This can include for instance a study on the efficacy of a particular method or treatment, with one population being a control group and the other having been subjected to the technique in question. In this case one generally refers to the null hypothesis and alternative hypothesis: the null hypothesis is the supposition that there is no difference between the groups, and the alternative hypothesis is that there is one. The researcher is generally interested in the question of whether the null hypothesis can be rejected, in other words whether we can conclude that a difference in measurement between groups is highly likely to be due to some factor other than chance alone.

In this context there are two possible types of errors that a study may make, even when the statistics are handled correctly. The first is that we may reject the null hypothesis when it is true, i.e. the study result is a false positive. This is known as a Type I error, and the likelihood of this occurring is known as the significance level of the study. This quantity is calculated theoretically, and the standard significance level considered acceptable for most applications is 0.05; that is, we will reject the null hypothesis when we should not less than 5% of the time.

The other type of error is a false negative, also known as a Type II error. This is when we should reject the null hypothesis (that is, there really is a difference between the populations) but are unable to from the data available. The robustness of a study to Type II errors is known as the power of the study. Unlike with the significance level, a higher power is considered better; this is because if *p* is the power, then the probability of making a Type II error is *1-p*. The power of a study generally increases as the sample size does, and ensuring that a study has sufficient power often equates to ensuring that the sample size is adequate. Performing power calculations by hand generally requires sophisticated knowledge, but statistical software such as G*Power or PANGEA is available, which can remove the burden from the researcher.

STATISTICAL TESTS

Acquiring basic knowledge of any statistical tests applied in a study is of value in this context. Many of the most common statistical tests rely on the Central Limit Theorem, which states that the means of independent measurements of the same quantity have a tendency to obey a certain pattern known as a normal distribution. The t-test and Chi-square test are examples of such tests. Each test of this type has rules of thumb that should be satisfied in order to run the test. As an example, the standard rule for the Chi-square test is that the expected number of subjects in each category should be at least five. The t-test can in principle be applied to any sample of size two or more; however, samples which are small depend on a normality assumption, which is that the population in question follows a normal distribution. There are ways of testing this assumption, and when it does not appear to be valid, then a t-test is not suitable for small samples. Most references that discuss these tests will also include this type of relevant information.

OPERATIONALISING VARIABLES

Variables can include the manipulations (e.g. easy versus difficult test materials) and expectations (e.g. pass/fail standards) of the empirical study. These should be set before data is collected and should not evolve over the course of the study. This is known as operationalising the variables of interest.

When manipulations and expectations are not explicitly defined from the outset, there is a possibility that definitions may change over time. This can result in a misrepresentation of the outcomes of the study and may give a false impression of the results. These manipulations and expectations should be set in light of the claim of interest and the use of pilot studies or manipulation checks may assist in providing a greater understanding of expected responses or results. Pilot studies are a smaller scale preliminary study performed to estimate the difference expected in the data (effect size) as well as any issues in the experimental design. Manipulation checks involve altering one variable in the study to see if it has any impact on the variable of interest. The latter may also be used to assess the difficulty/complexity of the test materials that have been used.

The types of manipulations and expectations will depend on the type of empirical study that is performed. Most claims will require a measure of sensitivity, accuracy and precision. For feature comparison disciplines it will likely be important to define a match and a non-match, while analytical disciplines usually require more quantitative measures such as limit of detection and limit of quantitation.[5] Studies that have not operationalised their variables in sufficient detail to prevent gradual shifts in expectations or the characterisation of results should be regarded with caution. Examples of correctly operationalised variables for studies may be:

- Cognitive task claim: DNA analysts can determine the number of contributors to a complex DNA profile.

  Manipulations: Number of contributors (3, 4 and 5); Mixture ratios (resolvable major vs non resolvable); Allele sharing (high degree of allele sharing vs low degree). Expectations: Accuracy of determinations (relative to ground truth) will decrease in a linear manner as number of contributors' increases, the ratio between contributors decreases, and allele sharing increases.

- Analytical task claim: Instrument A produces equivalent quantitation results to Instrument B.

  Manipulations: Quantity of sample (high vs low); Quality of sample (high vs low); Complexity (single source vs mixed). Expectations: Instruments are deemed equivalent if no statistical difference ($p < 0.05$) is found between samples of different quantities/qualities/complexities, and limits of detection/quantitation are statistically equivalent.

[5] NATA, "General Accreditation Guidance – Validation and Verification of Quantitative and Qualitative Test Methods," 2018, www.nata.com.au.

CONFOUNDING VARIABLES

A confounding variable is an extraneous (unintended or uncontrolled) variable that could plausibly account for observed patterns in the data. For example, if assessing the ability of a firearms examiner to accurately match bullets fired from the same weapon, it is important that the only variable that could account for their match accuracy is their discipline specific knowledge. Something as simple as using the same label for all matching bullets, which is different to the label used for non-matching bullets, could impact the decision making of participants.

Wherever possible it is important to control everything and only select one variable to test at a time. Where there is a risk of a confound, it may be useful to try:

- Counter-balancing – use of a controlled approach for the selection of participants and materials

  E.g. for signature analysis half of both the forged and genuine signatures are in blue pen while the other half are in red pen.

- Randomisation – use of a randomised process for selection of participants/materials

  E.g. for signature analysis a random pen colour is used for each signature regardless of whether it is genuine or forged.

