SHOW Audit Criteria

Certification scheme for SHOW for Platforms — the SHOW v2.1 Platform Addendum

Published: November 2026
Scheme owner: r8rly.org, governing body for the SHOW Standard
Applies to: independent certification bodies auditing platforms against SHOW for Platforms, and platforms seeking SHOW Certified™
Licence: CC BY-SA 4.0

Contents

1. Purpose
2. Scope
3. Normative references
4. Terms and definitions
5. Roles
6. Certification scheme
7. Requirements for certification bodies
8. Sampling, test accounts and assessor calibration
9. Audit domains and requirements
10. Nonconformities and certification decisions
11. Audit reports, right of reply, regulators and the public register
12. Use of the SHOW Certified™ mark
13. Changes to these criteria

- Annex A — Audit checklist summary
- Annex B — Minimum sample sizes

---

1. Purpose

1.1 These criteria set out how an independent certification body audits a platform against SHOW for Platforms (the SHOW v2.1 Platform Addendum, "the addendum") and decides whether the platform may claim SHOW Certified™.

1.2 They are published in full so that every certification body tests the same way, every platform knows what it will be tested against, and parents, regulators and the public can see what certification means.

1.3 SHOW Certified™ means that, at the time of the audit and within the certified scope, an independent certification body found that the platform's SHOW labels, flags, ceilings and youth account defaults worked as the addendum requires, to the thresholds in these criteria. It does not

mean that content on the platform is "safe". SHOW tells parents and platforms what is in content so that each family can decide.

---

2. Scope

2.1 **Services covered.** These criteria apply to any platform within the scope of the addendum: social media, video-sharing, livestreaming, messaging-with-feeds and any platform that serves user-generated content to minors.

2.2 **Phased scope.** Certification is granted by phase, matching addendum Part 1.7:

Phase	In scope	Domains audited (Section 9)
Phase 1 (core)	Public content surfaces, new uploads, youth accounts, ads, flags (addendum Part 2) and youth account defaults (addendum Part 3)	1, 2, 3, 4, 5, 6, 7, 8, 10
Phase 2	Comments, and back-rating existing content	Phase 1 domains plus 9A and 9B
Phase 3	Livestreams	Phase 1 and Phase 2 domains plus 9C

A platform is certified to Phase 1, to Phases 1–2, or to Phases 1–3. Phase 2 cannot be certified without Phase 1, and Phase 3 cannot be certified without Phases 1 and 2.

2.3 **Out of scope.** Private one-to-one and ephemeral messaging is out of scope and is not audited. The Part 3 youth account settings that govern direct messages (AC-5.1, AC-5.2) are audited as account settings only; message content is never audited. Where a link received through messaging opens content on a platform content surface, the content shown on that surface is in scope (see AC-3.3); the message itself is not.

2.4 **Learning Dividend.** The Learning Dividend (addendum Part 4) is a separate, optional module. It is **not a condition of SHOW Certified™**. Domain 11 applies only to a platform that has opted in, and its findings are reported separately and do not affect the certification decision.

2.5 **What these criteria do not assess.**

- Whether the platform meets its legal obligations in any jurisdiction. Certification does not replace those obligations and does not override any law, including Australia's social media minimum age law.
- The method the platform uses to establish that an account belongs to a young person (age assurance). The audit tests accounts that the platform treats as youth accounts.
- The quality of content, or whether content is suitable for any particular child.

---

3. Normative references

The following documents are indispensable to these criteria. Where a reference is undated, the edition current at the start of the audit applies.

- **The SHOW Standard v2.1** (July 2026), including its axis definitions, level names, WTF bands and Calibration Reference Set.
- **SHOW for Platforms — SHOW v2.1 Platform Addendum** (November 2026).
- **r8rly.org jurisdiction table** for flag age gates (addendum Part 2.1), as published at r8rly.org.
- **ISO/IEC 17065:2012** Conformity assessment — Requirements for bodies certifying products, processes and services.
- **ISO/IEC 17067:2013** Conformity assessment — Fundamentals of product certification and guidelines for product certification schemes.

Informative references:

- r8rly.org Separation Charter (r8rly.org and r8rly.com).
- r8rly.org Conflict of Interest Policy.
- Privacy Act 1988 (Cth), including the Australian Privacy Principles.
- Regulation (EU) 2016/679 (General Data Protection Regulation).
- Regulation (EU) 2022/2065 (Digital Services Act), Article 28.

---

4. Terms and definitions

4.1 SHOW value — the four-axis rating defined in SHOW v2.1: Spice (S, 0–6), Heat (H, 0–5), OMG (O, 0–5) and WTF (W, 0–10), together with the metadata in addendum Part 1.1.

4.2 WTF band — one of the five SHOW v2.1 WTF bands: 0–2 Vanilla, 3–4 Edgy, 5–6 Wild, 7–8 Unhinged, 9–10 Off the Rails.

4.3 In-scope item — any post, video, reel, story, image, ad, comment (Phase 2) or livestream (Phase 3) that the platform serves on a surface within the certified scope.

4.4 Surface — any route by which the platform serves an item to an account, including the feed, search, recommendations, "up next" and autoplay, notifications, profile and hashtag pages, shared links and embeds opened on the platform, ads, and (by phase) comments and livestreams.

4.5 Served — displayed, played or auto-played to an account, including as a thumbnail, preview or snippet that shows the item's content.

4.6 Youth account — an account the platform treats as belonging to a person under 18.

4.7 Age band — one of the four bands in addendum Part 1.5: Under 10, 10–12, 13–15, 16–17.

4.8 Default ceiling — the suggested starting ceiling for an age band, as set out in addendum Part 1.5:

Age	Starting ceiling
Under 10	S0 · H1 · O1 · W2
10–12	S0 · H2 · O1 · W2
13–15	S1 · H3 · O2 · W4
16–17	S1 · H4 · O2 · W4

For 16–17, a parent may raise Spice to 2, OMG to 3 or WTF to 6.

4.9 Ceiling — the SHOW ceiling applying to a youth account at a given time: the default ceiling, or the ceiling the parent or guardian has set.

4.10 Above the ceiling — an item whose SHOW value exceeds the ceiling on **any** axis.

4.11 Parent — the parent or guardian linked to a youth account through the platform's family controls.

4.12 Reference rating — the SHOW value for an item established by the certification body's assessors under Section 8.4. The reference rating is the benchmark against which the platform's label is measured.

4.13 Assessor — a person trained and calibrated by the certification body to rate content against SHOW v2.1 (Section 8.4).

4.14 Exact agreement — for S, H and O, the platform label and reference rating are the same level. For W, they fall in the same WTF band.

4.15 Under-rating / over-rating — the platform label is lower / higher than the reference rating on an axis. For W, measured in WTF bands.

4.16 Ceiling-crossing under-rating — an item whose platform label is within a given default ceiling but whose reference rating is above that ceiling.

4.17 Flag — one of the yes/no tags in addendum Part 2: `self-harm-ed`, `dangerous-challenge`, `bullying`, `gambling`, `scam`.

4.18 Test account — an account created and operated by the certification body for testing. No real child takes part in any test.

4.19 Probe item — an item with a known reference rating or known flag status, placed or identified by the certification body to test serving behaviour.

4.20 Seeded item — a probe item uploaded by the certification body under the safeguards in Section 8.3.

4.21 Leakage — serving to a youth account an item that is above that account's ceiling, or that carries a flag gated for that account.

4.22 Rollout content — in-scope content that falls within a phase the platform has not yet completed (for example, existing content not yet back-rated before Phase 2 is complete).

4.23 Audit window — the period of platform operation from which audit evidence is drawn: the 90 days before the start of Stage 2 for an initial or recertification audit, and the period since the previous audit (not less than 90 days) for a surveillance audit.

4.24 Certification cycle — the three-year period of validity of a certification.

4.25 Relevant regulator — the regulator or regulators with online-safety responsibility for the jurisdictions in the certified scope, as named in the certification agreement.

4.26 Nonconformity — failure to meet a requirement of these criteria. Graded under Section 10.

Unless stated otherwise, "must" denotes a requirement, "should" a recommendation and "may" a permission.

---

5. Roles

5.1 **r8rly.org** is the governing body for the SHOW Standard and the owner of this certification scheme. It publishes the SHOW Standard, the addendum, these audit criteria and the jurisdiction table; licenses use of the SHOW Certified™ mark to independent certification bodies; and keeps a public register of certification status built from certification bodies' notices. **r8rly.org does not audit platforms and makes no certification decision.**

5.2 **Independent certification bodies** take platforms on as clients, run audits at their own cost and risk, carry their own liability and insurance, hold or seek their own accreditation, and make the certification decision.

5.3 **Platforms** apply to a certification body, provide access and evidence under the certification agreement, and pay the certification body for the audit.

5.4 **r8rly.com** has no role in any audit or certification decision, receives no audit report, platform data or confidential audit information, does not publish ratings of certified platforms or of platforms being audited, and does not use SHOW Certified™ status, or any platform's certification, in its own marketing or sales, as set out in the Separation Charter.

5.5 **Relevant regulators** receive the final audit report under Section 11.

---

6. Certification scheme

6.1 Application

6.1.1 A platform applies to a certification body of its choice that holds a licence from r8rly.org to use the SHOW Certified™ mark.

