Age Estimation Test Report: Persona

Introduction

This report summarizes the results of the independent evaluation of the Age Assurance Software Solution (referred to as the Target of Evaluation, or ToE), performed as part of the Age Assurance Technology Trial. The evaluation focused on the core properties required by ISO 27566-1: functionality, performance, privacy, security and acceptability.

The objective of this test was to assess the readiness and effectiveness of the solution in real-world conditions, to inform regulatory, industry and public stakeholders.

Disclaimer

The inclusion of this test report in the suite of Age Assurance Technology Trial (AATT) documents does not constitute endorsement, certification or approval of any product, service or provider. The findings are based on self-declared information, interviews and test results submitted by participating organisations, and while evaluated under structured criteria, not all claims have been independently verified in full by the Trial team.

This report reflects the status of the technology at the time of testing and within the scope of the Trial. No guarantee is given as to the completeness, accuracy or continued applicability of the findings. The Trial was a technical evaluation only and did not assess legal or regulatory compliance. Inclusion of a test report does not imply market readiness or regulatory acceptance. Any person considering the use of the technology described in this test report should ask the provider for up-to-date evidence of independent conformity assessment of their products or services and should not rely on this test report.

The AATT, Age Check Certification Scheme, nor any of the AATT contractors do not accept any liability for any statements or assessments relied upon in this test report.

Funded by

Australian Government
Department of Infrastructure, Transport, Regional Development, Communications and the Arts

Project by

Age Check Certification Scheme

Amendment register

VERSION	VERSION DATE	AMENDMENT DESCRIPTION
0.9	30 June, 2025	Initial draft
1.0	4 July 2025	Approved for release
1.1	9 July, 2025	Minor Updates to referencing
1.2	21 July, 2025	Minor update to test schedule

Approvals:
This document has been approved for release by:
Mark Pedersen

References

Document	Location (URL address or Other)
ALM Octane – Test Cases	Test Case Execution Dashboard

Glossary

Term	Meaning
ToE	Target of Evaluation

Target of Evaluation

Product Name and Provider

- **Product Name**: Persona age assurance
- **Provider Name**: Persona
- **Version Number/Description**: Current version at time of testing.

Provider's Practice Statement

System Overview and Purpose
Persona is an **Age Assurance Provider** offering a comprehensive platform that supports:

- **Age verification** using government-issued IDs and databases.
- **Selfie-based age estimation** using AI models trained with ethically sourced data.
- **Orchestrated age assurance flows** configurable for relying parties across regulatory environments.

Age Assurance Methods

Age Estimation

- **Single selfie** input.
- Produces an **estimated age**, **image quality score**, and **presentation attack likelihood**.
- Includes **confidence levels and probability scores** for thresholds (e.g., "Likely over 18").

Age Verification

- Government ID verification with selfie match.
- Cross-checks against national databases (e.g. Australia's Document Verification Service).
- **NFC chip verification** of passports for high-assurance scenarios.
- Supports **re-usable digital identities** with passkey technology (with user consent).

Indicators of Confidence

- Uses **reason codes and risk levels** (High, Medium, Low, Minimal) to guide responses:
  - Allows for **step-up verifications**, **custom error messages**, and **manual reviews**.
- Tuning options allow organisations to balance **false positives and negatives**.
- Supports **configurable verification checks** and responses for diverse use cases.

Binding Process

- **Unique token binding** of results to the user's account within the relying party's system.
- **Additional authentication layers** supported:
  - Devices, emails, phone numbers, biometrics, and government IDs.
- Supports **re-verification** for high-assurance or repeated use cases.

Privacy and Data Protection

- **Minimal PII collection**, configured per organisation.
- Complies with **GDPR, Australian Privacy Principles, and ISO 27001**.
- **Purpose limitation**: data is used only for the defined age assurance purpose.
- **Data retention policies** are configurable; Persona supports **deletion and anonymisation**.
- Provides **user rights**: access, correction, erasure, and challenge mechanisms.

Accessibility and Ease of Use

- Interfaces dynamically adapt to:
  - **User age**, **jurisdiction**, and **regulatory context**.
- Features include:
  - Simplified flows for younger users.
  - Localised language and region-specific legal text.
  - Mobile, browser, and multi-device support.
- VPATs (Voluntary Product Accessibility Templates) used to improve accessibility.