If neither of these options address the problem, the claim or the inferences made from the results obtained may need to be limited. Attention should also be paid to a potential trial order effect, whereby the order in which the materials are assessed by participants can impact decision-making, instead of the stimuli or method.

WHITE BOX AND BLACK BOX DESIGN

When considering experimental design, it may be important to consider the benefits and limitations of black box and white box design. This is especially relevant in the assessment of human performance, and although these types of studies have been used in software development for a number of years, a helpful discussion on their application in forensic science is provided in the PCAST report.[6] A black box design assesses the final output of the system as a whole. This will investigate the claim as if it were a 'black box' in the examiner's mind and provides a measure of the accuracy. A white box design investigates the processes within a system that are involved in generating a result. This will provide an understanding of the factors that affect an examiner's decision-making. Where no studies exist, the measure of reliability and accuracy is the priority and often met through black box studies. Over time an examination of the thought processes involved is valuable, so as to identify opportunities for improvement, including error management. This will often only be achievable through white box studies.

[6] PCAST, "Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods," 2016.

OPEN SET AND CLOSED SET

As it can be difficult to collect enough test materials or to recruit a sufficient number of participants for an empirical study, it can be tempting to use a closed set experimental design strategy. This involves giving participants a limited number of test materials and asking them to perform pairwise comparisons, to determine which pairs match.

While this approach may provide a large number of comparisons, the individual comparisons are not independent of each other. The difficulty of the comparison decreases each time, until no comparison is needed at the last step. Generalising from these closed set studies to open set forensic casework is not possible, as the claims are not consistent. Open set design on the other hand is more appropriate to ensure that the accuracy rates achieved are realistic, given the independent nature of the comparisons. The set may also be partly open, whereby the questioned materials are compared to a collection of reference material, with some questioned materials that did not originate from the reference set also included.

Open set design does require more test material and may be harder to construct; however, it is critical that the test design reflects the claim and operational casework realities. If casework involves a one to one comparison of questioned and reference samples, the comparison in the study should be designed to reflect this. If part of the claim involves an understanding of the possibility of random matches, the likelihood of a particular event occurring, or the support the evidence provides for a particular hypothesis, then the study design should be broad and open enough to ensure participants are considering all potential possibilities.

'Examiners are provided with:

- *10 unknown fired bullets*
- *10 bullets fired from 10 known guns*

and the task is to determine which of the 10 known guns fired each of the 10 unknown bullets.'

This is a closed set:

The first bullet examined will have 10 possible answers that must be evaluated; the second will have only nine; the third eight and so on. Once the ninth bullet is compared to the two remaining guns, the tenth bullet must have been fired from the one remaining gun.

BLINDING OF PARTICIPANTS AND ASSESSORS

Research in cognitive psychology over the last several decades has revealed that people can be influenced by contextual information without awareness,[7] and that experiences and expectations can shape people's judgements.[8] The implications of these findings are not only important for how forensic science is conducted, but also for undertaking empirical studies. For example, participants might behave differently if they know they are participating in an experiment (referred to as 'subject-expectancy' effect). This can be minimised by blinding participants to the purpose of the study until after the experiment. An extreme scenario would be to insert test materials into routine casework with participants unaware that they are participating in a study, although this may not always be achievable, ethical or desirable.

It is also important to consider what information participants need to know to aid their decision making. For example, if the purpose of the exercise is to identify differences between two matching items that develop over time, consideration should be given to whether the participant needs to know that they do in fact match. Research shows that if examiners know two items match, they perform a different assessment of the similarities and differences, compared to if they assess the presence or absence of features independently of a comparator.

Experimenters are also prone to these cognitive biases (referred to as experimenter-expectancy effect), which can affect the way experimenters assign participants to conditions or interact with them given this knowledge.

[7] G Edmond, A Towler, B Growns, G Ribeiro, B Found, D White, K Ballantyne, RA Searston, MB Thompson, JM Tangen, R Kemp, K Martire. 2017. Thinking Forensics: Cognitive science for Forensic Practitioners, *Science and Justice* 57: 144-154.
[8] G Edmond, JM Tangen, RA Searston, IE Dror. 2015. Contextual bias and cross-contamination in the forensic sciences: the corrosive implications for investigations, plea bargains, trials and appeals, *Law, Probability and Risk* 14: 1-25.

As a result, it is also important to take steps to minimise their influence, such as remaining blinded to which condition each participant is in as an example.

USE OF CONTROLS

Analytical methods will often require standardised reference materials, as well as positive and negative controls to ensure that the results obtained are reliable. All claims that involve an element of human expertise should also involve human performance testing. This is often performed by comparing the performance of a trained practitioner to that of a lay person, with the view that expertise demonstrated by the practitioner, will exceed the level demonstrated by the lay person. If for example, a random group of people could distinguish between known matching and non-matching toolmark images just as well as trained practitioners, it would indicate that the current training and expertise does not distinguish a practitioner from a lay person on this task. It is the extent to which a trained practitioner exceeds a lay person that demonstrates competence and expertise, and comparing the two groups will highlight at what point during training this difference appears.

In cases where specific technical information is required to understand the test materials, such as for DNA or drug analysis, it may not be appropriate to include a true lay person in the assessment. In these cases, a more appropriate lay person might be a trainee scientist for example, who has an understanding of the background methodology but no experience in the actual task being performed.

