

# Age Assurance Technology Trial

Document Sensitivity: Public, Basic, High

## Age Estimation Test Report: YOTI

### 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.

Date: 10/06/2025

Doc. Version: 1

Page 1 of 23

Funded by



**Australian Government**

**Department of Infrastructure, Transport,  
Regional Development, Communications and the Arts**

Project by





## Amendment register

VERSION	VERSION DATE	AMENDMENT DESCRIPTION
0.9	30 June, 2025	Initial release
1.0	4 July, 2025	Approved for release
1.1	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	<a href="#">Test Case Execution Dashboard</a>

## Glossary

Term	Meaning
ToE	Target of Evaluation



## Target of Evaluation

### Product Name and Provider

- **Product Name:** N/A
- **Provider Name:** YOTI
- **Version Number/Description:** Current version at time of testing

### Provider's Practice Statement

**Yoti** is an *Age Assurance Provider* offering an age verification solution that:

- Ensures **no persistent storage** of personal data.
- Prevents **traceable digital footprints**.
- Provides **context-specific verification options** tailored to the user and their device/environment.
- Gives control to the **user and relying party** to choose the method of verification.
- **Facial Age Estimation:**
  - Uses a real-time facial image.
  - Includes **liveness checks**.
  - Approved by **NIST**, **ACCS**, and recognized by **FSM** and **KJM** in Germany.
  - Does **not** use or store biometric identifiers.
  - Does **not** perform facial recognition.
- **Digital ID:**
  - Shares only age-related data (e.g., "18+") with user consent.
  - Data source may include official ID documents or facial estimation.
  - Process involves secure, user-controlled sharing.
- **Age from Government-issued Documents:**
  - Users submit document images + selfies.
  - **Face-match and document authenticity checks** are performed.
  - All captured data is **destroyed after use**, only age result is shared.
- **Age Tokens:**
  - Tokens saved in the user's browser post-verification.
  - Help avoid repeated age checks while maintaining privacy.

### Indicators of Confidence

Yoti ensures high-confidence age verification by:

- Setting **risk-informed thresholds** for age estimation.
- Ensuring only **one face** is processed.
- Performing **image quality checks**.
- Using **liveness detection** and **document authenticity** validation.
- Implementing **anchoring** (face match against document photo).



## Binding Processes

- Binding methods vary based on the chosen verification approach:
  - **Facial Estimation:** Checks for single, live face using a genuine input device.
  - **Digital ID:** Verified identity bound to device + secure decryption tied to user control.
  - **ID Document:** Multi-step validation including face match, liveness, and document verification.

## Privacy & Security Protections

- **No data retained** post-verification.
- Compliant with **ISO 270001, ISO 277001, SOC2 Type 2, and PAS 1296:2018.**
- Regular **internal and external audits.**
- **Access controls and training** for staff with sensitive data access.
- **Automated monitoring** of internal systems for anomalies.
- **Bounty program** and **penetration testing** to strengthen security.

## Accessibility & Ethics

- Designed to meet **WCAG 2.2 AA** standards and **UK ICO's Age-Appropriate Design Code.**
- **Simple, user-friendly interface** using plain language.
- **Internal Ethics Committee** and **external Guardian Council** review ethical impacts.
- **Whistleblower policy** supports internal accountability.

## Certifications and Audits

- External audits by Big 4 accounting firms.
- Regular internal quality checks.
- Certifications include:
  - **ISO 270001, ISO 277001, ISO 9001**
  - **SOC2 Type 2, PAS 1296:2018**
- Active participation in **benchmarking studies** (e.g., NIST facial age analysis).
- White papers published on **accuracy of age estimation models.**

## Technology Readiness Assessment (TRL)

TRL	Definition
TRL 1	<b>Basic Research:</b> Initial scientific research has been conducted. Principles are qualitatively postulated and observed. Focus is on new discovery rather than applications.
TRL 2	<b>Applied Research:</b> Initial practical applications are identified. Potential of material or process to solve a problem, satisfy a need or find application is confirmed.
TRL 3	<b>Critical Function or Proof of Concept Established:</b> Applied research advances and early stage development begins. Studies and laboratory measurements validate an



