

Predicting Item Difficulty and Learner Performance on a Bilingual Gamified Quiz Platform

Aibolat Kulatay¹, Dinara Zhaisanova², Bakhytzhana Kulambayev³, Saken Mambetov^{4*}

Department of Artificial Intelligence and BIG DATA-Faculty of Information Technology and Artificial Intelligence, al-Farabi
Kazakh National University, Almaty, Kazakhstan^{1,2}

Higher School of Telecommunications, Turan University, Almaty, Kazakhstan³

Higher School of Information Technologies, Turan University, Almaty, Kazakhstan⁴

Abstract—Gamified quiz platforms need estimates of question difficulty before response data exist, and learner ratings that support personalized practice. Three related analyses were carried out on separate data sets from a Kazakh–Russian quiz platform. First, lexical and transformer models predicted the observed difficulty of 30,444 items, derived from 177.5 million recorded attempts. Second, item response theory (IRT) models were fitted to 67,990 tournament responses, and text-based predictions were compared with IRT difficulty for 327 new items. Third, online rating models were evaluated on 100,000 duels and 322,046 reconstructed question responses. A fine-tuned XLM-RoBERTa-large ensemble predicted observed difficulty with a Pearson correlation of 0.527, against 0.461 for a lexical baseline (difference 0.066; 95% confidence interval 0.032–0.099). For new tournament items, correlations with two-parameter logistic (2PL) IRT difficulty were 0.24–0.33; the predictions identified the hardest tenth of items better than chance (area under the ROC curve, AUC, 0.66–0.69), but in a cross-fitted calibration simulation they improved on a population prior only before any responses were observed. Observed difficulty correlated at 0.94 between 4,963 Kazakh originals and their Russian translations, which were harder on average. Information criteria did not separate the one- and two-parameter IRT models. Elo predicted duel outcomes with an AUC of 0.828, slightly above tuned TrueSkill and Glicko-2 in the configurations tested, and a learner-by-category term raised the AUC for response correctness from 0.684 to 0.697. Text-based estimates therefore suit coarse screening of new items, and the ratings predict performance but were not validated as measures of proficiency.

Keywords—Educational data mining; item difficulty estimation; item response theory; transformer models; Elo rating; Kazakh language; gamified learning

I. INTRODUCTION

Game-based and gamified learning applications have become a mass channel of informal and exam-oriented learning [1]. Studies of large-scale gamified EdTech platforms have combined psychometric modeling and machine learning to support learner personalization, retention analysis, and adaptive content delivery [2]. The content of such platforms is usually a large bank of multiple-choice questions that grows continuously through editors and user submissions, and their learners generate millions of responses. Two measurement questions arise in such systems. How difficult is a question before anyone has answered it? How well will a learner perform, overall and in particular subjects? Both are central to adaptive practice and testing, where items must be matched to learners [3].

Classical answers to these questions require calibration: item parameters are estimated with item response theory (IRT) from responses collected in pretesting, which is expensive and delays the use of new items. Research on question difficulty estimation from text tries to reduce this cost by predicting difficulty from the item's wording, and transformer language models have recently become the dominant approach [4], [5]. For learners, rating systems originally designed for games, above all Elo, have been adopted in adaptive educational systems because they update learner and item estimates online, need little data, and are easy to deploy [6], [7].

Most of this evidence comes from English-language content and from formal assessments. Kazakhstan offers a contrasting setting. Education is bilingual, with Kazakh and Russian as languages of instruction, and national testing drives great demand for self-study, while Kazakh remains a low-resource and morphologically rich language that presents specific challenges for NLP systems [8]. Evidence for Kazakh–Russian quiz platforms remains limited. This study therefore examines whether text-based difficulty estimates transfer to tournament items and whether Kazakh-specific encoders outperform multilingual models, compares observed difficulty across language versions, and evaluates rating models that predict learner performance.