Security Measures

- Built to **OWASP standards** with regular **penetration testing**.
- Complies with **ISO 27001** and **SOC 2 Type II**.
- Uses **TLS encryption**, **role-based access control**, and **secure APIs**.
- Includes **anti-spoofing**, **liveness detection**, and **presentation attack detection** aligned with ISO 27566.

Human Rights and Inclusivity

- Conducts **bias impact analysis** across gender, age, and skin tone.
- Training datasets sourced with **ethical protocols and diverse representation**.
- Supports multiple verification methods to cater to **users without biometrics or ID**.
- Designed to avoid **undue restriction** of access to legitimate services (e.g., health, education).

Certification and Audit

- Certified by:
  - **iBeta**: ISO/IEC 30107-3 Levels 1 and 2 for selfie liveness.
  - **ACCS**: Technical Requirements for Age Estimation Level 2.
- Undergoes **internal governance reviews**, external **audits**, and **risk assessments**.

Technology Readiness Assessment (TRL)

TRL	Definition
TRL 1	**Basic Research**: Initial scientific research has been conducted. Principles are qualitatively postulated and observed. Focus is on new discovery rather than applications.
TRL 2	**Applied Research**: Initial practical applications are identified. Potential of material or process to solve a problem, satisfy a need or find application is confirmed.
TRL 3	**Critical Function or Proof of Concept Established**: Applied research advances and early-stage development begins. Studies and laboratory measurements validate an
TRL 4	**Lab Testing/Validation of Alpha Prototype Component/Process**: Design, development and lab testing of components/processes. Results provide evidence that performance targets may be attainable based on projected or modelled systems.
TRL 5	**Laboratory Testing of Integrated/Semi-Integrated System**: System Component and/or process validation is achieved in a relevant environment.
TRL 6	**Prototype System Verified**: System/process prototype demonstration in an operational environment (beta prototype system level).
TRL 7	**Integrated Pilot System Demonstrated**: System/process prototype demonstration in an operational environment (integrated pilot system level).
TRL 8	**System Incorporated in Commercial Design**: Actual system/process completed and qualified through test and demonstration (pre-commercial demonstration).
TRL 9	**System Proven and Ready for Full Commercial Deployment**: Actual system proven through successful operations in operating environment and ready for full commercial deployment.

The vendor rates the ToE to be at TRL 9.

Testing Scope and Approach

The evaluation process followed principles defined in ISO/IEC 29119-2:2023, utilising two test levels to structure the test activities required:

- System testing
- Acceptance testing

System testing comprised the following activities:

Automated functional testing was used to evaluate the accuracy of each participating technology.

The test environment was integrated with the ToE so that the test environment could transmit HTTP requests with an image payload and receive an estimated age of the subject. This enabled automated testing of the system's effectiveness in producing age estimates using image-based input data.

To facilitate automated testing, a dataset comprising over 1,100 selfie portraits was assembled. These portraits represent individuals aged between 14 and 23 years. Additional images were sourced from the school-based component of the trial to expand the testing dataset. The test aimed to benchmark age estimation performance across multiple threshold categories relevant to policy implementation.

To facilitate bias detection, the dataset was divided into skin tone groups using a skin tone classifier trained on data labelled according to the Fitzpatrick skin tone scale. Skin tones I-II made up 44% of the total, skin tones III-IV made up 41% of the total, and skin tones V-VI made up 11%. Sample sizes for each subgroup were sufficient to meet the requirements for a 5% error margin at the target 95% confidence interval. Subgroup analysis was used to detect potential systemic disparities and to support broader fairness assessments in line with ISO/IEC 27566 and IEEE 2089.1 standards

Manual functional testing was used to test:

- the **interoperability** aspects of the ToE by a combination of manual tests, such as to confirm that a given technology works on various device platforms
- **the robustness** of the system with respect to variations in input quality and presentation attack detection features based on ISO/IEC 30107
- **privacy aspects** of the system in terms of revealing unnecessary Personally Identifiable Information in results.

Manual and automated functional testing was conducted in a laboratory setting. The test environment provided direct integration with the ToE, simulating the use of the system in a typical age assurance setting (e.g. a public-facing web application).

All lab testing executed through framework portal on multiple devices and browsers. Testing performed through the ppurpose built testing framework and connected to the specified vendor endpoint.