<b>TRL 4</b>	<b>Lab Testing/Validation of Alpha Prototype Component/Process:</b> Design, development and lab testing of components/processes. Results provide evidence that performance targets may be attainable based on projected or modelled systems.
<b>TRL 5</b>	<b>Laboratory Testing of Integrated/Semi-Integrated System:</b> System Component and/or process validation is achieved in a relevant environment.
<b>TRL 6</b>	<b>Prototype System Verified:</b> System/process prototype demonstration in an operational environment (beta prototype system level).
<b>TRL 7</b>	<b>Integrated Pilot System Demonstrated:</b> System/process prototype demonstration in an operational environment (integrated pilot system level).
<b>TRL 8</b>	<b>System Incorporated in Commercial Design:</b> Actual system/process completed and qualified through test and demonstration (pre-commercial demonstration).
<b>TRL 9</b>	<b>System Proven and Ready for Full Commercial Deployment:</b> 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 based on skin tone distribution in Australia. 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 purpose 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 24/03/2025 to 20/06/2025.

## Evaluation Results

### System Testing: Manual Functional Testing

The Age Estimation scenarios are listed in the table below.

Test Scenarios	Results
Valid case for 16+ age estimation	Pass
Age Estimation Correctly Identifies User in 13+ Bracket	Pass
Age Estimation for Aboriginal Individuals	Pass
User Partially Accepts Privacy Terms but Does Not Provide Consent for Age Estimation	Pass
User Partially Accepts consent but Does Not accept terms for Age Estimation	Pass
Age Estimation Fails Due to Camera Permissions Denied	Pass
User Attempts to Use a Photo Instead of Real-Time Face	Pass
Age Estimation Accuracy Across Different Ethnicities and Skin Tones	Pass
Facial Structure Altered Due to Medical Conditions	Pass
User's Distance from Camera Affects Age Estimation	Pass
Exaggerated Facial Expressions Impact Age Estimation	Pass
Age Estimation Correctly Identifies User in 16+ Bracket	Pass
Age Estimation Correctly Identifies User in 18+ Bracket	Pass
User Accesses Age Estimation Feature from Different Australian States	Pass
User Does Not Properly Position Face in Frame	Pass
Age Estimation Affected by Poor Lighting Conditions	Pass
Multiple Faces in Frame during Age Estimation	Pass
Verify System Deletes the Image Post-Age Estimation	Pass
Facial Hair or Beard Impacting Age Estimation	Pass
User wearing mask trying to use face estimation	Pass
User Attempts to Bypass Estimation with Deepfake or AI-Generated Image	Pass
User wearing hat during age estimation	Pass
Age estimation correctly works end-to-end using different devices. Laptop, PC, tablet, smartphone.	Pass
Age estimation correctly works end-to-end using different Browser (Chrome, Firefox, Microsoft Edge, Safari)	Pass
Solution does not show excessive and/or PII data in results.	Pass
Age Estimation Correctly Identifies Aboriginal User in 18+ Bracket	Pass
Age Estimation Correctly Identifies Aboriginal User in 16+ Bracket	Pass
Age Estimation Correctly Identifies Aboriginal User in 13+ Bracket	Pass
Age Estimation Correctly Identifies Torres Strait User in 13+ Bracket	Pass
Age Estimation Correctly Identifies Torres Strait User in 16+ Bracket	Pass
Age Estimation Correctly Identifies Torres Strait User in 18+ Bracket	Pass



