A PRE-REGISTERED TEST OF THE CURRENT GRADER

How the WarPaths grader performs

GRADER VERSION	FROZEN	GRADED	AUTHOR
Version 4	18 November 2026	19 November 2026	Thomas Nagle, founder

WarPaths scores written reasoning in individually played wargames. Before any game in this study was graded, the grader was frozen and 38 pass marks were written down. It met 29 of them and passed two of the six tests outright: repeat grading and a model update. It did not pass the other four.

WHAT HELD

- **Same game, same score.** Graded 20 times, a game's total moved 1.8 points at the median and 4.4 at most.
- **Steady under a model update.** Totals moved 0.97 points on average.
- **Strong play beat weak play** in 12 of 12 pairs, by 13 points on average.
- **Padding never raised a score.** 24 of 24 padded games scored no higher.
- **Missing reasoning is caught.** A removed dimension fell to the lowest level in 30 of 30 copies.

WHERE IT FELL SHORT

- **Agreement with the expert:** 155 of 210 judgments, 74 percent against a mark of 80.
- **It grades more strictly than the expert,** by about a third of a level per answer.
- **A single missing element** was noticed in 38 of 55 cases, and too rarely in three of five dimensions.
- **Padding lowered scores** by 3.05 points on average against a limit of 3.0.
- **Problem Framing** is read too low and barely separates strong play from weak.

WHAT AGREEMENT MEANS FOR A SCORE

- **Agreement counts judgments, not points.** Agreement of 80 percent means the same call on 80 judgments in every 100, not two scores 20 points apart. One level on one dimension is worth 5 of a game's 100 points.
- **The 80 percent mark came with two marks in score terms:** the grader's level within 0.6 of a level of the expert's on average, and 20 of 25 answers within one level. On the one scenario where agreement reached 79 percent, the grader's level sat within half a level of his on average and 7 of 8 answers were within one level.
- **Overall, at 74 percent:** 17 of 25 answers were within one level. The grader read lower: an estimated 8.8 points on a game's total, with wide uncertainty.

At a glance

#	TEST	KEY RESULT	PASS MARK	OUTCOME
1	**Agreement with expert**	**74%** of 210 judgments agree	80%	**Not passed**
2	**Planted defect**	Whole dimension **30 of 30**; element **38 of 55**	80%; 70% per dimension	**Not passed**
3	**Added text**	**24 of 24** scored no higher; fell 3.05 points	no rise; fall of 3.0 at most	**Not passed**
4	**Repeat grading**	**1.8 points** median spread over 20 gradings	3 points	**Passed**
5	**Unseen scenarios**	Depends on tests 1 to 3	tests 1 to 4 all met	**Not passed**
6	**Model update**	**0.97 points** mean shift in a game's total	2.5 points	**Passed**

Each test has several pass marks; the table shows the main one. A leading general-purpose AI model given the same answers agreed with the expert on 73 percent of judgments, no better than the grader.

READ THIS FIRST

- The games were played by AI players at set ability levels. They were not played by serving officers.
- The expert in test 1 is one person using a rubric against expected standards for each dimension. How consistent one expert is with himself, or with a second expert, has not been measured.
- WarPaths ran this study on its own product and set the pass marks. It has not been independently audited.
- Agreement percentages count individual judgments. They are not gaps between scores. Test 1 shows how far the grader's scores sit from the expert's.
- The tests show whether the grader applies its standard steadily and as an expert would. They do not show that the standard is the right one for every question, or that a score predicts later performance.
- WarPaths is used for practice, feedback and instruction. This note does not claim the grader is fit to decide anything about a career.

What was tested

The grader in brief

Each game is scored on five dimensions: Problem Framing, Decision Effectiveness, Risk Management, Strategic Cohesiveness and Context Management. Each dimension has a published standard with five levels. A game's total runs from 0 to 100; one level on one dimension is worth 5 points.

The grader makes a series of judgments on whether specific pieces of reasoning are stated in the player's own words, and an answer's level follows from those judgments. A separate relevance check stops an answer that does not address its question from earning credit.

The version tested here was frozen for this study. It is the next release of the grader and was not yet in service when tested.

How the study was run

1	2	3	4	5	6
Freeze the grader	Fix the pass marks	Play fresh games	Grade everything once, sealed	Expert judges a sample, blind	Compare and report

"Sealed" means no result was read by anyone until the expert's judgments were recorded. "Blind" means the expert saw no score.