Static reviews were used on evaluation of features relating to privacy, data security, compliance with human rights requirements and technology readiness assessment. Dynamic testing of these features for

each participating technology was beyond the scope of the current trial, however dynamic testing of these features is recommended for any technology being deployed.
Static reviews comprised a review of the provider's practice statement and interviews with the provider to clarify any additional details.

Acceptance testing comprised the following activities:

Field trials in schools:

As part of the School Testing initiative, participating students received a structured and age-appropriate briefing outlining the objectives of the Age Assurance Technology Trial. This introductory session explained the role of age estimation technologies in enhancing online safety and supporting regulatory compliance within digital environments.

Students were shown how to access the designated test platform via https://test.aatt.kjr.com.au. Participants provided the required demographic details needed for analysing the results and were instructed on how to access the assigned age assurance service provider service. Students then were left to complete their assigned age assurance tasks independently.

Accuracy metrics were calculated for across all three age gates by aggregating the true and false classification results. These metrics were used to quantify the system's overall performance, including true positive and true negative rates.

Field trials with mystery shoppers:

Mystery shopper testing is a real-world, scenario-based testing approach where testers simulate actual user interactions without revealing their identity as testers. In the context of the Age Assurance Project, it plays a crucial role in evaluating how the system performs in live or semi-live conditions, mimicking genuine user experiences across diverse scenarios.

- As a part of this testing, **Anonymous Testers** interact with the system as regular users.
- They go through the same flow as any public-facing user:
    - Logging in
    - Uploading ID documents or using selfie-based age estimation
    - Providing or declining consent
    - Encountering and reacting to system feedback

Observations are made on:

- Whether the system **correctly accepts or denies access, and estimates age (if appropriate)**
- **User experience** in terms of task completion rates, ease of use, response time, and satisfaction.

Test Schedule

The testing reported in this document occurred during the period 07/04/2025 to 24/06/2025

Evaluation Results

System Testing: Manual Functional Testing

The Age estimation and Age Verification scenarios are listed in the table below.

Test Scenarios	Results
User successfully verifies age via selfie using a desktop camera	Pass
User attempts verification with an image instead of live selfie	Pass
User attempts to use deepfake software or AI-generated image	Pass
Camera not accessible (permissions denied)	Pass
User has a hat, scarf, or hoodie covering part of face	Pass
Verify that the system prevents users wearing glasses from performing age estimation.	Pass
User falls into 18+ age bracket and system correctly estimates	Pass
User falls into 16+ age bracket and system correctly estimates	Pass
User falls into 13+ age bracket and system correctly estimates	Pass
Poor lighting conditions	Pass
Blurry or unclear selfie	Pass
User successfully verifies age via selfie using a mobile camera	Pass
Solution does not show excessive and/or PII data in results.	Pass
Aboriginal User falls into 18+ age bracket and system correctly estimates	Pass
Aboriginal User falls into 16+ age bracket and system correctly estimates	Pass
Aboriginal User falls into 13+ age bracket and system correctly estimates	Pass
Torres Strait User falls into 16+ age bracket and system correctly estimates	Pass
Torres Strait User falls into 13+ age bracket and system correctly estimates	Fail – System does not give the correct age and there is an error of 3-4 years
Torres Strait User falls into 18+ age bracket and system correctly estimates	Pass

Table 1: Aggregated Manual Functional Test Results

System Tests: Automated functional Testing

Evaluation Criteria

The estimated age of each subject was compared against their actual (ground-truth) age for three key policy-relevant thresholds: 13, 16, and 18 years. Classification outcomes were assigned as follows:

- **True Positive (TP):** Both estimated age and actual age are equal to or greater than the age gate.
- **False Positive (FP):** Estimated age is equal to or greater than the age gate, while the actual age is below the threshold.
- **True Negative (TN):** Both estimated and actual ages are below the age gate.
- **False Negative (FN):** Estimated age is below the age gate, while the actual age is equal to or greater.
- **Null Result:** Where a face was not detected were excluded and marked as null.

Where the system has high accuracy (greater than 80%), entries have been highlighted in green. Where the system has low accuracy (less than 50%) and a false positive rate less than 20%, these entries have been highlighted in amber as this performance indicates a significant number of subjects have been blocked incorrectly, but a low number of subjects passed the gate incorrectly. When users fail to pass an age gate, the ToE will conclude either a false positive or a false negative result.