<b>Solution does not show excessive and/or PII data in results.</b>	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	50.00%	N/A	N/A	50.00%	50.00%	2.03	1.48
10	28	64.29%	N/A	N/A	35.71%	35.71%	3.13	1.54
11	33	60.61%	N/A	N/A	39.39%	39.39%	2.12	1.06
12	184	86.41%	N/A	N/A	13.59%	13.59%	1.89	1.03
13	226	N/A	5.31%	94.69%	N/A	94.69%	1.41	1.00
14	242	N/A	1.65%	98.35%	N/A	98.35%	1.25	1.00
15	68	N/A	0.00%	100.00%	N/A	100.00%	1.11	0.94
16	133	N/A	0.00%	100.00%	N/A	100.00%	1.04	0.97
17	129	N/A	0.00%	100.00%	N/A	100.00%	0.95	1.15
18	308	N/A	0.00%	100.00%	N/A	100.00%	1.00	1.30
19	351	N/A	0.00%	100.00%	N/A	100.00%	1.40	1.12
20	193	N/A	0.00%	100.00%	N/A	100.00%	1.93	1.26
21	45	N/A	0.00%	100.00%	N/A	100.00%	2.59	1.30
22	31	N/A	0.00%	100.00%	N/A	100.00%	2.60	1.70
23	22	N/A	0.00%	100.00%	N/A	100.00%	2.58	1.65
>=25	128	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%	2.03	1.48
10	28	7.14%	N/A	N/A	92.86%	92.86%	3.13	1.54
11	33	3.03%	N/A	N/A	96.97%	96.97%	2.12	1.06
12	184	8.70%	N/A	N/A	91.30%	91.30%	1.89	1.03
13	226	15.93%	N/A	N/A	84.07%	84.07%	1.41	1.00
14	242	34.71%	N/A	N/A	65.29%	65.29%	1.25	1.00
15	68	57.35%	N/A	N/A	42.65%	42.65%	1.11	0.94
16	133	N/A	20.30%	79.70%	N/A	79.70%	1.04	0.97
17	129	N/A	4.65%	95.35%	N/A	95.35%	0.95	1.15
18	308	N/A	0.97%	99.03%	N/A	99.03%	1.00	1.30
19	351	N/A	1.14%	98.86%	N/A	98.86%	1.40	1.12
20	193	N/A	0.52%	99.48%	N/A	99.48%	1.93	1.26
21	45	N/A	0.00%	100.00%	N/A	100.00%	2.59	1.30
22	31	N/A	0.00%	100.00%	N/A	100.00%	2.60	1.70
23	22	N/A	0.00%	100.00%	N/A	100.00%	2.58	1.65
>=25	128	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%	2.03	1.48
10	28	0.00%	N/A	N/A	100.00%	100.00%	3.13	1.54
11	33	0.00%	N/A	N/A	100.00%	100.00%	2.12	1.06
12	184	1.09%	N/A	N/A	98.91%	98.91%	1.89	1.03
13	226	2.21%	N/A	N/A	97.79%	97.79%	1.41	1.00
14	242	3.31%	N/A	N/A	96.69%	96.69%	1.25	1.00
15	68	4.41%	N/A	N/A	95.59%	95.59%	1.11	0.94
16	133	15.79%	N/A	N/A	84.21%	84.21%	1.04	0.97
17	129	27.13%	N/A	N/A	72.87%	72.87%	0.95	1.15
18	308	N/A	52.60%	47.40%	N/A	47.40%	1.00	1.30
19	351	N/A	49.57%	50.43%	N/A	50.43%	1.40	1.12
20	193	N/A	31.61%	68.39%	N/A	68.39%	1.93	1.26
21	45	N/A	11.11%	88.89%	N/A	88.89%	2.59	1.30
22	31	N/A	0.00%	100.00%	N/A	100.00%	2.60	1.70
23	22	N/A	0.00%	100.00%	N/A	100.00%	2.58	1.65
>=25	128	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 most users who are eligible to pass a gate will be able to make use of estimation without having to proceed to a verification step. For Age gate 18 the False negative rate is for 18-year-olds is 55%, meaning more than half will have to proceed to an age verification step, which is slightly worse than 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 13 - 20, with a standard deviation of less than 1.3 years.

These values indicate strong 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	Sam ples	FPR	FNR	Accu racy	MAE	FPR parit y	FPR bias	FNR parit y	FNR bias	MAE parit y	MAE bias	Accu racy parit y	Accu racy bias	Ove rall bias
I-II	953	84.4 8%	1.19 %	88.6 7%	1.38	4.21	H	0.34	L	0.07	L	1.25	L	L
III-IV	975	78.6 4%	0.69 %	91.0 8%	1.45	1.63	M	0.16	L	0.00	L	1.16	L	L
V-VI	195	67.8 6%	0.00 %	90.2 6%	1.78	12.4 1	H	0.85	L	0.33	M	0.34	L	M

