

Assessing Moral Reasoning Without Rewarding Conformity: A Fairness Framework for Healthcare Ethics Education

Kristian Eriksen^{1*}, Solveig Berg¹, Magnus Dahl²

¹Department of Ethics Education and Moral Reasoning Assessment, Faculty of Medicine, University of Oslo, Oslo, Norway.

²Department of Fairness and Conformity in Ethics, Faculty of Medicine, University of Bergen, Bergen, Norway.

*E-mail ✉ kristian.eriksen@medisin.uio.no

Abstract

Assessment in healthcare ethics education carries an unusual validity problem: learners may reach different conclusions while reasoning carefully, yet conventional scoring can make agreement with an expected answer easier to reward than the quality of the reasoning that produced it. This scholarly perspective examines that problem as a question of construct definition and assessment fairness rather than as evidence that disagreement is inherently virtuous or that expert consensus is inherently unjustified. We distinguish moral reasoning from moral correctness, broader ethical competence, professionalism, and mere verbal sophistication. Building on contemporary work in ethics education, assessment validity, equity, rater judgment, and inclusive assessment, we propose a nonconformity-sensitive assessment framework in which conclusions remain substantively bounded but are not treated as proxies for reasoning quality. The framework emphasizes ethical salience, factual and conceptual accuracy, reason construction, engagement with competing considerations, coherence, uncertainty, and defensible responsiveness. It also requires assessors to distinguish failures of justification from legitimate disagreement and to examine whether scoring criteria privilege dominant styles, assumptions, or conclusions. The approach is intended to support later multimethod assessment and programmatic use rather than function as a stand-alone high-stakes instrument. Its principal limitation is evidential: existing literature supports the underlying fairness and validity concerns but does not establish the reliability, predictive validity, cross-cultural equivalence, or educational consequences of the framework proposed here. These properties require empirical testing before consequential use.

Keywords: Moral reasoning, Ethics education, Assessment fairness, Nonconformity, Validity, Medical education

Introduction

The conformity problem

Healthcare ethics assessment has a difficult object. Learners may encounter the same ethically contested case, recognize the relevant values, weigh competing obligations carefully, and nevertheless defend different

courses of action. Ethics assessment therefore cannot assume that every important task has one uniquely correct response or that one assessment format is optimal for every ethical learning objective [1]. The educational problem is not disagreement itself. It is the possibility that an assessment ostensibly designed to evaluate moral reasoning instead rewards proximity to the conclusion preferred by an examiner, institution, dominant professional culture, or answer key.

That possibility is best understood as a validity and fairness problem. Contemporary assessment scholarship distinguishes fairness, inclusion, and justice while also arguing that consequences, social impact, and equity matter to the defensibility of assessment systems [2, 3].

Access this article online

<https://smerpub.com/>

Received: 10 January 2026; Accepted: 06 April 2026

Copyright CC BY-NC-SA 4.0

How to cite this article: Eriksen K, Berg S, Dahl M. Assessing Moral Reasoning Without Rewarding Conformity: A Fairness Framework for Healthcare Ethics Education. Asian J Ethics Health Med. 2026;6(1):104-13. <https://doi.org/10.51847/OF4ndl6wl6>

Accordingly, an ethics assessment may be technically reproducible yet still underrepresent the intended construct if it systematically confuses conclusion agreement with reasoning quality. Conversely, an assessment that tolerates any conclusion merely because it is articulated confidently would also fail. The relevant aim is neither conformity nor unrestricted pluralism.

Assessment judgments are additionally produced within social systems. Measurement choices interact with language, expectations, identity, culture, and assessor interpretation, making fairness broader than the elimination of obvious psychometric error [4]. This matters particularly in ethics, where judgments about what sounds “reasonable,” “professional,” or “balanced” may themselves contain tacit normative expectations.

The problem also reaches construct measurement. Recent critique of the Moral Competence Test argues that apparently objective scoring may depend on contestable assumptions about how arguments are classified and what consistency signifies [5]. This does not invalidate moral-reasoning assessment as such. It strengthens the case for making the assessed construct explicit. We therefore define the conformity problem as a proposed assessment failure in which agreement with an expected ethical conclusion contributes more to a score than the defensibility of the reasoning supporting that conclusion.