INCONCLUSIVE RESPONSES

It can be tempting to place few or no restrictions on the judgements, decisions and answers permitted by participants, but this approach can often lead to difficulty in interpreting the data obtained. Restricted response options minimise noise and data loss and, if defined appropriately, also ensure that the data speaks directly to the claim being tested.

Where the claim relates to the accuracy of expert opinions in case-like scenarios, participants should be permitted to use the verbal, numerical or open conclusion scales used in casework. However, the interpretation of results may be limited by certain subjective conclusions such as inconclusive, which is difficult to mark as either correct or incorrect. This may be resolved by incorporating binary responses (e.g. match/no match) alongside a conclusion scale and has the benefit of providing additional information about:

- whether examiners are able to complete the task accurately
- what examiners infer
- whether examiners are biased towards conservative or non-conservative responses when using their scales
- whether different examiners use the same scale in the same or different ways.

Studies that only take a one or the other approach will often be limited in some respect.

RESULTS AND REPORTING

Methods of reporting statistics can generally be classified as being either descriptive or inferential. As the name suggests, descriptive statistics are quantities that describe the data. These include the data points themselves, calculations from the data such as mean, median and variance, as well as data visualisations such as graphs. Inferential statistics, on the other hand involve using the data from a sample in order to deduce properties of a population, or to predict future behaviour. Inferential statistics will generally involve applying one or more standard statistical tests, such as the t-test or Chi-square test, in order to test the claim. Generally, if the claim under test addresses differences between conditions/groups/participants, there must be inferential statistics. If no difference is claimed, descriptive may be sufficient, but this will depend on the study.

When presenting results, the use of data visualisations will assist by representing as much of the data and the variation as possible. Generally, the most informative way to present the information will be the use of a scatter plot to present all data points; however, if this is not possible using box plots or errors bars can provide a greater representation of data. Whenever variance in data is presented, it is important to use an appropriate measure such as confidence intervals or standard error. Simply presenting the mean will often be insufficient. While data visualisation is useful, the raw data should be available in some form, either as an appendix, open access or on request, to allow other groups to perform their own assessment.

It is also important to ensure that measures are put in place to avoid overgeneralising or presenting inferences that are not supported by the data. For example, if an error rate is presented it should be clear as to whether this relates to a single laboratory, a specific practitioner or an entire discipline. It is recommended that those who are unfamiliar with statistics contact a qualified statistician to seek advice as required.

LIMITATIONS

When considering the limitations of a study, it is important to distinguish those that relate to the design of the study and those that are limitations to the applicability of results. A good quality study will often outline the limitations and while these should be seriously considered, it is important to identify where additional limitations may arise because of a mismatch between the study parameters and the scope of the inferences made.

Examples of limitations that may occur include small sample sizes, the impact of fatigue or practice effects on large trials, changed response profiles due to participation in a trial, the use of unfamiliar testing platforms or software, or the inability to generalise to a wider population of examiners/methods/sample types.

CONCLUSIONS AND IMPLICATIONS

The conclusions of any study should summarise the findings and inferences within the bounds of the experimental design. It should be made clear what can be inferred from the data and these inferences should be directly related to the claim. When looking to apply the study, the data should be examined to ensure that it is presented in full and is suitable to draw conclusions and inferences from. Care should be taken to ensure that the study does not overgeneralise, and identification of what should not be inferred from the data will also assist in the application to forensic casework.

For example, a study testing an examiner's ability to match and interpret fibres, conducted in a single laboratory using a single method, should not be taken to provide validity to all laboratories, using all methods, across all trace evidence. Likewise, finding a high level of accuracy in the analysis of single source, high quantity DNA samples should not be used to conclude a similar level of accuracy on mixed, low quantity DNA samples.

REVIEW AND PUBLICATION

It is important to share the findings of a study to ensure that they can be considered by anyone aiming to address a similar claim. Ideally, the study should be published in a recognised peer-reviewed journal as this adds confidence that the study has been critically reviewed by experts in the field. However, not all peer-reviewed journals are equal, with many different practices in place to elicit critical review from the scientific community. While a high journal impact factor is informative, careful consideration should be given to the composition of the review board, the process for addressing reviewer comments and the general quality of material published in the journal. In 2013, the Department of Justice established the National Commission on Forensic Science in partnership with the National Institute of Standards Technology, to act as a Federal Advisory Committee. Throughout the term of the committee, a number of documents were approved that are of relevance to the publication of scientific literature:

- **Views of the Commission – Scientific Literature in Support of Forensic Science and Practice**
  https://www.justice.gov/archives/ncfs/file/786591/download – *accessed 25-01-2019.*
- **Views of the Commission – Identifying and Evaluating Literature that Supports the Basic Principles of a Forensic Science Method or Forensic Science Discipline**
  https://www.justice.gov/archives/ncfs/file/839716/download – *accessed 25-01-2019.*
- **Views of the Commission – Technical Merit Evaluation of Forensic Science Methods and Practices**
  https://www.justice.gov/archives/ncfs/file/881796/download – *accessed 25-01-2019.*
- **Recommendation to the Attorney General – Technical Merit Evaluation of Forensic Science Methods and Practices**
  https://www.justice.gov/archives/ncfs/page/file/905541/download – *accessed 25-01-2019.*

There will be occasions where publication in a peer-reviewed journal is not possible and if this is the case, appropriate internal review is vital. This review should be independent and include reviewers that are familiar with experimental design, research and statistics. A review that is only performed by internal staff including managers or a quality team who lack the required expertise, may not necessarily detect errors in design, analysis or interpretation. Having an independent reviewer evaluate the study using the checklist provided in this document could be beneficial.