## Age gate 16

Skin type	Sam ples	FPR	FNR	Accu racy	MAE	FPR parit y	FPR bias	FNR parit y	FNR bias	MAE parit y	MAE bias	Accu racy parit y	Accu racy bias	Ove rall bias
I-II	953	20.8 4%	2.73 %	89.6 1%	1.38	1.55	M	0.26	L	0.07	L	0.07	L	L
III-IV	975	20.9 6%	3.65 %	91.1 8%	1.45	1.43	L	0.66	L	0.00	L	1.50	L	L
V-VI	195	37.0 8%	0.94 %	82.5 6%	1.78	14.6 9	H	2.05	M	0.33	M	7.12	H	H

## Age gate 18

Skin type	Sam ples	FPR	FNR	Accu racy	MAE	FPR parit y	FPR bias	FNR parit y	FNR bias	MAE parit y	MAE bias	Accu racy parit y	Accu racy bias	Ove rall bias
I-II	953	5.37 %	41.2 0%	78.3 8%	1.38	1.84	M	3.66	M	0.07	L	0.80	L	L
III-IV	975	8.43 %	35.1 8%	76.2 1%	1.45	1.22	L	2.36	M	0.00	L	1.37	L	L
V-VI	195	10.0 9%	31.4 0%	80.5 1%	1.78	2.88	M	6.14	H	0.33	M	2.93	M	M

Table 3: Aggregated Automated Lab Test Bias Results

In assessing of bias across all the age gates, we observe Low to Medium bias in terms of MAE, indicating that while the ToE performs relatively consistently in terms of actual age estimation, it is not as consistent across for darker skin tones. The Accuracy disparity generally Low, however we see some points of variance: in Age gate 16, there is disparity of 9% for skin tones V-VI, where the ToE is less accurate compared to other skin tones, and a variance of 2.75% for skin tones III-IV, where the accuracy is slightly better than others. Similarly, we see some variance for skin tones V-VI in Age gate 13, where accuracy is slightly better than others.

In terms of False Negatives, disparity is generally Low except in Age gate 18, where the False Negative rate is higher for skin types I-II and lower for types V-VI. The variance in False Negatives for Age gate 16 is also because skin types V-VII have a lower FNR.

We see similarly mixed results in terms of False Positives. The High variance for skin types V-VI in Age gate 13 comes from a lower FPR compared to others, whereas for the same skin type in Age gate 16 the variance comes from slightly higher FPR.



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 4 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	9	100.00%	N/A	N/A	0.00%	0.00%	1.99	0.93
12	93	96.77%	N/A	N/A	3.23%	3.23%	1.63	1.03
13	27	N/A	7.41%	92.59%	N/A	92.59%	1.07	0.73
14	65	N/A	0.00%	100.00%	N/A	100.00%	0.87	0.68
15	14	N/A	0.00%	100.00%	N/A	100.00%	1.01	0.83
16	13	N/A	0.00%	100.00%	N/A	100.00%	0.58	0.35
17	1	N/A	0.00%	100.00%	N/A	100.00%	2.08	N/A

### Age gate 16

Subject age	Samples	FPR	FNR	TPR	TNR	Accuracy	MAE	Absolute error stdev
11	21	0.00%	N/A	N/A	100.00%	100.00%	1.85	0.87
12	105	10.48%	N/A	N/A	89.52%	89.52%	1.50	1.09
13	40	30.00%	N/A	N/A	70.00%	70.00%	1.60	1.24
14	101	39.60%	N/A	N/A	60.40%	60.40%	1.10	0.78
15	39	56.41%	N/A	N/A	43.59%	43.59%	0.97	0.86
16	45	N/A	15.56%	84.44%	N/A	84.44%	0.81	0.54
17	22	N/A	4.55%	95.45%	N/A	95.45%	1.57	1.38

### Age gate 18

Subject age	Samples	FPR	FNR	TPR	TNR	Accuracy	MAE	Absolute error stdev
11	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A
12	1	0.00%	N/A	N/A	100.00%	100.00%	2.42	N/A
13	1	0.00%	N/A	N/A	100.00%	100.00%	2.25	N/A
14	1	0.00%	N/A	N/A	100.00%	100.00%	1.67	N/A
15	2	0.00%	N/A	N/A	100.00%	100.00%	0.42	0.00
16	17	17.65%	N/A	N/A	82.35%	82.35%	0.68	0.78
17	4	0.00%	N/A	N/A	100.00%	100.00%	0.83	0.99

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 the Age 13 and Age 16 gates. Performance for Age gate 18 is consistently high. False Negative Rates are consistently low, less than 10% except for in the Age 16 gate where the FNR for 16-year-olds was 15%, which is acceptable. Overall, in schools testing, the performance of the ToE demonstrates good 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 in order to pass the age gate.



