

Level B Evaluation Assurance – Identity Check Bias Evaluation: Test Report

Prepared for
Te Tari Taiwhenua | Department of Internal Affairs

By



3/16 Bentham Street, Yarralumla,
ACT 2600, Australia

Attention: Grant Stark
grant.stark@dia.govt.nz

Item Under Test (IUT):

IdentityX-cloud platform with liveness method
seamless ID Face (video) v0.0.11

Document Code 24_BXL030_TR_12

Submit Date: 23th October 2024

Version: 3.0

Submitted by: Dr. Ted Dunstone

Contact: ted@bixelab.com



Te Tari Taiwhenua
Internal Affairs

**IDENTITY CHECK EVALUATION
FOR
TE TARI TAIWHENUA | DEPARTMENT OF INTERNAL AFFAIRS (DIA)**

Version History			
Version Number	Description of change	Author	Date
V1.0	Final report [Draft – to be approved by the Department]	BixeLab	10 th October
V2.0	Revised draft – changes include replacing pie charts with tables, analytical updates.	BixeLab	20 th October 2024
V3.0	Statistical analysis outcomes were added to support the outcomes	BixeLab	23 rd October 2024

BIXELAB PTY LTD PROPRIETARY NOTICE

BixeLab Pty Ltd has taken every care in preparing this document. Information contained within is accurate to the best of BixeLab's knowledge at the date of release. BixeLab cannot accept any liability to any person or company for any financial loss or damage arising from the use of this document. It should not be reproduced or made available in any form to persons outside the group/s directly responsible for evaluating its contents, without the written consent of BixeLab Pty Ltd. This report has been prepared collaboratively by the team at BixeLab Pty Ltd, and no individual member is solely responsible for its contents.

All brands and products referenced in this document are acknowledged to be trademarks or registered trademarks of their respective owners.

This report must not to be used by the client to claim certification, approval, or endorsement of a product, by NIST, NVLAP or any agency of the U.S government.

IDENTITY CHECK EVALUATION
FOR
TE TARI TAIWHENUA | DEPARTMENT OF INTERNAL AFFAIRS (DIA)

Table of Contents

Executive Summary.....	4
1. Introduction	5
Evaluation Scope	5
Evaluation Constraints and Risks	6
2. Evaluation Strategy.....	7
On-Site testing	7
3. Evaluation Plan	10
On-Site Evaluation	10
4. Scenario Evaluation	13
Log Data overview	13
Test Participants – Demographic Metadata	15
5. Front-end APP testing	17
Liveness Performance.....	17
Liveness-Matching Performance	19
6. Back-end Matching testing.....	22
Genuine (Mated) Comparisons	23
Impostor (Non-Mated) Comparisons	24
Statistical Analysis Overview.....	25
7. Conclusion	27
Appendix A: Evaluation Management	28
Data and Privacy Management	28
System Configuration Management	28
Test Readiness Criteria	28
Test Completion Criteria	29
Appendix B: Terms and Definitions	30

EXECUTIVE SUMMARY

Te Tari Taiwhenua | The Department of Internal Affairs (DIA) commissioned BixeLab Pty Ltd to conduct an evaluation of its Identity Check solution. The evaluation assessed the biometric system's fairness and performance for different demographic groups, to distinguish Māori¹ and non-Māori populations. The evaluation validated whether the solution's liveness detection and matching algorithms perform consistently for all users, regardless of demographic differences.

The testing measured the system's ability to correctly match genuine users and reject impostor attempts.

The evaluation confirms that the Identity Check solution is operating effectively and fairly in usability across the tested demographic groups. Consistent performance was measured in liveness detection and matching accuracy supported by statistical analyses.

Key Findings:

This consistent performance was confirmed through tests involving 148 diverse volunteers who submitted live selfie images via the Identity Check web application.

1. App Testing Results

The system accurately detected 100% of genuine live submissions, achieving a 0% Bona Fide Presentation Classification Error Rate (BPCER) across all participant groups, including both Māori and non-Māori individuals.

2. Liveness-Matching Performance

Liveness matching demonstrated consistent accuracy across all tested demographic groups, with a low rejection rate of 0.7% and no statistically significant differences observed in match accuracy among different groups.

3. Matching Engine Performance

The matching engine showed reliable performance across all demographic groups for usability. Variance in false accept was noted for Māori and Pasifika – a DIA review of images suggested this was at least in part due to image quality issues. A false match rate (FMR) of 0.09% and a false non-match rate (FNMR) of 0.9% were measured.

At the current system threshold, there are no significant disparities in liveness and matching engine performance among the tested ethnic groups. This conclusion is supported by the Chi-Squared Test conducted at a 95% confidence level, which showed consistent performance between Māori and non-Māori ('Others'). However, the Kolmogorov-Smirnov (KS) Test did identify some variations in the underlying distributions across the tested groups. While these variations do not impact on the user experience, they suggest that any future adjustments to system configurations should be made cautiously to ensure continued equitable performance.

It is noted that there is scope for improvement on false accepts which did show slightly higher error rates for Māori and Pasifika. It is our recommendation that ongoing evaluations further enhance system capabilities and ensure public transparency, reaffirming the importance of fairness and equity for all users.

¹ Sample set did not include any individuals with Tā moko.

I. INTRODUCTION

This project conducted comprehensive testing to measure the fairness of the DIA's Identity Check solution by undertaking tests to ensure that the solution's biometric performance is consistent across specified demographic groups. The testing included:

1. Measurement of liveness detection performance – measuring performance of the solution allowing live users distinguishing Māori and non-Māori populations.
2. Measurement of matching engine performance for genuine and impostor users – measuring performance of the solution of correctly matching genuine users and rejecting impostors distinguishing Māori and non-Māori populations.

The testing strategy validates the degree to which the DIA Identity Check's liveness detection and matching algorithms have "fair performance" for the identified demographic groups. "Fair performance" means that the algorithm's bona fide presentation classification error rate for liveness and FMR-FNMR for authentication, are consistent across tested demographics.

The evaluation involved multiple steps and activities:

1. Front-end testing.

- a. Set up the front-end web app to enable live user selfie submission (remotely),
- b. Test subject recruitment,
- c. Testing of live, demographically representative, volunteers,
- d. Liveness logs for selfies submitted to BixeLab,
- e. Anonymised demographic metadata of recruits submitted to BixeLab,
- f. Web app transmits data and saves selfie images of test subjects,
- g. Reference images for recruited subjects from DIA passport face image database obtained from DIA's provider DXC Technology.

2. Back-end testing.

- a. Back-end instance of biometric matching algorithm established.
- b. Instance interface of database containing the raw selfie images and reference images (DIA passport face image) tested for each recruit.
- c. 492 additional passport images used to check non-match outcomes.
- d. API to the Daon matching algorithm tested.
- e. Execution of run match scripts provided by BixeLab tested.
- f. Match logs provided to BixeLab.

Evaluation Scope

This evaluation:

1. Developed and executed the test design as outlined in this document.
2. Supported on-site testing with DIA volunteers.
3. Post-data-collection assistance in the DIA test environment, including test script provision.
4. Analysis of scores and additional data provided by DIA to BixeLab, including front-end (liveness detection logs) and back-end (matching logs) testing.
5. A detailed report summarising performance metrics and key insights from the test trials.

**IDENTITY CHECK EVALUATION
FOR
TE TARI TAIWHENUA | DEPARTMENT OF INTERNAL AFFAIRS (DIA)**

6. A document detailing findings, observations, and recommendations regarding the solution's demographic fairness and its impact on the target demographics in the test corpus.
7. Verification of similar error rates (BPCERs) and the balance between false match and false non-match rates (FMR-FNMR) distinguishing Māori and non-Māori groups.
8. Evaluation of other demographic variables (e.g. age, sex, and ethnicity) as specified by DIA.
9. Assessment of the solution's effectiveness in distinguishing between live presentations, including liveness and genuine/imposter match scores, using a test corpus representing the identified demographic groups.
10. Results on demographic fairness with a limited confidence level.

This evaluation was intended to determine the biometric system's fairness in processing bona fide presentations and in distinguishing between genuine and impostor comparisons at a specified confidence interval. This evaluation was not to establish whether the system was entirely unbiased.