The material

ITEM	IN THIS STUDY
Scenarios	3 published scenarios used in no earlier round of testing: After November Sixth, Debris Field and Suwalki Window. None was used to set any of the grader's settings (see test 5 for their history).
Games	18: on each scenario 2 weak, 2 middling and 2 strong. 330 answers.
Players	AI players at three ability settings, with a 1,200-character limit on each answer.
Gradings	1,774 gradings of a whole game: each of the 18 games 20 times, each altered copy 10 times, and the 18 games 3 times on a newer version of the underlying model.
Expert judgments	210 judgments on 25 answers, 5 from each dimension, by one expert.
Altered copies	73 with one element removed, 30 with a whole dimension removed, 24 with text added, and 9 poor-play games. Each copy was checked by independent readers before grading; of 347 attempts to remove a single element cleanly, 73 passed that check.

Results, test by test

VALIDITY

1 Agreement with an expert

NOT PASSED | 74% (mark 80%)

Does the grader score an answer the way an expert does?

HOW IT WAS RUN

One expert graded 25 answers in full, by hand and blind: every check on each answer, 210 judgments in all. The grader scored the same answers. Two things are compared: the level each answer receives, and each individual judgment.

PASS MARKS, SET IN ADVANCE

At least 80 percent of judgments agree; agreement beyond chance at least 0.55; the grader's level within 0.60 of the expert's on average; at least 20 of 25 answers within one level.

RESULT

74%

155 of 210 judgments agree (mark 80%)
agreement beyond chance 0.45 (mark 0.55)
average level difference 0.85 (mark 0.60)
17 of 25 answers within one level (mark 20)

WHAT THIS MEANS FOR A SCORE

- **Judgment agreement is not a score gap.** Agreement of 74 percent means the grader and the expert made the same call on 74 judgments in every 100. It does not mean their scores differ by 26 points.
- **Level by level.** Of the 25 answers, the grader gave exactly the expert's level on 7, was within half a level on 10 and within one level on 17. Three answers were two levels or more apart.
- **The grader is the stricter reader.** It was lower than the expert on 11 answers, higher on 7 and equal on 7: about a third of a level lower on average. It does not inflate.
- **Estimated effect on a game's total: 8.8 points lower than the expert would give.** This is an estimate built from 5 answers in each dimension, not a measured game: no complete game was graded by hand. It is uncertain; 95 percent of the estimate's range lies between 17.2 points lower and 0.1 lower.
- **For scale:** strong and weak play differed by 13.0 points on average, and the same game graded twenty times differed by at most 4.4.
- **What was not measured:** whether a difference of this kind changes the order in which players are ranked.

DIMENSION	JUDGMENTS AGREEING	BEYOND CHANCE	EXPERT CREDITED, GRADER DID NOT	GRADER CREDITED, EXPERT DID NOT	GRADER'S LEVEL MINUS EXPERT'S
Problem Framing	28 of 40 (70%)	0.40	10	2	-1.30
Decision Effectiveness	33 of 50 (66%)	0.09	9	8	-0.73
Risk Management	31 of 40 (78%)	0.55	4	5	+0.10
Strategic Cohesiveness	32 of 40 (80%)	0.60	5	3	-0.30
Context Management	31 of 40 (78%)	0.54	3	6	+0.47
All	155 of 210 (74%)	0.45	31	24	-0.35

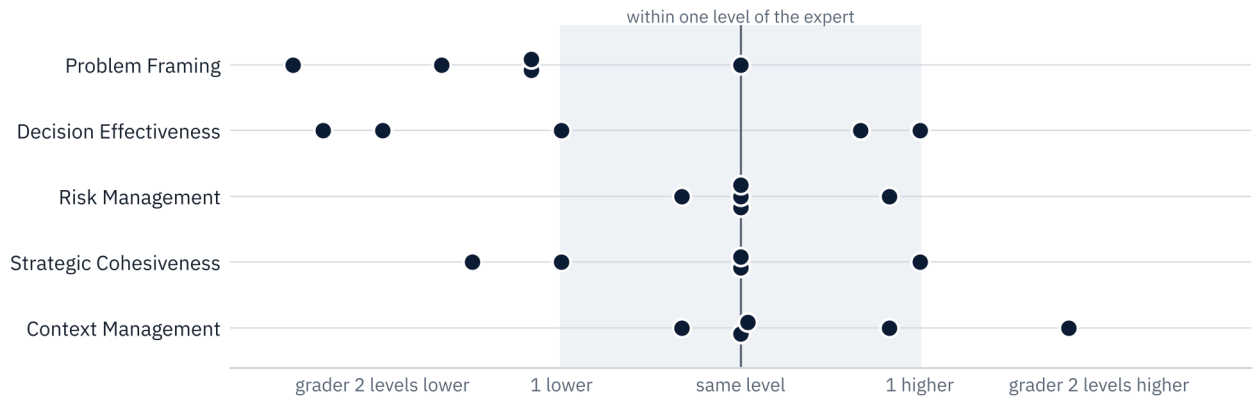


Figure 1. The grader's level minus the expert's. Each dot is one of the 25 answers. The shaded band is within one level.

