DOUBLE BLIND SYSTEM TESTING

A Model Framework for Forensic Science Laboratories

2019

ANZPAA
Australia New Zealand
Policing Advisory Agency

NIFS
National Institute of Forensic Science Australia New Zealand

Copyright Notice

© STATE OF VICTORIA 2019

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:	1.0
Date Distributed:	October 2019
Approved by:	ANZFEC
Status and Security:	Unclassified

CONTENTS

OVERVIEW

This document describes a proposed model framework for conducting a Double Blind System Test (DBST) program in forensic science laboratories. The program involves the preparation and submission of test samples in a manner that laboratory personal are not aware the exercise is a test.

A well-planned program can provide an effective means to check the performance of a forensic science laboratory. DBSTs can be used for proficiency testing of examiners, identifying opportunities for system improvement or for estimating casework relevant error rates.

Although logistically challenging to implement, the key benefit of double blind system testing is that it can give an accurate indication of a laboratory's true performance, unobtainable with conventional single blind proficiency testing. The model framework, based upon existing double blind programs and key literature on human performance testing, provides practical advice on how to design, implement and conduct a successful DBST program. At present, the framework should be considered a proposal only. Pilot testing will be completed in 2019-2020, which will be used to inform and revise the framework as necessary.

DOUBLE BLIND SYSTEM TESTING FRAMEWORK

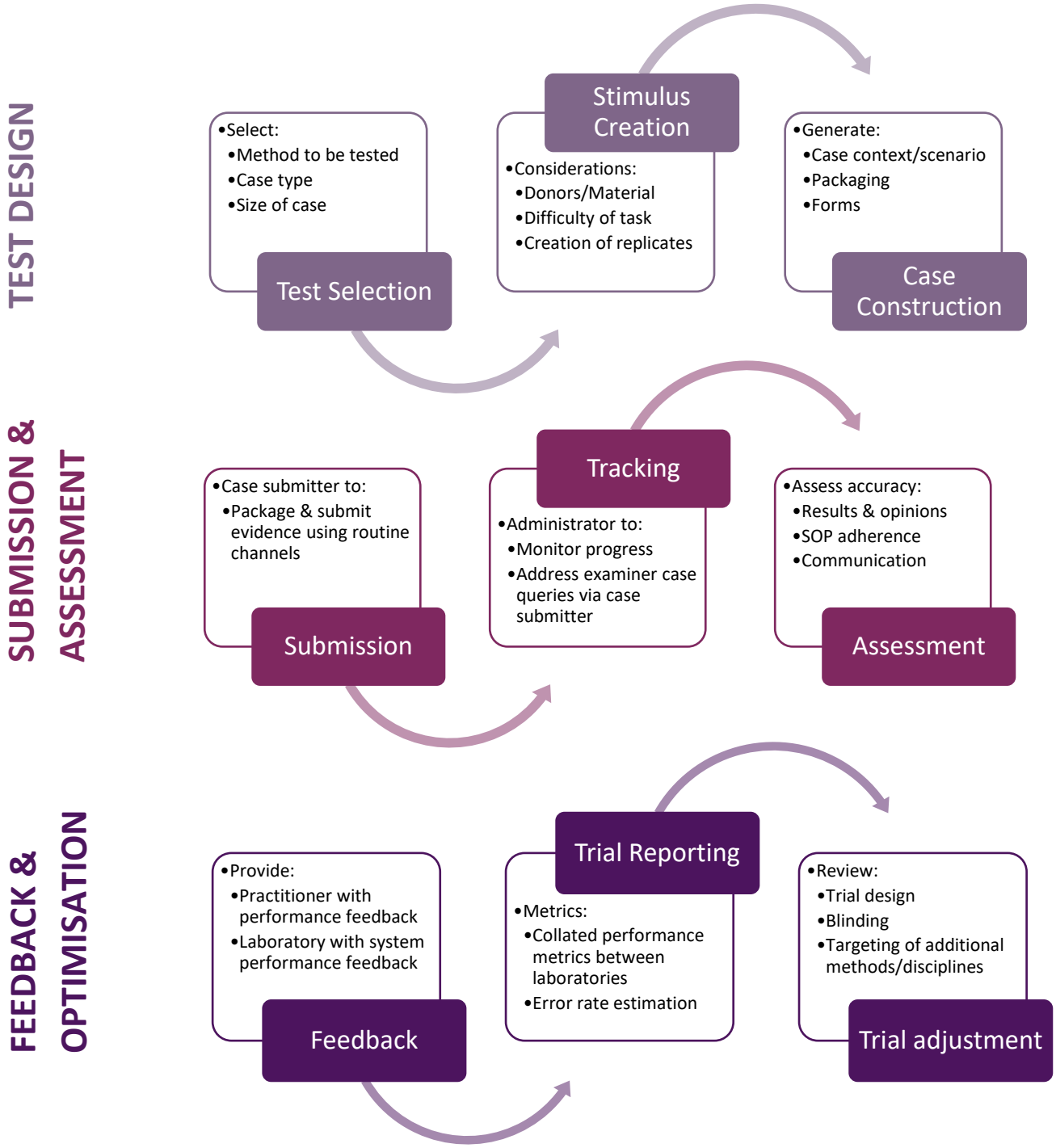
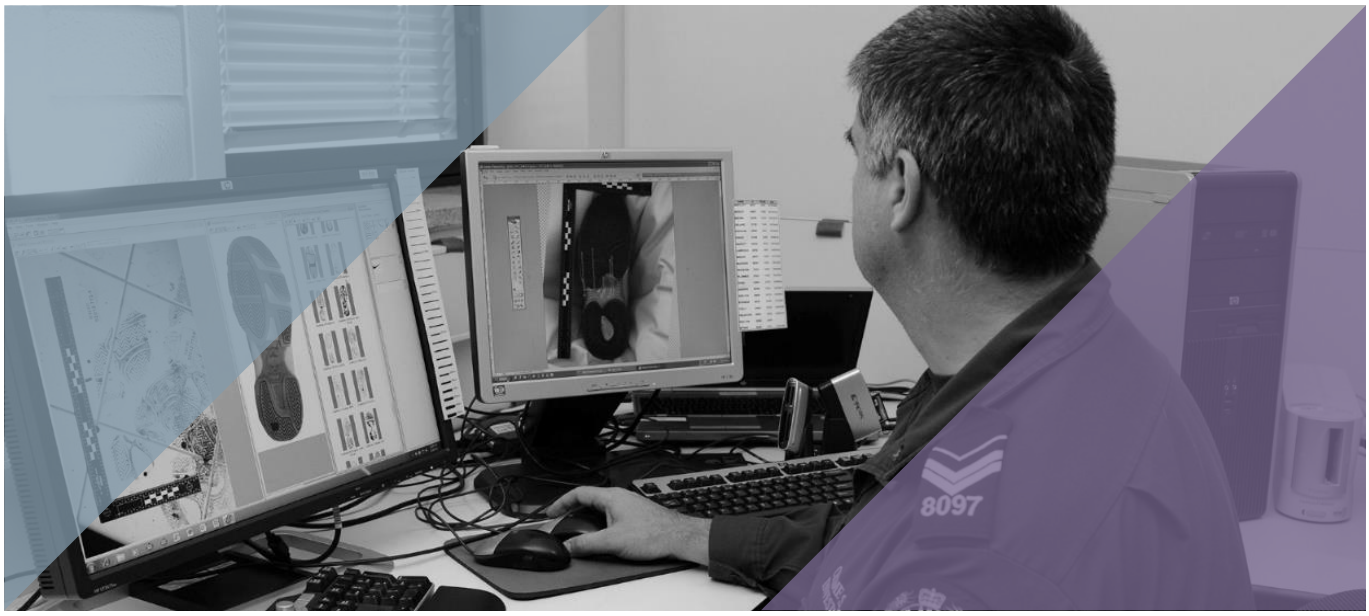