Evaluation Constraints and Risks

Constraints and risks were identified, managed and mitigated:

1. Availability of environment metadata (e.g. lighting levels) – BixeLab worked with Daon to identify and include relevant metadata to support testing outcomes.
2. Achieving sufficiently large sample sizes for identified subgroups. Due to operational constraints, the number of test subjects was less than 245, which is the specified number for a 95% confidence level when determining matching accuracy. In the context of this evaluation looking at fairness for different demographic groups, the sample size was sufficient to undertake a comparative bias analysis.
3. The evaluation did not involve any participants of Maori heritage that had Tā moko.
4. Assessment required temporary holding of biometric samples for offline testing to maximise sample size for non-mated tests
5. Test execution strategy/plan reflected the test criteria and vice-versa.
6. BixeLab provided a detailed test execution plan and strategy and support during test execution to ensure data captured in the evaluation was relevant for assessment and analysis. The accuracy of the results hinges upon test logs and data labelling. Measures were taken to ensure the data used for analysis was free from labelling errors.
7. The Daon backend system logs browser and device metadata were used to derive generalised performance metrics across all browsers and devices involved in the testing. The capture conditions for selfie submission were largely uncontrolled. Selfie images² may have introduced lenticular distortion, impacting matching results against DIA passport holding images. The volunteer test subjects received instructions on how to conduct tests – simulating the operational interface.
8. Performance metrics obtained from the post-data-collection test execution are within specific confidence intervals and levels, reflecting the inherent probabilistic nature of biometric systems.

² Selfie or selfie image refers to a face image captured through the front camera lens of the mobile phone or smart device, where the subject-to-capture distance is below the standards recommended in ISO/IEC 19794-5 2005 and 2011

2. EVALUATION STRATEGY

The evaluation involved the following key components:

1. Data collection
2. Execution
3. Analysis and reporting

On-Site testing

Data Collection

A test crew, comprising 148 demographically diverse volunteer test subjects, were briefed on the test process. The volunteers interacted with the DIA's Identity Check web application following the process specified for the test app.

The data collection process also involved front end testing, as described below:

1. Each test volunteer provided written consent, having read and understood the privacy and data usage form, and signed the consent form provided by DIA.
2. DIA ensured that a reference image for each test volunteer was available in the DIA database.
3. Post-consent, for each volunteer:
 - a. The database was queried to export the reference image for that volunteer (for later comparison).
 - b. A test account was created, generating export of credentials in CSV format.
 - c. A three-digit user ID was allotted to each test account (e.g. formatted as "DIA_123").
4. Each volunteer was provided with their three digital user ID/login credentials to test the web application.
5. For testing, the process flow was:
 - a. test volunteer received test instructions including the web application link and general instructions on performing the test and completing login.
 - b. test volunteer completed the test – a liveness check that generated a selfie image in the back-end database with test logs generated for each volunteer.
 - c. End.

**IDENTITY CHECK EVALUATION
FOR
TE TARI TAIWHENUA | DEPARTMENT OF INTERNAL AFFAIRS (DIA)**

Levels of effort and decision policy

Note: Volunteer test subjects were required to read the DIA privacy and data use information about the testing event, and accepted these provisions before being issued with the link to start the test.

The volunteer effort expected was:

Table 1 Test volunteer levels of effort recommendation

Levels of effort criteria	Recommended	Tested criteria
The number of allowed attempts per transaction	3	3
The amount of time allowed per transaction	<45 seconds	<45 seconds

Key outputs from data collection

The Data collection process yielded the following key outcomes:

Output 1: Test logs from bona fide testing (selfie submission) containing liveness pass/fail, quality data, user IDs, device/browser metadata, timestamps, re-attempt counts.

Output 2: A test corpus of biometric images:

1. Selfies images collected from live volunteers using the Identity Check web application remotely, across various mobile browsers and;
2. Matched reference images (DIA passport face image) for the same live volunteers.

Test Execution

After data collection (and front-end testing), the back-end test execution followed:

1. A local enrolled image dataset was compiled (containing reference images - DIA passport face images, exported from the DIA database for each test volunteer) and a verify image dataset (containing selfie images derived from the front-end testing completed by each test volunteer).
2. An IT environment for installation and execution of test scripts was established.
 - a. A cross-comparison script provided by BixeLab was tested to validate outcomes.
 - b. The technology evaluation (cross comparison) test script using comparison endpoint was applied: Source images (corresponding to three-digit user ID) were read and converted to a base64 format to be available for the comparison endpoint.
 - c. Selfie images (corresponding to three-digit user ID) read and converted to a base64 format to be available for the comparison endpoint.
 - d. Run cross comparison script, iterating over each source image and selfie image, calling the comparison endpoint until all comparison options are exhausted.
 - e. Output includes matching scores and metadata outputs from the matching subsystem.

After completion of all comparisons, the results were exported to a CSV log.

IDENTITY CHECK EVALUATION
FOR
TE TARI TAIWHENUA | DEPARTMENT OF INTERNAL AFFAIRS (DIA)

Key outputs from test execution

The test execution process yielded the following key outcomes:

Output 1: CSV with quality metrics associated with the source image dataset containing Daon face image quality metrics.

Output 2: CSV with three-digit user IDs and manually completed metadata information including age, sex, and ethnicity.

Output 3: CSV with source user ID, selfie user ID, match pass/fail, match score, other metadata outputted by the Daon matching subsystem (e.g., error codes) across the mated (genuine: same person in source and selfie image) and non-mated (impostor: different people in source and selfie image).

Analysis

After test execution, BixeLab conducted analysis in two steps:

1. Pre-analysis: Manual merging of back-end logs with metadata files.
2. Analysis:
 - a. Liveness detection log analysis with subject demographic metadata.
 - b. 1:1 matching log analysis with subject demographic metadata, and quality metadata for source images (DIA passport face image).

Key outputs from Analysis phase:

Output 1 *Liveness detection*:

1. Performance analysis against the tested demographic groups from bona fide (live) subject testing completed using front-end web app.
2. Inter and intra group variation analysis.
3. Provision of bona fide performance classification error rate (BPCER) for each demographic category tested and indication of performance variation. Methodology was aligned with the spirit of the WD 19795-10 standard.
4. Usability insights collected from post participation surveys and review of the re-attempt counts for specific demographic groups of interest were compiled.

Output 2 *1:1 matching*:

1. Performance analysis against tested demographic groups for genuine (mated) and impostor (non-mated) match trials completed for test crew images.
2. Inter and intra group variation analysis
3. Provision of false non-match rate (FNMR) for each demographic category tested and indication of performance variation. Methodology was aligned with the spirit of the WD 19795-10 standard.
4. Provision of false match rate (FMR) for each demographic category tested and indication of performance variation. Methodology was aligned with the spirit of the WD 19795-10 standard.

3. EVALUATION PLAN

On-Site Evaluation

1. System under test (SUT): front and back-end system overview.
2. Test crew: Subject demographic breakdowns.
3. Test methodology: Pre-test processes and testing procedures.

System Under Test (SUT)

SUT incorporated both the front-end test application and separate back-end system to support matching accuracy testing.

The front-end test application used for data collection was Daon's identity check web application (with selfie flow), which is accessible on the latest browsers running on mobile platforms and laptops with cameras.

Post data collection, the back-end matching subsystem was used to run off-line tests for deriving match results. The back-end process involved comparison of the following samples to complete matching transaction (i.e. not in real time):

1. Volunteer's reference passport image (obtained from DIA's provider DXC Technology) – tagged with three-digit user ID.
2. Corresponding selfie image (available from the front-end testing) – tagged with three-digit user ID.

Each reference image was compared in a genuine (mated) comparison trial against the selfie of the same volunteer, and in an impostor (non-mated) comparison trial against the selfies of all other volunteers. Match score data along with all other available outcomes were collated in a CSV. The following fields were received from the comparison trial:

1. Enrol ID
2. Verify ID
3. Match Outcome
4. Match Score
5. Error label/code
6. Other metadata (if available):
 - a. Timestamp
 - b. SDK version
 - c. Device model
 - d. Platform used

**IDENTITY CHECK EVALUATION
FOR
TE TARI TAIWHENUA | DEPARTMENT OF INTERNAL AFFAIRS (DIA)**

Test Crew

The test crew was comprised of volunteers recruited by DIA. Best efforts were made by DIA to align the demographic attributes to the recommended distributions in the table below.