## Acceptance Tests: Mystery Shoppers

### Results:

#### Age gate 16

Subject age	Samples	FPR	FNR	TPR	TNR	Accuracy	MAE	Absolute error stdev
<10	1	0.00%	N/A	N/A	100.00%	100.00%	0.08	N/A
13	1	100.00%	N/A	N/A	0.00%	0.00%	3.08	N/A
14	12	33.33%	N/A	N/A	66.67%	66.67%	1.54	2.06
15	5	60.00%	N/A	N/A	40.00%	40.00%	0.92	0.70
16	8	N/A	12.50%	87.50%	N/A	87.50%	1.11	1.04
17	1	N/A	0.00%	100.00%	N/A	100.00%	1.33	N/A
>=25	1	N/A	0.00%	100.00%	N/A	100.00%	N/A	N/A

Table 5: Aggregated Mystery Shopper Field Trial Results

In assessing accuracy, the number of mystery shopper subjects using this ToE was relatively low (under 10 samples for most ages) so it is difficult to draw any significant conclusions. We observe that performance is consistent with the school tests for ages on either side the Age 16 gate. The false positive rate is above 20% for ages within 3 years for the Age 16 gate. False Negative Rates are similar to that for the schools test (under 15%). Overall, in mystery shopper testing, the performance of the ToE is consistent with that of the schools testing.



## Usability & Response Time



Figure 1: Evaluation Response Time (s)

The response time is somewhat higher than typical for evaluation systems, which generally have a median response time of approximately 30s, however this ToE does demonstrate greater accuracy and consistency than some other systems.



Mystery shoppers provided the following usability feedback as part of the field trial:

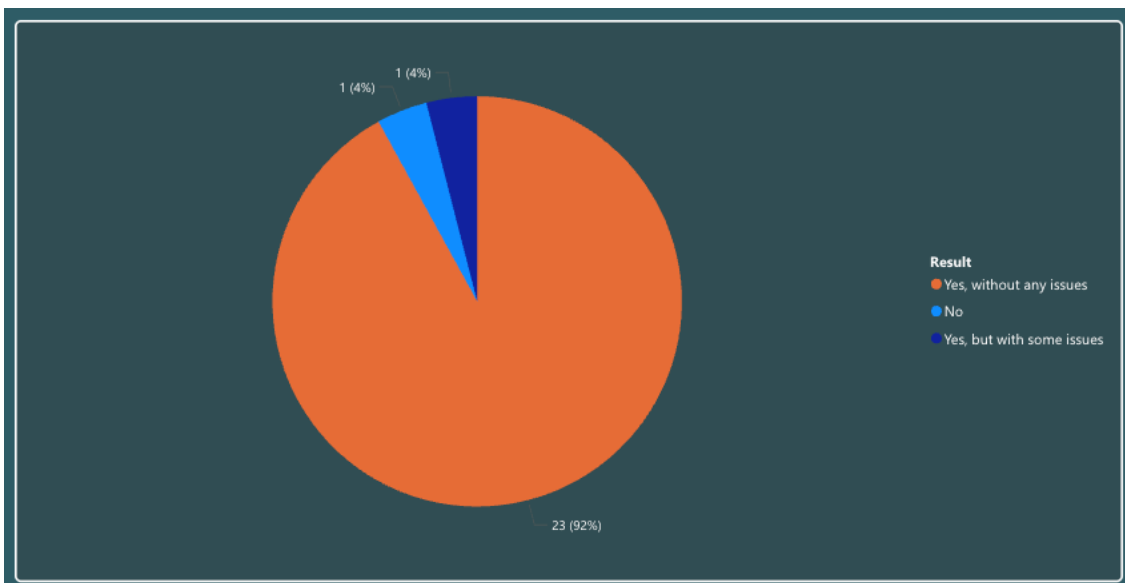


Figure 2: Did you complete the task?

