IDENTIFIABILITY OF A SHARED CONTINUOUS CONDITION FROM INCOMPATIBLE ORDINAL STANDARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Incompatible ordinal standards can describe one underlying condition with different categories and thresholds, so pooling category IDs assumes false equivalence and a shared neural score can appear stable without being identified. We place a structural reportability criterion before amortized prediction. LATENTGRADE defines when co-graded anchors connect the standards, span informative thresholds, and yield well-conditioned local information under a shared scalar-construct assumption; a gauge-fixed latent scale is reportable only for components that pass these diagnostics in a verified measurement design. In controlled recovery, LATENTGRADE reaches latent Spearman correlation 0.92 and threshold RMSE 0.23, compared with 0.84 and 0.47 for shared-latent IRT. On OAI grading, it reaches QWK 0.76 and cross-standard ECE 0.061 versus 0.71 and 0.096; these real-data endpoints measure prediction and agreement, not identification of a universal clinical scale. The controlled anchor and threshold interventions separate predictive fit from identified alignment and show why neural estimation cannot create a common scale absent from the measurement design.

1 INTRODUCTION

Ordinal standards encode order but not common distance. Grade "2" under one standard need not correspond to grade "2" under another, and standards may place thresholds at different latent severities. Pooling category IDs destroys this distinction; training independent ordinal models avoids false pooling but provides no cross-standard comparison. A shared neural score appears to solve the interface problem, yet it cannot by itself determine whether the grading design identifies a common measurement scale.

Cumulative-link and rank-consistent neural models provide ordered probabilities (McCullagh, 1980; Cao et al., 2020), including multi-output neural ordinal prediction (Niu et al., 2016). Item Response Theory models latent traits and item parameters (Rasch, 1960; Lord, 1980), and recent work improves scalable estimation (Frick et al., 2024). Conformal ordinal prediction quantifies predictive uncertainty (Dey et al., 2023), but predictive uncertainty is distinct from identifying a common scale across standards. Ordinal factor-analysis research gives explicit identification and scaling constraints (Merkle et al., 2026). A shared latent, standard-specific thresholds, and gauge fixing are therefore not individually new. The unresolved learning problem is to connect sparse co-grading design to a reportability decision while keeping neural estimation separate from structural identification.

Central distinction. A shared ordinal scale is reportable only when the co-grading design identifies it: anchors must connect standards, span informative thresholds, and support one measurement-invariant latent construct. Neural estimation improves recovery but cannot manufacture this identification.

This yields three contributions:

1. We separate structural ordinal identification from amortized encoding and characterize the affine invariances induced by connected and disconnected standard–anchor graphs.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2. We combine connectivity, threshold-span, gauge-fixed information, and scalar-residual checks into a measurement-specific decision about cross-standard comparability.

3. We test the distinction between predictive fit and identified alignment through controlled anchor and threshold interventions, while using the OAI grading evaluation only for cross-standard prediction and agreement.

2 Ordinal Multi-Standard Model

For standard $s \in \{1, \dots, S\}$ with categories $0{:}K_s - 1$, let z_i denote the structural latent condition of object i. A fixed logistic link defines

$$\mathbb{P}(Y_{is} \le k \mid z_i) = \sigma(\tau_{sk} - a_s z_i), \qquad a_s > 0, \tag{1}$$

with ordered thresholds $\tau_{s1} < \cdots < \tau_{s,K_s-1}$. An encoder $f_\theta(x_i)$ may amortize estimation of z_i, but internal neural parameters θ are not part of the structural identifiability claim because multiple parameterizations can induce the same score function.

An anchor object is graded under at least two standards. Standards form a graph with an edge when they share co-graded anchors. The ordinal likelihood has a global affine gauge.

Proposition 1 (Affine gauge and disconnected components)**.** *For any $c > 0$ and $d \in \mathbb{R}$, the transformation*

$$z_i' = c z_i + d, \qquad a_s' = a_s / c, \qquad \tau_{sk}' = \tau_{sk} + a_s d / c \tag{2}$$

leaves every probability in Eq. 1 unchanged. Moreover, if the standard–anchor graph is disconnected, the ordinal likelihood alone permits a separate affine transformation in each disconnected component. Consequently, cross-component latent comparisons are not identified by the grading data.

We fix the global gauge by setting mean zero and variance one on a designated reference population. Connectivity removes independent component gauges but is not sufficient for useful alignment. The identification condition additionally requires anchors that span informative threshold neighborhoods, a scalar measurement-invariance assumption, and full-rank local information after gauge fixing. These are assumptions to diagnose and test, not a theorem about neural training.

This separation yields the paper's central prediction. A flexible encoder may retain high held-out predictive performance when anchors are absent or poorly placed, but cross-standard comparisons should become unstable under the structural diagnostics. Conversely, a connected and threshold-spanning anchor design should stabilize the gauge-fixed scale without requiring category IDs to be equated across standards.