Where the system has low accuracy (less than 50%) or a false positive rate greater than 20%, these entries have been highlighted in red, as this performance indicates a significant number of subjects have passed the gate incorrectly. Note that **testing was performed without age buffers in place**, to see the actual performance of the ToE's age estimation method.

Bias was assessed using parity for a number of metrics across demographic subgroups based on skin tones. Skin tones were classified according to the Fitzpatrick skin tones, from Type I (pale white) to Type VI (very dark brown/black). Parity figures show the disparity in each metric (False Positive Rate, False Negative Rate, Accuracy and Mean Absolute Error) across demographic subgroups by looking at the difference between each group and the average performance across all skin types. FPR, FNR and Accuracy were labelled as follows:

FPR Bias	FNR Bias	Accuracy Bias	MAE Bias
L: < 1.5% points	L: < 1.5% points	L: < 1.5% points	L: < 0.25
M: 1.5 – 4% points	M: 1.5 – 4% points	M: 1.5 – 4% points	M: 0.25 – 0.75
H: > 4% points	H: > 4% points	H: > 4% points	H: > 0.75

Images in the test set which could not be reliably classified because of variance in lighting conditions have been excluded from the bias calculations.

Results:

Age gate 13

Subject age	Samples	FPR	FNR	TPR	TNR	Accuracy	MAE	Absolute error stdev
<10	2	0.00%	N/A	N/A	100.00%	100.00%	0.87	0.22
10	22	13.64%	N/A	N/A	86.36%	86.36%	1.53	1.69
11	33	18.18%	N/A	N/A	81.82%	81.82%	1.16	1.16
12	204	44.61%	N/A	N/A	55.39%	55.39%	1.28	1.09
13	211	N/A	27.96%	72.04%	N/A	72.04%	1.42	1.20
14	226	N/A	11.95%	88.05%	N/A	88.05%	1.68	1.14
15	72	N/A	5.56%	94.44%	N/A	94.44%	1.68	1.09
16	135	N/A	0.74%	99.26%	N/A	99.26%	1.38	1.06
17	167	N/A	0.60%	99.40%	N/A	99.40%	1.53	1.23
18	360	N/A	0.00%	100.00%	N/A	100.00%	1.41	1.15
19	353	N/A	0.00%	100.00%	N/A	100.00%	1.41	1.12
20	204	N/A	0.00%	100.00%	N/A	100.00%	1.65	1.22
21	48	N/A	0.00%	100.00%	N/A	100.00%	2.04	1.78
22	40	N/A	0.00%	100.00%	N/A	100.00%	2.20	1.65
23	24	N/A	0.00%	100.00%	N/A	100.00%	2.48	1.72
>=25	121	N/A	0.00%	100.00%	N/A	100.00%	N/A	N/A

Age gate 16

Subject age	Samples	FPR	FNR	TPR	TNR	Accuracy	MAE	Absolute error stdev
<10	2	0.00%	N/A	N/A	100.00%	100.00%	0.87	0.22
10	22	0.00%	N/A	N/A	100.00%	100.00%	1.53	1.69
11	33	3.03%	N/A	N/A	96.97%	96.97%	1.16	1.16
12	204	3.43%	N/A	N/A	96.57%	96.57%	1.28	1.09
13	211	15.17%	N/A	N/A	84.83%	84.83%	1.42	1.20
14	226	44.25%	N/A	N/A	55.75%	55.75%	1.68	1.14

15	72	61.11%	N/A	N/A	38.89%	38.89%	1.68	1.09
16	135	N/A	17.78%	82.22%	N/A	82.22%	1.38	1.06
17	167	N/A	10.78%	89.22%	N/A	89.22%	1.53	1.23
18	360	N/A	1.67%	98.33%	N/A	98.33%	1.41	1.15
19	353	N/A	7.08%	92.92%	N/A	92.92%	1.41	1.12
20	204	N/A	3.92%	96.08%	N/A	96.08%	1.65	1.22
21	48	N/A	0.00%	100.00%	N/A	100.00%	2.04	1.78
22	40	N/A	0.00%	100.00%	N/A	100.00%	2.20	1.65
23	24	N/A	0.00%	100.00%	N/A	100.00%	2.48	1.72
>=25	121	N/A	0.00%	100.00%	N/A	100.00%	N/A	N/A