Defining the construct of moral reasoning

A fairness framework requires clarity about what moral reasoning is—and what it is not. Moral reasoning can be understood as an activity of identifying morally relevant considerations, constructing reasons, examining their relationships, evaluating competing reasons, and determining what those reasons justify [6]. This process orientation matters because a conclusion alone provides limited evidence about how the learner interpreted the

case. Two learners may reach the same answer through very different reasoning, while two learners who disagree may each identify important considerations that the other has weighted differently.

Moral reasoning is nevertheless narrower than ethical competence. Reviews of healthcare ethics education describe ethical competence as multidimensional, incorporating knowledge, recognition, reflection, communication, judgment, attitudes, and action rather than a single reasoning score [7]. A moral-reasoning assessment should therefore not be interpreted as measuring whether a learner is a morally good clinician, possesses stable professional character, or will act ethically in practice. Those are larger inferences than the construct can sustain.

Nor should assessment silently collapse ethical reasoning into one philosophical vocabulary. Medical education draws on principle-based, virtue-oriented, care-oriented, humanistic, and other traditions that direct attention toward partly different features of ethical problems [8]. Programs may legitimately require learners to understand specified ethical frameworks, professional duties, or legal constraints. Yet when the intended construct is reasoning rather than doctrinal recall, the learner's ability to justify how relevant considerations bear on a conclusion should remain visible.

Educational studies also show why humility is necessary. Structured case discussion has been examined as a means of fostering moral competence, but pilot evidence does not establish a decisive or universally transferable effect [9]. The assessment construct proposed here therefore concerns observable reasoning performance in a defined task, not an inferred inner moral capacity.

These distinctions generate the construct boundaries summarized in **Table 1**.

Table 1. Boundaries of Moral Reasoning as an Assessment Construct in Healthcare Ethics Education

Construct dimension	What may legitimately be assessed	What should not be inferred	Evidence status
Ethical salience	Recognition of morally relevant facts, persons, values, duties, and uncertainties	Agreement with the assessor's preferred priority	Proposed operationalization informed by moral-reasoning analysis
Reason construction	Relevance and intelligibility of reasons offered for a position	Moral worth of the learner as a person	Reason-giving is construct-relevant
Competing considerations	Serious engagement with reasons pointing toward alternative judgments	Requirement to endorse every alternative	Proposed synthesis consistent with multidimensional ethical competence

Normative perspective	Appropriate use of concepts or traditions relevant to the task	Mandatory use of one unstated philosophical doctrine	Plural traditions are represented in medical ethics education
Coherence	Fit among stated facts, reasons, priorities, and conclusion	Proof that the conclusion is uniquely correct	Proposed assessment distinction
Construct boundary	Task-specific reasoning performance	Global ethical competence, professionalism, character, or future clinical behavior	Ethical competence is broader than reasoning alone
Developmental interpretation	Evidence that reasoning performance can be examined educationally	Assumption that one teaching method reliably produces moral development	Current educational evidence remains limited

Note: Elements labelled as proposed are conceptual specifications advanced in this article, not validated scoring dimensions.

How assessment rewards agreement rather than reasoning

The danger of conformity is easiest to see when a scoring system requires an externally determined key. Situational judgment tests, for example, can be built through expert judgments that determine the preferred ordering or acceptability of responses [10]. Such procedures may be useful and psychometrically defensible for particular purposes. The fairness concern arises when the key becomes a substitute for the construct: a learner who chooses the expected option may receive credit without demonstrating why it is justified, whereas a learner who reaches another defensible position may be penalized despite stronger reasoning. The problem is therefore not expert involvement but conclusion-dominant inference. Rater-based assessment creates a related difficulty. Variation among assessors is often treated as undesirable noise, yet some work demonstrates that raters can approach performance from systematically different points of view [11]. This observation does not imply that every disagreement is meaningful. Inconsistent judgment can have serious consequences, especially around borderline performance, and assessment systems still require defensible common standards.

The relevant evidence supports a more discriminating account: assessor disagreement is not uniformly random or uniformly remediable; it may reflect divergent viewpoints, vary in consequence according to performance level, and respond inconsistently to calibration [11-13]. Calibration should therefore seek shared understanding of the construct, evidence, and standards without assuming that identical judgments are always the only sign of quality.

