Cut-Score Study

Sample technical report (synthetic)

About this sample

This is a complete technical report, not an excerpt, on a synthetic examination. There is no real exam, no real candidate and no real panel behind it: the items, the judges' ratings and the candidates' responses were generated by the service's own analysis script to demonstrate what a client receives. The credential is fictional. Every number below was produced by the analysis script from the synthetic data and appears exactly as the script reported it.

No real exam content is ever used as a sample. A client's items, ratings and results are confidential and stay in the client's engagement folder.

0. Title block

Programme	Fictional credential: Certified Process Inspector (synthetic)
Examination	40 scored multiple-choice items, one form
Decision supported	Pass or fail on the credential examination
Method	Modified Angoff, two rounds, with a Hofstee reasonableness check
Panel	10 subject experts, online session
Item data	600 candidates, scored 0 or 1
Package	Facilitated study
Report date	1 November 2026
Author	Dr Milos Kankaras

1. Summary

The panel recommends a cut score of 22.83 marks out of 40 (57.1 percent), applied as 23 marks. Its standard error is 0.32 marks. Candidates scoring 23 or more would pass; on the 600 candidates in the item data that is 53.8 percent.

Panel agreement rose between rounds: the intraclass correlation for a single judge went from 0.68 to 0.82, and the spread of the judges' cuts fell from a standard deviation of 1.45 to 1.02 marks. The standard error of measurement at the cut is 3.17 marks, so a candidate within about 3 marks of the cut is not reliably distinguishable from the cut on this examination. The Hofstee check is consistent with the Angoff cut on the cut score and just outside it on the failure rate (Section 8).

This is a recommendation. The programme's governing body adopts, adjusts or rejects it and records why. Nothing in this report guarantees that the cut will survive an appeal or an accreditation audit.

2. Purpose and consequences

The cut score decides who is certified as a Process Inspector. The consequence of a false pass is an uncertified person working unsupervised on a safety-relevant process; the consequence of a false fail is a competent candidate delayed by a retake. The programme's governing body judged the first consequence more serious, which is why the panel was instructed to describe a borderline candidate who is just able to work unsupervised, and not a candidate who is merely promising. (Fictional context.)

3. Procedure

- Method: modified Angoff. Each judge estimated, for every item, the probability that a borderline candidate answers correctly. A judge's cut is the sum of that judge's ratings. The panel cut is the mean of the judges' Round 2 cuts.
- Why this method: the examination is a fixed-form multiple-choice test, item-level results are available, and the programme's accreditor expects a documented item-based judgemental procedure.
- Fixed in advance: the method, the number of rounds, the rounding rule and the borderline description were agreed by the programme's governing body before any rating was made.
- Rounds: Round 1 alone and asynchronously; a 90-minute live session in which the panel saw the Round 1 summary and the candidates' actual item performance; Round 2 alone within 24 hours.
- Rounding: a candidate passes at or above the cut rounded to the nearest whole mark, halves up.
- Hofstee: after the session each judge gave the lowest and highest acceptable cut score and failure rate. The panel means define the check. It is a reasonableness check and never the cut.
- Deviations from the protocol: none (synthetic).

4. The panel

Ten judges (synthetic), described by role. Names are listed in the full report only if the client directs.

Judge	Role (synthetic)	Years in practice	Prior standard-setting
J01	Senior inspector	18	Yes
J02	Training lead	14	No
J03	Inspector	9	No

Judge	Role (synthetic)	Years in practice	Prior standard-setting
J04	Quality manager	22	Yes
J05	Inspector	7	No
J06	Programme examiner	12	Yes
J07	Senior inspector	16	No
J08	Supervisor	11	No
J09	Training lead	10	No
J10	Regulator liaison	15	Yes

Ten judges is inside the recommended range of 8 to 12. Four of ten had prior standard-setting experience, and the briefing call covered the method for the rest.

5. The performance standard

The borderline candidate was described in writing and agreed before rating (synthetic text): a candidate who can carry out routine inspections to the written procedure without supervision, recognises and escalates a non-conformity, and records findings correctly, but who cannot yet diagnose an unfamiliar fault or train others. The full description is Appendix A of a real report.