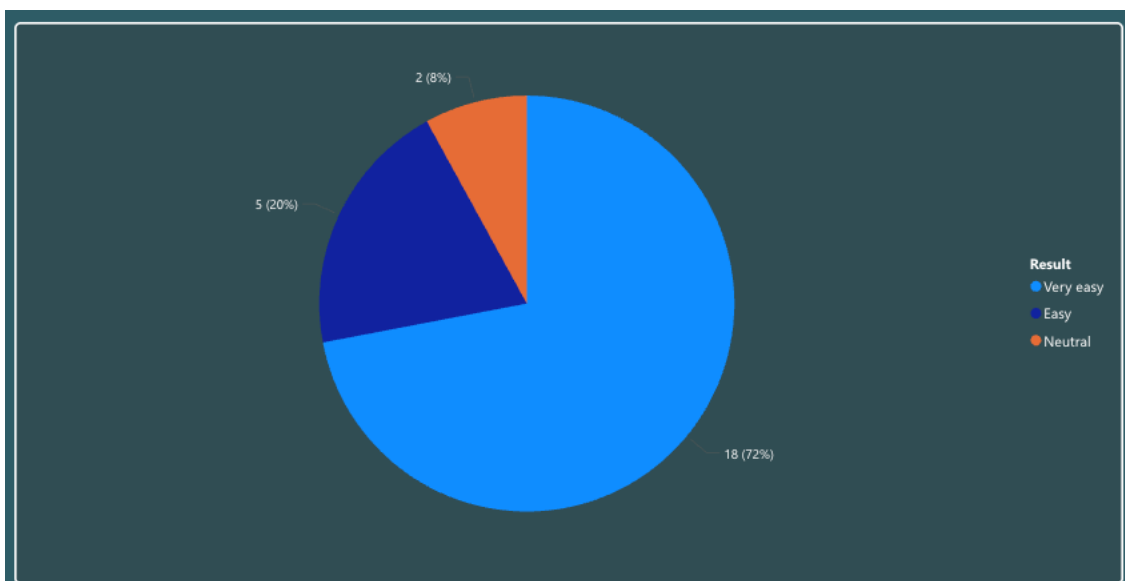


Figure 3: How easy was it to complete the task?

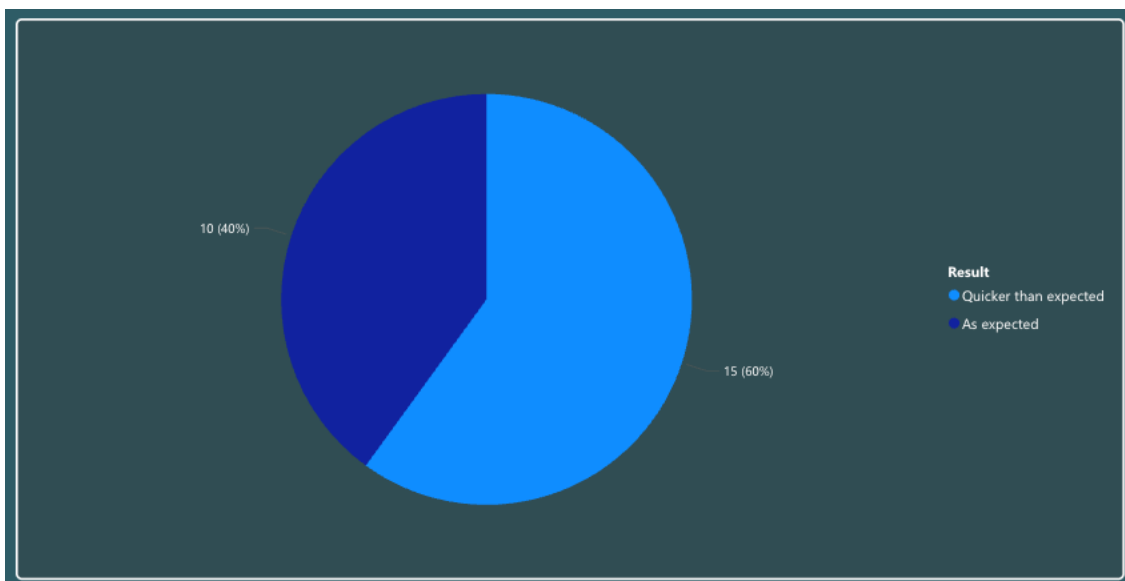


Figure 4: How would you rate the time it took?