In ethics assessment, this distinction becomes especially important. If evaluators are trained only to converge on a preferred conclusion, increased interrater agreement

could conceal construct underrepresentation. If, alternatively, evaluators are given unlimited discretion, scores may become unstable and inequitable. A fairer objective is disciplined interpretive consistency: assessors should apply common criteria for reasoning quality while retaining space to recognize more than one substantively defensible conclusion. This is a proposed design principle, not evidence that current ethics assessments routinely penalize dissent.

Fairness principles for ethics assessment

Fairness in ethics assessment should begin at the level of the assessment system rather than with an assumption that impartial individual examiners are sufficient. Assessment inequity has been described as a systemic problem because tasks, criteria, opportunities to demonstrate competence, assessor judgments, grading structures, and subsequent decisions interact [14]. For moral reasoning, these interactions can determine not only whose argument receives credit but also which forms of argument are readily recognizable as competent. Equity-oriented assessment therefore requires system-level attention to structural inequity, documented bias processes in observed assessment, and intersecting social and contextual influences [14-16]. These concerns are relevant even when ethical conclusions themselves are not scored directly. Linguistic style, assertiveness, cultural conventions of disagreement, disability, hierarchy, or familiarity with dominant professional discourse could affect whether the same underlying reasoning is perceived as sophisticated or deficient.

We propose four fairness commitments. Construct fidelity requires scoring the reasoning the assessment claims to measure. Defensible pluralism allows different conclusions when each remains within explicit substantive boundaries. Interpretive accountability

requires assessors to explain how evidence in a response satisfies shared reasoning criteria rather than relying on intuitive agreement. Consequential proportionality requires stronger evidence before a moral-reasoning judgment carries higher educational stakes.

These commitments do not require ethical neutrality. Programs may legitimately specify nonnegotiable constraints—for example, factual accuracy, respect for the task's stipulated patient interests, relevant professional obligations, or applicable legal boundaries. Nor does fairness demand equal credit for incompatible arguments. It demands that differences in scores track differences in relevant reasoning quality rather than unexplained preference for a particular conclusion.

The nonconformity-sensitive assessment framework

The proposed Nonconformity-Sensitive Assessment Framework operationalizes that distinction through a sequence of assessment questions. First, assessors ask whether the response recognizes the ethically salient features of the case and accurately represents the relevant facts and constraints. Second, they assess the reasons advanced: Are they relevant, intelligible, internally coherent, and appropriately connected to the proposed action? Third, they examine whether the learner has seriously engaged competing considerations rather than merely asserting a position. Only after these steps should the assessor consider whether the conclusion crosses an explicitly stated substantive boundary.

This ordering matters. Concordance approaches in professionalism and ethical decision-making demonstrate that expert judgment can be represented as a distribution rather than collapsed into a single response, making disagreement itself visible to learners [17]. That approach does not validate the present framework, but it shows that assessment need not hide professional plurality.

Nonconformity sensitivity must also remain context-sensitive. Inclusive assessment scholarship emphasizes the tension between common educational purposes and local contexts [18]. Thus, a conclusion defensible in one jurisdiction or professional setting may be constrained differently elsewhere. Context sensitivity, however, is not relativism: learners remain accountable to relevant facts, stated norms, affected persons, and the reasoning needed to connect them.

The framework is also intentionally programmatic rather than a single-rubric solution. Contemporary assessment design emphasizes multidimensional architectures rather

than isolated instruments [19]. We therefore propose that evaluators apply a conclusion-separation **test**: before finalizing a judgment, they should ask whether the same quality of reasons would receive a comparable reasoning score if attached to a different conclusion that remained within the task's defensible boundaries. A large unexplained score change would signal possible conclusion substitution.

Finally, the framework asks “valid for whom?” as well as “valid for what?” Inclusion-focused validity scholarship makes that question difficult to avoid [20]. Assessment should therefore examine whether learners have equitable opportunities to make their reasoning visible and whether supposedly neutral criteria privilege particular styles of ethical expression. These components define a proposed fairness architecture; they do not establish reliability, validity, or superiority over existing approaches.

A multi-method assessment blueprint

A nonconformity-sensitive framework should not depend on one task type. Complex reasoning is better sampled through complementary methods because different formats reveal different aspects of performance and different sources of error [21]. By analogy, moral reasoning should be assessed across tasks that vary in response mode, contextual ambiguity, opportunity for explanation, and assessor interaction. The purpose is not to accumulate instruments indiscriminately but to prevent a single format from monopolizing the inference.