6.1.2 The application must include a **scope statement** setting out:

a) the service(s), apps and client types (web, iOS, Android, smart TV, other) to be certified;
b) the jurisdictions in which the service is offered and, for each, the age bands for which youth accounts are offered;

c) the phase(s) for which certification is sought, with the date each phase was completed;
d) every surface through which in-scope items are served;
e) the rating architecture: creator rating, platform classifier, human review, and the logic that applies the highest credible rating;
f) the label mapping published to r8rly.org under addendum Part 1.6, if any, with its publication date;
g) the family controls through which ceilings are set, and any limit the platform places on the ceilings a parent may set;
h) the platform's band-transition rule for a youth account moving into an older age band;
i) the approximate volume of in-scope items, by content type and jurisdiction, for the audit window;
j) whether the platform has opted in to the Learning Dividend.

6.1.3 A phase may be certified only once the platform has completed it. Content in the scope of a later phase is treated as rollout content (Domain 8).

6.2 Certification agreement

Before the audit begins, the certification body and the platform sign an agreement covering, at minimum: scope; data access (including access to production systems, logs and sampling frames); confidentiality; the data-processing agreement (Section 7.6); the operation of undisclosed test accounts (Section 8.2); the platform's right of reply; delivery of the final report to the relevant regulator; notice of certification status to r8rly.org for the public register; the platform's consent to publication of its status on that register; the rules of use of the mark (Section 12); and a liability cap. Certification is not an admission of liability, and audit findings are given under the agreed audit terms.

6.3 Initial audit

6.3.1 **Stage 1 — readiness review.** The certification body reviews the scope statement, documentation, mapping, rating architecture, family controls and records, confirms the scope and phase, and identifies any issue that would prevent Stage 2. Stage 1 findings are reported to the platform. Stage 2 must begin no later than 90 days after Stage 1, or Stage 1 is repeated.

6.3.2 **Stage 2 — technical audit.** The certification body tests every applicable requirement in Section 9 using the evidence, sampling and test methods specified there and in Section 8. Stage 2 includes:

a) extraction and verification of sampling frames;
b) blind re-rating of the label-accuracy sample;
c) seeded and field flag testing;
d) live test-account testing across age bands, ceilings, jurisdictions and client types;
e) inspection of youth account defaults and family controls;
f) ad testing;
g) review of records, complaints handling and change control.

6.4 Certification decision

6.4.1 The certification decision is made by a person or committee within the certification body who took no part in the audit, applying the decision rules in Section 10.4.

6.4.2 r8rly.org takes no part in the decision. A platform's appeal against a decision is heard by the certification body under its own appeals process.

6.5 Surveillance

6.5.1 A surveillance audit is carried out within 12 months of the initial certification decision and within 12 months of each surveillance audit after that, so that there are two surveillance audits in each three-year cycle.

6.5.2 Each surveillance audit covers, at minimum, Domains 2, 3, 4 and 5 in full at surveillance sample sizes (Annex B), plus every other requirement where a nonconformity was raised at the previous audit or where a change has been notified under AC-10.4. Every requirement in Section 9 must be tested at least once in each certification cycle.

6.5.3 In addition, the certification body conducts at least one **unannounced gate-integrity check** (a subset of Domain 3 using undisclosed test accounts) between each pair of scheduled audits.

6.6 Recertification

A recertification audit, equivalent in scope to Stage 2 of the initial audit, is completed and decided before the end of the certification cycle. If it is not decided before expiry, the certification lapses and the platform is not certified until a decision is made.

6.7 Scope extension

A platform certified to an earlier phase may apply to extend to a later phase. The extension audit covers the new phase's domains in full and Domains 2 and 3 at surveillance sample sizes. The extension does not change the date of the certification cycle.

6.8 Special audits

The certification body may carry out a special audit, at short notice, where it has credible information that the platform no longer meets these criteria, including from complaints, regulator information, public reporting or a notified significant change.

6.9 Suspension, withdrawal and reinstatement

6.9.1 The certification body **suspends** certification where:

a) a pass threshold is found to be failed at a surveillance, special or unannounced audit (Section 10.4.3), subject to Section 10.4.5;
b) a major nonconformity is not closed within the time allowed;
c) the platform refuses or unreasonably delays a surveillance or special audit;
d) the platform fails to notify a significant change under AC-10.4;
e) the platform misuses the mark (Section 12).

6.9.2 During suspension the platform must not use the mark, and the public register shows the platform as not certified.

6.9.3 The certification body **withdraws** certification where a suspension is not resolved within 90 days, where the platform has provided false or misleading evidence, or where the platform requests withdrawal.

6.9.4 A suspended certification is reinstated only after the cause is corrected and re-audited. A platform whose certification has been withdrawn must complete a new initial audit.

6.10 Notice to r8rly.org

The certification body notifies r8rly.org in writing within **5 business days** of every grant, suspension, reinstatement, withdrawal, lapse or scope change. The notice contains only: the platform and service name, the certified phase(s), the status (certified or not certified) and the date of the status. r8rly.org keeps the certified phase(s) in its internal register record and publishes only the platform and service name, the status and the date (Section 11.4). The notice contains no audit findings or platform data.

---

7. Requirements for certification bodies

7.1 Accreditation

7.1.1 A certification body must operate in conformity with ISO/IEC 17065 and must hold, or have applied for and be actively seeking, accreditation to ISO/IEC 17065 for these criteria from an accreditation body that is a signatory to the IAF Multilateral Recognition Arrangement for product certification (for example, JAS-ANZ in Australia and New Zealand).

7.1.2 A certification body that is seeking but does not yet hold accreditation for these criteria must say so in every certification document it issues, and must give the accreditation body access to witness its audits.

7.1.3 A certification body must hold a current licence from r8rly.org to use the SHOW Certified™ mark. r8rly.org may end the licence of a certification body that does not apply these criteria.

7.2 Liability and insurance

The certification body carries its own liability and holds professional indemnity and public liability insurance adequate to the scope of its certification activities.

7.3 Impartiality

7.3.1 The certification body must identify, analyse and document risks to its impartiality for each platform client, and maintain a mechanism for safeguarding impartiality as required by ISO/IEC 17065.

7.3.2 The certification body, and any entity under common ownership or control with it, must not provide consultancy, implementation, design, classifier training, rating or moderation services to a platform it certifies, or to any entity under common control with that platform, for **24 months** before the start of the audit and for the duration of the certification.

7.3.3 No person may audit, assess or decide on a platform where, in the previous **24 months**, that person was employed by, contracted to or provided consultancy to the platform.

7.3.4 No part of the certification body's remuneration, or of any individual's remuneration, may depend on whether a platform is certified.

7.3.5 The certification body must not be owned or controlled by r8rly.com, by r8rly.org, or by any platform. No person acting for r8rly.org or r8rly.com may take part in an audit or certification decision.

7.4 Conflict-of-interest declarations

7.4.1 Every auditor, assessor, technical expert and decision-maker must declare in writing any interest in, or relationship with, the platform, its competitors, r8rly.org or r8rly.com: on appointment, at least once a year, and before each engagement.

7.4.2 A person with a declared conflict takes no part in the affected audit or decision. The certification body keeps a register of declarations and how each was handled, and makes it available to its accreditation body.

7.5 Competence

7.5.1 Each audit team must include, between its members:

a) a **lead auditor** with demonstrated competence in product or service certification auditing under ISO/IEC 17065;
b) a **technical expert** with demonstrated competence in content-serving, recommendation and advertising systems at scale;
c) a person competent in **statistical sampling and estimation**;
d) a person with competence in **child safety or child development**;
e) **assessors** meeting Section 8.4, with fluency in the languages of the sampled content.

7.5.2 The certification body must maintain records of the training, qualifications, calibration results and experience of every member of an audit team.

7.5.3 The certification body must provide assessors exposed to distressing content with welfare support, including limits on continuous exposure and access to counselling.

7.6 Data handling and privacy

7.6.1 Platform data handled during an audit must be covered by a data-processing agreement that complies with the Privacy Act 1988 (Cth), including the Australian Privacy Principles, with the GDPR for data relating to users in the EU, and with any other privacy law applying to the data.

7.6.2 The certification body must collect only the data it needs to test these criteria. Sampling frames are limited to the fields listed in Section 8.1.1: item identifier, SHOW value, `rated_by`, `rated_at`, flags, content type, jurisdiction of upload, language and reach.

7.6.3 **No individual tracking of children.** The audit must not identify, track or profile any real child user. Gate, default and parent-control testing is carried out with test accounts only. Any analysis of production data relating to youth accounts uses aggregated or de-identified data.

7.6.4 Audit data must be held in a secure environment with access limited to the audit team and decision-makers. Personal information must be deleted or de-identified within **90 days** of the final report, except where the certification body is required by law or by its accreditation to keep it. Audit records that contain no personal information are retained for the current and previous certification cycle.

7.6.5 The certification body must not share audit data with r8rly.org or r8rly.com, except the status notice in Section 6.10.

---

8. Sampling, test accounts and assessor calibration

8.1 Sampling frames

8.1.1 The certification body obtains from the platform a complete list of in-scope items for the audit window (the sampling frame), with item identifier, SHOW value, `rated_by`, `rated_at`, flags, content type, jurisdiction of upload, language and reach (impressions in the window).

8.1.2 The certification body verifies the frame's completeness by reconciling item counts against the platform's own operational totals for the window and by checking that a set of items the certification body has independently observed on the platform (at least 200) appear in the frame. The discrepancy is the larger of (a) the unexplained difference between the frame's item count and the platform's operational total, as a percentage of that total, and (b) the percentage of independently observed items missing from the frame without explanation. A discrepancy above 0.5% is a major nonconformity under AC-10.2.