Table 2 Demographic attributes Metadata breakdown and distribution

Subject Metadata/Input factors	Metadata Breakdown	Recommended Distributions	Achieved Distribution
Age group	< 18 y/o	0%	0%
	18 – 30 y/o	25-40%	17%
	31 – 50 y/o	25-40%	49%
	51 – 70 y/o	25-40%	33%
	> 70 y/o	0%	1%
Sex	Male	40-60%	35%
	Female	40-60%	65%
Ethnic Origins	European/East Asian/South Central Asian	30-40%	61%
	Māori and Pasifika	30-50%	39%
	Māori (with Tā moko)	10-30%	Not Available

Test Methodology

As described above in "On-Site testing", BixeLab received multiple CSV files with logs for the various testing subcomponents.

Log data was combined to measure key performance metrics for intra demographic groups:

Front-end

1. **Bona fide presentation classification error rate (BPCER):** proportion of bona fide transactions that result in a liveness fail outcome.
2. **Bona fide presentation non-response rate (BPNRR):** proportion of bona fide transactions that result in no response (failure to acquire).
3. Visual graphical plots illustrating performance variation across tested demographic groups.
4. Analysis to understand the effect of demographic parameters on liveness performance.

Back-end

1. **False match rate (FMR):** proportion of non-mated (impostor) comparison trials that result in a match outcome.
2. **False non match rate (FNMR):** proportion of mated (genuine) comparison trials that result in a non-match outcome.

**IDENTITY CHECK EVALUATION
FOR
TE TARI TAIWHENUA | DEPARTMENT OF INTERNAL AFFAIRS (DIA)**

3. Visual graphical plots illustrating performance variation across tested demographic groups.
4. Analysis to understand the effect of demographic parameters on matching performance.
5. Analysis to understand the effect of other image quality parameters on matching performance.

Test Environment

1. DIA data collection process was a test environment within Datacom service, allowing test volunteers to submit selfie images following the test instructions provided to them.
2. Back-end matching was completed in a secure, access-controlled test environment within Datacom's service. It incorporated a datasets folder containing submitted selfie images tagged with user IDs of test volunteers (test crew); their corresponding reference images (DIA passport face image), tagged with user IDs previously extracted from the DXC Technology database; and the Daon provided Matching end point API. BixeLab provided test scripts to SushLabs/Daon. These Test scripts invoked the Daon API to submit two images (selfie and passport held image) and receive match score outcomes.

4. SCENARIO EVALUATION

For the scenario testing, a test crew, comprising 148 volunteer test subjects, were informed about the test process and then selected to be demographically diverse. Volunteers were DIA employees and performed the tests remotely in New Zealand. This test crew completed tests using DIA's Identity Check's web application interacting with the web app via the user interface. The data collected informed analysis of front-end and back-end functionality.

Front-end testing of the end-to-end system included quality assessment, liveness detection and matching subsystems and usability insights based on participants' interactions with the IUT (Item Under Test). The front-end testing quantified bona fide presentation classification error rate (BPCER) and false rejection rate (FRR) metrics.

The back end (offline) testing of the matching subsystem tested true biometric matching accuracy, and the IUT's susceptibility to fraud. The back end testing quantified false match rate (FMR) and false non-match rate (FNMR).

Log Data overview

Test generated anonymised log data of genuine (mated) match attempts, quality assessments, liveness decision outcomes from test participants, and match outcomes from zero-effort impostor tests is summarised under. Demographic metadata for participants was cross-referenced with test artifacts to inform bias assessments.

Table 3 Log data overview

Test Phase	Title	Description	Number of Attempts	Number of unique USER IDs
Front-end testing	CorrelatedFaceData	Logs contained match decision outcomes related to genuine (mated) match attempts made by the test participants	486 162 attempts for the IDFACE_FACE check 162 attempts for the IDFACE_VIDEO_SAMPLE_1 Check 162 attempts for the IDFACE_VIDEO_SAMPLE_1 check	145
	CoorelatedQualityData	Logs contained Daon quality metadata associated with genuine attempts made by the test participants	617	146
	CorrelatedLiveness	Logs contained liveness decision outcomes related to bona fide liveness attempts made by the test participants	162	146

**IDENTITY CHECK EVALUATION
FOR
TE TARI TAIWHENUA | DEPARTMENT OF INTERNAL AFFAIRS (DIA)**

Test Phase	Title	Description	Number of Attempts		Number of unique USER IDs
Back-end testing	Round_scores	Logs contained match decision outcomes related to zero effort impostor tests done by way of comparing one test participant's attempt against attempts related to all other test participants.	Genuine (mated)	220	146
			Impostor (non-mated)	140140	638
Reference files and metadata	Bias Assessment Participant Demographics DIA-Bias-Artifacts-Key	Anonymised demographic metadata logs for the live test participants were provided. Additionally, consolidated cross references between IdentityX artifact information and client reference codes (User IDs as assigned by DIA to the test participants) were provided to support analysis. This included indexes for code to Daon user and id check references for selfie images, and passport inventory.			150 reference codes for Daon images (EXT_9999 was the same as EXT_0022, ext_3456, ext_5678, and EXT_9089 were not found in the anonymised demographic log data or front-end testing performance logs), these user IDs were found in the Round2 back-end testing logs.

Log Data Validation

Data was captured for 151 unique reference IDs with EXT_0001 being manually added later due to a missing log entry despite passing liveness and matching. EXT_0102 was not included in the logs and requires further investigation to determine the cause of omission. It was confirmed that EXT_0022 and EXT_9999 were the same user, so data related to EXT_9999 was removed. No demographic metadata was available for ext_3456, ext_5678, and EXT 9089, hence data pertaining to these IDs was not used for analysis purposes. Therefore, liveness data for 147 unique participants was used for analysis purposes.

Liveness match logs from genuine tests contained score data for 146 unique test participants. The face_type field included three distinct checks per user: IDFACE_FACE, IDFACE_VIDEO_SAMPLE_1, and IDFACE_VIDEO_SAMPLE_2. Based on feedback from the Department, it was noted that only the IDFACE_FACE check results are relevant to the system's liveness matching performance. Consequently, the log data was filtered to assess performance solely based on outcomes related to the IDFACE_FACE check.

Match logs from offline (full cross-comparison) tests contained score data for 146 unique test participants. Upon further investigation, it was determined that user ID EXT_9999 was the same as EXT_0022. ext_3456, ext_5678, and EXT_9089 were removed because they were not present in

**IDENTITY CHECK EVALUATION
FOR
TE TARI TAIWHENUA | DEPARTMENT OF INTERNAL AFFAIRS (DIA)**

the anonymised demographic log data—understood to represent the recruited test group provided by DIA. The analysis was conducted based on log data for 146 unique users' selfie and 638 users' passport images.

Test Participants – Demographic Metadata

To facilitate scenario testing, the Department of Internal Affairs (DIA) recruited a test crew, providing demographic metadata that included 45 unique self-declared ethnicities. This metadata was crucial in evaluating the performance of the system under test (SUT) across different ethnic groups.

The evaluation assesses and compares system performance between Māori and non-Māori individuals. To facilitate a streamlined analysis and ensure consistency with the evaluation's objectives, the ethnicity metadata provided was consolidated into three categories. This approach enabled any potential disparities in system performance across different ethnic groups to be identified. The categories Table 4 High-level breakdown of the test crew by ethnic background were defined as follows:

- Māori: Participants who self-identified as being of Māori descent.
- Pasifika: Participants who self-identified as being of Pacific Island descent other than Māori.
- Others: Participants who self-identified as being of descent other than Māori or Pasifika.

Table 4 High-level breakdown of the test crew by ethnic background

Ethnicity Category	Percentage of the test crew
Māori	28%
Pasifika	11%
Others	61%

The demographic metadata relating to sex are reported in Table 5 Breakdown of the test crew by sex under. While there were more females more in the test crew than males, meaningful results for both sex categories can be derived from the data.

Table 5 Breakdown of the test crew by sex

Sex Category	Percentage of the test crew
Female	65%
Male	35%

An age cohort breakdown is reported in Table 6 Breakdown of the test crew by age group. The 31 – 50 age group makes up the largest cohort by a significant margin. Any age cohort specific factors are not adequately represented for the 70+ cohort and may not be adequately represented for the under 30 cohorts.

**IDENTITY CHECK EVALUATION
FOR
TE TARI TAIWHENUA | DEPARTMENT OF INTERNAL AFFAIRS (DIA)**

Table 6 Breakdown of the test crew by age group

Age group Category	Percentage of the test crew
18 – 30 y/o	17%
31 – 50 y/o	49%
51 – 70 y/o	33%
70 +	1%

In addition to the primary test data, 492 passport images were randomly selected from the Department of Internal Affairs (DIA) database to support expanded non-match testing. The selection of these images was random, with no associated demographic metadata available. Demographic information was only accessible for participants in the live testing crew, who provided selfies as part of the evaluation.