We propose a blueprint with three complementary functions: eliciting an initial judgment, making the learner's justification inspectable, and testing whether that justification remains defensible when counterarguments or new facts are introduced. These functions may be sampled through written cases, structured oral examination, concordance-style prompts, or observed discussion, provided each method is aligned with the construct rather than treated as interchangeable. Stakes also matter. Evidence from progress testing shows that summative consequences can alter learner motivation and feedback use [22]. Early implementation should therefore rely on repeated, lower-stakes sampling before major progression decisions.

Aggregation also requires restraint. A strong performance on one task should not automatically erase a serious failure on another when the failures concern different components of the construct. Conversely, a single atypical judgment should not dominate a

longitudinal interpretation. We propose retaining domain-level information long enough to identify patterns across cases, methods, and occasions before collapsing evidence into an overall judgment. This preserves the distinction between broad reasoning weakness, context-specific difficulty, and assessor-

specific disagreement without pretending that an optimal weighting scheme has already been established.

Table 2 compares assessment approaches according to what each can reveal and where conformity risk may enter.

Table 2. Comparing Assessment Methods for Nonconformity-Sensitive Moral-Reasoning Assessment

Method	Distinctive evidence obtained	Principal conformity risk	Proposed fairness safeguard	Evidence anchor
Selected-response case	Rapid judgment across many scenarios	Expert key substitutes for reasoning	Use only for bounded judgments; pair with justification	Consensus-keyed formats
Written justification	Explicit reasons and prioritization	Stylistic fluency mistaken for quality	Score reasoning criteria separately from prose polish	Proposed synthesis from construct definition
Structured oral probe	Clarification, counterargument, uncertainty	Examiner preference shapes questioning	Standardize prompts, not conclusions	Rater-perspective evidence
Concordance-style task	Response under expert disagreement	Panel consensus becomes hidden authority	Show distribution and rationale; keep use developmental	Professional-judgment concordance
Repeated multimethod sampling	Cross-context pattern of performance	One weak task dominates a decision	Triangulate methods and occasions proportionately	Multimethod reasoning principle

Note: The safeguards are proposed design specifications, not validated scoring procedures.

Scoring justification, revision, and responsiveness

Scoring should make reasoning visible without converting openness to revision into obedience. We propose three separable scoring objects: justification, meaning the quality and relevance of reasons offered; revision, meaning whether a learner appropriately changes or retains a position after receiving new information; and responsiveness, meaning whether the learner can understand and address a serious counterargument. These objects should not be collapsed into whether the final answer matches the assessor's own conclusion.

Assessor development is therefore central. Contemporary work on rater training argues for developing judgment rather than treating raters as instruments that should merely reproduce a predetermined score [23]. In this framework, assessors require shared exemplars of strong and weak reasoning, explicit boundary conditions, and practice distinguishing a disagreement about conclusions from a defect in reasoning. Standardization remains important for task interpretation and criteria, but identical moral

conclusions are not themselves the target of standardization.

Feedback adds a temporal dimension. Formative-assessment scholarship treats feedback as useful when learners can interpret information and act on it [24]. A learner who revises after recognizing a factual error, neglected stakeholder, or stronger counterargument may demonstrate responsiveness; a learner who retains a position while directly answering the challenge may do so as well. By contrast, changing an answer solely to mirror examiner preference should not automatically earn additional credit.

For consequential decisions, the score should be accompanied by a brief rationale identifying the response evidence on which the judgment rests. This makes disagreement auditable: another assessor can ask whether the criterion was applied differently or whether the raters genuinely interpreted the learner's reasoning differently. Where the distinction cannot be resolved, uncertainty should remain visible rather than being converted into false precision. The framework therefore favors transparent judgment over the appearance of objectivity produced by unexplained numerical agreement.

The resulting analytic rubric is summarized in **Table 3**.