8.1.3 The certification body draws all samples itself, using a random seed it does not disclose to the platform before selection.

8.2 Test accounts

8.2.1 Test accounts are created through the platform's ordinary sign-up and family-control flows, as a family would create them. The certification agreement authorises their creation. **The platform is not told which accounts are test accounts until testing is complete**, after which identifiers are disclosed so the platform can provide serving logs.

8.2.2 For each age band the platform offers in a jurisdiction in scope, the certification body operates at minimum:

Configuration	Accounts per age band
Default ceiling	4
Parent-lowered ceiling (at least one axis lowered)	2
Parent-raised ceiling (at least one axis raised; for 16–17, one all-raises account at S2 · H4 · O3 · W6)	2

Accounts are spread across every client type in scope and across the jurisdictions in scope (at least the three with the largest youth-account volume, or all if fewer). Each parent account linked to a test youth account is also a test account. Adult test accounts are used for scam and gambling testing (Domain 4). Each single-axis raise for 16–17 (Spice to 2, OMG to 3, WTF to 6) is tested separately under AC-6.3. At surveillance, the 4 accounts per age band are split 2 default, 1 parent-lowered and 1 parent-raised (for 16–17, the all-raises account).

8.2.3 **Scripted use.** Each test youth account is used for at least **7 days** and accumulates at least **1,000 served items**, following a script that exercises every surface in scope: scrolling the feed, running a fixed list of search queries (including queries likely to surface content above the ceiling), following accounts, watching through autoplay and "up next", opening shared links and embeds, viewing profile and hashtag pages, receiving notifications and viewing ads. Scripts must include behaviour signalling interest in borderline content, to test that recommendations do not drift above the ceiling.

8.2.4 Every served item is recorded by screen capture with timestamp and matched to its item identifier and SHOW metadata.

8.2.5 The sample sizes, test-account numbers, durations and probe numbers in Sections 8 and 9 are for initial and recertification audits. At surveillance audits, the surveillance figures in Annex B apply and prevail.

8.3 Seeded items and safeguards

8.3.1 Seeded items are uploaded by certification-body test creator accounts with visibility restricted to test accounts, where the platform supports this, or otherwise identified and removed immediately after the test. Seeded items above any youth ceiling must never be made available to real users.

8.3.2 Seeded test sets must never contain child sexual abuse material or any material that is unlawful to possess or distribute. Items depicting real minors in any sexualised context are prohibited.

8.3.3 Self-harm and disordered-eating test items are drawn from material already held under the platform's safety review processes, or are staged items designed with advice from a clinician with relevant expertise. They are used only in restricted-visibility tests.

8.3.4 Where the platform supplies a test environment that runs the production classifier and serving logic, the certification body may use it for seeded flag recall testing (Domain 4) only after verifying, on at least 50 items, that the test environment produces the same flags and SHOW values as production. Gate-integrity testing (Domain 3) is always carried out in production.

8.3.5 Seeded items that contain links, including scam and gambling probes, use only inert links or links to destinations controlled by the certification body for testing. No seeded item links to a real gambling, betting or scam site, or to any live payment or credential-collection page.

8.4 Reference ratings and assessor calibration

8.4.1 **Blind double rating.** Each sampled item is rated by two assessors working independently. Assessors do not see the platform label, the creator rating, the classifier output, the item's reach or

each other's rating. Assessors rate the item as served, applying the SHOW v2.1 level definitions and the addendum's short-form rules: the peak rule (rate the strongest moment), audio counts (lyrics, sound effects, voiceover), clips inherit the source work's rating unless re-rated by a human reviewer, and coarse language is recorded in content-warning tags, not on the four axes.

8.4.2 **Adjudication.** Where the two assessors do not reach exact agreement on an axis, a senior assessor who has not seen either rating rates the item, and the median of the three ratings on that axis is the reference rating. For W, the median value is taken and its band used.

8.4.3 **Initial calibration.** Before rating audit samples, each assessor must:

a) rate the twelve works in the SHOW v2.1 Calibration Reference Set, achieving exact agreement (W: same band) on each axis for at least 10 of the 12 works; and
b) rate the certification body's **platform calibration set** — at least 200 short-form platform items covering every content type, every level of each axis and every flag, with reference ratings established by adjudication — achieving exact agreement of at least **85%** on each axis (W: same band) and a quadratic-weighted kappa of at least **0.80** on each axis.

8.4.4 **Recalibration.** Each assessor recalibrates against the platform calibration set at least every **6 months** and before each audit in which they rate.

8.4.5 **Drift control.** At least **5%** of the items in every rating batch are calibration items inserted without the assessor's knowledge. An assessor whose exact agreement on these items falls below **85%** on any axis in a batch has that batch re-rated by another assessor.

8.4.6 **Inter-assessor reliability.** The certification body reports inter-assessor quadratic-weighted kappa per axis for each audit. If it falls below **0.75** on any axis, the certification body must investigate and recalibrate before the reference ratings for that audit are used.

8.4.7 The certification body reviews its platform calibration set at least once a year to reflect current platform content, including current coded language.

8.5 Estimation and reporting

8.5.1 Every metric is reported with its point estimate, sample size and 95% confidence interval: Wilson score intervals for proportions, and bootstrap intervals (at least 1,000 resamples) for kappa, harm-weighted error and bias.

8.5.2 Pass thresholds in these criteria apply to the point estimate, at or above the minimum sample sizes in Annex B. Where a sample falls short of the minimum, the certification body draws a top-up sample; a threshold cannot be judged on an undersized sample.

8.5.3 Label-accuracy results are reported both **item-weighted** (each item counts once, after weighting back to the population for any stratum oversampling) and **impression-weighted** (each item weighted by its reach in the audit window). Pass thresholds apply to both.

---

9. Audit domains and requirements

Each requirement states what is required, the evidence, the test method, the sampling, the metric, the pass threshold and the grading on failure. "Major" and "minor" are defined in Section 10. Where the grading says "major (threshold)", failure is a threshold failure under Section 10.4. The sample sizes and test figures in this section are for initial and recertification audits; at surveillance audits the surveillance figures in Annex B apply and prevail (Section 8.2.5).

Domain 1 — Content field and labelling

AC-1.1 Presence of SHOW values

- **Requirement.** Every in-scope item carries a SHOW value (addendum Part 1.1). In Phase 1 this applies to every item uploaded after the Phase 1 completion date and to every ad.
- **Evidence.** Sampling frame; item metadata for sampled items.
- **Test method.** Inspect metadata for a random sample drawn from the frame, and for all items served to test accounts.
- **Sampling.** Annex B, presence sample.
- **Metric.** Share of items carrying a SHOW value.
- **Pass threshold.** At least **99.5%** of in-scope non-ad items; at least **99.5%** of ads.
- **Grading.** Major (threshold).

AC-1.2 Field format

- **Requirement.** The SHOW field conforms to addendum Part 1.1: `version` is "2.1"; `s` is an integer 0–6; `h` 0–5; `o` 0–5; `w` 0–10; `rated_by` is one of `creator`, `platform`, `reviewer`; `rated_at` is an ISO 8601 timestamp with a UTC offset; `flags` is an array containing only `self-harm-ed`, `dangerous-challenge`, `bullying`, `gambling` or `scam`.
- **Evidence.** Field schema; sampled metadata.
- **Test method.** Automated validation of the presence sample.
- **Metric.** Share of present fields that are valid.
- **Pass threshold.** At least **99.5%** valid.
- **Grading.** Major (threshold).

AC-1.3 Display and web metadata

- **Requirement.** Wherever the platform displays a SHOW value to users, it uses the inline format `SHOW S1·H0·O2·W3`. On web pages that serve an in-scope item, the rating is also published as `<meta name="show-rating" content="S1·H0·O2·W3; v=2.1">`.
- **Evidence.** Screen captures; page source.
- **Test method.** Inspect at least 100 item displays across client types and at least 100 web pages.
- **Metric.** Share conforming.
- **Pass threshold.** At least **99%** conforming.
- **Grading.** Minor; major if the meta field is absent from web pages altogether.

AC-1.4 Highest credible rating

- **Requirement.** Ratings are set in three layers (creator, platform classifier, reviewer), and the highest credible rating wins. Where the classifier rates higher than the creator on an axis, the classifier's rating stands (addendum Part 1.2).
- **Evidence.** Rating architecture documentation; rating event logs.
- **Test method.** Upload at least 50 seeded items with creator ratings deliberately set below the content's reference rating on at least one axis. Compare the final label with the classifier output (from logs) and the creator rating.
- **Metric.** Share of seeded items whose final label, on every axis, is not lower than the classifier output.
- **Pass threshold. 100%**.
- **Grading.** Major.

AC-1.5 Edits re-rate

- **Requirement.** Any edit to content triggers a fresh rating before the edited version is served to youth accounts.
- **Test method.** Make at least 30 edits to seeded items, including edits that raise content on one axis. Check `rated_at` and the label, and whether the edited version reaches youth test accounts before re-rating.
- **Metric.** Share of edits re-rated before the edited version is served to any youth account.
- **Pass threshold. 100%**.
- **Grading.** Major.

