From Puzzle to Psychometric Instrument: A Forced-Choice Adaptation of Bongard Problems for Standardized Fluid-Reasoning Assessment

Jean-François Parent
jfparent.dev@pm.me
https://www.supersharp.quest/bongard-long-test
March 2022

Abstract

Bongard problems—pairs of small diagram sets separated by a single abstract rule—are widely regarded as a uniquely demanding probe of visual abstraction and rule induction; Hofstadter placed the skill of solving them "very close to the core of 'pure' intelligence." Yet their classical response format, in which the solver must *verbalise* the distinguishing rule, is subjective and not automatically scorable, and has kept them out of standardized ability testing. This paper describes an adaptation, implemented in the *super-sharp* web application, that reformulates the task as a forced-choice completion: several panels of a Bongard problem are withheld, and the test-taker must fill each from a pool of candidate tiles so that both sides remain internally consistent. The reformulation preserves the inductive demand—correct placement is impossible without having inferred the rule—while yielding an objective, automatically scored response. We situate the instrument within the Cattell–Horn–Carroll account of fluid reasoning, specify a measurement model (item response theory for these dichotomously-scored items, with a guessing lower asymptote), and set out a concrete standardization protocol: sampling, reliability targets, a convergent/discriminant validity battery, form equating, and fairness analysis. We make no empirical claims: no norms, reliability, or validity data exist yet. This is a design and a research programme, not a validated test.

Keywords: Bongard problems, fluid intelligence, inductive reasoning, item response theory, test standardization, forced-choice assessment, construct validity, culture-reduced testing.

1. Introduction

A Bongard problem (BP) presents twelve small diagrams arranged as two groups of six. Every diagram on the left shares an abstract property that every diagram on the right lacks, and the solver's task is to discover and articulate the single rule that separates the two sides [1, 2]. The diagrams are visually simple, but the space of candidate rules—shape, count, size, orientation, topology, symmetry, convexity, relative position—is vast, and the intended abstraction is usually one the solver has never been told to look for. Hofstadter, who popularised BPs in *Gödel, Escher, Bach*, argued that this act of perceiving the right concept out of an open field "lies very close to the core of 'pure' intelligence" [2].

This combination of properties—minimal perceptual load, maximal conceptual demand, and near-total independence from language or acquired knowledge—is exactly what a culture-reduced measure of fluid reasoning aspires to. And yet BPs are essentially absent from the standardized-testing literature. The obstacle is not the stimulus but the *response*: the classical task asks for a

verbal statement of the rule, which is subjective to score, impossible to grade automatically, and difficult to place on a common scale across items and persons.

This paper describes an adaptation that removes that obstacle without diluting the task. In the *super-sharp* web application, a BP is presented with several of its twelve panels withheld; the test-taker completes them by selecting from a pool of candidate tiles, and an answer is correct only if every placement keeps the two sides rule-consistent (Section 3). Because a tile can be placed correctly only once the distinguishing rule has been induced, the reformulation retains the cognitive core of the original while producing an objective, automatically scored outcome. Our contributions are: (i) a description of the forced-choice mechanism and its rationale; (ii) a psychometric framework locating the instrument among established ability constructs and specifying a measurement model (Section 4); and (iii) a standardization protocol for turning the instrument into a normed subtest (Section 5). We are explicit throughout that this is a proposal: no norms or validity data have yet been collected.

2. Background: the classic Bongard problem

BPs originate with Mikhail Bongard, who introduced them in his 1967 monograph on pattern recognition (English translation, 1970) as a challenge for both human and machine perception [1]. Each problem is a self-contained induction task: six positive instances, six negative instances, and an implicit predicate that holds of the former and fails of the latter. Crucially, the discriminating feature is abstract and context-dependent—the same shape may be positive in one problem and negative in another—so the solver cannot rely on a fixed feature vocabulary but must construct the relevant abstraction on the fly. Figure 1 illustrates the structure with a simple rule.