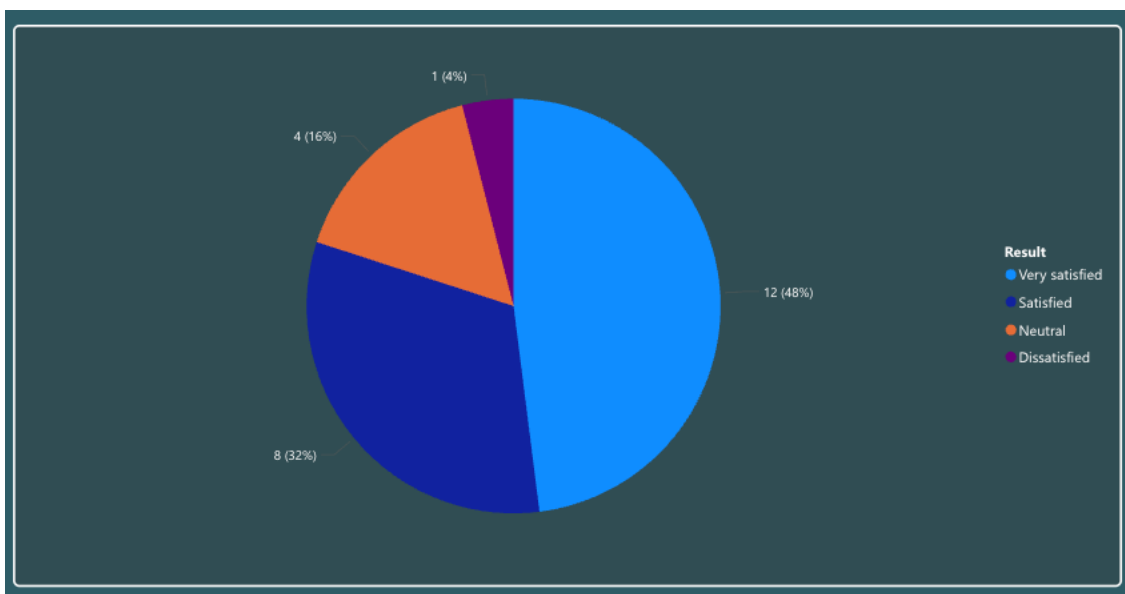


Figure 5: How would you rate overall experience?