AC-1.6 Clips inherit

- **Requirement.** Where the platform identifies an item as a clip from a film, series, game or book trailer that has a SHOW rating, the clip carries the source work's rating unless a human reviewer re-rates it (`rated_by: reviewer`).
- **Evidence.** Documentation of how source works are identified; label logs.
- **Test method.** Upload at least 30 seeded clips from SHOW-rated source works that the platform's content matching identifies.
- **Metric.** Share carrying the source rating, or a reviewer re-rating.
- **Pass threshold.** At least **95%**.
- **Grading.** Minor.

AC-1.7 Mapping of existing labels

- **Requirement.** Where the platform maps existing content labels and maturity settings onto SHOW values (addendum Part 1.6), the mapping in production is the mapping published to r8rly.org; it maps every existing label to a SHOW value on all four axes; and it does not alter SHOW levels.
- **Evidence.** Published mapping; production configuration; change history.
- **Test method.** Compare the production mapping with the published mapping. Trace at least 50 items carrying each mapped label to confirm the mapped value is applied.

- **Metric.** Correspondence of production to published mapping; mapped-label stratum accuracy (AC-2.1 to AC-2.5).
- **Pass threshold.** Production mapping identical to the published mapping; the mapped-label stratum meets every Domain 2 threshold.
- **Grading.** Major (threshold).

Domain 2 — Label accuracy

Domain 2 measures how closely the platform's SHOW labels match reference ratings. Because the harm of a label lies in under-rating, under-rating is measured separately and weighted more heavily than over-rating. Over-rating is also measured, so that blanket high rating is not rewarded.

Sample. A stratified random sample of rated in-scope items from the audit window (Annex B), stratified by:

- **content type** — long video, short-form video, image, text post, public story, ad (and, by phase, comment and livestream recording);
- **jurisdiction** — each jurisdiction in scope that accounts for at least 5% of in-scope items, with at least 150 items each; smaller jurisdictions pooled;
- **reach** — top 1% of items by impressions, next 9%, remaining 90%, with equal allocation across the three tiers;
- **rater layer** — `creator`, `platform`, `reviewer`, and mapped labels where AC-1.7 applies.

At least half of the sample is drawn from items whose platform label is within the 16–17 default ceiling, since these are the items that can reach youth accounts. The sample must include at least **600** items whose label is within the Under 10 default ceiling and at least **600** items whose label is within the 16–17 default ceiling but outside the 10–12 default ceiling. Oversampled strata are weighted back for population estimates. Items in a language for which no calibrated assessor is available may be excluded only up to 2% of the sample, and the exclusion is reported.

AC-2.1 Exact agreement

- **Requirement.** Platform labels agree with reference ratings.
- **Test method.** Blind double rating with adjudication (Section 8.4).
- **Metric.** Exact agreement per axis (W: same band), item-weighted and impression-weighted.
- **Pass threshold.** At least **80%** on each axis.
- **Grading.** Major (threshold).
- **Justification.** S, H and O have a distinct named level at every point, and ceilings turn on single levels, so exact agreement on those axes is required. SHOW v2.1 defines W by band, with one label and description per band and default ceilings at the top of a band, so band agreement is the meaningful test for W. Exact W values are reported but not thresholded.

AC-2.2 Agreement within one step

- **Metric.** Share of items within one level (W: within one band) of the reference rating, per axis.
- **Pass threshold.** At least **97%** on each axis.

- **Grading.** Major (threshold).

AC-2.3 Weighted kappa

- **Metric.** Quadratic-weighted Cohen's kappa between platform labels and reference ratings, per axis (W computed on bands).
- **Pass threshold.** At least **0.75** on each axis.
- **Grading.** Major (threshold).

AC-2.4 Harm-weighted error and bias

- **Metric 1 — harm-weighted error (HWE).** Per axis: HWE = (3 × sum of under-rating differences + 1 × sum of over-rating differences) ÷ number of items, with differences in levels (W: bands).
- **Pass threshold.** HWE at most **0.50** on each axis.
- **Metric 2 — directional bias.** Mean signed difference (platform label minus reference rating) per axis, in levels (W: bands).
- **Pass threshold.** Not lower than **–0.10** on any axis (no systematic under-rating).
- **Grading.** Major (threshold).

AC-2.5 Ceiling-crossing under-rating

- **Requirement.** Items labelled within a youth ceiling are not in fact above it.
- **Metric.** For each default ceiling, the share of sampled items labelled within that ceiling whose reference rating is above it.
- **Pass threshold.** At most **1.0%** for each of the four default ceilings, item-weighted and impression-weighted. In addition, no more than **0.2%** of items labelled within any default ceiling may have a reference rating two or more steps above it on any axis (a step is one level on S, H and O, and one WTF band on W).
- **Grading.** Major (threshold).

AC-2.6 Stratum results

- **Metric.** AC-2.1 and AC-2.5 reported for each stratum.
- **Pass threshold.** Exact agreement of at least **70%** per axis in every content-type, jurisdiction and rater-layer stratum of at least 150 items (at surveillance, the reviewer stratum is judged at 100 items), except that the mapped-label stratum (AC-1.7), the ads stratum (AC-7.1) and the back-rated stratum (AC-9.5) must each meet every Domain 2 threshold.
- **Grading.** Minor for each stratum below the 70% threshold; major (threshold) if the mapped-label, ads or back-rated stratum fails any Domain 2 threshold.

Domain 3 — Ceiling enforcement and gate integrity

AC-3.1 Above-ceiling probes not served

- **Requirement.** Content above a youth account's ceiling on any axis is not served on any surface (addendum Part 1.5).
- **Evidence.** Test-account screen captures; serving logs after disclosure.

- **Test method.** For each test-account configuration, place at least **40 above-ceiling probe items**, including for every axis an item exceeding the ceiling on that axis alone by one level (W: one band), so that the "any axis" rule is tested axis by axis. Make each probe reachable by every surface route: a unique search term, a followed account, a direct link, a shared link, an embed and, for ads, a test campaign (AC-7.2).
- **Metric.** Number of above-ceiling probe items served.
- **Counting.** Only probe items whose final platform label is above the account's ceiling, confirmed in the platform's rating logs, count towards this metric. A probe whose final label is within the ceiling is recorded as a labelling finding under Domain 2.
- **Exclusion.** Rollout content handled under AC-8.2 is excluded from this requirement and reported separately.
- **Pass threshold. Zero**.
- **Grading.** Major (threshold).

AC-3.2 Within-ceiling probes served

- **Requirement.** The ceiling is not applied more strictly than set, so that the parent's choice is respected and blanket blocking is not rewarded.
- **Test method.** For each configuration, place at least **20 probe items at or within the ceiling**, including items exactly at the ceiling on each axis, reachable by the same routes.
- **Metric.** Share reachable by the intended route.
- **Pass threshold.** At least **95%**.
- **Grading.** Major (threshold).

AC-3.3 Label-based leakage in live use

- **Requirement.** In ordinary scripted use (Section 8.2.3), no item whose own platform label is above the account's ceiling is served, on any surface: feed, search, recommendations, "up next" and autoplay, notifications, profile and hashtag pages, shared links and embeds opened on a platform content surface, ads, and (by phase) comments and livestreams.
- **Metric.** Share of served items whose platform label is above the ceiling, per age band and per surface.
- **Exclusion.** Rollout content handled under AC-8.2 is excluded from this requirement and reported separately.
- **Pass threshold.** At most **0.05%** of served items in each age band. The 0.05% tolerance recognises sampling of live recommendation behaviour; any instance is still recorded as a nonconformity. Any instance on an Under 10 or 10–12 account, and any instance of an item labelled S3 or higher or O4 or higher on any youth account, is a major nonconformity regardless of rate (zero-tolerance instances).
- **Grading.** Major (threshold); an instance within the tolerance that is not a zero-tolerance instance is a minor nonconformity.

AC-3.4 Reference leakage in live use

- **Requirement.** Content actually served to youth accounts is within their ceiling, combining label accuracy and gate integrity.

- **Test method.** Draw a random sample of at least **400 served items per age band** from test-account captures and have them reference-rated (Section 8.4).
- **Metric.** Share whose reference rating is above the account's ceiling.
- **Exclusion.** Rollout content handled under AC-8.2 is excluded from this requirement and reported separately.
- **Pass threshold.** At most **1.0%** per age band.
- **Grading.** Major (threshold).

AC-3.5 Ceiling changes take effect

- **Requirement.** A change to the ceiling made by the parent is enforced on every surface.
- **Test method.** For each age band, make at least **10** ceiling changes (5 lowering, 5 raising), then test with probes on every surface.
- **Metric.** Time from change to enforcement on every surface.
- **Pass threshold.** A lowered ceiling enforced within **15 minutes** in every test; a raised ceiling within **15 minutes** in at least 90% of tests.
- **Grading.** Major for lowering; minor for raising.

AC-3.6 Client and jurisdiction consistency

- **Requirement.** Gate behaviour is the same on every client type and in every jurisdiction in scope.
- **Metric.** AC-3.1 and AC-3.2 results, and zero-tolerance instances under AC-3.3, for each client type and jurisdiction tested. The rates in AC-3.3 and AC-3.4 are judged on the pooled sample and reported by client type and jurisdiction without separate thresholds.
- **Pass threshold.** AC-3.1 and AC-3.2 thresholds met, and no AC-3.3 zero-tolerance instance, on each client type and in each jurisdiction tested.
- **Grading.** Major (threshold).