5.FRONT-END APP TESTING

Front-end testing evaluated the end-to-end system assessing liveness and matching performance using data gathered from a test crew of live participants. The system's effectiveness in detecting genuine attempts (bona fide trials) and to assess its ability to match users accurately across various demographic categories was tested. This section provides a detailed analysis of the test outcomes, including liveness performance, match scores, and the potential challenges faced by certain users during the testing process.

Liveness Performance

The test crew consisted of 148 unique test subjects. Out of these, 146 users were recorded in the liveness log. To understand the discrepancy, further investigation was conducted by the DIA for users *EXT_0001* and *EXT_0102*. The investigation revealed that user *EXT_0001* had successfully passed both the liveness and matching checks but was not included in the initial exported logs, and therefore was manually added later for analysis. User *EXT_0102* was missing from the logs and so did not count towards the number of unique test participants logged.

A total of 162 liveness attempts were recorded. Out of these, 21 attempts failed the liveness check, resulting in an 87% attempt level success rate. These failed attempts were made by 19 unique participants. 17 participants were able to successfully pass liveness on their second attempt, with two requiring a third attempt. This resulted in an overall user level bona fide presentation classification error rate (BPCER) of 0%. Figure 1 Distribution of liveness scores across all attempts and Table 7 Attempt level liveness performance, displayed below, present the total distribution of liveness scores across all attempts.

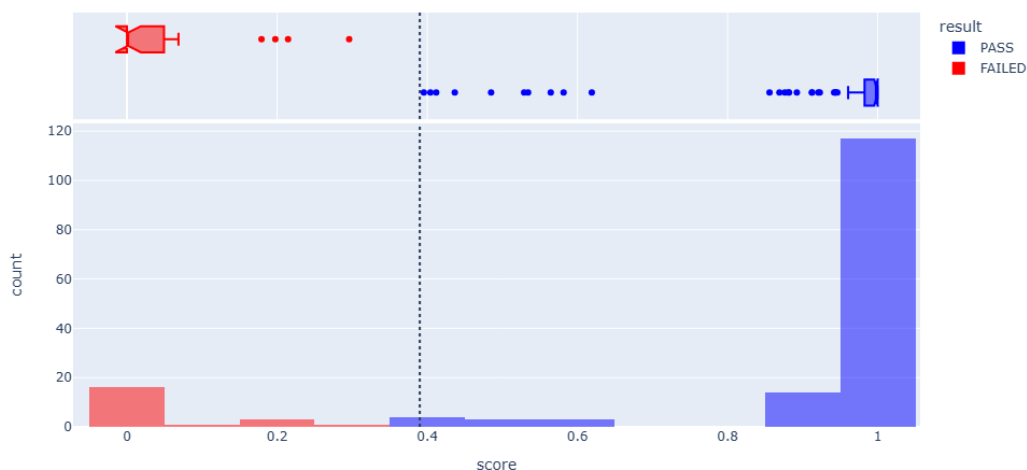


Figure 1 Distribution of liveness scores across all attempts

**IDENTITY CHECK EVALUATION
FOR
TE TARI TAIWHENUA | DEPARTMENT OF INTERNAL AFFAIRS (DIA)**

Table 7 Attempt level liveness performance

Total liveness attempts	Attempts that passed liveness	Attempts that failed liveness
162	141	21

Table 8 Bona fide presentation classification error rate across test crew

Number of unique test participants logged	Number of test participants who could not pass liveness (even after re-trying)	Bona fide presentation classification error rate (BPCER)
146	0	0%

Table 9 High-level overview of the reasons for failed liveness attempts provides further details on the reasons for failed liveness attempts. The most common causes were blurry images, poor lighting, and suboptimal poses by test participants. These issues were reported to DIA, for system refinement.

Table 9 High-level overview of the reasons for failed liveness attempts

Reason for failed liveness attempt	Number of occurrences	Comment from DIA staff
Captured photo is blurry	5	
Issues with test subject's device	1	
Issues with lighting conditions	5	
Issues with test subject's pose	6	
Issue unclear	2	"EXT_0059, I couldn't see anything wrong with this image. Maybe it is slightly dark but not more so than others."
		"EXT_0005: This user used the service a few different times. I remember talking to them about it and the issue seemed related to the quality of their device. The algorithm had trouble detecting their face, when I helped them, it worked."

**IDENTITY CHECK EVALUATION
FOR
TE TARI TAIWHENUA | DEPARTMENT OF INTERNAL AFFAIRS (DIA)**

Figure 2 Distribution of liveness scores for broad (grouped) ethnic categories, as seen below, shows the distribution of liveness scores for the broad ethnic categories assessed across all bona fide attempts. Overall, there is no significant variance in performance between the 'Māori' and 'Others' demographic groups in terms of liveness detection. While there are observable differences in score distribution, particularly among the 'Others' group due to its diversity, the overall detection outcomes remain consistent, indicating that the system performs comparably across both Māori and Non-Māori populations.

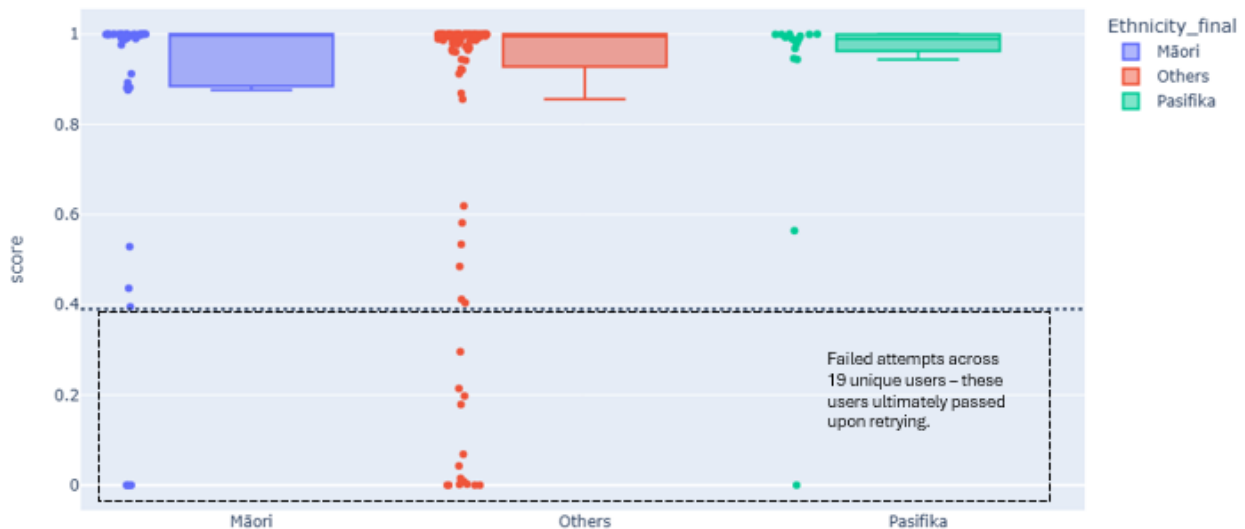


Figure 2 Distribution of liveness scores for broad (grouped) ethnic categories

For adverse outcomes in terms of user attempts being falsely rejected, no statistically significant impact was found for any demographic groups. However, the range of liveness results was much lower for Non-Māori (Others), and therefore any changing of system thresholds must be done carefully since it may disproportionately affect Māori users.

Liveness-Matching Performance

A user is required to pass the liveness check followed by a liveness matching check before they can proceed for authentication to the passport image. Liveness matching is designed to ensure consistent presentation by an individual and is matching the acquired selfies against each other (not the passport image) at two points during the enrolment process.

The liveness-match log recorded 146 unique users. Two users, *EXT_0102* and *EXT_0138*, were found to be missing from the log data. Note that only bona fide data is discussed in this section, as discussions relating to impostors is addressed later, considering the outcomes of cross-comparison testing.

Of the 162 total liveness match attempts, 161 were successful, with one "NO_MATCH" outcome.