There are standard conventions for reporting the results of a controlled empirical study that should always be followed. Although the exact format and content requirements may differ depending on the nature of the study, the type of report and the place of publication, all reports of scientific experiments need to include enough information so that reader:

- could reproduce the study exactly as performed
- can interpret the data independently
- can evaluate the conclusions against the stated claims and predictions.

Most scientific reports follow the IMRD format, where the Introduction outlines the claim and the background to the claim, the Methods state the hypotheses and describes the testing procedure, the Results describe what was found, and the Discussion evaluates the results against the hypotheses, as well as the implications of the results for the claim under test. Commonly, limitations associated with the study are provided in the Discussion section, as are comparisons with previous studies. Additional data, materials and analyses may also be placed into Appendices, particularly for online journals.

Within the report, the methods used to carry out the study must be detailed exactly and precisely. This includes describing all stimuli used during the test, the testing conditions, locations and timing. If the study involves human participants, details should be given regarding their recruitment and demographic information, how the test was administered to them (including instructions given), how participants were compensated (if at all), and how they reported their results. All statistical tests used should be detailed, as should any data cleansing or elimination of anomalous results. Ideally, all materials used should be openly available, including pictures of stimuli, questionnaires given to participants, and raw data from all experiments. If this is not possible, examples of representative samples and results should be provided. All assumptions made regarding the stimuli, test or participants should also be detailed, as these may impact on the validity of the results and conclusions.

It is also important that any potential conflicts of interest are declared, to allow the reader to make an assessment when determining the applicability of results to the claim of interest.

SUPPLEMENTARY MATERIAL

STUDY DESIGN EXAMPLES

APPROPRIATE STUDY DESIGN

Example 1

- To test the claim that forensic pathologists are able to correctly interpret injuries caused by motor vehicle accidents, a collection of X-rays, CT scans and pathological case notes are collated from a number of jurisdictions. The ground truth was established in each case via CCTV and dash-cam footage. Clinical findings were obtained from case notes, and combined with images to create test stimuli. Forensic pathologists were recruited to participate in the study, and provided with a selection of the cases from other jurisdictions in randomised order. They were informed that it was a study of injuries from motor vehicles. Participants were instructed to review the case materials without consulting anyone else, and form conclusions regarding the nature of the vehicle, the impact and any other potential information able to be derived. Responses were marked according to a predetermined key by two independent analysts.
  - Strengths: well defined claim, ground truth for materials inferred from reliable sources, marking criteria defined before study commenced, randomisation of variables.
  - Limitations: no indication of the number of cases involved, restricted to cases where pathologist is aware that the injuries are the result of a motor vehicle.

Example 2

- The ability for toolmark examiners to recover and correctly interpret obliterated serial numbers on firearms was investigated. Two hundred confiscated firearms of various makes with unobscured and unaltered serial numbers were obtained, and classified on the basis of method of serial number marking (i.e. stamping, engraving or casting) and base metal composition. Firearms of common stamping method and metal were then grouped, and randomly assigned to an obliteration method (drilling, over-stamping, or grinding), such that each of the fifteen groups represented contained at least five firearms with obliterated serial numbers. Due to differences in serial marking method and metal type, it was not possible to equalise sample numbers in each group. Forty participants were recruited via professional networks, with full knowledge that their ability to restore and interpret the serial number was being tested. Each participant was provided with five firearms, selected at random from each group, such that all fifteen groups had at least five firearms in the test population. Participants were instructed to select and use the most appropriate method of restoration for each, documenting the process photographically during and following the restoration. An answer sheet was provided for participants to transcribe the recovered serial number, along with their justification for the interpretation and details of the method used. Answers were marked by two researchers against the known serial number for each firearm, with predefined keys for the completeness of recovery, accuracy and reasoning. Statistical tests were used to compare the accuracy of each recovery method, stamping method and metal type to determine the most reliable method for each type.
  - Strengths: well defined claim, incorporation of human interpretation component, sufficient sample size given the examiner population, ground truth known material used, marking criteria defined before study commenced, randomisation of variables, statistical testing of results.
  - Limitations**:** incomplete testing (examiners did not receive test items for all groups), interpretation of recovered numbers not independently re-interpreted to determine possible source of errors.

Example 3

- An internal validation study was conducted to determine the accuracy and reliability of the detection of methamphetamine using GC-MS. The sensitivity and specificity of the system was investigated by analysing commercially obtained methamphetamine and internally synthesised samples, as well as closely related compounds such as phetamine, with a total of 50 samples analysed. In addition, 20 sample matrices containing common diluents and impurities, as well as the target methamphetamine, were created through the mixing of controlled amounts of known standards of each component. Results were then checked to ensure that all compounds were well-separated and identifiable, with three analysts unaware of the matrix composition interpreting the chromatogram independently. All compounds were identified correctly in all samples. The precision of the GC-MS method was investigated by examining the area ratios of the methamphetamine peak to the internal standard peaks in a total of 50 replicate injections, with replicates conducted on the same day (n = 5), on different days (n = 5 days), and on different instruments (n = 2 instruments). F tests showed no significant variation in precision. The limits of detection and quantification were determined using a dilution series of methamphetamine standards and blank injections. The LOD was calculated using a 5 x signal to noise threshold, with the appropriateness of this confirmed via repeated injections of the lowest quantity of methamphetamine deemed to be detectable. The LOQ was calculated from the lowest quantity of methamphetamine that could reliably be quantified after 10 repeat injections of the dilution series. This dilution series was also used to determine the linearity of measurement, with quantification results examined for normality and linear correlation to concentration.