Domain 4 — Flags

Flag gates (addendum Part 2):

Flag	Default gate
`self-harm-ed` — self-harm or eating-disorder promotion	Hidden from all youth accounts; support resources shown instead
`dangerous-challenge`	Hidden under 18
`bullying` — harassment or pile-on	Hidden from youth accounts and held for review
`gambling`	Legal gambling age of the viewer's country
`scam`	Blocked for all ages

Domain 4 measures missed flags (under-blocking) and false flags (over-blocking) for each flag, so that hiding everything is not rewarded.

AC-4.1 Detection recall on seeded positives

- **Test method.** For each flag, run a seeded set of known positives through the production pipeline (or a verified test environment, Section 8.3.4). At least **25%** of positives use coded language or evasion techniques observed in the 6 months before the audit. Positives cover images, video frames, captions and audio, as the addendum's "detected from" column describes.
- **Sampling.** At least **200 positives per flag** (initial and recertification); 100 per flag (surveillance).
- **Metric.** Recall: share of positives flagged.
- **Pass threshold.** See table below.
- **Grading.** Major (threshold).

AC-4.2 False flags on seeded negatives

- **Test method.** For each flag, run a seeded set of negatives, of which at least half are **hard negatives**: content close to the flag but not within it — recovery and support content, news reporting about a challenge, harm-prevention material about gambling, anti-bullying content, and legitimate commercial links.
- **Sampling.** At least **200 negatives per flag** (initial and recertification), at least 100 of them hard negatives; 100 and 50 respectively at surveillance.
- **Metric.** False-flag rate: share of negatives flagged.
- **Pass threshold.** See table below.
- **Grading.** Major (threshold).

AC-4.3 Field precision

- **Test method.** Draw a random sample of production items carrying each flag in the audit window and have assessors confirm or reject the flag.
- **Sampling.** At least **200 items per flag** (initial and recertification); 100 per flag (surveillance); or all flagged items if fewer.
- **Metric.** Precision: share of flagged items confirmed.
- **Pass threshold.** See table below.
- **Grading.** Major (threshold).

Flag	Recall (AC-4.1) at least	False-flag rate (AC-4.2) at most	Field precision (AC-4.3) at least
`self-harm-ed`	95%	10%	80%
`dangerous-challenge`	90%	10%	80%
`bullying`	85%	15%	75%
`gambling`	95%	10%	85%
`scam`	95%	5%	85%

The scam false-flag ceiling is tighter because a scam flag blocks content for all ages, including adults.

AC-4.4 Field missed-flag check

- **Test method.** Draw a risk-enriched sample of unflagged production items (items reported by users, items matching current code-word and link lists, items from accounts previously actioned) together with a random sample of unflagged items, and have assessors check for missed flags.
- **Sampling.** At least **1,000** enriched and **1,000** random unflagged items (initial and recertification); 500 and 500 (surveillance).
- **Metric.** Missed-flag rate per flag, reported by sample type.
- **Pass threshold.** In the random sample, no more than **one** missed `self-harm-ed` item and no more than **one** missed `scam` item; two or more misses of either flag fails the threshold. At surveillance, no missed item of either flag is permitted, or one miss only after the random sample is extended to 1,000 under the Annex B note. The Wilson 95% upper bound on each missed-flag rate is reported. Enriched-sample results are reported; a recurring pattern of missed coded language not addressed in the platform's detection lists is a minor nonconformity.
- **Grading.** Major (threshold) for the random-sample threshold; minor as stated.

AC-4.5 Timing of flags

- **Requirement.** Flags are applied automatically on upload (addendum Part 2).
- **Metric.** For seeded positives detected under AC-4.1: share flagged before first availability to any youth account (for `scam`, before first availability to any account). For `bullying`: time from the triggering signal (comment pattern, reports or reply velocity) to the item being hidden from youth accounts and held for review.
- **Pass threshold.** At least **99%** flagged before first availability to any youth account for `self-harm-ed`, `dangerous-challenge` and `gambling`, and before first availability to any account for `scam`; for `bullying`, hidden from youth accounts and held within **60 minutes** in at least 95% of tests.
- **Grading.** Major (threshold).

AC-4.6 Flagged content hidden from youth accounts

- **Test method.** Place flagged probe items (at least **10 per flag** per age band offered) reachable through every surface route, and test with youth test accounts.
- **Metric.** Number of flagged probe items served to a youth account for which the flag's gate applies.
- **Counting.** Only probe items carrying the gated flag, confirmed in the platform's rating logs, count towards this metric. A probe the platform did not flag is recorded as a detection finding under AC-4.1.
- **Pass threshold. Zero**.
- **Grading.** Major (threshold).

AC-4.7 Support resources

- **Requirement.** Where `self-harm-ed` content is hidden, support resources are shown instead.

- **Test method.** From youth test accounts in each jurisdiction, open links to `self-harm-ed` probe items and run at least 20 search queries that include current code words.
- **Metric.** Share of attempts that display support resources relevant to the jurisdiction.
- **Pass threshold.** At least **95%**.
- **Grading.** Major (threshold).

AC-4.8 Gambling gates by jurisdiction

- **Requirement.** The `gambling` gate applies the legal gambling age for the viewer's country or state, read from the r8rly.org jurisdiction table, which currently includes AU 18, NZ 18, GB 18 and US-NV 21. Platforms may apply stricter local rules. The platform implements any change to the reference table within **30 days** of its publication. For a jurisdiction in scope that is not in the reference table, the platform applies the legal gambling age of that jurisdiction and documents it, with its legal source, in the scope statement; tests in that jurisdiction use the documented age.
- **Evidence.** Production jurisdiction configuration; change records.
- **Test method.** Compare the production configuration with the reference table. In each jurisdiction in scope, test with accounts one year below and at the gate age (for example, 17 and 18 in AU; 20 and 21 in US-NV), using at least 10 gambling probe items each.
- **Metric.** Configuration match; probes served to below-age accounts.
- **Pass threshold.** Configuration matches the reference table or is stricter; **zero** gambling probes served to below-age accounts.
- **Grading.** Major (threshold).

AC-4.9 Scam blocking for all ages

- **Test method.** Place at least 20 scam probe items and test with adult and youth test accounts.
- **Metric.** Scam probes served to any account.
- **Pass threshold. Zero**.
- **Grading.** Major (threshold).

AC-4.10 Dangerous challenges hidden under 18

- **Test method.** Test `dangerous-challenge` probe items with youth test accounts in every age band offered, including 16–17 accounts with the maximum parent-raised ceiling.
- **Metric.** Probes served to any account under 18.
- **Pass threshold. Zero**.
- **Grading.** Major (threshold).

AC-4.11 Flags independent of the ceiling

- **Requirement.** Flag gates apply whatever ceiling is set. Raising the ceiling does not unlock flagged content.
- **Test method.** Covered by AC-4.6 and AC-4.10 using parent-raised configurations.
- **Grading.** Major.

Domain 5 — Youth account defaults

AC-5.1 Defaults on at account creation

- **Requirement.** Every youth account defaults to (addendum Part 3):
  a) no direct messages from adults the child is not connected to, unless the parent allows it;
  b) no appearing in "people you may know" suggestions to unconnected adults;
  c) location sharing off;
  d) livestreaming off for under-13s; for 13–17, streams viewable only by connections unless the parent allows otherwise;
  e) the SHOW ceiling for the child's age (addendum Part 1.5) until the parent changes it.
- **Default ceiling.** The default ceiling is the ceiling in Section 4.8 for the account's age band. A lower default is allowed only if it is documented in the scope statement and shown to parents in the family controls; tests then use the documented default.
- **Features not offered.** Where the platform does not offer a feature to youth accounts (for example, direct messages, location sharing or livestreaming), the corresponding default is recorded as not applicable, and the certification body verifies that the feature is not available to youth accounts.
- **Evidence.** Settings screens; configuration; test results.
- **Test method.** Create new youth test accounts in every age band offered, in every jurisdiction and on every client type in scope. Inspect each setting, and test each functionally: an unconnected adult test account attempts to message the youth account; the youth account is checked for in suggestions shown to unconnected adult test accounts; location is checked on posts and profile; livestream start and audience are tested; the ceiling is tested with probes.
- **Metric.** Share of accounts on which every default is present and effective.
- **Pass threshold. 100%**.
- **Grading.** Major (threshold).

AC-5.2 Changes to defaults follow the addendum

- **Requirement.** Direct messages from unconnected adults are enabled only with the parent's permission. For 13–17, a wider livestream audience is enabled only with the parent's permission. Livestreaming stays off for under-13s. The youth account holder cannot raise their own ceiling.
- **Test method.** From each youth test account, attempt to change each of these settings without the parent; then make each change through the parent account.
- **Metric.** Changes possible without the parent; changes effective with the parent.
- **Pass threshold. No** change possible without the parent; every permitted change effective when made by the parent.
- **Grading.** Major (threshold).

AC-5.3 Defaults survive updates and band transitions

- **Requirement.** Defaults are not reset or weakened by app updates, settings migrations or a youth account moving into an older age band, and the platform applies its documented band-transition rule (Section 6.1.2 h).
- **Evidence.** Change records; release notes; band-transition rule.

- **Test method.** Re-inspect test accounts after any app update in the audit period; inspect test accounts configured at the boundary of a band.
- **Pass threshold.** No weakening of a default without the parent's action; band transitions follow the documented rule.
- **Grading.** Major.