Table 3. Proposed Analytic Rubric for Justification, Revision, and Responsiveness

Scoring domain	Strong evidence	Concerning evidence	What must not determine the score
Justification	Relevant reasons linked coherently to facts and values	Assertions, contradiction, irrelevant reasons	Agreement with assessor conclusion
Counterargument engagement	Identifies and answers a serious competing reason	Ignores or caricatures alternatives	Mere number of alternatives mentioned
Revision	Changes position when reasons materially warrant change	Changes without explanation or despite unchanged reasons	Compliance with feedback itself
Reasoned retention	Maintains position while addressing new challenge	Repeats conclusion without engagement	Refusal to conform
Uncertainty	States material uncertainty and its effect on judgment	False certainty or indiscriminate hedging	Confidence style alone

Note: This rubric is an original proposed synthesis. Rater-development principles are informed by [23], and the emphasis on feedback-responsive learning is informed by [24].

Validity, bias, and consequential fairness

The framework should be evaluated not only by score reproducibility but by the interpretations and consequences produced by its use. Small assessment differences can amplify through grading and award systems into much larger educational consequences [25]. An ethics assessment that modestly favors one style of expression, one interpretation of professionalism, or one preferred conclusion may therefore have disproportionate effects when embedded in progression, honors, or remediation decisions. This possibility is a reason for surveillance, not evidence that such amplification has already been demonstrated in ethics assessment.

Fairness evaluation should also ask whether performance patterns differ across groups and whether assessment decisions reproduce disparities unexplained by the

intended construct. Contemporary clerkship research demonstrates that grading patterns and prediction errors can vary across demographic groups [26]. Such findings do not establish discriminatory intent and should not be transferred mechanically to ethics assessment, but they support routine examination of subgroup consequences rather than assuming that formally identical criteria are automatically equitable.

For the proposed framework, validity review should therefore remain cyclical: define the construct; inspect tasks for construct-irrelevant demands; examine assessor reasoning; analyze score patterns and consequential decisions; solicit learner experience; and revise tasks, criteria, or governance when credible threats emerge. **Figure 1** represents this manuscript-derived quality-assurance logic.