Age gate 18

Subject age	Samples	FPR	FNR	TPR	TNR	Accuracy	MAE	Absolute error stdev
<10	2	0.00%	N/A	N/A	100.00%	100.00%	0.87	0.22
10	22	0.00%	N/A	N/A	100.00%	100.00%	1.53	1.69
11	33	0.00%	N/A	N/A	100.00%	100.00%	1.16	1.16
12	204	1.47%	N/A	N/A	98.53%	98.53%	1.28	1.09
13	211	4.74%	N/A	N/A	95.26%	95.26%	1.42	1.20
14	226	7.52%	N/A	N/A	92.48%	92.48%	1.68	1.14
15	72	9.72%	N/A	N/A	90.28%	90.28%	1.68	1.09
16	135	34.07%	N/A	N/A	65.93%	65.93%	1.38	1.06
17	167	52.69%	N/A	N/A	47.31%	47.31%	1.53	1.23
18	360	N/A	26.11%	73.89%	N/A	73.89%	1.41	1.15
19	353	N/A	36.54%	63.46%	N/A	63.46%	1.41	1.12
20	204	N/A	25.49%	74.51%	N/A	74.51%	1.65	1.22
21	48	N/A	4.17%	95.83%	N/A	95.83%	2.04	1.78
22	40	N/A	2.50%	97.50%	N/A	97.50%	2.20	1.65
23	24	N/A	4.17%	95.83%	N/A	95.83%	2.48	1.72
>=25	121	N/A	0.00%	100.00%	N/A	100.00%	N/A	N/A

Table 2: Aggregated Automated Lab Test Results

In assessing accuracy, we observe that for Age gates 16 & 18 there is high degree of accuracy for ages outside the typical buffer zone of 2 years. For Age gate 13, the false positive rate is high for all ages beneath the gate. For Age gate 16, the False Negative rate is less than 20%, indicating that the majority of users could use age estimation without having to proceed to a verification step. For Age gate 13 the False negative rate is 30% for 13-year-olds and for Age gate 18, the FNR is between 25%-36% for 18–20-year-olds, meaning many will have to proceed to an age verification step to pass the gate. This is similar to human accuracy for binary threshold classification of approximately 60-70%.

In lab testing we observe a Mean Absolute Error of less than 2 years for ages 11 - 20, with a standard deviation of less than 1.5 years.

These values indicate good performance for commercial systems used in a retail context. As expected, a buffer age of at least 2-3 years is required for reliable performance when used for age gating.

Age gate 13

Skin type	Samples	FPR	FNR	Accuracy	MAE	FPR parity	FPR bias	FNR parity	FNR bias	MAE parity	MAE bias	Accuracy parity	Accuracy bias	Overall bias
I-II	989	37.31%	5.15%	90.50%	1.39	1.16	L	0.45	L	0.11	L	0.86	L	L
III-IV	1027	40.00%	4.23%	92.11%	1.52	1.53	M	0.47	L	0.02	L	0.75	L	L
V-VI	206	36.36%	4.89%	91.75%	1.95	2.11	M	0.19	L	0.45	M	0.39	L	M

Age gate 16

Skin type	Samples	FPR	FNR	Accuracy	MAE	FPR parity	FPR bias	FNR parity	FNR bias	MAE parity	MAE bias	Accuracy parity	Accuracy bias	Overall bias
I-II	989	22.17%	6.42%	87.26%	1.39	1.42	L	0.80	L	0.11	L	0.81	L	L
III-IV	1027	21.16%	5.18%	90.26%	1.52	2.43	M	0.44	L	0.02	L	2.19	M	M
V-VI	206	42.50%	3.97%	81.07%	1.95	18.91	H	1.65	M	0.45	M	7.00	H	H

Age gate 18

Skin type	Samples	FPR	FNR	Accuracy	MAE	FPR parity	FPR bias	FNR parity	FNR bias	MAE parity	MAE bias	Accuracy parity	Accuracy bias	Overall bias
I-II	989	13.62%	27.81%	79.88%	1.39	2.50	M	3.43	M	0.11	L	0.13	L	L
III-IV	1027	17.17%	23.99%	78.87%	1.52	1.05	L	0.39	L	0.02	L	0.88	L	L
V-VI	206	22.86%	9.90%	83.50%	1.95	6.74	H	14.48	H	0.45	M	3.75	M	M

Table 3: Aggregated Automated Lab Test Bias Results

In assessing of bias across all the age gates, we observe Low bias in terms of MAE for lighter and medium skin tones, and Medium bias for darker skin tones, indicating that while the ToE performs relatively consistently, it is less consistent for darker skin tones. The average Accuracy disparity is under 7% across all skin types and ages gates, however we see some points of greater disparity, e.g for Age gate 16 and 18, with disparity of around 5-6% for skin type V-VI. Note that the variance in accuracy goes both ways: for Age gate 16, the accuracy for Types V-VI is lower compared to others, but for Age gate 18, the accuracy is higher.