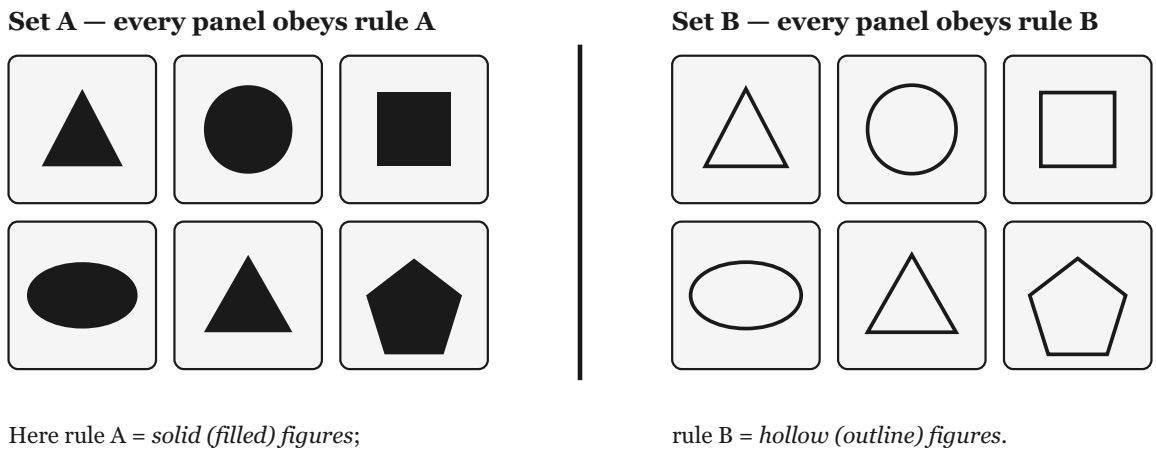

Figure 1. A classic Bongard problem. Every panel in Set A satisfies one rule (here, *solid figures*) and every panel in Set B another (*hollow figures*); the solver must induce and state what distinguishes the sets. Irrelevant attributes are varied deliberately so that only the intended abstraction is diagnostic. Redrawn schematically after the classic format [1, 2].

Hofstadter's treatment in *Gödel, Escher, Bach* [2] framed BPs as a window onto human concept formation and made them a touchstone for artificial intelligence: a solver must flexibly generate,

test, and revise candidate concepts, sliding between perception and conceptualisation until a description "clicks." The most sustained scientific treatment is Foundalis's *Phaeaco*, a cognitive architecture built expressly to solve BPs through interacting perceptual and conceptual processes; the accompanying research and the curated *Index of Bongard Problems* remain the standard reference collection [3, 4]. Chapman, in a complementary vein, uses BPs to argue that real-world categorisation is *nebulous*—rules are often underdetermined and purpose-relative—a "meta-rational" caution we return to in the Discussion [5]. More recently BPs have re-entered machine-learning research as a benchmark for human-level concept learning, precisely because they resist brute-force pattern matching [13].

Cognitively, solving a BP recruits visual abstraction, inductive rule discovery, analogical mapping, hypothesis generation and testing, and cognitive flexibility—the abilities that factor-analytic models place at the heart of *fluid reasoning* (Gf) [7, 8]. This is the construct our instrument targets.

3. The multiple-choice reformulation

The *super-sharp* adaptation keeps the twelve-panel structure—six left-hand panels governed by one rule, six right-hand panels by another, the two rules necessarily different—but *withholds* a subset of the panels. The test-taker is shown the remaining panels and a strip of candidate tiles, and must place a tile into each empty position. In the interface the currently active empty slot is highlighted; the other empty slots remain marked as pending; and each candidate may be used at most once. Figure 2 shows four items drawn directly from the running application.