Figure 3 Distribution of liveness-match scores from front end testing below provides the distribution of liveness match scores for genuine trials based on front-end testing undertaken by live test participants. For this system, low scores are consistent with successful matches. The histogram demonstrates that most genuine attempts result in low match scores, primarily clustering below 0.05. This indicates that the system generally performs well when matching genuine users, successfully verifying them with low scores. The isolated instance of a "NO_MATCH" is indicated by

**IDENTITY CHECK EVALUATION
FOR
TE TARI TAIWHENUA | DEPARTMENT OF INTERNAL AFFAIRS (DIA)**

the red bar, while the "FAILED" attempts did not receive match scores and are therefore not reflected in the plot.

One specific case involved a Non-Maori ('Others') male aged between 31 and 50 years who initially failed the liveness check, successfully re-attempted, but later failed the liveness-matching process. Upon manual intervention, it was determined that device quality issues were the likely root cause of the failures, with successful detection occurring after assistance was provided (as confirmed by test operators).

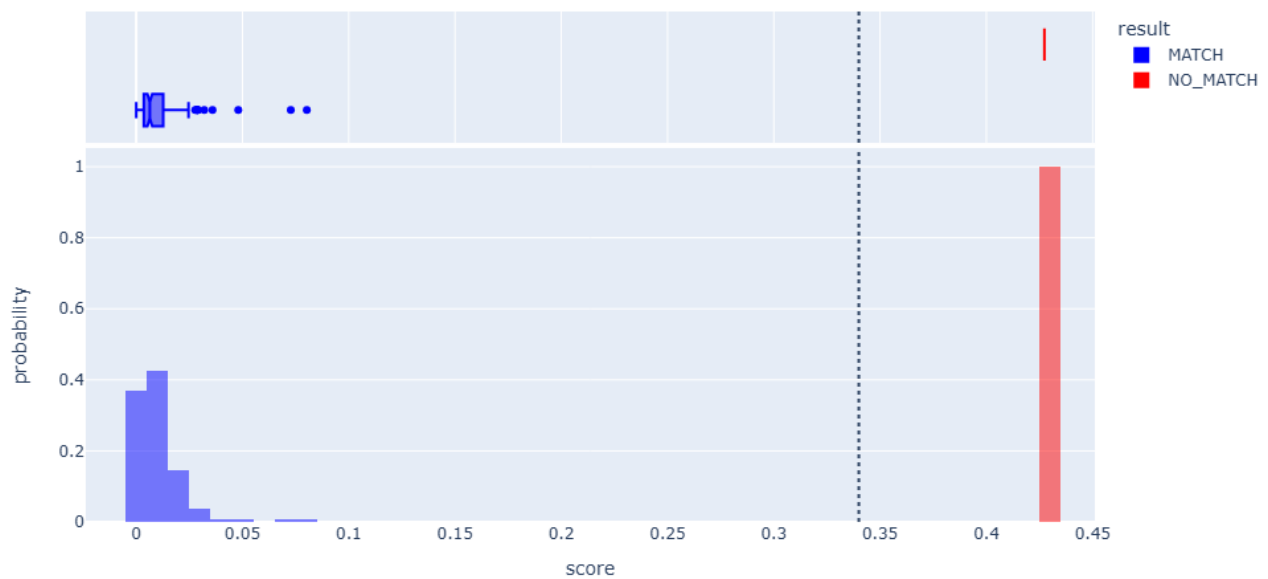


Figure 3 Distribution of liveness-match scores from font end testing

Table 10 Attempt level liveness-matching performance below provides a breakdown of the attempt level performance.

Table 10 Attempt level liveness-matching performance

Liveness match attempts	Attempts that correctly matched	Attempts that received a 'NO_MATCH' outcome
162	161	1 'NO_MATCH' attempt

Table 11 User level liveness matching performance below presents the user-level matching performance, with a liveness match reject rate of 0.7% . This was calculated based on the one user who could not pass the liveness matching process.

**IDENTITY CHECK EVALUATION
FOR
TE TARI TAIWHENUA | DEPARTMENT OF INTERNAL AFFAIRS (DIA)**

Table 11 User level liveness matching performance

Number of unique test participants logged	Number of test participants who could not pass liveness matching (even after re-trying)	Liveness match reject rate
146	1	0.7%

Figure 4 Distribution of liveness matching scores for broad (grouped) ethnicity categories illustrates the distribution of liveness-match scores for genuine trials across broad ethnic categories (Māori, Pasifika, and Others). The match scores are largely concentrated in the lower range, indicating consistent performance across all categories. No significant disparities in performance were observed between the Māori, Pasifika, and Non-Māori ('Others') groups.

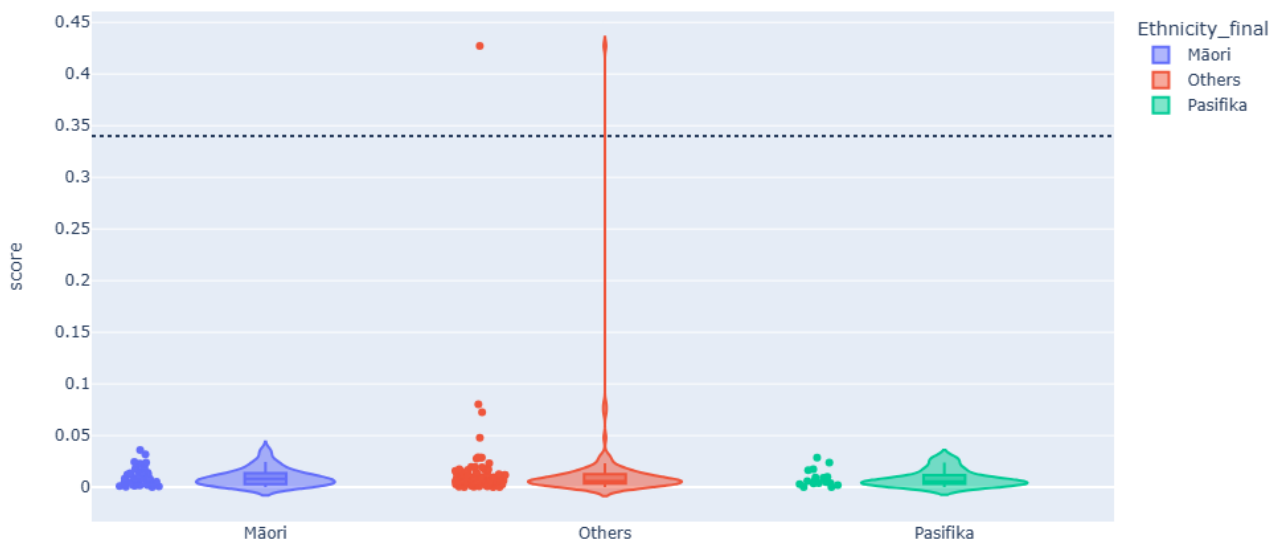


Figure 4 Distribution of liveness matching scores for broad (grouped) ethnicity categories

6. BACK-END MATCHING TESTING

Back-end testing was conducted in DIA's secure test environment, utilising offline processes to measure the true biometric matching accuracy of the IdentityX matching engine. The system's False Match Rates (FMRs) were derived by performing zero-effort impostor comparisons. This involved using the passport database images of test volunteers, along with the selfie images collected during front-end testing.

In these zero-effort impostor tests, each participant's selfie image was compared against the passport images of all other participants, simulating non-genuine (impostor) attempts. This allowed the system's ability to reject these non-mated comparisons to be evaluated and ensured that genuine biometric data was not mistakenly matched with unrelated users.

Figure 5 distribution of match scores for mated and non-mated comparison trials – includes re-attempts - below provides a distribution of mated (genuine) and non-mated (impostor) comparisons across all attempts. The genuine comparisons (mated) represented in blue show a tight clustering around 0.2 with very little spread, the bulk of the true matches fall below 0.3. Impostor comparisons (non-mated) represented in red show wider spread of match score with the median score around 0.5. Some outliers can be seen where impostor attempts have lower scores.