In terms of False Negatives, parity is under 4% except for skin tone V-VI in the Age 18 gate, which sits at 14%. In this case, the system is actually performing **better** for that skin tone group, with a False Negative rate of 12% compared to 26% or 31% for medium and light skin tone groups respectively.

Similarly, False Positives, the parity varies from Low to High. The high disparity in False Positives for Skin Types V-VII for Age gate 13 come from better performance compared to others, whereas the high disparity for the same group in Age gate 16 comes from worse performance.

Overall, the ToE appears to have less consistent performance for the darker skin tone groups (V-VI), although this does not equate to bias against that group – in some cases performance in terms of accuracy is better than for other groups.

Acceptance Tests: Schools

Table 3 below shows the performance of the system for each age gate, including:

- the actual subject age and the number of subject (samples) for that age
- the false positive rate: the percentage of subjects who passed the gate incorrectly
- the false negative rate: the percentage of subjects who were blocked incorrectly
- the true positive rate: the percentage of subjects who correctly passed the gate
- the true negative rate: the percentage of subjects who were correctly blocked
- the mean absolute error: for systems that provide an age estimate, the average difference (both higher and lower) between the subjects' actual age and estimated age in years
- the absolute error standard deviation: the margin of error in years

Note: In general age estimation systems perform better when the subject's age is further away from the actual age gate being assessed. Accuracy will tend to decline as the subject's actual age approaches the gate being assessed. Where the number of samples for a given age is less than 30, the performance cannot be said to be statistically significant. Entries with less than 30 samples have been included for completeness but should not be taken as definitive indicators of performance for that age.

Where the system has high accuracy (greater than 80%), entries have been highlighted in green. Where the system has low accuracy (less than 50%) and a false positive rate less than 20%, these entries have been highlighted in amber as this performance indicates a significant number of subjects have been blocked incorrectly, but a low number of subjects passed the gate incorrectly. When users fail to pass an age gate, the ToE will refer them to an age verification step.

Where the system has low accuracy (less than 50%) or a false positive rate greater than 20%, these entries have been highlighted in red, as this performance indicates a significant number of subjects have passed the gate incorrectly.

Results:

Age gate 13

Subject age	Samples	FPR	FNR	TPR	TNR	Accuracy	MAE	Absolute error stdev
11	20	30.00%	N/A	N/A	70.00%	70.00%	0.92	0.72
12	175	46.86%	N/A	N/A	53.14%	53.14%	1.25	1.19
13	136	N/A	21.32%	78.68%	N/A	78.68%	1.49	1.52
14	121	N/A	12.40%	87.60%	N/A	87.60%	1.70	1.25
15	22	N/A	9.09%	90.91%	N/A	90.91%	2.38	1.50
16	40	N/A	0.00%	100.00%	N/A	100.00%	1.46	1.07
17	22	N/A	0.00%	100.00%	N/A	100.00%	1.44	1.28

Age gate 16

Subject age	Samples	FPR	FNR	TPR	TNR	Accuracy	MAE	Absolute error stdev
11	20	0.00%	N/A	N/A	100.00%	100.00%	0.92	0.72
12	175	6.29%	N/A	N/A	93.71%	93.71%	1.25	1.19
13	136	17.65%	N/A	N/A	82.35%	82.35%	1.49	1.52
14	121	38.02%	N/A	N/A	61.98%	61.98%	1.70	1.25
15	22	81.82%	N/A	N/A	18.18%	18.18%	2.38	1.50
16	40	N/A	12.50%	87.50%	N/A	87.50%	1.46	1.07
17	22	N/A	0.00%	100.00%	N/A	100.00%	1.44	1.28