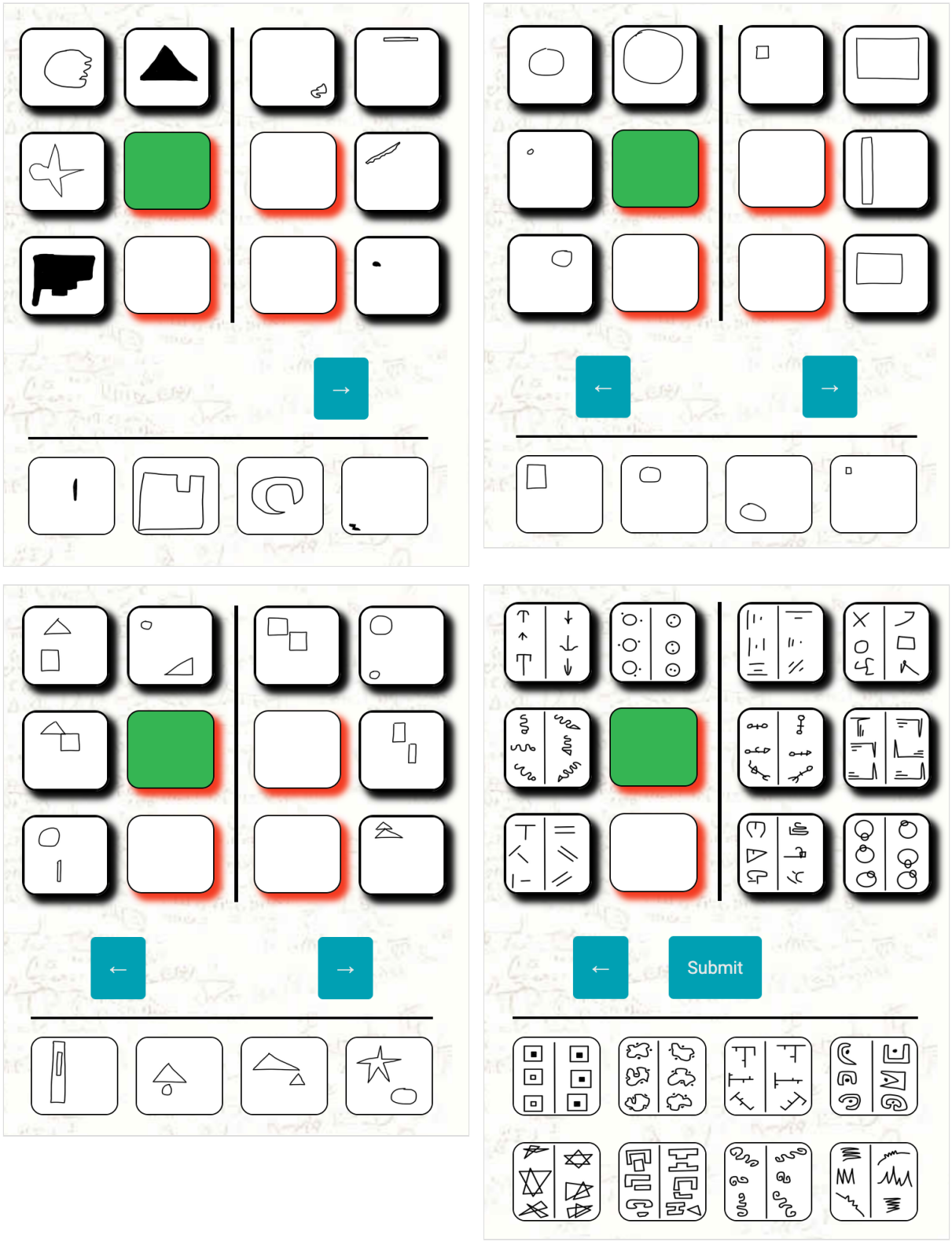

Figure 2. Four forced-choice Bongard items as rendered in the *super-sharp* web application. Each shows the twelve-panel problem (with withheld panels highlighted) above the strip of candidate tiles. The fourth item is the higher-ceiling *Meta-Bongard problem*, in which every panel is a complete Bongard problem in miniature.

Concretely, the twelve panels are indexed as six left cells and six right cells; a per-item configuration names which cells are withheld and supplies a candidate pool of up to eight tiles. Selecting tiles assigns one to each withheld cell; the item is scored by comparing the assignment against the key. Figure 3 details the mechanic.