Three things stand out. Problem Framing is read well below the expert. Risk Management, Strategic Cohesiveness and Context Management sit close to him. And agreement falls as judgments become more demanding: 84 percent on the least demanding checks, 77 on the middle ones and 65 on the most demanding.

The disagreements are consistent, not random: on 53 of the 55 judgments where the grader and the expert differ, all 20 gradings of that game gave the same verdict.

A yardstick. A leading general-purpose AI model was given the same 25 answers and the same judgments to make, after the expert's were recorded. It agreed with the expert on 153 of 210 (73 percent), no better than the grader, and it shared 32 of the grader's 55 differences from him. It gave credit more freely than the expert; the grader gave it less freely. Where two different readers part from one expert in the same places, the standard being applied and the consistency of a single expert are in question as well as the grader.

2 Planted defect

NOT PASSED | 38 of 55 single removals noticed

When reasoning is removed from a strong answer, does the score fall where it should?

HOW IT WAS RUN

Two kinds of copy were made from strong games. In one, a single required element was removed from one answer and nothing else. In the other, all of one dimension's answers were removed. Independent readers confirmed every copy before any was graded.

PASS MARKS, SET IN ADVANCE

One element: the grader stops crediting it in at least 70 percent of copies, in each dimension, and no more than 10 percent of its other judgments on those answers change. Whole dimension: that dimension falls furthest in at least 80 percent of copies.

WHAT IT MEANS

The grader reliably notices when a dimension's reasoning is absent. It is less reliable at noticing one missing element inside an otherwise complete answer, and that part did not pass.

RESULT

30 of 30

copies with a whole dimension removed were caught
38 of 55 single-element removals were caught
other judgments changed: 7.6% (limit 10%)

DIMENSION	COPIES	GRADER HAD CREDITED THE ELEMENT	REMOVAL NOTICED	RATE (MARK 70%)	
Problem Framing	12	10	9	90%	met
Decision Effectiveness	16	11	6	55%	not met
Risk Management	16	12	8	67%	not met
Strategic Cohesiveness	15	10	4	40%	not met
Context Management	14	12	11	92%	met
All	73	55	38	69%	

In 18 of the 73 copies the grader had not credited the element in the first place, although two independent readers had found it; those copies could not test a removal. Risk Management missed its mark by one copy. Whole-dimension removal passed on all five dimensions, 6 of 6 each; in those copies almost nothing of the answer was left, so that part shows what the grader does with an absent answer.

3 Added text

NOT PASSED | by 0.045 of a point

Does adding words that carry no reasoning raise a score?

HOW IT WAS RUN

Text of four kinds was added to otherwise identical answers, inside the answer length limit: a restatement of the question and situation; confident filler; bare claims without reasons; and the vocabulary of the scoring dimensions. 24 padded games, half weak and half strong.

PASS MARKS, SET IN ADVANCE

A mean rise of no more than 2.0 points overall and 3.0 for any one kind; no more than 10 percent of copies gain half a level on any dimension; and a mean change no lower than -3.0 points.

WHAT IT MEANS

Padding does not pay. Every mark against a rise was met. The one mark missed is the limit on how much padding may cost, by 0.045 of a point: added text makes the grader find slightly less in an answer, most of all on one scenario (-6.1).

RESULT

24 of 24

padded games scored no higher
mean change -3.05 points (limit -3.0)
by kind: -2.7, -4.1, -2.8, -2.6

REPEATABILITY AND COMPARABILITY

4 Repeat grading

PASSED | median spread 1.8 points

Does the same game get the same score each time it is graded?

HOW IT WAS RUN

Each of the 18 games was graded 20 times. The spread is the difference between the highest and lowest total a game received.

PASS MARKS, SET IN ADVANCE

Median spread of no more than 3 points; 95th percentile no more than 6; largest under 12; reliability of a single grading at least 0.95.