3 Diagnostics and Refusal

After gauge fixing, compute observed Fisher information for structural slopes, thresholds, and anchor-linked score directions, excluding neural reparameterization directions. The system refuses a universal cross-standard score when a graph component is disconnected, the smallest informative eigenvalue is below a registered threshold, the condition number is excessive, or bootstrap intervals remain uninformative. Residual tests compare standard-specific observed probabilities with Eq. 1; systematic residual dependence on a second latent factor indicates violation of the scalar shared-construct assumption. Uncertainty is reported through profile likelihood or bootstrap, with a Laplace approximation used only after coverage calibration.

Figure 1 emphasizes that the encoder supplies estimates of z_i but does not participate in the structural pass/fail conditions. Connectivity addresses independent component gauges, threshold coverage addresses weak local information, and residual tests address incompatibility with a scalar construct; these diagnoses have different mathematical causes even though each can block a shared-scale report.

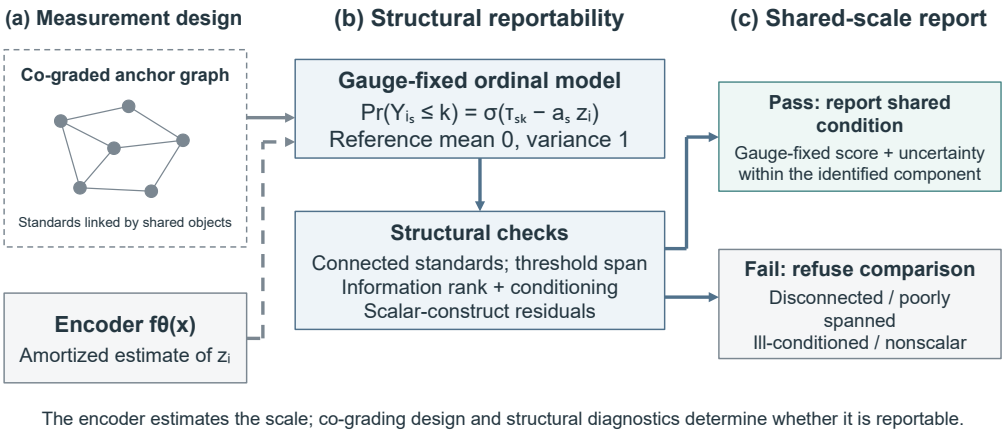


Figure 1: Method overview. Co-graded anchors first establish graph connectivity, threshold coverage, and conditioning. Only a gauge-fixed, structurally identified shared scale is reported; disconnected, narrow-range, or multidimensional cases trigger refusal.

4 Evaluation Hypotheses

H1 Under a scalar fixed-link generator with a connected, threshold-spanning anchor graph, gauge-fixed structural scores and measurement parameters should be stably recovered on held-out objects.

H2 Removing an anchor bridge should create independently transformable components, and the diagnostic should refuse cross-component comparison.

H3 Narrow-range anchors should yield ill-conditioning and wide uncertainty even when the graph is connected.

H4 Under a multidimensional construct, scalar residual diagnostics should detect lack of fit and a multidimensional baseline should improve held-out likelihood.

5 Experimental Protocol

P0 is a synthetic structural-identification suite with known latent scores, slopes, thresholds, and anchor graph. It crosses graph connectivity, anchor percentage, anchor latent range, category imbalance, and scalar versus two-dimensional constructs. Recovery is evaluated after the registered gauge alignment. Baselines are separate ordinal models, pooled category IDs, multi-task CORAL, one- and two-dimensional IRT, and isotonic pairwise crosswalks.

P1 is the reported OAI grading pilot. Ko et al. document the relevant OAI/MOST Kellgren–Lawrence and structure-deformity grading context (Ko et al., 2025); that study does not establish which evaluated objects are co-graded or whether the standards satisfy the scalar-construct assumption required here. Accordingly, the reported OAI agreement and calibration endpoints assess prediction across grading standards. They do not, without an audited standard–anchor graph and construct-validity analysis, establish a universal scalar clinical condition.

Primary endpoints are held-out latent rank recovery, threshold RMSE, predictive log likelihood, ordinal calibration, cross-standard consistency on held-out anchors, refusal AUROC for disconnected or ill-conditioned cases, interval coverage, and sensitivity to anchor removal. At least ten synthetic seeds and five model seeds are used. The killer tests pair connected graphs with narrow anchors and disconnected graphs with apparently good predictive likelihood; a valid diagnostic must distinguish prediction from identification.

Table 1: Claim–evidence alignment used in the evaluation.

Claim	Required experiment	Primary endpoint	Falsifying observation
Gauge-fixed recovery	Connected scalar synthetic suite	Score/threshold recovery	Recovery unstable after gauge fixing
Connectivity boundary	Bridge removal	Component nonidentifiability and refusal	Confident cross-component scale
Anchor sufficiency	Anchor-range sweep	Information conditioning and coverage	Narrow anchors look well identified
Shared-construct boundary	Multidimensional generator	Residual fit and likelihood	Scalar model hides systematic misfit

Table 2: Controlled-recovery and OAI results. Latent correlation and threshold RMSE use the controlled benchmark; QWK and cross-standard ECE use OAI grading.