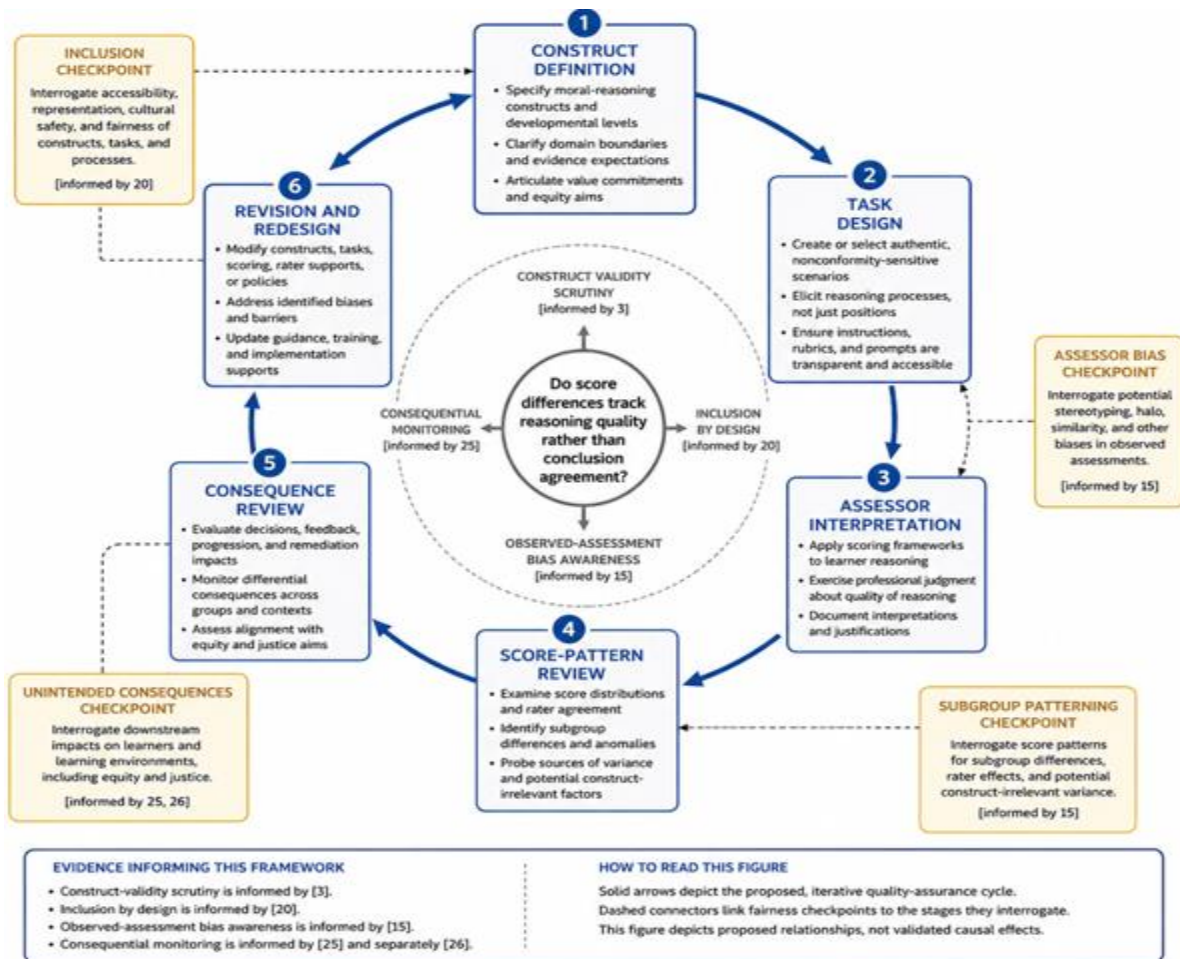


Figure 1. Closed-Loop Fairness Assurance for Nonconformity-Sensitive Moral-Reasoning Assessment

Implementation in health-professions programs

Implementation should begin as a developmental change to an assessment program, not as immediate replacement of established high-stakes decisions. Programmatic-assessment experience shows the feasibility of accumulating longitudinal observations, triangulating evidence, and matching the strength of decisions to the strength of the evidence [27]. Applied here, an institution might initially introduce nonconformity-sensitive tasks in formative ethics courses, compare judgments across methods and assessors, and examine where disagreement reflects criteria ambiguity rather than learner deficiency. Faculty development would need to focus on construct recognition, explicit substantive boundaries, and disciplined interpretation of reasons. Programs should make clear to learners that defensible disagreement does not mean immunity from criticism: factual error, omission of relevant persons or duties, incoherence, or disregard of explicit legal and professional constraints

may still justify lower scores. Equally, educators should be able to explain why a response was weak without relying on “that is not the answer we expected.”

Governance should include learners and other affected stakeholders. A justice-oriented residency assessment redesign demonstrates that co-design can be incorporated into assessment-system development [28]. For ethics assessment, co-design may expose hidden assumptions about what counts as respectful disagreement, articulate reasoning, or appropriate professional tone. It does not guarantee fairness, and participation can itself become tokenistic; its value is as a method for identifying otherwise invisible design assumptions.

Before consequential use, programs should document intended interpretations, assessor preparation, appeal processes, monitoring responsibilities, and conditions triggering review. Implementation is therefore best understood as staged validation under governance rather than deployment of a finished instrument. Programs

should also specify review and appeal routes for cases in which learners plausibly argue that a score reflected disagreement with their conclusion rather than a defect in reasoning. Appeals need not grant learners authority over standards; they provide a procedural check on whether the stated standard was actually applied.

Limitations and validation priorities

The framework is conceptual and remains unvalidated. Existing moral-competence studies illustrate why instrument dependence matters: longitudinal findings obtained with the Moral Competence Test remain contingent on what that instrument measures and how its scores are interpreted [29]. Likewise, recent measurement of medical ethics reasoning using localized intermediate-concept methods demonstrates the feasibility of contextual assessment but also the importance of institutional and cultural localization [30]. Validation should therefore test content representation, learner and assessor response processes, interrater and cross-case generalizability, relations with other measures, subgroup fairness, and downstream consequences. Comparative studies should examine whether conclusion-separated scoring improves construct representation without unacceptable loss of decision consistency. Cross-cultural and multi-institutional work should test whether ostensibly neutral criteria privilege particular languages, traditions, or professional norms. The framework should be revised or rejected if it rewards sophisticated rationalization of indefensible positions, is easily gamed, produces persistent inequity, or cannot support decisions with adequate consistency. Validation should also test an adverse possibility central to this article: criteria intended to protect nonconformity may themselves become a new conformity script if learners discover formulaic language that predictably signals “good reasoning.”