Figure 1. Proposed Model Framework for Double Blind System Testing in Forensic Science Laboratories

Figure 1 depicts the nine phases of the DBST model framework, which are outlined in subsequent sections. The test design, administration, conduct and analysis will need to be tailored to each participating laboratory, as considerations around submission, vetting, case acceptance and communication will differ depending on agencies and service models. The document provides advice on the key aspects that need to be altered or addressed to fit particular disciplines, operating models or desired goals of the DBST program. Examples of each phase are provided for two disciplines within each section, to illustrate possible designs of a DBST within operational laboratories.



Administration of DBST

The overall co-ordination, delivery and assessment of DBST could be performed under a variety of models, depending on the level of independence required, the funding available, the desire for inter-laboratory benchmarking and the internal resources available. As the construction, administration and assessment of proficiency tests can be complex and time consuming, many laboratories currently purchase externally administered single blind proficiency tests to reduce the resource burden. This is considered the best model for double blind proficiency testing as well, as it maximises independence of testing, reduces the probability that the existence of a test will be revealed to staff, and allows for benchmarking between groups sitting the same test. However, external administration of DBST requires additional considerations beyond those of single blind testing. The administrators must have knowledge of a laboratory's standard case type, typical types of exhibits, routine case size and type of information that normally accompanies a case, such that a believable case can be constructed. The administrator must also be able to recruit and work with a police or other case submitter to get the test into the laboratories standard workflow, and must be aware of laboratory protocols and procedures to enable accurate and fair assessment. Multi-agency trials, such as would be required under a combined Australia New Zealand DBST program, would therefore require the administrator to possess significant levels of knowledge about each participating agency, for each discipline being tested. While not insurmountable, particularly given the relatively low number of agencies across the region, the costs of facilitating such trials would, at least initially, be considerable.

Alternatively, agencies could administer DBST internally, creating, distributing and assessing tests through (for example) quality managers. It is not recommended that disciplines/teams administer DBST themselves, as this vastly increases the likelihood that the test would be revealed to participants, even inadvertently, and is less

likely to result in useful, independent feedback on subjective aspects of the test if assessors have developed or routinely use the same system as is being examined. Internal facilitation of tests would decrease the complexity of compiling and assessing tests, as internal laboratory knowledge should enable both the creation of cases that model typical instances, and the assessment of casefiles and reports against laboratory expectations. While reducing costs, this approach would increase the internal resources substantially, which may not be achievable for smaller agencies or those without large quality management teams.

A DBST program across the forensic sciences may also combine both models of administration. Disciplines which require relatively little administrative burden may be administered internally – such as those that receive little case context/surrounding documentation, typically test low numbers of exhibits or where known stimuli are easy to obtain. Disciplines that require greater experimental set-up, or where considerable expertise is required to create or assess tests, may be easier when facilitated centrally. An exchange program between laboratories could also be utilised, where experts from laboratory "A" create and assess tests for laboratory "B", who do the same for "C" *etc*, or where laboratories rotate the administration of DBST annually.