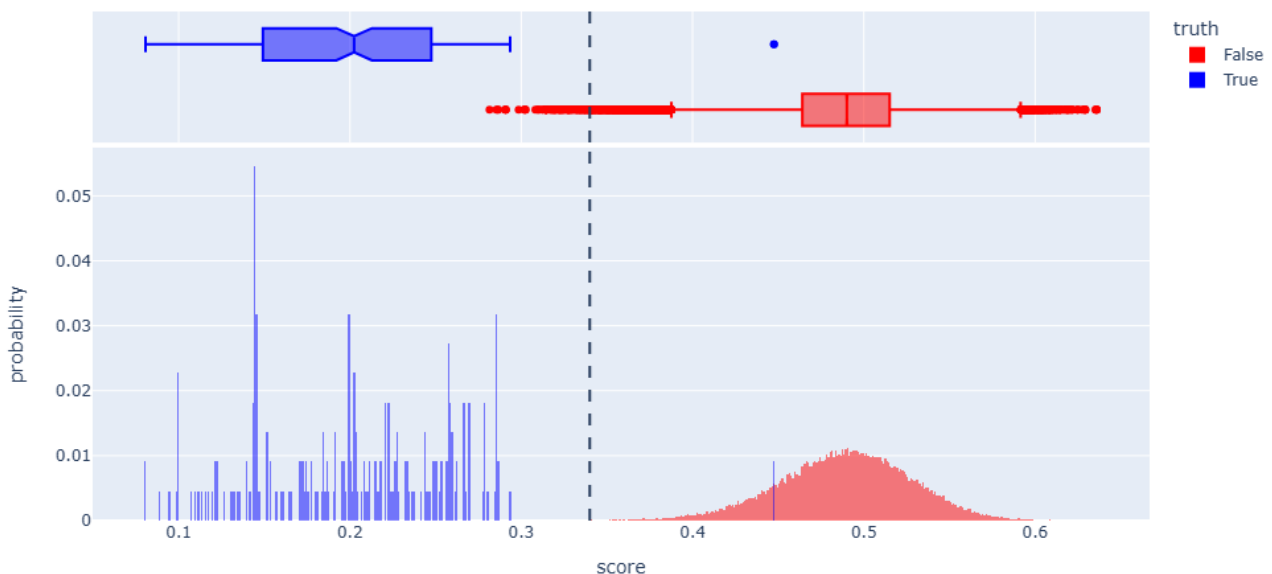


Figure 5 distribution of match scores for mated and non-mated comparison trials – includes re-attempts

Table below provides a breakdown of the false non-matches (false rejections).

Table 12 breakdown of false non-match rate (FNMR)

Total mated comparisons	Falsely non-matched attempts	False Non-Match Rate (FNMR)
220	2	0.9%

**IDENTITY CHECK EVALUATION
FOR
TE TARI TAIWHENUA | DEPARTMENT OF INTERNAL AFFAIRS (DIA)**

Table below provides a breakdown of the false matches.

Table 13 breakdown of false match rate (FMR)³

Total non-mated comparisons	Falsely matched attempts	False Match Rate (FMR)
140140	123	0.09%

The non-mated comparisons were performed by matching each of the 220 probe images (selfies) against a candidate set of 638 passport images. The candidate set contained passport images from the same 146 individuals represented in the probe set, along with an additional 492 passport images from random individuals. To ensure that the comparisons are non-mated, the passport image corresponding to the same user as each probe image is excluded. As a result, each probe image is compared against 637 non-mated candidate images (638 total candidates minus the one mated candidate).

Genuine (Mated) Comparisons

Figure 6 Match score distributions for genuine (mated) comparisons across broad ethnicity categories . The score threshold is represented by the dotted line at 0.34. While some variability in match scores is observed across different groups, all distributions remain below the threshold apart from one transaction, suggesting that these variations do not significantly impact on the system's ability to perform correct matches (and hence affect usability). Despite the physiological differences among the tested groups, reliable performance is observed across the tested participants of Māori, Non-Māori ('Others'), and Pasifika groups.

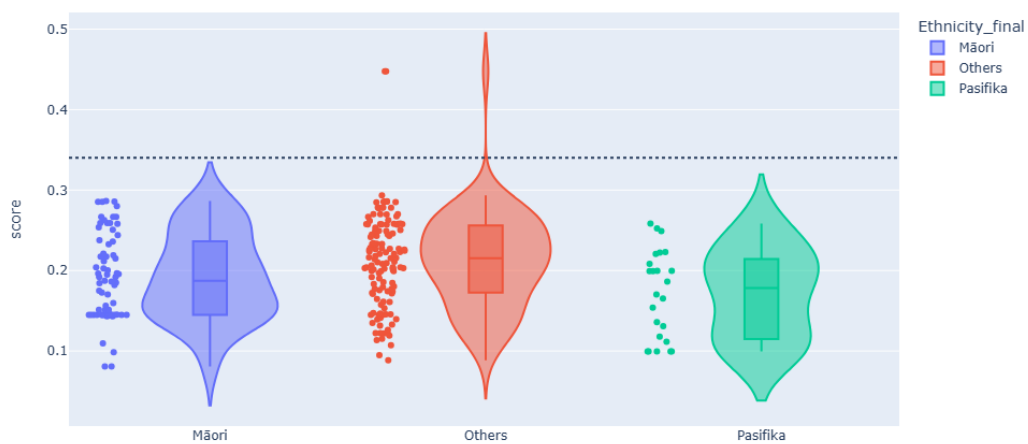


Figure 6 Match score distributions for genuine (mated) comparisons across broad ethnicity categories

³ Includes additional 492 randomly selected passport images.

**IDENTITY CHECK EVALUATION
FOR
TE TARI TAIWHENUA | DEPARTMENT OF INTERNAL AFFAIRS (DIA)**

Impostor (Non-Mated) Comparisons

Figure 7 Match score distributions for impostor (non-mated) comparisons across broad ethnicity categories - presents the match score distributions for non-mated (impostor) comparisons across demographic categories, highlighting the system's ability to distinguish between genuine and impostor attempts. The threshold score, set at 0.34, marks the decision boundary below which false matches are identified. The dashed line represents the mean score across all groups, providing a benchmark for average system performance.

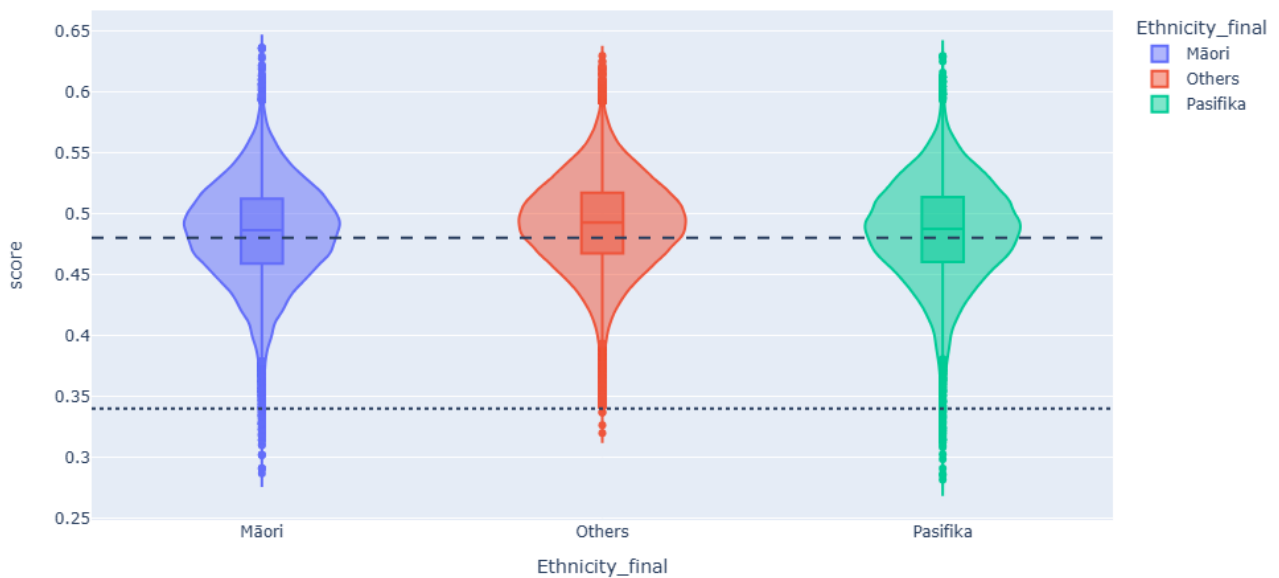


Figure 7 Match score distributions for impostor (non-mated) comparisons across broad ethnicity categories

Whilst all group have some outliers below the threshold the Pasifika and Māori groups clearly have a higher proportion.

Based on manual review undertaken by the Department, some of these might be attributed to factors such as blurry images, acute side angles, lighting issues, shadows, and reflections on face.

Statistical Analysis Overview

This section describes the statistical tests applied to evaluate the performance and fairness of the SUT.