Domain 6 — Parent controls

AC-6.1 Ceiling set by the parent

- **Requirement.** The ceiling is set by the parent or guardian through the platform's family controls. Where no parent is linked, the default ceiling for the age band applies.
- **Test method.** Inspect family controls; test with linked and unlinked youth test accounts.
- **Pass threshold.** Requirement met on every client type offering family controls.
- **Grading.** Major.

AC-6.2 Each axis adjustable at any time

- **Requirement.** The parent can raise or lower any axis at any time, each axis independently, subject only to limits documented under AC-6.4.
- **Test method.** From parent test accounts, change each axis independently in both directions; verify with probes (AC-3.5). For age bands other than 16–17, raises are tested by raising each axis one level (W: one band) above the default ceiling, where the documented limits permit; for 16–17, raises are tested under AC-6.3.
- **Pass threshold.** Every axis adjustable in both directions, independently.
- **Grading.** Major.

AC-6.3 Raises for 16–17

- **Requirement.** For 16–17, the parent can raise Spice to 2, OMG to 3 or WTF to 6.
- **Test method.** Make each raise singly on a 16–17 test account, and all three together on the all-raises account (Section 8.2.2), and verify with probes.
- **Pass threshold.** Each raise available and effective.
- **Grading.** Major.

AC-6.4 Documented limits

- **Requirement.** Any limit the platform places on the ceiling a parent may set is documented in the scope statement and shown to parents in the family controls, and does not prevent the raises in AC-6.3.
- **Test method.** Compare the documented limits with the controls.
- **Grading.** Minor.

AC-6.5 The ceiling set is the ceiling enforced

- **Requirement.** The ceiling shown to the parent in the family controls is the ceiling applied by the serving logic.

- **Test method.** For every parent-adjusted test account, compare the displayed ceiling with serving behaviour (Domain 3).
- **Pass threshold.** Match in every case.
- **Grading.** Major.

AC-6.6 Visibility to the young person *(observation only)*

- **Good practice.** Consistent with the child's right to have their views given due weight (UN Convention on the Rights of the Child, Article 12), the platform should show the youth account holder their current ceiling and tell them when it changes.
- **Test method.** Inspect youth test accounts.
- **Grading.** Observation only; not grounds for a nonconformity.

Domain 7 — Ads

AC-7.1 Ads carry SHOW values

- Covered by AC-1.1 (at least **99.5%** of ads carry a SHOW value) and AC-1.2. Ads form a content-type stratum in Domain 2 of at least **300 ads**, which must meet every Domain 2 threshold.
- **Grading.** Major (threshold).

AC-7.2 Ads respect the ceiling

- **Requirement.** An ad may not be served to anyone whose ceiling it exceeds (addendum Part 1.3).
- **Test method.** From certification-body test advertiser accounts, run test campaigns of ads with known reference ratings above each default ceiling, targeted by country and broad age range only. Monitor delivery to youth test accounts. Ads served to test accounts during scripted use are included in AC-3.3 and AC-3.4.
- **Metric.** Above-ceiling test ads served to youth test accounts.
- **Pass threshold. Zero**.
- **Grading.** Major (threshold).

AC-7.3 No profiling-based ads to youth accounts

- **Requirement.** Where the platform treats an account as belonging to a person under 18, ads are not selected by profiling that account holder's personal data (cf. Article 28(2), Digital Services Act).
- **Evidence.** Ad-serving configuration for youth accounts; targeting logs for youth test accounts (after disclosure).
- **Test method.** Inspect configuration. Run test campaigns targeted by interest, behaviour, lookalike and custom audiences built to include youth test accounts, and monitor delivery. Inspect targeting logs for every ad served to youth test accounts.
- **Metric.** Profiling-targeted ads served to youth accounts.
- **Pass threshold. Zero**.
- **Grading.** Major (threshold).

Domain 8 — Unrated content and rollout content

AC-8.1 Unrated items within a completed phase are hidden

- **Requirement.** Once a phase is complete, unrated content in that phase's scope is hidden from youth accounts (addendum Part 1.3). For Phase 1, this covers unrated items uploaded after the Phase 1 completion date and unrated ads.
- **Test method.** Identify unrated in-scope items from the sampling frame and, where the certification body can create them, unrated seeded items (for example, uploads made while a rating service is unavailable, where the platform's test procedures allow). Test reachability from youth test accounts. Record any unrated item served during scripted use.
- **Sampling.** At least **50** unrated probe items, or all available if fewer.
- **Limitation.** If fewer unrated probe items are available than the sample size in Annex B, the certification body records this as an audit limitation in the report, and the platform must demonstrate, through rating-service failure logs and serving configuration, that unrated in-scope items are not served to youth accounts.
- **Metric.** Unrated in-scope items served to youth accounts.
- **Pass threshold. Zero** probes served; unrated in-scope items at most **0.05%** of items served in scripted use. The 0.05% tolerance recognises sampling of live recommendation behaviour; any instance is still recorded as a nonconformity.
- **Grading.** Major (threshold); an instance within the tolerance is a minor nonconformity.

AC-8.2 Rollout content falls back to teen settings

- **Requirement.** During the phased rollout, unrated rollout content (for example, existing content not yet back-rated, or comments before Phase 2 is complete) falls back to the platform's existing teen content settings.
- **Evidence.** Documentation of the existing teen content settings; serving configuration.
- **Test method.** Test at least **50** unrated rollout items, including items the existing teen content settings would restrict, with youth test accounts in every age band.
- **Metric.** Share handled according to the existing teen content settings.
- **Pass threshold. 100%**.
- **Grading.** Major (threshold).

AC-8.3 Rollout plan

- **Requirement.** The platform has a documented plan and timetable for completing any uncertified later phase it intends to adopt, and reports progress at each surveillance audit.
- **Grading.** Observation where absent; not a condition of certification for the certified phase.

Domain 9 — Phase 2 and Phase 3

Domain 9 applies only where the platform is certified, or seeking certification, to the phase concerned.

9A — Comments (Phase 2)

AC-9.1 Comments rated

- **Requirement.** Comments are rated on their own (addendum Part 1.3).
- **Metric.** Share of comments posted after the Phase 2 completion date carrying a valid SHOW value.
- **Sampling.** Presence sample of at least **2,400** comments.
- **Pass threshold.** At least **99.0%**.
- **Grading.** Major (threshold).

AC-9.2 Comment label accuracy

- **Test method.** Domain 2 methods applied to a stratified sample of at least **1,000** comments.
- **Pass threshold.** AC-2.1 to AC-2.5 thresholds.
- **Grading.** Major (threshold).

AC-9.3 Comments above the ceiling hidden

- **Requirement.** A comment above a viewer's ceiling is hidden from that viewer even if the post is within it.
- **Test method.** Place at least **20 above-ceiling probe comments** on within-ceiling posts per test-account configuration, and **10 within-ceiling probe comments**.
- **Pass threshold. Zero** above-ceiling comments served; at least **95%** of within-ceiling comments visible.
- **Grading.** Major (threshold).

9B — Back-rating (Phase 2)

AC-9.4 Back-rating complete or hidden

- **Requirement.** Existing content is back-rated. Once Phase 2 is complete, unrated existing content is hidden from youth accounts.
- **Metric.** Share of existing items servable to youth accounts that carry a SHOW value; unrated existing items served to youth accounts in scripted use.
- **Pass threshold.** At least **99.5%** of existing items servable to youth accounts rated; unrated existing items at most **0.05%** of items served to youth test accounts.
- **Grading.** Major (threshold).

AC-9.5 Back-rating accuracy

- **Test method.** Back-rated items form a stratum of at least **600** items in the Domain 2 sample.
- **Pass threshold.** Every Domain 2 threshold.
- **Grading.** Major (threshold).

9C — Livestreams (Phase 3)

AC-9.6 Rating declared before going live

- **Requirement.** The streamer declares a rating before going live.

- **Test method.** Attempt at least **30** stream starts from test creator accounts, including attempts to skip the declaration on every client type.
- **Pass threshold. 100%** of streams require a declaration before going live.
- **Grading.** Major (threshold).

AC-9.7 Declared rating enforced

- **Requirement.** A stream declared above a youth viewer's ceiling is not served to that viewer, as for any other item.
- **Test method.** Run declared-rating test streams above and within each default ceiling, restricted to test audiences.
- **Pass threshold. Zero** above-ceiling streams served; at least **95%** of within-ceiling streams reachable.
- **Grading.** Major (threshold).

AC-9.8 Live monitoring, age-gating and warning

- **Requirement.** If live monitoring finds the stream above a youth viewer's ceiling, the stream is age-gated for that viewer with a warning, rather than viewers being ejected automatically.
- **Test method.** Run at least **30** scripted test streams, restricted to test audiences, declared within a youth ceiling and then escalated above it on one axis using licensed or staged material that is lawful and within Section 8.3 safeguards. Youth test viewers are present for each stream.
- **Metric.** Time from escalation to age-gate for youth test viewers; share of age-gates accompanied by a warning; share of trials in which youth test viewers were age-gated with a warning rather than ejected.
- **Pass threshold.** Age-gate applied within **60 seconds** in at least **90%** of trials and within **180 seconds** in every trial; a warning shown in **100%** of age-gates; youth test viewers age-gated rather than ejected in **100%** of trials.
- **Grading.** Major (threshold).

AC-9.9 Livestream recordings

- **Requirement.** Where a stream is kept as a recording, the recording carries a SHOW value reflecting the strongest moment of the stream (peak rule).
- **Test method.** Inspect recordings of the AC-9.8 test streams.
- **Pass threshold.** At least **95%** rated at or above the escalated level.
- **Grading.** Minor.