TEST DESIGN

The first, critical step in conducting a DBST is to decide upon which method is to be tested. While theoretically all forensic methods can be tested, practical, legal and ethical considerations limit the scope of what can be assessed through creation of ground-truth known stimuli. It must be viable and ethical to create stimuli, a case must be able to be created that will unlikely be detected as fake by examiners, and exhibits and case documentation must be able to stimulate regular casework items. Some methods and disciplines, such as fingerprint comparison, shoeprint impression evidence or DNA profiling, are relatively simple to create realistic scenarios and items for. Others may present considerably more difficulty. Forensic medicine disciplines such as pathology or clinical forensic medicine may be unethical or extremely difficult to obtain ground-truth known patients for analysis. Creating mock crime scenes is considered out-of-scope for the present framework, although it might be feasible under certain circumstances with particular case or exhibit types. Other types of analysis may be predominately associated with case types that would in real situations attract high media attention or police priorities, and therefore may be difficult to simulate. Examples of such disciplines include post-blast explosive analysis and disaster victim identification tasks. Given the high-profile nature of such cases, it is likely that examiners would question the lack of media reporting, and would deduce the existence of the test. Consideration may also need to be given to the need to create false records in external systems. For example, if examiners routinely confirm court dates or coronial inquest details, a believable test will need to have this information present in the relevant systems or databases. Attention must also be paid to acceptance rules for processing to ensure that evidence is submitted successfully for processing. Such rules may differ between laboratories, and therefore may create difficulty for central administration of DBSTs in designing tests that will be accepted at all participating agencies.

It may also be necessary to consider the level of specificity desired for the test. In some disciplines, different methodologies can be used to process the same samples and obtain similar overall conclusions such as in illicit drug analysis, latent fingerprint development or biological sample collection. It may be possible through the careful selection of case scenario and sample type to direct which methodologies are used by analysts, to ensure that particular methods are tested in the double blind manner.

Biometric Methods

Biometric methods may require special reflection regarding donor protections, legislative prohibitions on volunteer sample searching and the implications of matching to a crime sample on databases. If multiple laboratories or examiners are participating in a test at the same time, different donors will have to be used to prevent DNA or fingerprints from different blind tests matching to each other on national or jurisdictional databases. From a legal perspective, restrictions on how a volunteer sample can be uploaded to and/or searched against a database must be considered, and may force careful attention of the case scenario to prevent illegal searches occurring. Likewise, it may be necessary to develop procedures to address instances where a system test sample hits on an existing sample on the database. In a previous study piloting a double blind DNA testing process in the USA, Peterson *et al*. obtained legal assurances from the U.S. Department of Justice that the program would be able to resist any judicial or legislative proceeding process for obtaining identities of donors (Peterson & Gaensslen, 2002). Such assurances may be required from relevant authorities if biometric samples are searched against databases.

Example: Fingerprint Comparison

TEST DESIGN

Methods to be tested: Cyanoacrylate fuming, powdering, NAFIS searching, comparison

Typical case scenario: Theft

Typical case size: 1 – 5 items

Possible outcomes: Identification, Exclusion, Inconclusive

Modelling casework trends

Designing non-detectable tests for methods that are deemed feasible requires a certain level of knowledge of base rates and evidence types within casework. For example, it would be considered extremely unusual to receive a case with only one item for analysis, then tests should contain multiple items. Case scenarios should be modelled on those received in casework for standard cases, and for feature comparison disciplines, items should contain both matches and non-matches in frequencies similar to those encountered in casework. If such base rates are unknown, care should be taken to vary case types, opinion types and case details such that no pattern is immediately detectable.

Example: Illicit Drug Analysis

TEST DESIGN

Methods to be tested:
Qualitative: Colour testing, GC-MS, FTIR
Quantitative: UPLC or GC-FID

Typical case scenario: Possession of a controlled substance

Typical case size: 1-5 items

Possible outcomes: Identification of a controlled substance, no controlled substances present, quantitation/purity estimate

STIMULUS CREATION

Once the test is designed, exhibits containing the evidentiary material will need to be created. Careful attention will need to be given towards the inclusion of features that are always present in evidence, and features that should never be present, or would be considered highly unusual. Examples include an item that is normally heavily handled such as a wallet or phone containing only a single latent fingerprint without any smudged or partial prints, the use of different brands of swabs to normal casework, or a perfect shoe impression without any indication of movement or wear.

Difficulty of Examination

The desired difficulty of examination or analysis should, if possible, be estimated when designing stimuli. Ideally, at least some trials in a program should be created to emulate difficult tasks. Examples may include depositing extremely low levels of traces, producing damaged, fragmented or partial items, or using multiple sources or overlapping traces. Case scenarios that have been known to produce challenging evidence in the past may provide useful guidance for creating difficult stimuli. Although stimuli should preferably be pre-tested by qualified examiners to ensure that the desired outcome is achievable, this may not be possible for all methods due to destructive testing methods, or a low number of qualified examiners. For initial DBSTs however, all efforts including pre-trial testing should be made to ensure that stimuli are realistic and representative of casework.

Example: Fingerprint Comparison

STIMULUS CREATION

Items: Beer Bottle
Ten-print set from main donor

Donor: 1 x main donor, 1 x secondary donor. Neither donor should be on NAFIS.

Method of Deposition: Secondary donor to handle bottle as if stocking shelves/scanning bottle. Main donor to simulate drinking from bottle, holding by neck. Both donors should thoroughly wash hands ~5 minutes prior to handling to reduce deposition.

Replication

Multi-jurisdictional trials will require the creation of multiple replicates to enable benchmarking between laboratories. If possible, replicates should be of a similar nature, difficulty and level of completeness. While this may be relatively simple in some disciplines, such as a fired bullet, drug or fibre analysis, it may be more difficult in areas such as touch DNA or fingerprints, where every deposition of material changes slightly. If complete replication is not possible, assessment of results and benchmarking between participants should be adjusted to account for variation between replicates.

Example: Illicit Drug Analysis

STIMULUS CREATION

Item: Plastic bag containing white powder

Material: Sub-samples from known illicit substance such as methamphetamine. May be cut with a common cutting agent.