  - Strengths: well defined claim, ground truth known samples, control of variables, replication of case work conditions, sufficient sample numbers relative to variation of system, blind interpretation of results, relevant variables identified and investigated, thresholds and limits of system identified.

  - Limitations: human interpretation component not sufficiently examined – intra-analyst variation of interpretation is unknown.

Example 4

- A study was conducted to determine if bloodstain pattern analysts could accurately analyse stains from photographs, compared to attendance at the scene. A vacant house was utilised for testing, with five mock scenes set up across five separate rooms. Scenarios were constructed and enacted that resulted in the deposition of swipes and wipes, expirated blood, impact stains, cast off, drip patterns, altered stains and flow patterns, with each scene containing two to three different deposition methods. 50 authorised examiners and 13 trainees were recruited to participate in the study, and split randomly into two groups. Group 1 participants were given access to a room at a time, and requested to provide their opinion on each identified stain in accordance with their standard reporting practice. Following documentation of this opinion, the examiner was then asked to choose the most likely mechanism of deposition, based on a restricted and standardised set of options, which was fully explained to the participant. Estimations of confidence in accuracy were also elicited, on a scale of 1-10 (no confidence to complete confidence). Group 2 participants were provided with complete sets of photographs of each scene taken by a qualified crime scene photographer, and asked to provide the same information as the scene participants. Following completion of this testing, the house was fully cleaned, and five new scenes were created. The mechanism of stain deposition was kept constant from the first set, but combinations and locations were randomised across the five scenes to provide different scenes. Group 1 then examined and reported on the photographic set, and Group 2 examined and reported from the scenes. Responses were anonymised (stripped of participant number), and checked independently by three authorised BPA examiners, with the non-standardised responses interpreted for accuracy against ground truth and standardised responses against pre-determined correct answers. The accuracy of each determination was correlated against the confidence rating provided by each participant. Error rates were calculated for each participant in each condition, with t-tests used to compare the two conditions.

  - Strengths: well defined claim, moderate sample size, use of control group, counter-balancing of variables, ground truth known stimuli, defined response scales, statistical testing of results.

  - Limitations: large number of variables (stain types, scenarios).

Example 5

- The accuracy of the identification, interpretation and matching of paint fragments was studied through an inter-laboratory controlled trial. To obtain samples with sufficient complexity and range, a collaboration was developed with local trade schools that taught automotive spray painting and house painting. As part of the teaching curriculum, students repeatedly repainted the same items, keeping records of the date painted, the type and colour of the paint, and an estimation of the thickness of the layer applied. Students were instructed which paints to use in which order, to ensure that some items had paint profiles that differed only slightly from each other. Fragments were then chipped off the surfaces, with each fragment containing between 3 and 7 layers of paint. Two experienced examiners then selected chips to make up test kits for participants, with each kit containing 20 paired samples of varying size, difficulty and complexity. Within each kit, 10 samples came from the same origin, and 10 from different origins. The different source pairs were matched as closely as possible for paint type, colour and layer number to create difficult comparisons. Participants were instructed to sequentially analyse the samples, beginning with microscopy, followed by FTIR, then Pyrolysis GC/MS. At each stage, they were to document their observations, and complete a multiple choice questionnaire regarding their interpretations, including questions about their opinion on the type of paint, number of layers and possible origin. 35 participants completed 20 samples each, resulting in 700 opinions. These opinions were compared to the known origin of the paint, and the accuracy of interpretation was marked for each stage of the analysis procedure, with false positives and false negatives calculated.
  - Strengths: well defined claim, ground truth known samples, inclusion of difficult/complex samples, open set design, large sample number, controlled testing regime, standardised documentation of results.
  - Limitations: different samples provided to different participants meaning that practitioner proficiency cannot be easily inferred.

INAPPROPRIATE STUDY DESIGN

Example 1

- Fingerprint examiners were recruited to participate in an expertise study. Participants were provided with 10 x latent fingerprints and 10 x ten-print reference sets, and were required to nominate the donor of each of the latent fingerprints. The fingerprints were all of good quality had been obtained from staff within the laboratory and displayed a wide range in pattern types.
  - Limitations: closed set design, small sample numbers and easy stimuli (i.e. lack of close non-matches and potential familiarity with impressions).

Example 2

- The uniqueness of handwriting was examined by collecting a large number of requested handwriting samples from multiple classes from a local high school. Each student wrote a standard paragraph on a lined sheet, along with their age, preferred writing hand and whether they had moved schools in their life on the same piece of paper. A trained document examiner involved in this study then compared the writing of all students to determine if any characteristics or letter formations was the same between writers.
  - Limitations: human bias effects (i.e. context information provides an indication of identity, only a single examiner involved who also participated in other parts of the study).