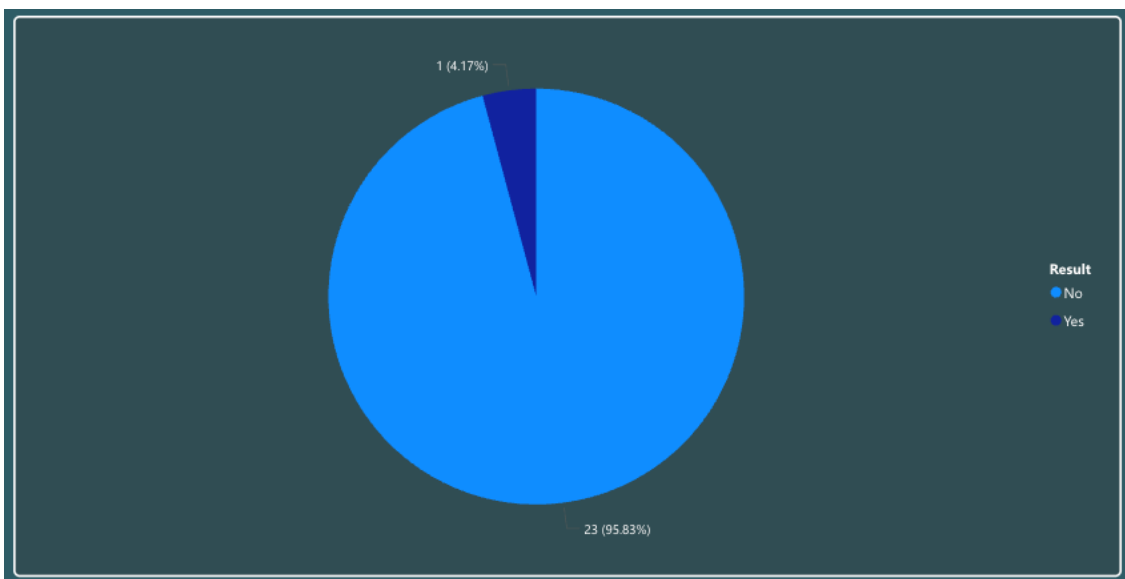


Figure 6: Did you have any concerns about privacy?

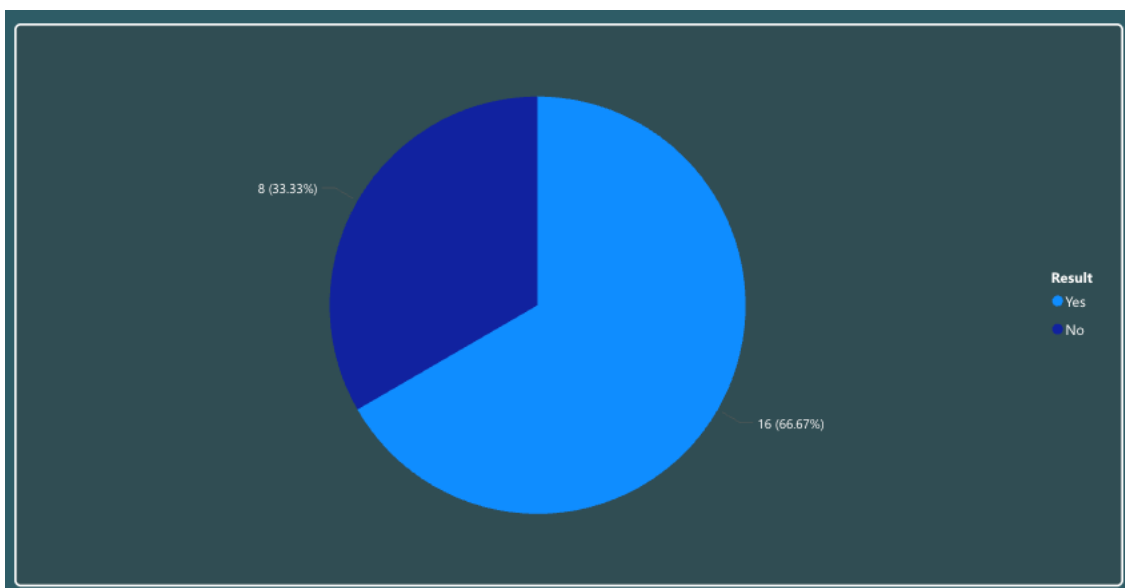


Figure 7: Did the technology accurately guess your age?



## Evaluator Observations

- Manual functional testing confirmed interoperability, robustness and privacy aspects of the system.
- Automated lab testing showed high accuracy for Age Gates 13 and 16, with True Positive Rates consistently above 94% from age 13 upwards and Mean Absolute Error (MAE) values under 2 years for ages 13–20.
- There were very few usability issues arising from Mystery Shopper trials. While response time is not as fast as some systems tested, but users reported that the system met or exceeded their expectations in terms of task time. The majority of users had no issues in completing the evaluation task and reported the system to be either easy or very easy to use.
- 80% of users reported being either satisfied or very satisfied with the experience, and the majority had no concerns about privacy. 66% reported accurate age estimation results, which is consistent with the overall average accuracy of 70% for the mystery shopper estimation task. We note that some users had an expectation of the system estimating their age exactly (to within a few months) and thus reported “no” for the accuracy question without allowing for any margin.
- Overall, we assess the vendor’s Target of Evaluation (ToE) to be at Technology Readiness Level (TRL) 9, in alignment with the vendor’s self-assessment.



## Vendor Comments on Evaluation Results

TBC