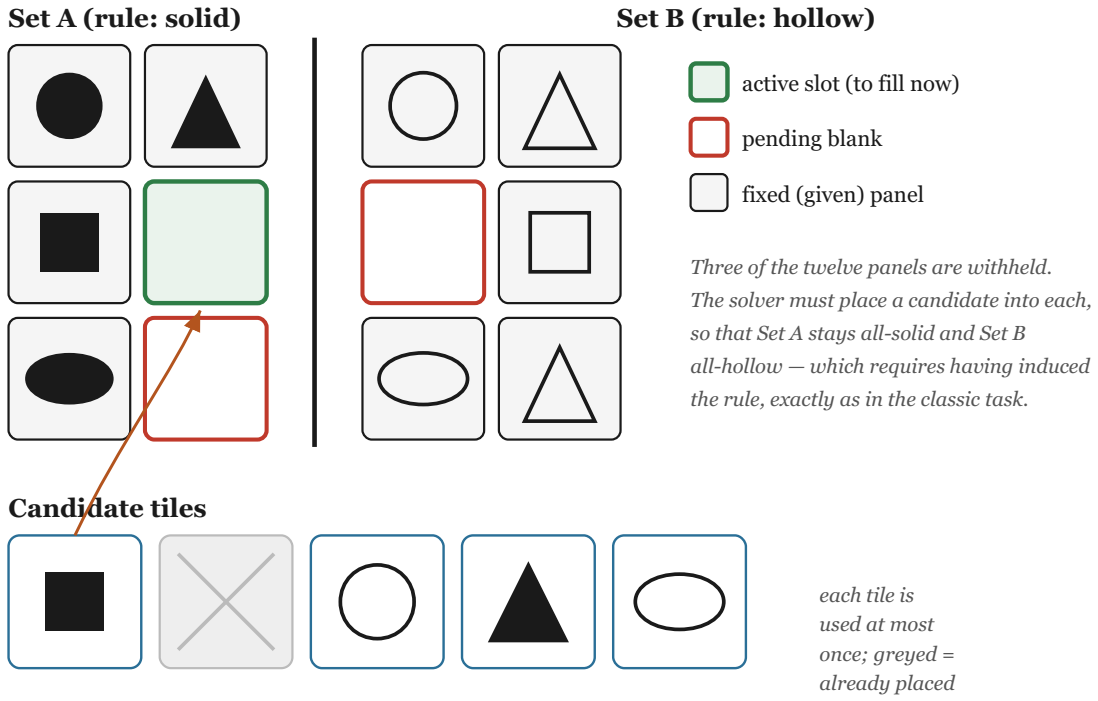

Figure 3. The completion mechanic. Several panels are withheld (one active, the others pending); the solver draws from a candidate pool—containing both rule-consistent tiles and distractors—placing each tile at most once. A placement is correct only if it keeps Set A and Set B internally rule-consistent, which is possible only after the distinguishing rule has been induced.

Why the reformulation preserves the construct. The concern with any multiple-choice version of an open-ended task is that it might convert *construction* into mere *recognition*. Here it does not, for a structural reason: the candidate pool is not a list of paraphrased rules but a set of diagrams, and a diagram can be placed correctly only if the solver has represented the abstract predicate well enough to test a novel instance against it. Distractors are chosen to be rule-*violating* in ways that are visually unremarkable, so elimination by surface features does not suffice. The solver still performs the full induction; only the final report changes, from a verbal statement to a placement.

Distractor design and the answer key. For each withheld cell the pool contains at least one tile that satisfies that side's rule and several that do not (for example, tiles that would satisfy the *opposite* side, or neither). Because tiles are placed without replacement and several cells are filled from a shared pool, the placements are interdependent—a property with consequences for guessing, addressed in Section 4.

A tunable ceiling. Item difficulty is controlled by the abstractness of the rule, the number of withheld panels, and the similarity of distractors to correct tiles. The *Meta-Bongard problem* (Figure 2, fourth item)—the established term for the nested construction—raises the ceiling further: each panel is itself a Bongard problem, so the governing rule is a rule *about rules*, engaging a second, meta-level of abstraction.