Domain 10 — Platform governance and records

AC-10.1 Mislabel complaints and appeals

- **Requirement.** Viewers, parents and creators can dispute a SHOW value or a flag. Disputes and flagged items go to human review, using the SHOW suggestion sliders described in addendum Part 1.2. Creators can appeal a rating or flag applied to their content. Outcomes are communicated to the person who raised the dispute.

- **Evidence.** Complaint and appeal procedures; case records.
- **Test method.** Review the procedure. Lodge at least **30** test disputes from test accounts, including disputes that an item is above a youth ceiling. Review a random sample of at least **100** production cases from the audit window.
- **Metric.** Share resolved within the time limit.
- **Pass threshold.** At least **90%** of disputes resolved within **7 days**; at least **90%** of disputes alleging that an item is above a youth ceiling resolved within **72 hours**.
- **Grading.** Minor; major if no dispute route exists.

AC-10.2 Rating records

- **Requirement.** The platform keeps, for every in-scope item, a history of every rating event (SHOW value, flags, `rated_by`, `rated_at`) for at least **3 years**, and provides complete, verifiable sampling frames and serving logs for the audit (Section 8.1).
- **Test method.** Trace the rating history of at least 50 sampled items; reconcile frames (Section 8.1.2).
- **Pass threshold.** Complete histories for every traced item; frame discrepancy no more than **0.5%**.
- **Grading.** Major.

AC-10.3 Platform reviewer competence

- **Requirement.** The platform's human reviewers are trained on SHOW v2.1 and the addendum's short-form rules, and the platform monitors their agreement.
- **Evidence.** Training materials; reviewer agreement records.
- **Test method.** Reviewer-rated items form a stratum of at least **150** items in the Domain 2 sample, topped up if necessary.
- **Pass threshold.** Reviewer stratum meets AC-2.6. At surveillance, the reviewer stratum is judged at 100 items.
- **Grading.** Minor.

AC-10.4 Change notification

- **Requirement.** The platform notifies the certification body in writing of any significant change at least **30 days** before deployment where the change is planned, and otherwise within **10 business days** of deployment. Significant changes include: changes to the label mapping (which must also be republished to r8rly.org before deployment); classifier changes that alter SHOW values or flags for more than 1% of items; changes to serving logic, ceilings, family controls, youth defaults or ad targeting for youth accounts; new surfaces, client types or jurisdictions; and changes to the scope statement.
- **Evidence.** Change records; notifications.
- **Test method.** Compare the platform's change log for the audit window with notifications received.
- **Pass threshold.** Every significant change notified.
- **Grading.** Major.

AC-10.5 Internal monitoring

- **Requirement.** The platform monitors its own label accuracy, flag accuracy and gate integrity between audits, and acts on what it finds.
- **Evidence.** Monitoring reports; actions taken.
- **Grading.** Minor where absent or not acted on.

Domain 11 — Learning Dividend reporting (optional module)

Domain 11 applies only to a platform that has opted in to the Learning Dividend. **It is not a condition of SHOW Certified™.** Findings are reported in a separate section of the audit report and do not affect the certification decision.

AC-11.1 Verification. Curriculum tags earn nothing until confirmed by teachers or education reviewers. *Test:* trace at least 50 tagged items. *Threshold:* no payment or algorithmic priority on any unverified tag.

AC-11.2 Pool and payments. The education pool is funded by the platform, not by profiling-based ads to minors, and payments to creators and school regions reconcile with the pool. *Test:* reconcile pool records and a sample of at least 50 payments. *Threshold:* full reconciliation.

AC-11.3 Privacy. Regional counts come from aggregated viewing data, with no individual student tracking. No regional count covering fewer than **50** viewers is reported or used for payment. *Test:* inspect the aggregation method and outputs. *Threshold:* requirement met.

---

10. Nonconformities and certification decisions

10.1 Grades

- **Major nonconformity** — failure to meet a requirement or pass threshold that Section 9 grades Major (including "major (threshold)"); absence of a required control or process; a systemic failure; any zero-tolerance event named in Section 9; or false or misleading evidence.
- **Minor nonconformity** — failure to meet a requirement or pass threshold that Section 9 grades Minor; or an isolated lapse or weakness that does not cause a requirement or threshold graded Major to be failed and does not undermine the effectiveness of a required control.
- **Observation** — an opportunity for improvement, or a risk that could become a nonconformity. Observations do not affect the certification decision.

A pattern of related minor nonconformities that together undermine a required control is graded as a major nonconformity.

10.2 Threshold failures

A major nonconformity graded "major (threshold)" in Section 9 is a **threshold failure**. Threshold failures are treated under Section 10.4 and cannot be closed by a corrective-action plan alone: they are closed only by re-audit showing the threshold is met on a fresh sample.

10.3 Corrective action

Grade	Correction and root-cause plan due	Implementation verified
Major (threshold)	14 days	By re-audit of the affected requirements on a fresh sample, at surveillance sample sizes or larger
Major (other)	14 days	Within 60 days, by document review or re-test as the certification body decides
Minor	30 days	At the next scheduled audit, or within 90 days if the certification body decides
Observation	No response required	Followed up at the next audit

Periods run from the date of the final audit report.

10.4 Decision rules

10.4.1 **Initial certification and recertification.** Certification is granted only when there is no open major nonconformity and every minor nonconformity has an accepted corrective-action plan. A platform below a threshold is not certified until the issues are fixed and re-audited.

10.4.2 **Major (other) at surveillance.** Certification continues while the nonconformity is corrected within 60 days. If it is not corrected and verified within 60 days, certification is suspended.

10.4.3 **Threshold failure at surveillance, special or unannounced audit.** Certification is suspended from the date of the certification decision on the audit, and reinstated only when the issues are fixed and re-audited.

10.4.4 **Minor not corrected.** A minor nonconformity not corrected by the next scheduled audit is raised to a major nonconformity.

10.4.5 **Scope.** A threshold failure in a Phase 2 or Phase 3 domain does not affect certification to the earlier phase(s) if every earlier-phase requirement is met. The certification body records the reduced scope and notifies r8rly.org under Section 6.10.

---

11. Audit reports, right of reply, regulators and the public register

11.1 Report structure

The audit report contains:

1. Platform, service, scope statement summary, certified or requested phase(s), client types and jurisdictions.
2. Audit team, with confirmation of impartiality and conflict-of-interest declarations.
3. Audit dates, audit window and any limitation on scope or access.

4. Methods: sampling frames and their verification, sample designs and sizes, test-account configurations, seeded-item safeguards, assessor calibration results and inter-assessor reliability.
5. Results for every requirement tested: metric, point estimate, 95% confidence interval, sample size, pass threshold, and pass or fail.
6. Stratum results (content type, jurisdiction, reach tier, rater layer, client type, surface).
7. Nonconformities and observations, each with grade, requirement reference and evidence.
8. The platform's reply (Section 11.2), reproduced in full.
9. Conclusion and recommendation to the certification decision-maker.
10. Learning Dividend findings, in a separate section, where Domain 11 applies.

11.2 Right of reply

The certification body gives the platform a preliminary report. The platform has **15 business days** to reply. The certification body corrects factual errors the platform demonstrates, may change a finding only on evidence, and reproduces the platform's reply in full in the final report.

11.3 Delivery

The certification body delivers the final report to the platform and to the relevant regulator, as the certification agreement provides. The report is not given to r8rly.org or r8rly.com.

11.4 Public register

r8rly.org keeps a public register built from certification bodies' notices (Section 6.10). The register shows **certification status only** — the platform and service name, certified or not certified, and the date — released under the platform's contractual consent. It does not show audit findings, metrics or nonconformities.

---

12. Use of the SHOW Certified™ mark

12.1 Only a platform holding current certification from an independent certification body licensed by r8rly.org to use the mark may display SHOW Certified™.

12.2 The mark may be used only for the certified service and, wherever it appears, must be accompanied by the certified phase(s), for example "SHOW Certified™ — Phase 1".

12.3 The mark must be reproduced as supplied by r8rly.org, without alteration.

12.4 The platform must not use the mark, or describe its certification, in a way that:

a) states or implies that content on the platform is "safe" or approved;
b) states or implies that r8rly.org or r8rly.com certified, audited or endorses the platform;
c) extends to services, client types or jurisdictions outside the certified scope;
d) appears on individual content items.

12.5 On suspension or withdrawal, the platform must stop all new use of the mark immediately, remove it from digital surfaces within **7 days**, and remove it from other materials within **30 days**.

12.6 The certification body monitors use of the mark by its certified clients and treats misuse under Section 6.9.

12.7 SHOW Certified™ and SHOW Standard™ are marks of r8rly.org. The SHOW Standard itself is free to use under CC BY-SA 4.0; no certification is needed to use SHOW labels, flags or ceilings.

---

13. Changes to these criteria

13.1 r8rly.org publishes every change to these criteria for public comment before it takes effect. The comment period is at least **60 days** for a change that adds or tightens a requirement or threshold, and at least **30 days** for any other change.

13.2 An advisory panel of parents, educators and child-safety specialists will be formed to review changes to these criteria before they take effect. Once formed, the panel reviews each change and its members declare their interests before each review. The panel advises; r8rly.org remains the governing body for the standard.

13.3 r8rly.org publishes its response to the comments received with each final change.