Confounding Factors

Care needs to be taken that stimuli are presented to examiners in a way that will not elicit suspicion. For example, small marks or residue on a fired bullet may indicate that it has previously been mounted and examined microscopically. In a DNA blind study, examiners noticed that microscope slides were not streaked in the manner typical to the jurisdiction (Peterson & Gaensslen, 2002). Firearms examiners suspected a test in part due to remarkably clean cartridge cases (Kerkhoff, et al., 2015). Likewise, receiving exhibits from two or more different firearms of the same calibre and type was noted as a reason for detecting tests, as it is rare in casework (Kerkhoff, et al., 2018).

CASE CONSTRUCTION

Realistic case scenarios will need to accompany stimuli to ensure that examiners do not detect the test, and that it can pass through the laboratory without excessive questioning back to the submitting officer. Even if context control procedures are in place, information should be available if case questions arise during the examination process. The amount and nature of information will be dependent on the case and stimulus type, as well as the normal amount of information received by the laboratory.

Case Scenario

The scenario should be consistent with the method used to create the stimuli, and similar to cases that the laboratory or police have previously encountered. High-profile type scenarios should be avoided, as a lack of media reporting may raise suspicion. Witness statements and police reports, if required, should be of a similar level of detail to real reports. Details must be consistent with real situations – for example, if particular areas have low crime rates/drug use, addresses within these areas may indicate a test. Scenarios indicating evidence may have been exposed to rain must obviously only be used when rain has occurred within the time period specified. Even minor, seemingly non-consequential features may be signals to examiners, such as overly neat handwriting on forms or packaging, or the use of specialist phrases or words in requests for examination.

Example: Fingerprint Comparison

CASE CONSTRUCTION

Scenario: Burglary and theft from residential property. Probable entry through broken window. Property to the value of $10,000 taken. Empty beer bottle found in garden; homeowner reports it is not from house. Suspect arrested with stolen property, fingerprinted

Information required: Suspect name and CNI; homeowner name and address

Identifying Information

Where suspect and/or victim particulars are required, fictitious information should be pre-checked through systems to ensure that it will not match with existing persons. For example, fingerprints are recorded within NAFIS[1] with a unique identifying number, which should be provided and correct for any suspect or victim reference prints provided, or not match any existing individual already on the system.

As each agency varies in the amount and nature of information received, specific guidance regarding case construction cannot be provided here. However, the most efficient method of constructing a case is likely to be simply copying a previous case of sufficient age, altering details such as item and person particulars to suit the stimuli created.

Example: Illicit Drug Analysis

CASE CONSTRUCTION

Scenario: Suspect spoken to by Police in relation to a stolen vehicle. The suspect's vehicle was searched by Police who located 5 bags of crystal methamphetamine

Information required: Relevant police data from linked systems

[1] National Automated Fingerprint Identification System, Australia.

SUBMISSION

A true double blind system test will need to enter the laboratory in a manner indistinguishable from true cases. In many instances, this will mean that submitting individuals will need to be recruited from the laboratory's normal client base such as police investigators. Particular evidence types may require the recruitment of specialised police squads, such as arson investigation or sexual assault teams, while others may be submitted by general duties or property officers. Submitting officers should be fully briefed as to the nature and purpose of the system test, with the importance of not revealing the test emphasised.

Submitter Tasks

Submitters should be provided with the stimuli and case information, along with any additional information that may be required to address laboratory acceptance questions or examiner questions. To maximise the reality of the test, exhibits should be packaged into the jurisdiction specific evidence packaging, with all labels, tapes and signatures used as normal. This should occur regardless of whether the DBST is centrally or locally administered, as forensic staff may not replicate all the details of packaging and submission that are second nature to a submitting officer. If photographs or lifts of impression evidence are normally submitted, it may be necessary to have submitters perform these actions. Forensic examiners may detect photographs being taken with different scale rulers, unusual models of camera, or non-conventional lifting mediums for that jurisdiction. Jurisdictions with crime scene officers located remotely from the main forensic laboratory may be ideal candidates to produce and submit evidence in a realistic manner.

Example: Fingerprint Comparison

SUBMISSION

Submitter: Crime Scene Officer

Tasks:
- Develop, photograph and lift latent fingerprints from bottle
- Create case record in police system
- Submit photographs, ten-prints and forms to forensic laboratory
- Receive completed identification report

Potential tasks required by the submitting individual include:

- Package prepared test items as per standard operating procedures (SOP) for jurisdiction
- Create case/evidence numbers in police property or case tracking system (if required)
- Translate prepared case information into jurisdiction forms/laboratory submission portals
- Request forensic analysis on items, and lodge items with laboratory (in person or via post)
- Handle communications and queries from analysts/examiners
- Liaise with test administrators regarding any queries, passing on results and reports
- Purge information from systems at completion of testing (if required).

Example: Illicit Drug Analysis

SUBMISSION

Submitter: Drug Squad Member/general duties police member

Tasks:
- Prepare relevant laboratory information management system (LIMS) entry and associated paperwork
- Sample packaging
- Delivery to site
- Provision of court date
- Answer queries
- Receive report

Legal, Organisational Policy and Resource Considerations

Jurisdictional legislation, regulations or agency policies will need to be considered before entering simulated case data onto government, police or laboratory systems. Legal or organisational clearance may be required to allow such tests to be concealed within systems, as well as removing information after completion of the test. Submitters may also be required to invest a significant period of time in preparing the system test, particularly if they are required to create, photograph and package the evidence, as well as submit it in person to the laboratory. The time commitment required should be estimated and disclosed when recruiting submitters, noting that it is likely to differ substantially between exhibit/case types and jurisdictions.

An ongoing DBST program would require multiple submitters, with regular rotation between work areas (e.g. squads, stations) and individuals to ensure that examiners do not detect a pattern, and therefore deduce that any case from a particular work area may be a test. Centrally co-ordinated tests may be created and shipped either directly to submitters, or to nominated liaisons within each laboratory such as quality management staff who could liaise with submitters.