Conclusion

Healthcare ethics education needs assessment that can distinguish good reasoning from agreement without pretending that every conclusion is equally defensible. The framework proposed here treats moral reasoning as a bounded, inspectable process: learners must recognize ethically salient features, construct relevant reasons, engage competing considerations, respond to challenge, and remain accountable to explicit substantive constraints. Fairness therefore requires neither moral

neutrality nor abandonment of common standards. It requires that standards attach to the quality of reasoning rather than silently to preferred conclusions. The framework is best understood as an assessment-design hypothesis for programmatic testing, not a validated instrument. Its value will depend on whether future studies show that it improves construct representation, preserves defensible consistency, identifies bias and consequential inequity, and remains workable across diverse educational and jurisdictional contexts.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Moorlock G, Deans Z, Trimble M. Assessing ethics and law in medical schools: There is no single best answer. *J Med Ethics*. 2025;51(10):684-8. doi:10.1136/jme-2024-110298
2. Anderson HLK, Govaerts M, Abdulla L, Balmer DF, Busari JO, West DC. Clarifying and expanding equity in assessment by considering three orientations: Fairness, inclusion and justice. *Med Educ*. 2025;59(5):494-502. doi:10.1111/medu.15534
3. Kinnear B, St-Onge C, Schumacher DJ, Marceau M, Naidu T. Validity in the next era of assessment: Consequences, social impact, and equity. *Perspect Med Educ*. 2024;13(1):452-9. doi:10.5334/pme.1150
4. Hauer KE, Park YS, Bullock JL, Tekian A. “My assessments are biased!” Measurement and sociocultural approaches to achieve fairness in assessment in medical education. *Acad Med*. 2023;98(8S). doi:10.1097/ACM.0000000000005245
5. Katzer M. The Moral Competence Test in healthcare education: A critique based on informal logic. *Adv Health Sci Educ Theory Pract*. 2026. Online ahead of print. doi:10.1007/s10459-026-10552-w
6. Wienmeister A. Moral reasoning skills: What they are and how they can be furthered in health professions education. *Med Health Care Philos*.