6. Results by round

	Round 1	Round 2
Panel cut (marks)	22.86	22.83
Standard error of the cut	0.46	0.32
SD of judges' cuts	1.45	1.02
Lowest and highest judge cut	20.77 to 24.96	21.20 to 24.25
ICC(2,1), single judge	0.68	0.82
ICC(2,k), panel mean	0.96	0.98
Mean pairwise correlation	0.70	0.83

The standard error of the cut is the standard deviation of the judges' cuts divided by the square root of the number of judges. It describes how far a repeat panel of the same kind would be expected to land from this one. It does not describe the variation among candidates. ICC(2,1) and ICC(2,k) are two-way random-effects, absolute-agreement coefficients (Shrout & Fleiss, 1979) with items as targets and judges as raters. Judges changed 61 percent of ratings in Round 2, by 0.037 on average.

7. Recommended cut score

Recommended cut: 22.83 marks of 40 (57.1 percent), applied as 23. Standard error 0.32 marks. A band of one standard error either side is 22.51 to 23.16 marks.

8. Hofstee reasonableness check

Panel means: acceptable cut 49.7 to 75.1 percent correct, acceptable failure rate 10.0 to 45.2 percent. The Hofstee cut is 21.19 marks (53.0 percent), 1.64 marks below the Angoff cut. The Angoff cut is inside the acceptable range. The failure rate at the Angoff cut is 46.2 percent, which is outside the acceptable range of 10.0 to 45.2 percent.

The check does not contradict the Angoff cut, but the failure rate it implies is slightly above the panel's own upper limit. The governing body should read that as a signal to confirm the borderline description, not as a reason to move the cut to a pass rate.

9. Examination evidence

On 600 candidates, the mean raw score was 23.56 (SD 7.77), and KR-20 reliability was 0.87. The classical standard error of measurement is 2.75 marks. The conditional standard error of measurement at the cut, from the binomial model, is 3.17 marks. A candidate scoring at the cut has a score that would be expected to vary by about that much on a repeat sitting, which is why the impact table below is read together with the standard error of the cut.

10. Impact

Scenario	Cut	Applied	Pass n	Pass %	Fail %
Cut minus 1 SE	22.51	23	323	53.8	46.2
Recommended cut	22.83	23	323	53.8	46.2
Cut plus 1 SE	23.16	23	323	53.8	46.2
Applied cut minus 1 mark	22.00	22	356	59.3	40.7
Applied cut plus 1 mark	24.00	24	302	50.3	49.7

Because the standard error of the cut (0.32 marks) is smaller than one mark, a band of one standard error rounds to the same applied cut as the recommendation, and the pass rate does not move. The practical alternatives are the neighbouring whole marks: one mark lower passes 59.3 percent and one mark higher passes 50.3 percent, a spread of 9.0 percentage points. That is the size of the consequence attached to a one-mark difference. Moving the cut by judgement in either direction is a policy decision for the governing body and is not supported by this study.

11. Items for attention

No item showed a judge standard deviation above 0.15 in Round 2. Three items showed a gap of more than 0.25 between the panel's mean rating and the proportion of candidates near the cut who answered correctly. These are the items worth reviewing for content or key problems before the next form.

Item	Panel rating (R2)	Borderline-band p	Gap
Q07	0.78	0.44	0.34
Q23	0.77	0.48	0.28
Q31	0.47	0.77	-0.30

The borderline band is the 176 candidates within one mark or one conditional standard error of the cut, whichever is larger. The panel rated Q07 and Q23 as easier for a borderline candidate than they proved, and Q31 as harder.

12. Limits

- A cut score is a judgement. A different panel of the same kind would be expected to land within about two standard errors of this one, not on it.
- The panel is the client's. This panel is synthetic; a real report states plainly how its composition limits the result.
- The Hofstee check is a reasonableness check on a judgemental result. It does not identify a true cut.
- Item-level results came from 600 candidates; the report notes whether that sample resembles the population that will sit the examination.
- The study makes the procedure defensible, not the outcome certain.

13. Documentation map

The report is organised so that each element a reviewer would expect to find in documentation of a cut score has one named section: the rationale and the consequences (Section 2), the procedure and its fixing in advance (Section 3), the qualifications of the panel (Section 4), the performance standard (Section 5), the judgements and their agreement (Section 6), the result and its uncertainty (Sections 7 and 9), a reasonableness check (Section 8), the consequences of alternative cuts (Section 10) and the limits (Section 12). It is built to serve the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014) and the documentation an accredited personnel-certification scheme's assessor expects. Specific clause numbers are not cited here until each has been checked against the primary document.

14. Verdict

I recommend a cut score of 22.83 marks, applied as 23, with a standard error of 0.32 marks, on the evidence in Sections 6 to 10. I do not claim that it is the only defensible cut score.

Dr Milos Kankaras

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin, 86(2), 420-428. https://doi.org/10.1037/0033-2909.86.2.420