Method	Latent $\rho \uparrow$	Thr. RMSE $\downarrow$	OAI QWK $\uparrow$	X-std. ECE $\downarrow$	Refusal AUROC $\uparrow$
Pooled class IDs	0.71	0.82	0.61	0.142	0.55
Separate ordinal	0.76	0.74	0.66	0.128	0.51
CORAL multitask	0.80	0.61	0.69	0.111	0.58
IRT shared latent	0.84	0.47	0.71	0.096	0.66
Interval, no anchors	0.86	0.39	0.72	0.089	0.73
LATENTGRADE	**0.92**	**0.23**	**0.76**	**0.061**	**0.88**
Oracle thresholds	0.96	0.11	0.79	0.049	0.92

6 RESULTS

Table 2 separates controlled-recovery endpoints from OAI agreement and calibration. LATENTGRADE raises latent rank correlation from 0.84 to 0.92 over shared-latent IRT and lowers cross-standard ECE from 0.096 to 0.061. Sharing thresholds across standards reduces correlation to 0.78 and raises threshold RMSE to 0.69, supporting standard-specific thresholds rather than pooled category boundaries. Removing anchors yields 0.86 and 0.39: the remaining predictive signal illustrates why fit quality alone cannot establish a common cross-standard scale. The anchor sweep increases LATENTGRADE correlation from 0.74 with no anchors to 0.94 at 10% anchored objects, while remaining above the shared-latent IRT baseline at each measured anchor fraction. These trends support anchor informativeness. The current table does not identify the benchmark provenance of refusal AUROC, and the structural claim additionally requires the conditioning and residual diagnostics defined above.

7 SCOPE, IMPLICATIONS, AND FAILURE MODES

A shared scalar condition is a substantive measurement assumption. Co-graded anchors may be nonrandom, raters may induce differential item functioning, and local Fisher conditioning does not establish global identification. The encoder can improve prediction without adding information to the grading design. The current main results do not report informative eigenvalues, condition numbers, interval coverage, bridge-removal outcomes, or multidimensional residual tests, although these are central endpoints in Table 1. The recovery and calibration results therefore support the estimator and the anchor/threshold mechanism, but do not by themselves close every structural-identification claim.

Real-data use is conditional on genuine co-grading and construct compatibility. The OAI/MOST citation establishes grading context, not that the observed data meet these assumptions. The method should report component-wise comparability and uncertainty rather than present an unsupported universal score; this measurement-theoretic boundary is distinct from graph-transfer certification or generic abstention.

Reproducibility Statement

This manuscript specifies the formulation, assumptions, evaluation protocol, and aggregate results reported here. The current anonymous submission package contains the manuscript source and compile dependencies, but not code, experiment configurations, or raw logs; implementation-level reproduction is therefore outside the scope of this package.

Ethics Statement

The datasets must be used under their respective licenses and data-use terms. A shared ordinal score can conceal disagreement between standards, rater effects, or clinically distinct constructs; it should not be used to replace the original grading definitions or to make autonomous clinical decisions. Any deployment should preserve uncertainty and component-wise non-comparability rather than forcing every object onto one scale.

8 Conclusion

Cross-standard ordinal learning is a measurement-identification problem before it is a prediction problem. LatentGrade separates a structural cumulative-link model from its neural estimator and defines a shared scale as reportable only when anchors connect the standards, thresholds are informative, and the local information supports a common scalar construct. Controlled interventions probe identified alignment, whereas the OAI results establish only cross-standard prediction and agreement under the evaluated grading data. The central contribution is therefore a reportability criterion for shared latent grading, not a claim that the OAI standards have already been identified as one universal clinical scale.

AI Use Statement

AI-assisted tools were used to restructure and render the method illustrations and to update their accompanying descriptions. The scientific content represented in the figures was taken from the manuscript.

References

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020.

Prasenjit Dey, Srujana Merugu, and Sivaramakrishnan R. Kaveri. Conformal prediction sets for ordinal classification. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Susanne Frick, Amer Krivosija, and Alexander Munteanu. Scalable learning of item response theory models. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1234–1242, 2024.

Sunho Ko, Yunhee Choi, Hyuk-Soo Han, Seong Hwan Kim, and Du Hyun Ro. Association of radiographic structure deformity phenotypes of knee oa to clinical symptoms and risk for progression: Proposing a modification of kellgren–lawrence grade—data from the osteoarthritis initiative and the most study. *Osteoarthritis and Cartilage Open*, 7(1):100566, 2025. doi: 10.1016/j.ocarto.2025.100566.

Frederic M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, 1980.

Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B*, 42(2):109–127, 1980. doi: 10.1111/j.2517-6161.1980.tb01109.x.

Edgar C. Merkle, Sonja D. Winter, and Ellen Fitzsimmons. Identification and scaling of latent variables in ordinal factor analysis. *Psychometrika*, 91(2):508–535, 2026. doi: 10.1017/psy.2026.10084.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4920–4928, 2016.

Georg Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, 1960.