1. The Chi-Squared Test⁴ is used to examine whether two categorical variables are independent in influencing a test statistic. This test was used to determine if there was a significant association between tested demographic groups and matching outcomes (both for mated and non-mated comparisons).
2. The Kolmogorov-Smirnov (KS)⁵ Test is a test to determine how likely is it that two sets of samples are drawn from the same (but unknown) probability distribution. This was used to identify any differences in the distribution of match scores across tested demographic groups.

Chi-Squared Test

- The Chi-Squared result for mated comparisons was 1.48 at P-value 0.48. No significant association was found between ethnicity and matching outcomes. This suggests that, for mated scores, the system performs consistently across tested demographic groups.
- The Chi-Squared result for non-mated comparisons was 199.5 with P-value 4.8×10^{-44} . Tested demographic groups and non-mated matching outcomes were associated, indicating that variations do exist across tested demographic groups when the system is required to reject impostor match attempts.

Kolmogorov-Smirnov (KS) Test

- The KS result for mated comparisons was 0.25 with P-value 0.007. The mated score distributions for tested demographic groups are likely drawn from different distributions. Since this variation did not result in significant disparities in system performance (e.g., False Reject Rate) it is not of concern in the system as currently configured.
- The KS result for non-mated comparisons was 0.08 with P-value 1.6×10^{-148} . The non-mated score distributions for the tested demographic groups are different, however, the Chi-Squared result confirms that the current configuration adequately manages the difference.

⁴ https://en.wikipedia.org/wiki/Chi-squared_test

⁵ https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

**IDENTITY CHECK EVALUATION
FOR
TE TARI TAIWHENUA | DEPARTMENT OF INTERNAL AFFAIRS (DIA)**

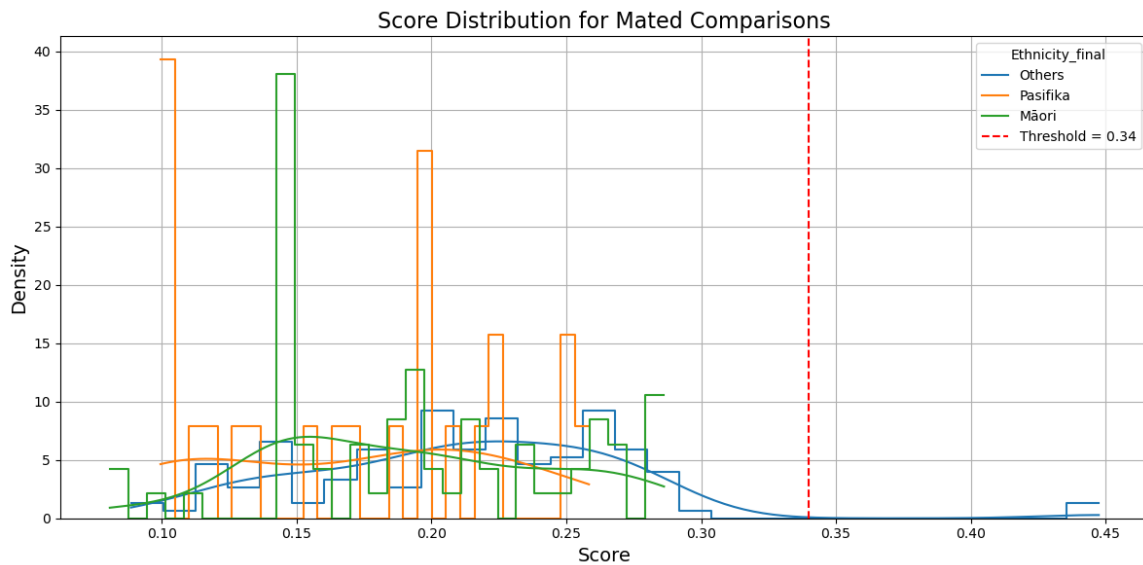


Figure 8 Score distribution for mated comparisons

Figure 8 Score distribution for mated comparisons, shows the score distributions for mated comparisons across the tested demographic groups, with the kernel density estimate (KDE) providing a smoother view of the distributions. Some variations in the score distributions across the tested groups are visible, consistent with the KS Test results. While differences are observed, they do not impact the system's performance as currently configured, as evidenced by the Chi-squared Test results.

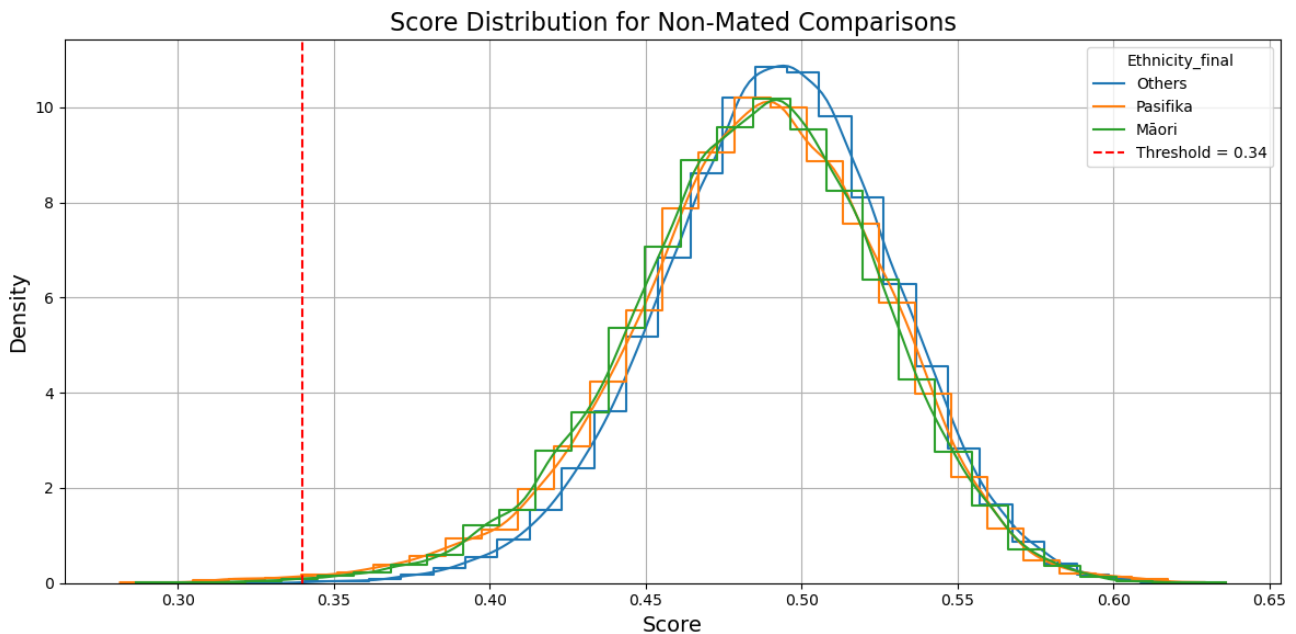


Figure 9 Score distribution for non-mated comparisons

Figure 9 Score distribution for non-mated comparisons shows the score distributions for non-mated comparisons across ethnic groups, with the kernel density estimate (KDE) providing a smoother view of the distributions. The score distributions for non-mated comparisons are generally above the threshold of 0.34 and are therefore correct rejections. However, while not revealed by the scale in Figure 9, variations across the tested groups are more pronounced below the threshold (false accepts) for Pasifika and Maori groups.

7. CONCLUSION

The following were the key findings of the report:

Liveness detection performance

- Out of 162 liveness attempts by 146 participants, 141 attempts passed liveness, resulting in an 87% attempt-level success rate for liveness.
- All participants ultimately passed the liveness check after a maximum of three attempts, leading to a bona fide presentation classification error rate (BPCER) of 0% at the user level
- Failed attempts were primarily due to blurry images, poor lighting conditions, and face alignment and positioning.
- No statistically significant variance in liveness performance was observed between Māori and other participants, indicating consistent performance across these groups.

Liveness Matching Performance

- Out of 162 liveness match attempts, 161 were successful, with one "NO_MATCH" outcome for a non- Māori ('Other') male user aged between 31-50 years.
- The liveness rejection rate was calculated at 0.7% at the user level.
- Liveness match score distribution for genuine trials was consistent within acceptable limits across 'Māori', 'Pasifika', and 'Others' groups, with no significant disparities observed.