4. Psychometric framework

4.1 Construct definition and validity

We define the target construct as **fluid reasoning (Gf)**, and within it the narrow abilities of **induction** and **visual abstraction**, as delineated by Cattell's fluid/crystallised distinction [7] and Carroll's three-stratum model [8]. Non-verbal inductive tasks of this family—most famously Raven's Progressive Matrices—are among the purest markers of the general factor *g*, sitting at the centre of the ability "radex" [9, 10]. We therefore predict the pattern of relations shown in Figure 4: strong *convergent* correlations with other Gf markers (Raven's Advanced Progressive Matrices, the matrix-reasoning subtests of the Wechsler scales, the Cattell Culture-Fair scales) and weak *discriminant* correlations with crystallised, language-bound, or schooling-dependent measures (vocabulary, reading). Confirming this pattern is the central validity question for the instrument; it is a prediction, not a result.

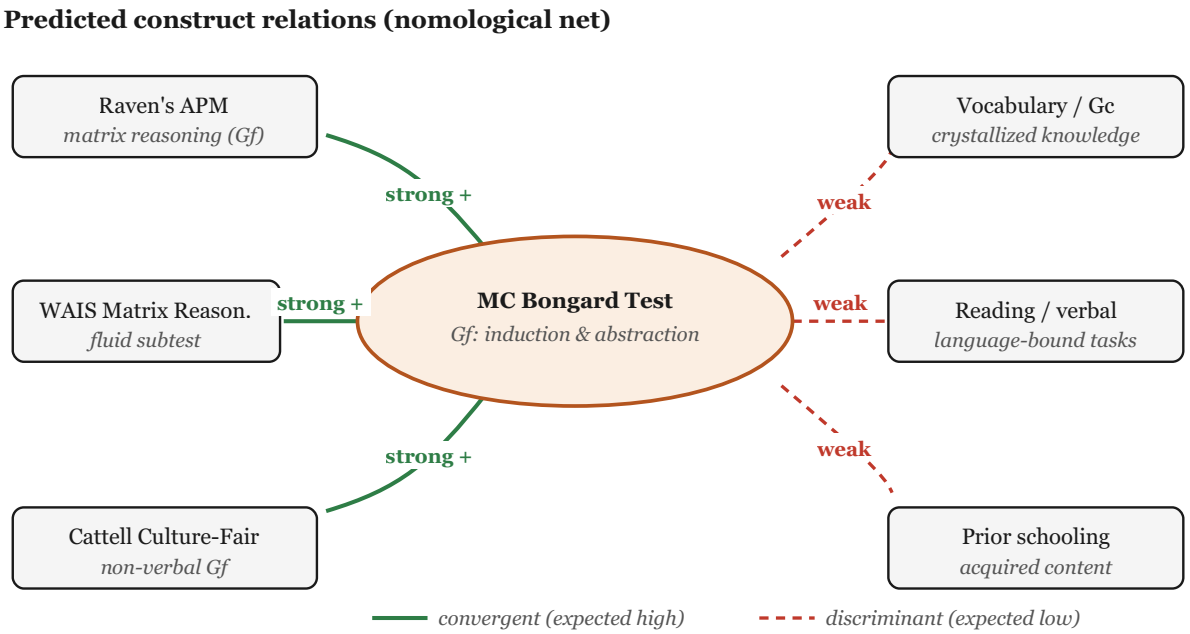

Figure 4. The predicted nomological net. The instrument is hypothesised to load on fluid reasoning: high convergent correlations with non-verbal Gf markers and low discriminant correlations with crystallised and language-bound measures. These are predictions to be tested by the validity study of Section 5.

4.2 Scoring, guessing, and the measurement model

Each item is scored *dichotomously*: the test-taker earns a single point only if every withheld panel is filled with a tile belonging to the correct side—the tile must satisfy that side's rule, though its exact cell within the side is immaterial and the order of placement does not matter—and no point otherwise. There is no partial credit for placing only some panels correctly, so the item score is binary and the appropriate measurement model is a dichotomous one. Because an item is credited only when *all* withheld panels are placed on the correct side, and tiles are drawn from a finite pool, success by chance is possible but small—essentially the probability that a random assignment sorts every withheld tile to its correct side. We therefore favour a model with a non-zero lower asymptote (a three-parameter logistic model) so that this residual guessing is estimated rather than assumed away. Item response theory then yields, per item, a difficulty and a discrimination parameter and an information function [11]; these support item selection, adaptive

administration, and equating of the short and long forms. Figure 5 traces the response from tile placements to an ability estimate, contrasting it with the unscorable open-ended format.