CASE TRACKING

It may be necessary to track the progress of the test through the system, particularly for case types with longer processing times or where analysis will not progress unless information or nominated suspects are received within a specific time period. Previous international trials encountered instances where testing was not progressed due to a lack of reference samples, without the administrating team being aware of the situation. Therefore, it may be beneficial to have either the submitting officer or a quality manager from within the forensic laboratory track the case, provided this can be accomplished without compromising the double blind nature of the test.

Context Information

Submitting officers may need to be briefed and prepared to handle any questions that may arise from examiners. Some disciplines, laboratories or evidence types require little context information, or only rarely require contact with submitting officers, while others heavily utilise case information to decide on acceptance, set propositions and priorities, clarify unexpected results or decide on analysis strategies. The need for and amount of contact should be predictable from previous knowledge regarding the laboratory's standard processing and case handling, and therefore the submitter should be briefed, as far as possible, with information that may be requested. If unanticipated questions are asked of the submitter, the administration team should be contacted for guidance as to the answers to provide that will be consistent with the scenario and desired direction of the system test.

Example: Fingerprint Comparison

TRACKING

Case tracking: Not required due to fast turn-around times

Case file request: Full notes to be provided, including all mark-ups of prints and comparison charts created during ACE-V.

Case File Requests

Following completion of the test and reporting of results to the submitter or relevant party, it may be necessary to follow up with the reporting scientist to obtain copies of case files or details of the examination. Short certificates, laboratory reports or statements may not contain sufficient detail for assessors to be able to ascertain which protocols were used, how analysis and interpretation were performed, and whether any quality issues were detected and rectified during the process. Such details should be contained within casefiles or electronic records, and so may be able to be provided on request. Depending on the structure of the laboratory, such a request could occur through the quality management system, or directly between administrators and examiners. However, such a request in the absence of a court subpoena would indicate to the examiner the existence of a test. As such, processes should be put in place to prevent any additional checking/detail being added to the casefile prior to submission to administrators, or to prevent the perception that this could occur.

Example: Illicit Drug Analysis

TRACKING

Case tracking: Dependant on laboratory system

Case file request: May require a subpoena; DBST administrators to liaise with quality management staff to obtain case notes if possible after existence of test is revealed.

ASSESSMENT

The assessment of a DBST can be carried out at a range of levels, depending on both the goals of the program and the ability to access information regarding how the system has been used in the particular case. Normally, commercially available single blind proficiency tests mark only the overall conclusion, with no assessment of methods used, documentation of results or mode of communication. Although this may occur as part of standard audit processes on casework, having such assessments conducted either by external individuals, or on cases where the ground truth is known and decision making and opinions can be viewed in light of the known answer, would be beneficial. Whilst the DBSTs can certainly be assessed only in light of the accuracy of the opinion relative to the known ground truth, limiting review in this manner would limit the potential benefit of the program, and the improvement opportunities that could be identified. The level of assessment may be tailored to a laboratory, or an individual discipline's requirements. If full assessment of all parts of the system is required, then **Table 1** indicates what is required for assessment and a suggested marking criteria.

Table 1. Potential assessment aspects and information requirements for DBST

Aspect	Possible Assessment Criteria	Information required for assessment
Vetting/Triage	- Appropriate application of acceptance criteria - Appropriate application of examination sequencing guidelines	- Laboratory acceptance guides/SOPs - Examination sequencing guides/SOPs
Evidence Collection	- Appropriate examination performed - Appropriate development/collection performed	- SOPs for examination, development & collection - Examination notes/photographs
Analysis	- Appropriate analysis method utilised - Analysis conducted according to laboratory SOPs/validated protocol	- Analysis notes/records - SOPs for analysis methods
Interpretation	- Interpretation of results conducted in line with analysis results and investigation needs	- Case notes detailing reasoning
Reporting	- Opinion and examination reported in line with SOPs and client/legal requirements	- Report/Statement - Reporting SOPs - Legal codes
Overall opinion	- Accuracy of opinion relative to ground truth	- Overall conclusion
Communication	- Timeliness, appropriateness and quality of communication with submitter - Lay comprehension of expert opinion	- Conversation records - Survey of submitter perceptions - Final report & supporting information (appendices/annexures)
Quality system performance	- Detection of errors/omissions during verification or review - Completion of all required quality procedures in line with SOPs	- Case notes including review documentation - Quality SOPs

Evaluation of Results

Detailed marking rubrics should be created for each test prior to initiation of the DBST, to avoid or minimise any bias in assessment. Some aspects require subjective assessment, as there are no empirically derived metrics in (for example) communication. Others will potentially require provision of SOPs and sensitive laboratory information to external individuals, if administration and assessment is being conducted externally. Some of the criteria will require expert, or at a minimum, discipline-specific knowledge, to assess whether a particular analysis protocol is correct, or if the interpretation and final opinion is justified based on the analysis results. Thus, comprehensive assessment may need to be performed by a team of individuals. Alternatively, a higher level assessment could be conducted where external parties or non-specialists assess the presence/absence of particular features of documentation, note any detections of errors or omissions in the case file or analysis, sense-check the overall process being used and provide feedback on the functioning of the overall system, rather than individually scrutinising the scientific results in each instance. The level of assessment should be informed by the overall desired goals of the DBST program, the resources available for both provision of the information and assessment, and the feasibility and/or appetite for improvements to occur as a result of feedback provided. It should be noted however, that the greater the extent of external, independent assessment there is from disinterested individuals, the greater the potential for identification of improvement opportunities will be.