Example 3

- Forensic biologists were tested for their ability to determine the manner of DNA deposition, based on a summary of case details and the DNA profile obtained. 12 participants were recruited after attending a training course on DNA transfer. To minimise the time required from each participant, one DNA profile was used throughout the experiment, with case information progressively revealed across six scenarios. At each scenario, the participant was required to rate whether they believed the evidence supported or strongly supported either primary or secondary transfer.

  - Limitations: small sample size with non-independent samples and judgements between stimuli meaning errors could be compounded across trials, lack of a control group (i.e. trainees) means the difficulty of the test is unknown, unclear if ground truth was known for the single sample used.

Example 4

- A study was conducted to test the accumulation of wear and Randomly Acquired Characteristics (RACs) on shoe outsoles. 10 pairs of shoes of the same brand and type were purchased and provided to volunteers. A series of group walks were organised, where the volunteers completed a 10 kilometre hike across a variety of terrains (bush trails, roads, running tracks etc). Test impressions and photographs were taken of each outsole before and after each walk, and clearly marked with the participant's ID, cumulative distance covered and date of walk on the impression and photograph. Shoes were not worn outside these controlled walks. An authorised examiner, who was involved in the design of the study, then scrutinised and recorded all impressions and photographs, noting areas of wear and identifying RACs. The average amount of wear and number of RACs was calculated across the twenty individual shoes, and used to conclude that increasing use of shoes creates increased wear and damage to outsoles.

  - Limitations: small sample size, single examiner performing the interpretation, interpretation not performed blind, no statistical testing performed, results over-generalised (from a single type of outsole to all outsoles).

Example 5

- A case required the assessment of speed of a vehicle from CCTV footage. Photogrammetry was used to estimate the speed as 92km/h, but the method required validation prior to reporting. To confirm the accuracy of the method employed, an experiment was devised where a car was driven past the camera in question at three speeds – 20km/h, 50km/h and 100km/h. The footage was then processed using the custom method developed for the case, and the error assessed. Based on the testing conducted, tolerance of the system was ±4km/h, and percentage error was 4-8% for various speeds.

  - Limitations: inadequate sample size (n = 3), variables not controlled or matched to case circumstances (weather, time of day, car etc), non-blinding of experimenters.

DATA VISUALISATION EXAMPLES

In the examples provided, each graph displays different aspects of the data:

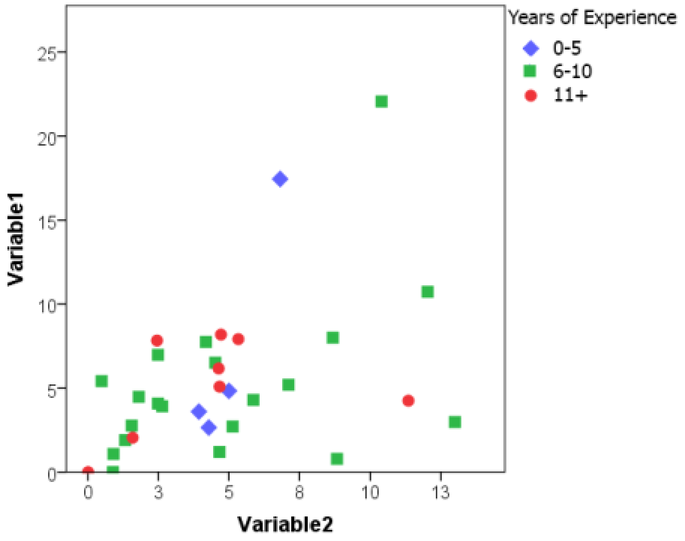
- The Scatter Plot represents the variation present amongst the whole sample set, and within each category of experience, as well as the relationship between variables 1 and 2.

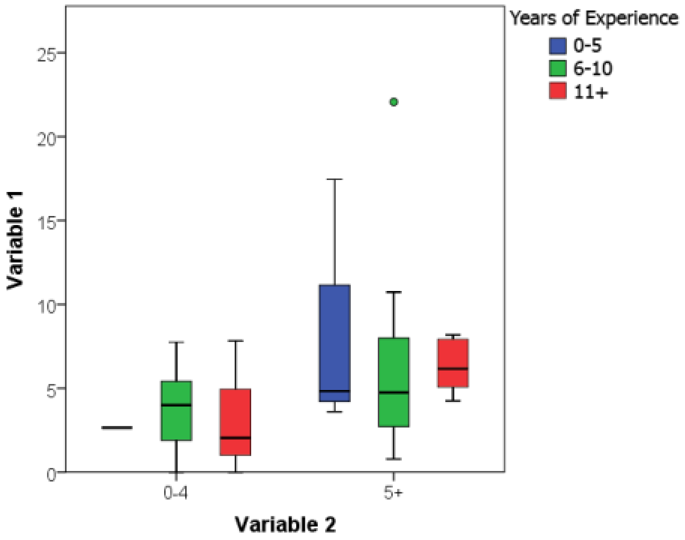
- The Boxplot shows the median score and variation within each group for variable 1, but reduces the information displayed for variable 2.