- 2025;28(4):763-74. doi:10.1007/s11019-025-10289-0
7. Andersson H, Svensson A, Frank C, Rantala A, Holmberg M, Bremer A. Ethics education to support ethical competence learning in healthcare: An integrative systematic review. *BMC Med Ethics*. 2022;23(1):29. doi:10.1186/s12910-022-00766-z
8. Doukas DJ, Ozar DT, Darragh M, de Groot JM, Carter BS, Stout N. Virtue and care ethics & humanism in medical education: A scoping review. *BMC Med Educ*. 2022;22(1):131. doi:10.1186/s12909-021-03051-6
9. Friedrich O, Hemmerling K, Kuehlmeier K, Nörtemann S, Fischer M, Marckmann G. Principle-based structured case discussions: Do they foster moral competence in medical students? A pilot study. *BMC Med Ethics*. 2017;18(1):21. doi:10.1186/s12910-017-0181-1
10. Hosseinpour A, Keshmiri F. Professionalism assessment of students in clinical education through Situational Judgment Test (SJT). *BMC Med Educ*. 2024;24(1):1291. doi:10.1186/s12909-024-06205-4
11. Gingerich A, Ramlo SE, van der Vleuten CPM, Eva KW, Regehr G. Inter-rater variability as mutual disagreement: Identifying raters' divergent points of view. *Adv Health Sci Educ Theory Pract*. 2017;22(4):819-38. doi:10.1007/s10459-016-9711-8
12. Schaubert SK, Olsen AO, Werner EL, Magelssen M. Inconsistencies in rater-based assessments mainly affect borderline candidates: But using simple heuristics might improve pass-fail decisions. *Adv Health Sci Educ Theory Pract*. 2024;29(5):1749-67. doi:10.1007/s10459-024-10328-0
13. Thampy H, Callanan N, Ahmed A, Taromsari S. Examiner training and calibration for simulated clinical examinations: A scoping review. *Med Educ*. 2026. Online ahead of print. doi:10.1111/medu.70253
14. Lucey CR, Hauer KE, Boatright D, Fernandez A. Medical education's wicked problem: Achieving equity in assessment for medical learners. *Acad Med*. 2020;95(12S). doi:10.1097/ACM.0000000000003717
15. Ismaeel R, Pusic LJ, Gottlieb M, Chan TM, Oyedokun TO, Thoma B. Bias in observed assessments in medical education: A scoping review. *Acad Med*. 2024;99(12):1438-50. doi:10.1097/ACM.0000000000005794
16. Lam JTH, Coret M, Khalil C, Butler K, Giroux RJ, Martimianakis MA. The need for critical and intersectional approaches to equity efforts in postgraduate medical education: A critical narrative review. *Med Educ*. 2024;58(12):1442-61. doi:10.1111/medu.15425
17. İş-Kara T, Kıyak YS. Large language models in professionalism and ethical decision-making education: A transferable process for generating concordance of professional judgment items. *Acad Med*. 2026. Online ahead of print. doi:10.1093/acamed/wvag147
18. Finn GM, Tai J, Nadarajah VD. Inclusive assessment in health professions education: Balancing global goals and local contexts. *Med Educ*. 2025;59(1):88-96. doi:10.1111/medu.15535
19. Harden RM, Lilley P. A traditional or an innovative approach to assessment: The Assessment PROFILE. *Med Teach*. 2025;47(1):50-7. doi:10.1080/0142159X.2024.2436443
20. Gray L, Sharpe SA. Validity for whom? Unpacking assessment validity from an inclusion perspective. *Med Educ*. 2026. Online ahead of print. doi:10.1111/medu.70279
21. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, et al. Clinical reasoning assessment methods: A scoping review and practical guidance. *Acad Med*. 2019;94(6):902-12. doi:10.1097/ACM.0000000000002618
22. van Wijk EV, van Blankenstein FM, Donkers J, Janse RJ, Bustraan J, Adelmeijer LGM, et al. Does 'summative' count? The influence of the awarding of study credits on feedback use and test-taking motivation in medical progress testing. *Adv Health Sci Educ Theory Pract*. 2024;29(5):1665-88. doi:10.1007/s10459-024-10324-4
23. Tavares W, Kinnear B, Schumacher DJ, Forte M. "Rater training" re-imagined for work-based assessment in medical education. *Adv Health Sci Educ Theory Pract*. 2023;28(5):1697-709. doi:10.1007/s10459-023-10237-8
24. Lipnevich AA, Mattern K, Feddock C. Formative assessment and feedback in medical education: A practical guide: AMEE Guide No. 189. *Med Teach*. 2026;48(6):921-40. doi:10.1080/0142159X.2025.2569623

25. Teherani A, Hauer KE, Fernandez A, King TE Jr, Lucey C. How small differences in assessed clinical performance amplify to large differences in grades and awards: A cascade with serious consequences for students underrepresented in medicine. *Acad Med.* 2018;93(9):1286-92. doi:10.1097/ACM.0000000000002323
26. Mejias PP, Lara G, Duran A, Musci R, Hueppchen NA, Ziegelstein RC, et al. Disparities in medical school clerkship grades associated with sex, race, and ethnicity: A person-centered approach. *Acad Med.* 2024;99(9):1007-15. doi:10.1097/ACM.0000000000005690
27. Emke AR, Torre D, Aagaard EM. Competency-based education and programmatic assessment in undergraduate medical education: Describing the gateway experience of successful design and implementation. *Acad Med.* 2025;100(11):1268-71. doi:10.1097/ACM.0000000000006141
28. Anderson HK, Cabral P, Gerstenzang E, Liverpool C, Gonzalez M, Weiss A, et al. Co-designing a justice-oriented assessment system in a pediatric residency program: Report from the Designing for Equity in Medical Education Project. *Perspect Med Educ.* 2025;14(1):141-8. doi:10.5334/pme.1541
29. Nadolny S, Bruns F, Nowak A, Schildmann J. Moral competency of students at a German medical school—a longitudinal survey. *BMC Med Educ.* 2024;24(1):691. doi:10.1186/s12909-024-05674-x
30. Feng C, Zhang X, Zhang L. Research on medical students' medical ethics reasoning ability—analysis based on the intermediate concept measurement method. *BMC Med Ethics.* 2026;27(1):117. doi:10.1186/s12910-026-01444-0