Assessment will also be dependent on the amount of information recorded during each stage within the process. For example, some areas may not record reasoning within case notes, but only analytical findings. Some may not produce annotated charts of evidence unless requested to for court purposes, or may not record every instance of quality checking and outcomes resulting from those checks. The amount recorded may depend on the discipline, the laboratory or the individual, and as such there will be a need for flexibility in the assessment process, with non-assessable aspects noted.

Example: Fingerprint Comparison

ASSESSMENT

Vetting/Triage: Not Applicable

Evidence Collection: Not Applicable

Analysis: Check of quality/sufficiency decisions, markup of latent and ten print

Interpretation: Check of correspondence of features between prints; level of information in evidence compared to opinion provided

Reporting: Confirmation that report is compliant with SOPs and legal expectations regarding amount and nature of information provided

Overall opinion: Marked against ground truth

Communication: Assessment of any communication records with submitter; assessment of language used in reporting for suitability for non-expert audience

Example: Illicit Drug Analysis

ASSESSMENT

Vetting/Triage: Dependant on laboratory system/SOPs

Evidence Collection: Dependant on laboratory system/SOPs

Analysis: Check if appropriate methods of analysis used

Interpretation: Check if correct substance identified; correct weight given; correct legislative assignment; purity estimate

Reporting: Confirmation that report is compliant with SOPs and legal expectations regarding amount and nature of information provided

Overall opinion: Marked against ground truth

Communication: Assessment of any communication records with submitter; assessment of language used in reporting for suitability for non-expert audience

FEEDBACK

Performance improvement, whether of an individual or a system, requires timely, appropriate and accurate feedback. In general, it is difficult to obtain such feedback within forensic science, where the ground truth is unknown (and unknowable) within casework. In the absence of known truths, improper or incomplete proxies are commonly used, such as guilty verdicts, acceptance by police investigators or courts, and performance information from stakeholders who may not be qualified or knowledgeable to assess the scientific accuracy of opinions. Feedback from only internal stakeholders can result in groupthink – where the status quo is not challenged, or there is an overreliance on existing methods and ways of performing tasks. Independent review from individuals outside the system being tested can provide valuable alternative points of view, clarify aspects such as communication or documentation, or confirm that the system is performing well from an independent perspective.

Learning Opportunities

Commercially available single blind proficiency tests are generally marked as 'consistent with expectations', rather than as correct or incorrect relative to ground truth. While providing feedback in this manner can allow for inconclusive opinions, it also overlooks a valuable opportunity to give practitioners feedback about their accuracy in a constructive manner. The feedback 'All opinions were correct relative to the ground truth' is likely to provide better positive reinforcement, as it is made explicit that the outcome matched ground truth. Furthermore, the DBST program provides an opportunity to provide feedback on performance of teams and individuals across the breadth of the system, which is generally rarely performed. Obtaining positive feedback is an important factor in the maintenance of correct skills, correct use of appropriate processes and building a culture where correct actions and behaviours are rewarded, rather than a culture of punishing incorrect behaviour. Conversely, in the absence of feedback, incorrect or sub-optimal actions and behaviours will persist without correction, potentially leading to stakeholder dissatisfaction, inefficient or inappropriate processes, and, in worst case scenarios, inaccurate results and opinions.

An example of a suggested feedback protocol is provided in **Table 2. It should be noted that the** structure and content will be dependent on the aspects being assessed the ability for test assessors to provide constructive and appropriate suggestions for improvement, and the desire for participating laboratories to receive such information. Feedback may also need to be tailored to different levels depending on the task, as aspects may be performed by a single individual or by a team. Given the system-wide nature of a DBST, an emphasis should be made on providing feedback on the performance of the system, rather than on a single individual. Where necessary and appropriate, this may require creation of separate feedback mechanisms for individuals and for teams.

Table 2. Example of tailored DBST feedback for a fictional trial

Aspect	N/A	Needs Improvement	Acceptable	Good	Excellent	Comment
Accuracy of opinion					✓	The opinion provided (Identification) was **correct** relative to ground truth.
Appropriateness of investigation					✓	The development technique selected was appropriate for the item, and correctly identified traces deposited on the item. Analysis, comparison, evaluation and verification procedures were performed as per SOPs, and industry standards.
Documentation				✓		Documentation of the examination phase was **excellent**, with appropriate photographs and notes taken. Documentation of the ACE-V process, although compliant with SOPs, did not fully note the reasoning of the experts involved, and did not list assumptions associated with the evidence.
Communication				✓		Communication with the investigator was rated as **excellent**, with all exchanges timely, appropriate and showing a high degree of professionalism. Communication of results and opinions was rated as **good**. Examination requests, methods, and overall conclusions were well documented and appropriate for non-experts. However, assumptions, limitations and reasoning were not documented in the final report.
Quality System					✓	All required checks were performed. Technical review detected an issue with documentation during the submission phase, which was rectified prior to reporting the information to the submitter. Blind verification was performed, with both examiners reaching the same conclusion.
Overall					✓	The system performance was rated as **excellent**: - The overall conclusion(s) was correct - Investigations performed were appropriate to address submitter's requests, and were carried out in accordance with SOPs - Documentation and communication was generally excellent, although some factors expected by legal codes were omitted - Opinions were communicated in a manner that was accessible and understandable to lay decision makers.

TRIAL REPORTING

If the DBST is being run for multiple laboratories, or highly similar tests are completed by different individuals within the same laboratory, results may be collated to provide benchmarking information, or, in the longer term, indicate error rates. Care needs to be taken to avoid inferring levels of performance (either good or bad) from limited data, but trends and improvement opportunities may be visible from scrutinising both longitudinal and latitudinal data.