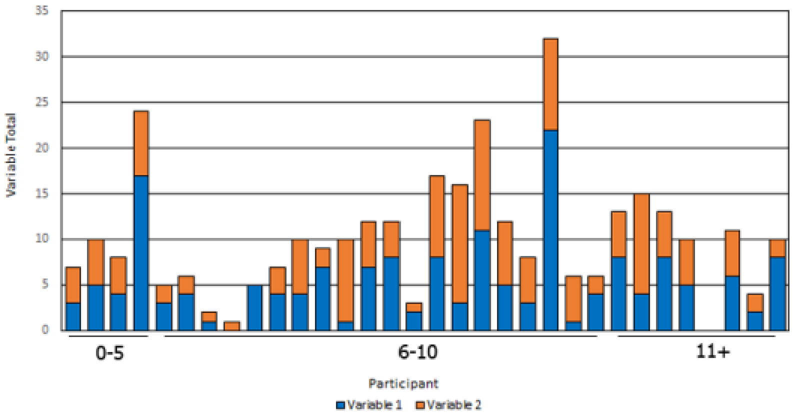
- The bar graph demonstrates the variation between participants in the sum of the scores of variables 1 and 2, but it is more difficult to see the variation within each variable, and the relationships between them.

Simply presenting the mean and standard deviation values in a table removes many of the important messages from the data.

Years of Experience	Mean Variable 1 (SD)	Mean Variable 2 (SD)
0-5	7.1 (6.9)	5.0 (3.9)
6-10	5.1 (4.8)	5.0 (4.4)
11+	5.2 (3.0)	4.3 (3.4)







MODEL ANSWERS TO VALIDITY QUESTIONS

It can be difficult to answer questions in court about validity, particularly when there are limited empirical studies, or where studies only address a part of the method in question. Possible answers are provided below:

- There are numerous good quality studies that address the claim:
  - I have [the relevant group has] critically reviewed empirical studies that are relevant to my claim. There are a number of well-constructed studies that provide data supporting my claim that I/people/procedures/technology can do x to x level of accuracy within the following parameters. My extensive search did not reveal any well-constructed studies that provide qualifying or contradictory evidence regarding my claim.
  - *Theory*: Multiple empirical studies have demonstrated that the underlying scientific theory behind my claim is accurate. These independent, rigorously designed studies were directly applicable to my claim, and have demonstrated with sufficient statistical power that the scientific assumptions behind my claim are experimentally supported.
  - *Performance*: The ability for trained practitioners/the method to perform the analysis/claim has been tested under controlled conditions in a number of independent empirical studies. All these relevant, well-designed studies have shown that there is a high level of accuracy, and that the method/trained practitioners are able to reliably obtain correct answers in situations that are relevant to casework.
- There are some intermediate level studies that address the claim:
  - I have [the relevant group has] critically reviewed empirical studies that are relevant to my claim. This is an area of developing science/active investigation and at the moment the evidence supporting my claim is quite mixed. There are several of high-quality studies that provide conflicting evidence regarding my claim/there are several studies of moderate-quality that provide mixed evidence regarding my claim. As a consequence the court needs to be aware that [state relevant limitations]. We are actively conducting research in this area and hope to have clearer evidence about the claim as the evidence-base develops.
  - At present, there are a small number of studies that provide empirical support to the validity of my discipline/analysis. The initial evidence that has been collected provides limited information/conflicting information regarding the accuracy of the method/my ability, particularly surrounding x (give examples of limitations/applicability of the studies). Research is currently being conducted/will be conducted to obtain additional information on these issues.
- There are no quality studies that address the claim:
  - I have [the relevant group has] actively sought out empirical studies relevant to my claim. This is an emerging area of science and at the moment there is no high-quality evidence supporting my claim. We realise the importance of making evidence-based claims and we [me/my laboratory/my discipline] are actively undertaking research to address this limitation, but the research is in its preliminary stages. It will be some time before there will be evidence available that speaks to the claim that I make.
  - At present, no controlled, empirical studies have been conducted to investigate the validity of my claim/method. This means that the performance of trained practitioners/the method on cases such as these is unknown, as are the error rates. While this does not necessarily mean the technique is inaccurate, there is currently no information to demonstrate accuracy. Research is being undertaken/will be undertaken to obtain this information, but will take some time to be completed.

- There are studies that address the claim but they have some important elements missing:
  - I have [the relevant group has] actively sought out empirical studies relevant to my claim. This is an area of developing science/active investigation and at present the studies available that speak to my claim have some important limitations. Based on the evidence we can say that part x of my claim is supported, however it is important for the court to be aware that x, y, z. We are actively conducting research in this area to address the limitations in past empirical studies and improve the quality of the evidence base for the claim that I am making. However, this work is in progress and it will be some time before the limitations are addressed.
  - Empirical studies have shown that trained examiners show significantly higher accuracy rates than untrained lay people. The largest international study conducted to date on the accuracy of conclusions by fingerprint examiners produced an erroneous identification rate of 0.17%. However, this figure does not incorporate the Verification step or any quality control measures that are in place within my organisation. It is unknown how the study results represent error rates for my organisation and current casework procedures.

FREQUENTLY ASKED QUESTIONS

- Can I use proficiency tests to infer validity?

  In general, proficiency tests are designed to address the accuracy of a system, rather than the competence of a practitioner or the accuracy of a method. The tests may not address the full range of difficulty experienced in casework, be in a format that represents the conventional format of casework materials, or represent a single practitioner's abilities and methods. As such, proficiency tests should not be used to infer validity, or be used to test a claim.