Age gate 18

Subject age	Samples	FPR	FNR	TPR	TNR	Accuracy	MAE	Absolute error stdev
11	20	0.00%	N/A	N/A	100.00%	100.00%	0.92	0.72
12	175	1.71%	N/A	N/A	98.29%	98.29%	1.25	1.19
13	136	6.62%	N/A	N/A	93.38%	93.38%	1.49	1.52
14	121	8.26%	N/A	N/A	91.74%	91.74%	1.70	1.25
15	22	31.82%	N/A	N/A	68.18%	68.18%	2.38	1.50
16	40	32.50%	N/A	N/A	67.50%	67.50%	1.46	1.07
17	22	59.09%	N/A	N/A	40.91%	40.91%	1.44	1.28

Table 4: Aggregated School Field Trial Results

In assessing accuracy, we observe that the false positive rate is above 20% for ages within 2-3 years for all gates. False Negative Rates are less than 13% except for in the Age 13 gate where the FNR for 13-year-olds was 21%. Overall, in schools testing, the performance of the ToE demonstrates acceptable levels of accuracy, with the understanding that users within a buffer zone of 2-3 years of a given age gate would need to "step-up" to an age verification method to pass the age gate.

Response Time

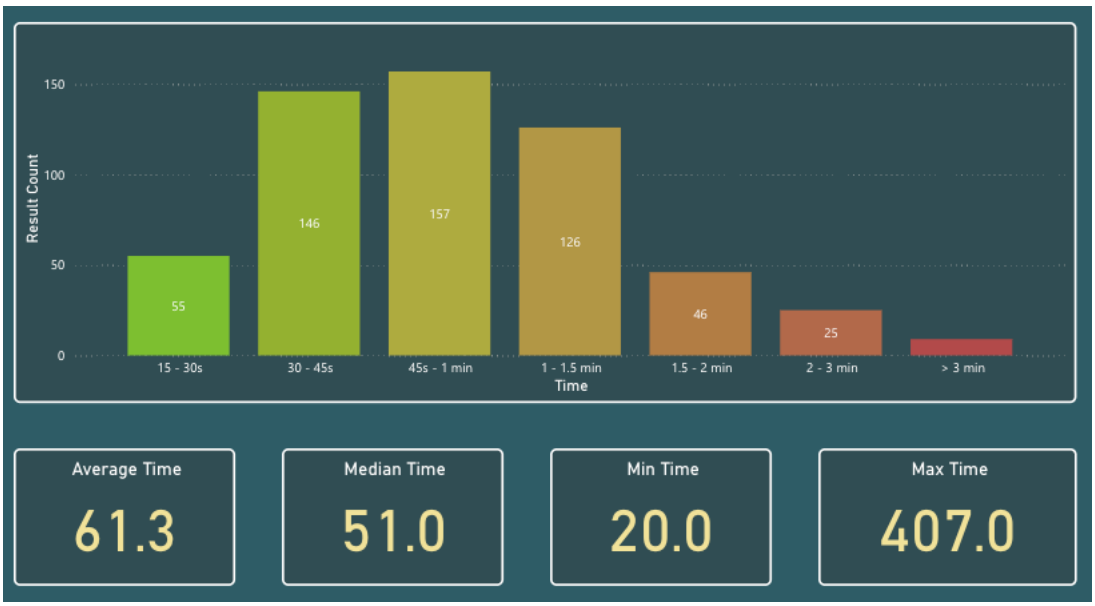

Figure 1: Task Completion Times

The system displayed a slightly longer median response time compared to the 30 seconds typical of age estimation systems.

Acceptance Tests: Mystery Shoppers

No mystery shopper tests were conducted for Persona

Evaluator Observations

- Manual functional testing confirmed the system's robustness, interoperability across devices and browsers, and privacy compliance. One edge case for Torres Strait Islander users (age 13+) failed with a consistent 3–4-year error.
- In assessing accuracy, we observe that for Automated lab tests, Age gates 16 & 18 there is high degree of accuracy for ages outside the typical buffer zone of 2 years. For Age gate 13, the false positive rate is high for all ages beneath the gate. Overall, there appears to be slight bias against darker skin tone, although performance is consistent across all age gates
- School testing showed acceptable overall accuracy, though performance declined near gate thresholds (especially Age 16 and 18 gates). False negative rates were mostly below 13%, but false positives exceeded 20% for borderline ages.
- Mystery shopper testing was not conducted, and the system had a slightly longer than typical response time.
- While the vendor assessed their ToE at **TRL 9**, independent evaluation concluded it to be at **TRL 8**, suggesting further validation is needed for full commercial deployment.

Vendor Comments on Evaluation Results

TBC