Matching Performance

- Based on back-end testing, a total of 140,140 non-mated comparisons were conducted, resulting in 123 false matches. The false match rate was calculated at 0.09%. A total of 220 mated comparisons were present in the log data, resulting in two false non-matches. The false non-match rate (FNMR) was at 0.9%.
- For the match score distributions for genuine (mated) comparisons some variation was seen across the different groups. However, this did not impact on the system's ability to consistently match genuine users of varying ethnicities accurately.
- Distribution for impostor (non-mated) comparisons showed some variations for Māori and Pasifika compared to non- Māori . Based on manual review undertaken by DIA, some of these may be attributed to the impact of influencing factors such as adverse quality. Further investigation of the causes of this higher false accept rate are recommended.

APPENDIX A: EVALUATION MANAGEMENT

Effective test management is critical to ensuring usable data is produced for analysis. Test management involved factors such as:

- Ensuring that all data collected is accurate and relevant.
- Accurate association of test volunteer IDs with reference images obtained from DIA databases to maintain ground truth in presentation attack test cases.
- Effective data logging at the completion of each test.
- Evaluation conducted so that the IUT/SUT is tested to applicable standards.

Data and Privacy Management

Testing process requires the exposure of a limited amount of Personally Identifiable Information (PII). To the maximum extent possible privacy of the test volunteers must be protected by:

- DIA ensuring they obtain explicit consent from volunteer test subjects to confirm their understanding of the personal data obtained that is required to perform testing (DIA)
- DIA ensuring that only the minimum required personal data is collected.
- For off-site testing, test conditions must be monitored as closely as possible.
- Only results relevant to the evaluation are recorded
- DIA and BixeLab delete and destroy all transacted data as soon as possible after completion of the transaction.
- DIA and BixeLab's reporting of references to test volunteers is only by deidentified tester ID's and demographic groupings.
- DIA, Daon and Sushi Labs statement of compliance that data is not retained or used other than in accordance with the test processes, including complete removal of all test data upon completion of the tests.

System Configuration Management

The hardware and software configurations associated with the system being tested must be monitored throughout the evaluation period.

IT Audit logs produced for each day of testing record data including equipment ID, manufacturer and model name, operating system names and versions, identified anomalies if any, system version, date, time, and any other relevant specifications for all software/hardware utilised to conduct the review.

Test Readiness Criteria

A test readiness review is performed to assess whether the IUT/SUT is ready for formal testing by determining whether the test procedures are complete and confirming compatibility with the test plan. This readiness review is conducted at the start of the evaluation process and includes:

1. Review of the proposed test plan
2. A plan for data collection to ensure high data quality and consistency for test analysis.
3. Ensuring the planned test cases are fit for purpose.
4. Ensuring the selected artefacts for evaluation are fit for purpose.

**IDENTITY CHECK EVALUATION
FOR
TE TARI TAIWHENUA | DEPARTMENT OF INTERNAL AFFAIRS (DIA)**

Test Completion Criteria

The evaluation was considered complete as the data collected during the evaluation enabled effective measure of IUT/SUT in terms of the following criteria as applicable to the evaluation scope:

1. Measuring likelihood of a live customer failing the liveness detection mechanism
2. Measuring biometric matching performance of the matching subsystem
3. Measuring performance variation across the tested demographic groups for liveness detection of genuine (live) users and accuracy in matching of genuine and impostor attempts.

IDENTITY CHECK EVALUATION
FOR
TE TARI TAIWHENUA | DEPARTMENT OF INTERNAL AFFAIRS (DIA)

APPENDIX B: TERMS AND DEFINITIONS

Table 14 Terms and definitions

Term	Abbreviation	Definitions
Ground Truth		Ground truth is the truth of whether a given match is genuine or an imposter.
Liveness detection		Measurement and analysis of anatomical characteristics or involuntary or voluntary reactions, to determine if a biometric sample is being captured from a living subject (volunteer) present at the point of capture
Mobile device		Small, compact, handheld, lightweight, standalone computing device, typically having a display screen with digitizer input and/or a miniature keyboard
Normal presentation		Interaction of the biometric capture subject (volunteer) and the biometric data capture subsystem in the fashion intended by the policy and design of the biometric system
System under test	SUT	A biometric system which implements the standard(s) being tested
Target of evaluation	TOE	Within Common Criteria, the IT product that is the subject (volunteer) of the evaluation. Note: The TOE in Common Criteria evaluations is the equivalent of IUT in biometric evaluations.
Test approach		Totality of considerations and factors involved in evaluation
Test Subject (volunteer)		A person who has been recruited to participate in a biometric evaluation.
Test Crew		Set of <i>test subject(s)</i> utilised in an evaluation.
Selfie		Face image captured through the front camera lens of the mobile phone or smart device where the subject to capture distance is below the standards recommended in 19794-5 2005 and 2011.
biometric candidate		biometric reference identifier of a biometric reference in the biometric reference database determined to be sufficiently similar to the biometric probe to warrant further analysis

**IDENTITY CHECK EVALUATION
FOR
TE TARI TAIWHENUA | DEPARTMENT OF INTERNAL AFFAIRS (DIA)**

Term	Abbreviation	Definitions
biometric probe		biometric sample or biometric feature set input to an algorithm for biometric comparison to a biometric reference(s)
comparison score		numerical value (or set of values) resulting from a comparison
mated		of or having to do with a paired biometric probe and biometric reference that are from the same biometric characteristic of the same biometric data subject (volunteer)
non-mated		of or having to do with a paired biometric probe and biometric reference that are not from the same biometric characteristic of the same biometric data subject (volunteer)
authentication		the act of proving or showing to be of undisputed origin or veracity
biometric mated comparison trial		Comparison of a biometric probe and a biometric reference from the same biometric capture subject (volunteer) and the same biometric characteristic as part of a performance test
biometric non-mated comparison trial		comparison of a biometric probe and a biometric reference from different biometric data subjects (volunteers) as part of a performance test
failure to acquire	FTA	failure to accept for subsequent comparison the output of a biometric capture process, a biometric sample of the biometric characteristic of interest
failure to enrol	FTE	failure to create and store a biometric enrolment data record for an eligible biometric capture subject (volunteer), in accordance with a biometric enrolment policy
false match rate	FMR	proportion of the completed biometric non-mated comparison trials that result in a false match
false non-match rate	FNMR	proportion of the completed biometric mated comparison trials that result in a false non-match
Subject (volunteer)		The person from whom the biometric enrolment was taken. The target of the attack

**IDENTITY CHECK EVALUATION
FOR
TE TARI TAIWHENUA | DEPARTMENT OF INTERNAL AFFAIRS (DIA)**

Term	Abbreviation	Definitions
Target of evaluation	TOE	Within Common Criteria, the IT product that is the subject of the evaluation. Note: The TOE in Common Criteria evaluations is the equivalent of IUT in biometric evaluations.
test approach		Totality of considerations and factors involved in technology evaluation
Aggregate equitability measure		A performance measure that combines multiple measures of differential performance into an aggregate metric of overall differential performance.
Categorical demographic variable		Demographic characteristics of an individual that is nominally or ordinally described
Comparison score differential measures		Differences in system measures between different demographic groups
Confidence Interval		an interval estimator (T_0, T_1) for the parameter θ with the statistics T_0 and T_1 as interval limits and for which it holds that $P[T_0 < \theta < T_1] \geq 1 - \alpha$
Continuous demographic variable		Demographic characteristics of an individual that is observable, measurable, and not necessarily constrained to discrete categories.
Demographic reference database		Database comprising biometric references annotated with demographic characteristics.
Differential performance		Difference in biometric system metrics across different demographic groups.
Differential treatment		Taking a set of actions for a biometric enrollee or biometric capture subject (volunteer) based on their demographic characteristics.
Effect magnitude		Statistical measure of the size of an observed differential.
False negative differential performance		Difference in false negative error rates calculated across multiple demographic groups.

IDENTITY CHECK EVALUATION
FOR
TE TARI TAIWHENUA | DEPARTMENT OF INTERNAL AFFAIRS (DIA)

Term	Abbreviation	Definitions
False positive differential performance		Difference in false positive error rates calculated across multiple demographic groups.
Intersectional demographic variable		Demographic variable that is the combination of multiple, individual, categorical demographic variables.
Mated comparison score differential measure		Difference in the statistics of mated score distributions observed for different demographic groups.
Non-mated comparison score differential measure		Difference in the statistics of non-mated score distributions observed for different demographics groups.