Benchmarking

Comparison of individual trials could examine similar metrics to the end-to-end projects that have been conducted on forensic volume crime analysis across Australia (Brown, et al., 2014) (Bruenisholz, et al., 2019) with additional accuracy and quality metrics added. An example of potential benchmarking is provided in **Table 3**, although the nature of the data will depend on the level of assessment that has been performed in each trial for each laboratory. Comparison of aspects performed well and areas of improvement identified may enable laboratories to share resources to rectify any issues, to benchmark turn-around-times against other laboratories, or to identify areas where their staff are performing particularly well.

Table 3. Collated performance information for a single fictional DBST in fingerprint analysis and comparison

Aspect	Metric	Lab A	Lab B	Lab C
Analysis	Time	<1 day	<1 day	2 days
	Assessment	Acceptable	Excellent	Excellent
	Notes	No markup performed	Full markup performed; reasoning documented	Full markup performed; reasoning documented
Interpretation	Time	<1 day	<1 day	<1 day
	Assessment	Acceptable	Excellent	Excellent
	Notes	No markup performed	Full markup performed; reasoning documented	Full markup performed; reasoning documented
Reporting	Time	<1 day	<1 day	<1 day
	Assessment	Excellent	Acceptable	Good
	Notes	Certificate with appendix describing process, limitations & assumptions provided	Certificate containing opinion provided; no additional information provided	Certificate with opinion & expert qualifications provided; no information regarding method, limitations or assumptions.
Accuracy	Assessment	Correct	Correct	Correct
Communication	Assessment	Excellent	Excellent	Excellent
Quality system	Assessment	Excellent	Excellent	Excellent

Error Rates

If multiple trials are run for the same discipline over a sustained period of time, it may be possible to begin calculating error rates for the accuracy of opinions being provided. Data may be collated for such purposes on a variety of levels, depending on the number of trials and suitability for combination. Statistically, there is no minimum number that is required to produce an error rate, although the greater the number of trials the lower the uncertainty will be around the accuracy, and the more reliance can be placed on the estimate. Importantly, the inferences that are drawn from any DBST should be appropriate for the amount of data used – for example, claims of 100% accuracy should not be made if low numbers of trials have not uncovered any errors in opinion

relative to ground truth. Estimates of uncertainty (for example, in the form of 95% confidence intervals) of any accuracy measure should always be included. An example of potential reporting of accuracy across various levels of measurement is provided in **Table 4**.

Table 4. Fictional error rate information collated across multiple DBST - example of possible data presentation

	Discipline	Lab A	Lab B	Lab C	Lab D
Number of tests completed	230	60	20	15	25
Number reported	226	60	20	15	25
True Positives	83	29	10	7	12
False Negatives	9	1	0	0	1
True Negatives	56	17	5	4	8
False Positives	1	0	0	0	0
Inconclusive	81	13	5	4	4
False Positive Rate (95% confidence interval)	1.8% (0.4 – 9.4%)	0 (0 – 19.5%)	0 (0 – 52.2%)	0 (0 – 60.2%)	0 (0 – 36.9%)
False Negative Rate (95% confidence interval)	9.8% (4.6-17.8%)	3.3% (0.8 – 17.2%)	0 (0 – 30.9%)	0 (0 – 41.0%	7.7% (0.2% - 36.0%)

Data could likewise be collated for individual performance metrics, such as the turn-around-times, the extent of compliance with SOPs, the detection and correction of errors by quality systems or selection and use of specific methods. The collation of data across laboratories and trials may provide useful information for experts when questioned in court about accuracy, for business or process improvement opportunities, for refinement of methods or documentation to ensure compliance, or streamlining of quality systems. Publication of outcomes in a peer-reviewed journal may also be considered, as a means of ensuring transparency, demonstrating the accuracy and competency of the laboratory, and contributing to the knowledge base of the forensic sciences, answering repeated calls for data regarding forensic process accuracy (Stoel, et al., 2016).

.

TRIAL ADJUSTMENT

Regular reviews of the DBST program should occur to ensure that the process is functioning as intended, fulfilling the goals of the program, and keeping pace with changes in service delivery processes and scientific advancement. If possible, practitioners that have been tested should be surveyed after completion of the case to check for successful blinding, and to obtain feedback on the realism of the case, the difficulty of the examination and the completeness of documentation. The Netherlands Forensic Institute (NFI) have added a questionnaire to all cases (both test and real) asking examiners to indicate whether they believe the case to be a test and why (Kerkhoff, et al., 2018). This has enabled assessment of features of successful blinded tests, and adjustment of the test process to prevent examiners detecting tests in the future. Consideration must be given to the timing of when such feedback is requested, and when results are released to practitioners and laboratories. Seeking individual feedback on tests immediately after completion will alert practitioners to which case is a test, and will likely result in the detection of patterns of set-up over time. Conversely, seeking feedback a long time after completion will lessen the accuracy and value of the feedback to the program. Thus, having a system where practitioners are surveyed on all cases (not just tests), as occurs at the NFI and the Houston Forensic Science laboratories (Augenstein, 2018), would seem to be most beneficial, although may have resourcing/efficiency implications.

Beyond reviewing the success of blinding, the nature of the testing should be reviewed at regular intervals. Considerations may include whether tests represent the diverse nature of casework, whether exhibits reflect the range of items normally encountered, if the variations in difficulty are building a representative picture of the performance of the system across the full range experienced by practitioners, and whether all aspects of the system are being tested in equal measure. Such reflections may assist in refining future trials, designing specific tests to challenge particular parts of the system, or to provide targeted training/assessment opportunities in potential areas of concern. Information should be fed back into test design considerations for forthcoming trials.