RESULT

1.8

points, median spread (mark 3)
95th percentile 4.4 (mark 6); largest 4.4 (mark 12)
reliability of a single grading 0.991 (mark 0.95)

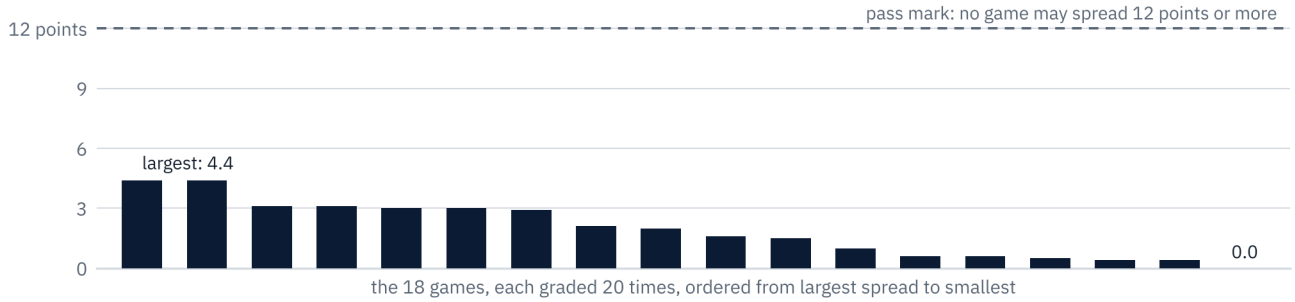


Figure 2. Spread of each game's total over its 20 gradings. A typical game's totals have a standard deviation of half a point.

5 Unseen scenarios

NOT PASSED | depends on tests 1 to 3

Does the grader apply the standard, or what it learned about one scenario?

HOW IT WAS RUN

All three scenarios were new to the study and none was used to set any of the grader's settings, so this test is the run as a whole: it asks that every mark of tests 1 to 4 be met.

PASS MARKS, SET IN ADVANCE

Every pass mark of tests 1 to 4 is met on the three scenarios together.

WHAT IT MEANS

This test cannot pass while tests 1 to 3 do not. It adds no new failure of its own.

RESULT

79%

agreement with the expert on the newest scenario alone (74% on all three)
18 of the 26 pass marks of tests 1 to 4 met; the eight not met are those reported under tests 1, 2 and 3

MEASURE	ALL THREE SCENARIOS	AFTER NOVEMBER SIXTH ALONE
Test 1: judgments agreeing with the expert	155 of 210 (74%)	52 of 66 (79%)
Test 1: agreement beyond chance	0.45	0.57
Test 1: average level difference	0.85	0.46
Test 1: answers within one level	17 of 25	7 of 8
Test 2: single-element removals noticed	38 of 55	14 of 18
Test 2: whole-dimension copies caught	30 of 30	10 of 10
Test 3: mean change from added text	-3.0 points	-2.0 points
Test 4: median and largest spread	1.8 and 4.4	1.1 and 3.0
Strong play minus weak play	13.0 on average	11.4

The scenarios' history. Debris Field and Suwalki Window were among five scenarios whose earlier plays were looked at in October, when the scale was designed; nothing in the grader was fitted to them. After November Sixth has no such history, so every test is also reported on it alone. Its counts are small (6 games, 8 of the expert's 25 answers) and are a report, not a verdict.

6 Model update

PASSED | 0.97-point average shift

Do scores hold when the underlying AI model is updated?

HOW IT WAS RUN

The same 18 games were graded again on a newer version of the underlying model, with no other change.

PASS MARKS, SET IN ADVANCE

Mean shift in a game's total of no more than 2.5 points; the order of the games preserved (rank correlation at least 0.90); no game moves by 8 points or more.

WHAT IT MEANS

A move to the newer version would shift scores by about a point. For scale, two sets of gradings on the same version differ by 0.3 points on average.

RESULT

0.97

points, mean shift in the total (mark 2.5)
order of the 18 games 0.96 (mark 0.90)
largest shift for one game 2.9 (mark 8)

Further checks

These are not among the six tests. They are reported because a grader could pass all six and still fail to tell good play from bad.

CHECK	PASS MARK, SET IN ADVANCE	RESULT	
Stronger play scores higher. Weak, middling and strong AI players on the same scenarios.	Strong exceeds weak by at least 10 points on every scenario, and in at least 11 of 12 pairs	Gaps of 11.4, 17.5 and 10.1 points; 12 of 12 pairs. Average totals: weak 56.2, middling 59.4, strong 69.2	met
Thin play scores low. One or two generic sentences per answer.	No dimension above level 2.5 and a total of 40 or less, each game	Totals 8.2, 8.1 and 6.3 of 100	met

CHECK	PASS MARK, SET IN ADVANCE	RESULT	
Wrong-crisis play scores low. Strong answers written for a different scenario.	The same, each game	Totals 0.0, 1.9 and 12.4. On the third, one dimension reached 2.65. The relevance check refused 50 of the 55 answers	**met on 2 of 3 games**
Genuine answers are not refused by the relevance check.	No more than 2 percent of 330	0 of 330	**met**
Wrong-question play. Each answer moved under another question of the same crisis.	Reported; no pass mark	25 of 54 moved answers refused; these games keep about 35 of 100	-

Where the grader falls short

Every mark not met, and every weakness the results show, whether or not a mark covered it.