- What is the distinction between proficiency and competency?

  Within forensic science, proficiency tests are usually used to check the correct operation of a system as a whole, while competency usually refers to an individual practitioner's ability to perform a technique accurately and reliably. A competent examiner may still obtain an incorrect result if the system is not operating appropriately, while an incompetent examiner may obtain a correct result due to checks and balances within a system. Therefore, both aspects need to be investigated to show overall foundational validity, and validity as applied within a single laboratory.

- Does accreditation mean my methods are valid?

  Accreditation does not automatically confer validity on disciplines, methods or practitioners. The process of accreditation should include thorough scrutiny of validation reports and the extent of empirical support for a technique, but the level of scrutiny will depend on how conversant the assessors are with principles of good experimental design and performance testing. Assessment by individuals who are not aware of these areas will not detect deficits in validation testing, and thus accreditation should not be inferred to mean validity.

- My technique has been accepted by courts for a long time – doesn't that mean it's valid?

  Unfortunately, no. Although many (but not all) courts have reliability standards on expert evidence, these generally have little effect on the assessment and admissibility of evidence on the basis of validity. As can be seen from this guideline, the assessment of the validity of a claim is complex, and requires specialised knowledge in experimental design, statistics, and cognitive psychology as well as the discipline itself. Many legal practitioners do not have the required knowledge, and have therefore relied on the forensic science practitioners to perform this assessment, and to disclose if there are issues with validity. In many cases, validity has been assumed in the absence of knowledge (or disclosure), and it is only recently that courts have begun questioning this assumption.

- My discipline has numerous published case reports – what can I use these for?

  In general, case reports cannot be used to infer validity as the ground truth is not known, the sample size is one, and there is generally only a single participant (the examiner). Case reports can be useful for identifying potential issues to test, providing information on methods and reasoning used, and on exploring the possible range of evidentiary variation that may be encountered. However, they should not be used to show that the claim is valid, or that methods validated on other, more routine questions can be expanded to be used on the scenario described.

- I'm designing a study, but I'm stuck on the statistics – what can I do?

  Consult a statistician wherever possible. Statistical departments within universities may be able to assist with both experimental design and analysis, or may be able to guide you to another source of advice. If it is not possible to involve external individuals, previous well-designed studies may provide guidance – published reports should provide full details of sampling numbers, strategies and statistical tests used. Even if there are no reports addressing your specific claim, reports from other disciplines with similar claims can be used to guide your experimental design and statistical approach.

ACKNOWLEDGEMENTS

This document would not be possible without the valuable contribution of each of the working group members:

- Greg Markowsky
  Senior Lecturer, School of Mathematical Sciences, Monash University
- Jason Tangen
  Associate Professor, School of Psychology, The University of Queensland
- Jennifer Raymond
  Research Coordinator, New South Wales Police Forensic Evidence and Technical Services Command
- Kate Sloan
  Team Leader Chemical Criminalistics, Australian Federal Police Specialist Operations
- Kaye Ballantyne
  Senior Research and Development Officer, Victoria Police Forensic Services Department
- Kristy Martire
  Associate Professor, School of Psychology, University of New South Wales
- Robert Morgan
  Senior Forensic Project Officer, ANZPAA National Institute of Forensic Science
- Tracie Gould
  Senior Forensic Project Officer, ANZPAA National Institute of Forensic Science.

RESOURCES

- International Fingerprint Research Group (IFRG), Guidelines for the Assessment of Fingermark Detection Techniques, *Journal of Forensic Identification,* 2014, 64 (2), pp 174-200.
- Forensic Science Regulator. Guidance: validation. 2014, FSR-G-201, Issue 1. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/375285/FSR-G-201_Validation_guidance_November_2014.pdf
- Kaye D.A., Freedman D.A. 2010. Reference Guide on Statistics. Chapter 5. In: National Research Council. 2011. The Reference Manual on Scientific Evidence: Third Edition. Washington, DC: The National Academies Press.
- Martire K.A., Kemp R.I., Considerations when designing human performance tests in the forensic sciences, *Australian Journal of Forensic Sciences*, 2018, 50 (2), pp 166-182.
- SWGDAM guidelines for the validation of probabilistic genotyping systems 2015 https://docs.wixstatic.com/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515.pdf
- SWGDAM validation guidelines for forensic DNA analysis methods 2016. https://docs.wixstatic.com/ugd/4344b0_813b241e8944497e99b9c45b163b76bd.pdf
- Tangen J.M., Thompson M.B., McCarthy D.J., Identifying fingerprint expertise, *Psychological Science*, 2011, 22 (8), pp 995-997.
- Thompson M.B., Tangen J.M., McCarthy D.J., Expertise in fingerprints identification, *Journal of Forensic Science*, 2013, 58 (6), pp 1519-1530.

ANZPAA
Australia New Zealand
Policing Advisory Agency

NIFS
NATIONAL INSTITUTE OF FORENSIC SCIENCE AUSTRALIA NEW ZEALAND

Level 6, Tower 3, World Trade Centre
637 Flinders Street, Docklands Victoria 3008
DX 210096 Melbourne

T +61 3 9628 7211 F +61 3 9628 7253
E secretariat@anzpaa.org.au

www.nifs.org.au