Example: Fingerprint Comparison

ADJUSTMENT

Case Realism: Blinding failed on one test due to incomplete information provided during submission
Action: Ensure all documentation is complete to jurisdictional standards prior to submission

Stimulus Realism: Practitioners did not report any issues with stimuli
No action required

Assessment utility: Feedback provided to practitioners viewed as useful. Laboratories report corrective actions taken on items with performance rated as “good” or “acceptable”
No action required

Example: Illicit Drug Analysis

ADJUSTMENT

Case Realism: Blinding successful on all cases
No action required

Stimulus Realism: Weight assessments varied across laboratories
Action: Investigate if packaging/shipping has affected quantities

Assessment utility: Feedback provided to practitioners viewed as useful. Laboratories report corrective actions taken on items with performance rated as “good” or “acceptable”
No action required

IMPLEMENTATION OF DBST PROGRAMS

The above information provides a detailed framework for the design, administration and assessment of DBST in forensic science laboratories. Implementation of the framework as a program of works requires several additional considerations. Firstly, agencies wishing to participate in DBST will need to address the additional resources, both in terms of cost and casework samples to be processed. Depending on the level of testing desired, these may not be inconsequential. Secondly, laboratories must decide upon an administration model – central or local. It is anticipated that central administration (i.e. trials involving multiple laboratories and/or jurisdictions) will increase the overall program cost substantially, but may decrease costs for each participating laboratory individually. An existing proficiency test provider could be used to create samples, but an independent group will likely be required to create case information, liaise with submitters, and assess results. University groups active in forensic science may have the required skills to administer a trial, but may not have the resources to administer sufficient numbers of trials without significant reimbursement. Local administration is likely to be lower in cost, but will create a much greater impact on resourcing within the forensic agency, as staff will be required to perform all of the nine key phases within the framework. It is possible that a mix of local and central administration may be most efficient and beneficial, depending on the discipline being tested, resourcing abilities and difficulty of creating tests.

Once a program is agreed upon and resourced, staff within the agency should be notified that DBST is occurring. This creates what has been termed “part-declared double blind” (Kerkhoff, et al., 2018), but is deemed necessary for ethical reasons, and may also create psychological benefits – knowledge of being observed can stimulate improved performance and greater adherence to protocols, particularly where the culture empowers workers to actively seek continuous improvement and take ownership of refining processes (Wickstrom & Bendix, 2000). Open and transparent communication with staff regarding the goals and desired outcomes of the project should assist in allaying any fears regarding punitive actions arising from the DBST program, and promotion of the system aspects, and potential benefits, should be used to create a positive attitude towards the program. In this regard, laboratories may need to examine their existing systems for feedback regarding performance testing to staff, and whether their quality management system is suitable for addressing system issues that may be raised by the DBST program. Finally, individual laboratories should implement a regular review of the entire DBST program, to ensure that issues raised are being addressed, that performance is meeting the high standards expected within forensic science, and that the program is meeting the goals outlined.

CONCLUSION

The framework presented provides a model for forensic science agencies to base a DBST program upon, informed by recognised best practices for performance testing, existing programs and previous research in DBST. Practitioners drawn from across Australia and New Zealand as members of the working group view such a framework as feasible and a valuable additional tool to monitor performance within forensic science laboratories. There is also considerable scope to expand the nature of testing beyond that covered by conventional testing, through the inclusion of multi-disciplinary exhibits or incorporating aspects of scene processing such as examination and collection of evidence from a vehicle. Multi-agency collaboration could also be tested, particularly where evidence recovery and initial processing occurs in a separate agency to analytical processing. However, given the complexity of the framework, commencing DBST with relatively simple tests is advisable, to allow logistical issues to be addressed and benefits to be demonstrated. As experience with DBST increases, the complexity of test design can be increased to fully capture all aspects of forensic science service provision.

ACKNOWLEDGEMENTS

This document is the result of significant contributions from each of the following working group members:

- Allison Hewitt, ChemCentre Western Australia
- Anna Petricevich, Institute of Environmental Science and Research
- Roslyn Wilson, New South Wales Health Pathology Forensic and Analytical Science Service
- Scott Harris, Victoria Police Forensic Services Department
- Brad Mason, Victoria Police Forensic Services Department
- Kaye Ballantyne, ANZPAA National Institute of Forensic Science.

BIBLIOGRAPHY

Augenstein, S., 2018. *Houston Forensic Science Center Slides Blind Testing Into Workload.* [Online]
Available at: https://www.forensicmag.com/news/2018/03/houston-forensic-science-center-slides-blind-testing-workload
[Accessed 16 July 2019].

Brown, C., Ross, A. & Attewell, R., 2014. Benchmarking Forensic Performance in Australia - Volume Crime. *Forensic Science Policy & Management: An International Journal,* Volume 5, pp. 91-98.

Bruenisholz, E., Brown, C. & Wilson-Wilde, L., 2019. Benchmarking forensic volume crime performance in Australia between 2011 and 2015. *Forensic Science International: Synergy,* Volume 1, pp. 86-94.

Kerkhoff, W. et al., 2015. Design and results of an exploratory double blind testing program in firearms examination. *Science and Justice,* Volume 55, pp. 514-519.

Kerkhoff, W. et al., 2018. A part-declared blind testing program in firearms examination. *Science and Justice,* 58(4), pp. 258-263.

Peterson, J. & Gaensslen, R., 2002. *Developing Criteria for Model External DNA Proficiency Testing,* 96-DN-VX-0001: U.S. Department of Justice.

Stoel, R., Mattijssen, E. & Berger, C., 2016. Building the research culture in the forensic sciences: Announcement of a double blind testing program. *Science and Justice,* Volume 56, pp. 155-156.

Wickstrom, G. & Bendix, T., 2000. The "Hawthorne effect" - what did the original Hawthorne studies actually show?. *Scand J Work Environ Health,* Volume 26, pp. 363-367.

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.nifs@anzpaa.org.au

www.nifs.org.au