- **It does not yet agree with the expert closely enough:** 74 percent of judgments against a mark of 80, and 17 of 25 answers within one level against a mark of 20.
- **It reads lower than the expert,** by about a third of a level per answer: an estimated 8.8 points on a game's total, with wide uncertainty.
- **Problem Framing is the weakest dimension.** It is read 1.3 levels below the expert and separates strong play from weak by at most 0.2 of a level. On one scenario all six games received the same level.
- **Decision Effectiveness agrees with the expert little better than chance would** (0.09), although it is the dimension that most separates strong play from weak.
- **A single missing element often goes unnoticed:** caught in 55 percent of cases in Decision Effectiveness, 67 in Risk Management and 40 in Strategic Cohesiveness, against a mark of 70. In 18 of 73 copies the grader had not credited the element to begin with.
- **Agreement falls as judgments become more demanding:** 84, 77 and 65 percent.
- **Added text lowers the score,** by 3.0 points on average and 6.1 on one scenario.
- **Middling play is barely separated from weak play:** 3.2 points on average, and 1.0 on one scenario.
- **Results are uneven across scenarios.** Suwalki Window is weakest on agreement with the expert (65 percent), on added text, and on the strong-to-weak gap (10.1 against a mark of 10).
- **The relevance check has limits.** It stops an answer written for an unrelated crisis, but let through three of 18 answers written for a related one, and it stops fewer than half of answers placed under the wrong question of the right crisis.
- **Not tested:** an answer that begins on the question and then drifts away from it.

What comes next

The grader is being revised in light of these results. Two commitments are public.

- **A measure of the expert.** The expert will re-grade a sample blind after an interval, and a second expert will be sought, so that the grader can be compared with how closely people agree.
- **A new pre-registered run** on scenarios used in no earlier testing, with pass marks fixed before any game is graded, reported in a new version of this note.

Until then, the results above stand as the record of this version of the grader.

Limits of this study

- **The players were AI players.** Answers written by serving officers may differ in length, style and content.
- **One expert.** Test 1 rests on one person's judgments, and his consistency has not been measured. There is no human benchmark to compare the grader against.
- **Self-run.** WarPaths designed, ran and reported the study, and set the pass marks.
- **Small counts.** 25 answers were hand-graded, 5 in each dimension; per-dimension figures are indications, not measurements.
- **Scope.** Three scenarios, all strategic-level crisis scenarios played by one person.
- **The standard itself was not tested.** The study asks whether the grader applies its standard as an expert does, not whether the standard is right for every question.
- **Not a prediction.** Nothing here shows that a score predicts performance in a later role.

How these tests were developed

The grader and the tests were both revised before this run. Earlier rounds of testing were used to find defects in the grader and in the tests themselves, and some early test designs proved unsound and were replaced. One of those rounds showed that answers written for a different crisis kept most of their score, and the relevance check was added in response. This note reports the tests as run on the frozen grader, on games used in no earlier round. A fuller record of that development is available on request through the contact form at warpaths.com. A dated record of progress, beginning with the first grader, is kept at warpaths.com/ai-grading-wargames.

What is not published

The standard for each dimension is public. The individual checks, their settings and the prompts are not, so that answers cannot be written to the grader. How the grader is being revised is not published either.

Disclosure

Thomas Nagle founded WarPaths, which sells individually graded wargames. He ran this study and was its expert rater. WarPaths has no current contract for work on the Joint Warfighter Evaluation; it is applying to a department marketplace for AI tools, for use in professional military education. No independent evaluation is available at this time.

Terms used

TERM	MEANING HERE
Dimension	One of the five things a game is scored on.
Level	The rung reached on a dimension's standard, from 1 to 5. One level on one dimension is worth 5 points of the total.
Check	One judgment the grader makes about an answer.
Judgment agreement	The share of checks on which the grader and the expert made the same judgment. It is not a difference between two scores.
Agreement beyond chance	Agreement after removing what two raters would agree on by chance (Cohen's kappa). 0 is chance; 1 is complete agreement.
Spread	Highest total minus lowest total across repeat gradings of one game.
Reliability of a single grading	The share of the variation in totals that comes from differences between games and not from the grader (intraclass correlation).
Order of games	How closely two sets of scores put the same games in the same order (Spearman rank correlation). 1 is identical order.

Questions about this note: the contact form at warpaths.com.