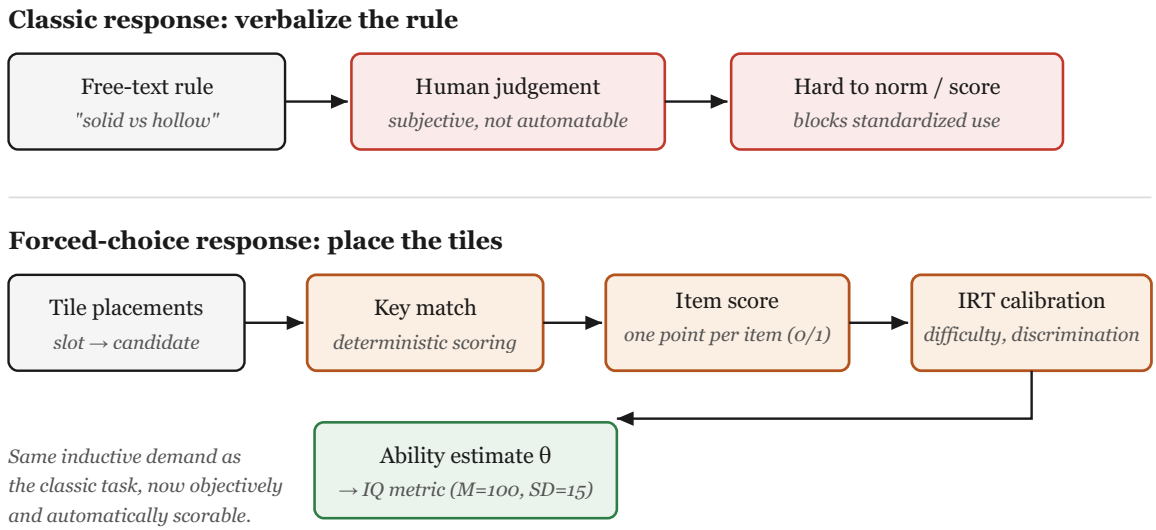

Figure 5. From response to score. The classical verbal response resists automatic grading and norming; the forced-choice response is matched against a key and scored dichotomously (one point per item), calibrated under item response theory, and mapped to an ability estimate and an IQ metric—all deterministically, with the same inductive demand on the solver.

4.3 Reliability and test forms

The application currently offers two untimed forms—a 10-item short form and a 40-item long form. Test length bounds internal-consistency reliability, so the long form is expected to be the more reliable and the appropriate basis for individual scores; the short form suits screening. Target reliabilities and the choice between timed and untimed administration are parameters to be fixed empirically during standardization, not asserted here.

5. Proposed standardization protocol

Turning the instrument into a normed subtest requires a staged programme, summarised in Table 1. None of these steps has yet been carried out.

Stage	Purpose	Key methods
1. Item bank & pilot	Assemble items; screen for clarity and defects	Expert review; small-sample pilot; flag ambiguous keys
2. Calibration	Estimate item difficulty & discrimination	Dichotomous IRT (2PL / 3PL); item-fit and information analysis
3. Norming	Build population reference	Large, demographically stratified sample; derive IQ metric (M = 100, SD = 15)
4. Reliability	Quantify measurement precision	Internal consistency; test–retest; conditional standard error from IRT
5. Validity	Test the construct claim	Convergent/discriminant battery (APM, matrix reasoning vs. vocabulary)
6. Fairness	Detect bias across groups	Differential item functioning; measurement-invariance testing
7. Form equating	Put short and long forms on one scale	Common-item / IRT linking

Table 1. A staged standardization programme. Every stage is future work; the current release provides only the instrument and objective scoring, not the evidence that would license interpretive claims.