13.4 **Transition.** A change that adds or tightens a requirement or threshold takes effect no earlier than **6 months** after the final change is published. Certification bodies apply it from the first audit that begins after the effective date. Certifications granted under the earlier criteria remain valid until that audit.

13.5 Changes to the r8rly.org jurisdiction table are not changes to these criteria. Platforms implement them under AC-4.8.

13.6 These criteria do not change the SHOW Standard v2.1, its axes, levels, WTF bands or the ceilings in the addendum.

Annex A — Audit checklist summary

Ref	Requirement	Phase	Pass threshold	On failure
AC-1.1	SHOW value present	1	≥ 99.5% items; ≥ 99.5% ads	Major (threshold)
AC-1.2	Field format valid	1	≥ 99.5%	Major (threshold)
AC-1.3	Inline format and web meta field	1	≥ 99%	Minor (major if meta absent)
AC-1.4	Highest credible rating wins	1	100%	Major
AC-1.5	Edits re-rate	1	100%	Major
AC-1.6	Clips inherit	1	≥ 95%	Minor

Ref	Requirement	Phase	Pass threshold	On failure
AC-1.7	Mapping matches published; accurate	1	Identical; Domain 2 thresholds	Major (threshold)
AC-2.1	Exact agreement per axis (W: band)	1	≥ 80%	Major (threshold)
AC-2.2	Within one step per axis	1	≥ 97%	Major (threshold)
AC-2.3	Weighted kappa per axis	1	≥ 0.75	Major (threshold)
AC-2.4	Harm-weighted error; bias	1	HWE ≤ 0.50; bias ≥ −0.10	Major (threshold)
AC-2.5	Ceiling-crossing under-rating	1	≤ 1.0% per ceiling; severe ≤ 0.2%	Major (threshold)
AC-2.6	Stratum agreement	1	≥ 70% per stratum; mapped, ads and back-rated strata meet all Domain 2 thresholds	Minor (major (threshold) for mapped, ads, back-rated strata)
AC-3.1	Above-ceiling probes not served (labels confirmed in logs)	1	Zero	Major (threshold)
AC-3.2	Within-ceiling probes served	1	≥ 95%	Major (threshold)
AC-3.3	Label-based leakage in live use	1	≤ 0.05%; zero-tolerance instances zero	Major (threshold); minor within tolerance
AC-3.4	Reference leakage in live use	1	≤ 1.0% per band	Major (threshold)
AC-3.5	Ceiling changes enforced (10 per band)	1	Lowered ≤ 15 min (all); raised ≤ 15 min (90%)	Major / minor
AC-3.6	Consistent across clients and jurisdictions	1	AC-3.1, AC-3.2 and AC-3.3 zero-tolerance met for each; rates pooled	Major (threshold)
AC-4.1	Flag recall (seeded)	1	Per flag table	Major (threshold)
AC-4.2	False-flag rate (seeded)	1	Per flag table	Major (threshold)
AC-4.3	Field precision	1	Per flag table	Major (threshold)
AC-4.4	Field missed flags	1	Self-harm-ed, scam: at most one miss each (random); two or more fails (surveillance: none, or one after extending to 1,000)	Major (threshold) / minor

Ref	Requirement	Phase	Pass threshold	On failure
AC-4.5	Timing of flags	1	≥ 99% before youth availability (scam: any account); bullying ≤ 60 min (95%)	Major (threshold)
AC-4.6	Flagged probes hidden from youth (flags confirmed in logs)	1	Zero	Major (threshold)
AC-4.7	Support resources shown	1	≥ 95%	Major (threshold)
AC-4.8	Gambling gates by jurisdiction (incl. jurisdictions not in table)	1	Match or stricter; zero below-age	Major (threshold)
AC-4.9	Scam blocked for all ages	1	Zero	Major (threshold)
AC-4.10	Dangerous challenges hidden under 18	1	Zero	Major (threshold)
AC-4.11	Flags independent of ceiling	1	—	Major
AC-5.1	Youth defaults on (default ceiling per 4.8 or documented lower)	1	100%	Major (threshold)
AC-5.2	Changes to defaults need parent	1	None without parent	Major (threshold)
AC-5.3	Defaults survive updates and transitions	1	No weakening	Major
AC-6.1	Ceiling set by parent	1	Met	Major
AC-6.2	Each axis adjustable at any time, subject to AC-6.4 limits	1	Met	Major
AC-6.3	16–17 raises available	1	Met	Major
AC-6.4	Limits documented	1	Met	Minor
AC-6.5	Ceiling set is ceiling enforced	1	Match in every case	Major
AC-6.6	Ceiling visible to young person	1	Good practice	Observation only

Ref	Requirement	Phase	Pass threshold	On failure
AC-7.1	Ads carry SHOW; ad accuracy	1	≥ 99.5%; Domain 2 thresholds	Major (threshold)
AC-7.2	Ads respect ceiling	1	Zero	Major (threshold)
AC-7.3	No profiling-based ads to youth	1	Zero	Major (threshold)
AC-8.1	Unrated in completed phase hidden	1	Zero probes; ≤ 0.05% in use	Major (threshold); minor within tolerance
AC-8.2	Rollout content falls back to teen settings	1	100%	Major (threshold)
AC-8.3	Rollout plan	1	Documented	Observation
AC-9.1	Comments rated	2	≥ 99.0%	Major (threshold)
AC-9.2	Comment label accuracy	2	Domain 2 thresholds	Major (threshold)
AC-9.3	Above-ceiling comments hidden	2	Zero; ≥ 95% within-ceiling visible	Major (threshold)
AC-9.4	Back-rating complete or hidden	2	≥ 99.5% rated; ≤ 0.05% unrated served	Major (threshold)
AC-9.5	Back-rating accuracy	2	Domain 2 thresholds	Major (threshold)
AC-9.6	Rating declared before live	3	100%	Major (threshold)
AC-9.7	Declared rating enforced	3	Zero; ≥ 95% within-ceiling reachable	Major (threshold)
AC-9.8	Live age-gate with warning, no ejection	3	≤ 60 s (90%), ≤ 180 s (all); warning 100%; no ejection 100%	Major (threshold)
AC-9.9	Recordings rated at peak	3	≥ 95%	Minor
AC-10.1	Mislabel disputes and appeals	1	≥ 90% in 7 days; ≥ 90% youth-ceiling disputes in 72 h	Minor (major if no route)
AC-10.2	Rating records and frames	1	Complete; ≤ 0.5% discrepancy	Major
AC-10.3	Platform reviewer competence	1	AC-2.6 for reviewer stratum (≥ 150 items)	Minor
AC-10.4	Change notification	1	All notified	Major

Ref	Requirement	Phase	Pass threshold	On failure
AC-10.5	Internal monitoring	1	In place and acted on	Minor
AC-11.1–11.3	Learning Dividend (optional)	Opt-in	As stated	Reported separately; no effect on certification

Annex B — Minimum sample sizes

Sample	Initial and recertification	Surveillance	Notes
Presence and format (AC-1.1, AC-1.2)	2,400 items + 600 ads	1,200 items + 300 ads	Worst case ±2 percentage points at 95% confidence (initial); at the 99.5% threshold, about ±0.3 points for 2,400 items and ±0.6 for 600 ads (surveillance ±0.4 and ±0.8)
Label accuracy (Domain 2)	3,000 items	1,500 items	≥ 50% within 16–17 default ceiling; ≥ 600 (surveillance 300) within Under 10 ceiling; ≥ 600 (300) within 16–17 but outside 10–12 ceiling; ≥ 300 (150) ads; ≥ 400 (200) mapped-label items where mapping used
Flag recall (AC-4.1)	200 positives per flag	100 per flag	≥ 25% coded language
Flag false-flag (AC-4.2)	200 negatives per flag (≥ 100 hard)	100 (≥ 50 hard)	
Field precision (AC-4.3)	200 flagged items per flag	100 per flag	Or all if fewer
Field missed flags (AC-4.4)	1,000 enriched + 1,000 random	500 + 500	
Test accounts (Domain 3)	8 per age band offered (Section 8.2.2)	4 per age band offered (for 16–17, including the all-raises account)	Spread across clients and jurisdictions
Scripted use per account	7 days, 1,000 served items	5 days, 600 served items	
Above-ceiling probes	40 per configuration	20 per configuration	Every axis tested singly
Within-ceiling probes	20 per configuration	10 per configuration	

Sample	Initial and recertification	Surveillance	Notes
Reference leakage (AC-3.4)	400 served items per age band	200 per age band	
Ceiling changes (AC-3.5)	10 per age band (5 lowering, 5 raising)	6 per age band (3 lowering, 3 raising)	
Unrated probes (AC-8.1, AC-8.2)	50 each	25 each	AC-8.1: limitation procedure if fewer available
Reviewer stratum (AC-10.3)	150	100	Within the Domain 2 sample
Comments (AC-9.1, AC-9.2)	2,400 presence; 1,000 accuracy	1,200; 500	Phase 2 only
Back-rated items (AC-9.5)	600	300	Phase 2 only
Livestream trials (AC-9.6, AC-9.8)	30 each	15 each	Phase 3 only
Assessor gold items	5% of each rating batch	5%	Section 8.4.5

Where a threshold is judged at surveillance on a smaller sample and the 95% confidence interval of the point estimate includes the threshold, the certification body extends the sample to the initial-audit size before deciding.

---

r8rly.org · Governing body for the SHOW Standard · The SHOW Standard is free and open under CC BY-SA 4.0 · SHOW Certified™ is a mark of r8rly.org