All stages should follow the *Standards for Educational and Psychological Testing* [12], which make reliability, validity, and fairness evidence preconditions for interpretive use. In particular, the culture-reduced ambition of a non-verbal task must be *demonstrated*, through invariance and DIF analyses across language and educational backgrounds, not merely assumed from the absence of words.

6. Discussion

The meta-rationality caveat. Chapman's analysis warns that categorisation rules are often nebulous and purpose-relative [5]. Forced choice cuts both ways here. By committing to an answer key it buys objectivity and eliminates the ambiguity of grading free-text rules; but a key also legislates a single "intended" abstraction where a thoughtful solver might defend another. Good item design must therefore ensure that the key rule is the *most parsimonious* discriminator and that no distractor placement satisfies an equally simple alternative rule—a constraint that item review (Stage 1) must enforce explicitly.

Test-wiseness. Any forced-choice format invites strategies—elimination, consistency-checking across the shared pool—that can inflate scores independently of the target ability. The interdependence of placements makes such strategies partially effective, which is one reason guessing must be modelled (Section 4) and item difficulty calibrated empirically rather than assumed.

Item exposure and generation. As with any fixed bank, public circulation erodes validity. A natural extension is a generator that synthesises fresh, calibrated BPs on demand—connecting this work to procedural item generation and to the AI-benchmark literature that treats BPs as a test of genuine concept learning [13]. This is noted as future work, not claimed.

7. Limitations and future work

The overriding limitation is empirical: *no norms, reliability coefficients, or validity data exist*. Every psychometric statement here is a hypothesis or a target. The construct claim—that the test measures fluid reasoning with little crystallised contamination—is plausible on theoretical grounds but untested. The two forms have not been equated; the answer keys have not been vetted for alternative-rule ambiguity at scale; and the fairness of the format across populations is assumed rather than shown. The immediate next step is Stage 1–2 of Section 5: a pilot and an IRT calibration on a modest sample, sufficient to estimate item parameters, prune defective items, and justify a full norming study.

8. Conclusion

Bongard problems have long been admired as an unusually direct probe of abstraction and rule induction, yet their open-ended response format has confined them to the status of a puzzle. Reformulating the task as a forced-choice completion—fill the withheld panels so that both sides stay rule-consistent—preserves the inductive core while making the response objective and automatically scorable, the essential precondition for standardized measurement. On that basis we have set out a psychometric framework and a concrete programme to normalise the instrument as a culture-reduced measure of fluid reasoning. What remains is the empirical work: piloting, calibration, norming, and the validity and fairness studies without which no score should be interpreted. The instrument is ready; the evidence is the next chapter.

References

1. Bongard, M. M. (1970). *Pattern Recognition*. Rochelle Park, NJ: Hayden Book Co. (Spartan Books). English translation of the 1967 Russian original.
2. Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An Eternal Golden Braid*. New York: Basic Books.
3. Foundalis, H. E. (2006). *Phaeaco: A Cognitive Architecture Inspired by Bongard's Problems*. Doctoral dissertation, Indiana University, Bloomington. https://www.foundalis.com/res/diss_research.html
4. Foundalis, H. E. *Index of Bongard Problems*. https://www.foundalis.com/res/bps/bpidx.htm
5. Chapman, D. *A first lesson in meta-rationality*. https://metarationality.com/bongard-meta-rationality
6. Wikipedia contributors. *Bongard problem*. Wikipedia, The Free Encyclopedia.
7. Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54(1), 1–22.
8. Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge: Cambridge University Press.
9. Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlates. In R. J. Sternberg (Ed.), *Advances in the Psychology of Human Intelligence* (Vol. 2, pp. 47–103). Hillsdale, NJ: Erlbaum.
10. Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology*, 41(1), 1–48.
11. Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
12. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

13. Nie, W., Yu, Z., Mao, L., Patel, A. B., Zhu, Y., & Anandkumar, A. (2020). Bongard-LOGO: A new benchmark for human-level concept learning and reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)* 33.