This study addresses these questions using the production database of Aqyl Battle, a Kazakh mobile quiz-based learning game. Because the available export did not allow learners to be linked across data sources and time periods, the study consists of three related analyses on separate data subsets rather than a single integrated pipeline: Parts 1 and 2 are linked directly through new tournament items, whereas Part 3 is linked to them only at the level of subject categories. The contributions are as follows:

- A benchmark of lexical, embedding-based, and fine-tuned transformer models (including a Kazakh RoBERTa) for predicting the observed difficulty of 30,444 bilingual items from their text, where observed difficulty is derived from cumulative counters of 177.5 million recorded attempts, with leakage-safe splits, a near-duplicate audit, and an ablation of answer options.
- A comparison of 1PL, 2PL and 3PL models on 67,990 tournament responses to 416 items, with checks of connectivity, dimensionality and local dependence, followed by a validation of text-based predictions against

*Corresponding author.

IRT difficulty for 327 new tournament items, a screening task and a cross-fitted calibration simulation that tests whether these predictions reduce the number of responses needed to estimate the difficulty of a new item.

- A descriptive comparison of observed difficulty across 4,963 Kazakh–Russian translation pairs.
- A comparison of Elo, Glicko-2 and TrueSkill with machine-learning classifiers for predicting duel outcomes and response correctness, with sensitivity analyses of the rating-model settings, and an exploratory category-level comparison of difficulty estimated from learner interaction with difficulty estimated from item content.

The aim of the study is to evaluate how well the difficulty of new quiz items can be predicted from their text, to test whether these predictions agree with IRT difficulty estimated from learner responses, and to compare models that predict learner performance. Three research questions are addressed, one for each part of the study:

RQ1. How accurately can the observed difficulty of items be predicted from the item text, and how do lexical, embedding-based, and fine-tuned transformer models compare?

RQ2. To what extent do text-based predictions agree with IRT difficulty estimates for new tournament items, can they be used to screen new items, and do they reduce the number of responses needed to calibrate a new item?

RQ3. How well do rating models (Elo, Glicko-2, TrueSkill) and machine-learning models predict duel outcomes and the correctness of learners' responses?

The remainder of the study is organized as follows. Section II reviews related work. Section III describes the data and Section IV the methods. Section V presents the results, Section VI discusses them together with the limitations, and Section VII concludes.

II. RELATED WORK

A. Question Difficulty Estimation from Text

Estimating item difficulty from text has been studied with increasingly expressive representations, from readability and linguistic features to word embeddings and neural networks [9]. Benedetto et al. proposed R2DE, which regresses IRT parameters of newly generated questions from TF-IDF features [10], and later showed that fine-tuned transformers such as BERT and DistilBERT outperform earlier approaches on multiple-choice questions [11]. In language assessment, Settles et al. estimated item difficulty with machine-learning models to build an adaptive English test without conventional pilot testing [12], and McCarthy et al. combined text-based estimates with incoming responses to jump-start the parameters of new items [13]. Recent surveys conclude that transformer-based models [14] are the strongest general approach, that answer options matter for multiple-choice items, and that most datasets are English and small [4], [5]. Recent work on morphologically rich languages has further shown the value of combining Transformer-based contextual representations with local and sequential feature extraction [15]. Multilingual encoders such as

XLM-RoBERTa [16] make low-resource languages accessible, while recent studies on Kazakh have shown that language-specific transformer models such as KazRoBERTa can outperform multilingual alternatives in morphology-intensive NLP tasks [17]. However, evidence on text-based item difficulty estimation for Kazakh educational content remains limited.

B. Item Response Theory and Online Calibration

IRT models the probability of a correct response as a function of learner ability and item parameters: difficulty in the Rasch or one-parameter logistic model [18], plus discrimination and pseudo-guessing in the 2PL and 3PL models [19]. Parameters are commonly estimated by marginal maximum likelihood, integrating over the ability distribution [20]. Calibrated items are the basis of computerized adaptive testing [3]. Text-based difficulty prediction can therefore be seen as a way to obtain informative starting values for items before calibration data exist. A systematic review of assessment in adaptive learning identified IRT as one of the main theoretical bases of the algorithms that adapt assessment difficulty to the learner [21].

C. Rating Systems and Learner Modeling

The Elo system rates players from pairwise outcomes [22]. Glicko adds a rating deviation that reflects uncertainty, Glicko-2 adds a volatility parameter [23], and TrueSkill formulates skill estimation as Bayesian inference in a factor graph [24]. In education, Elo has been adapted to learner–item interactions, where the learner “plays” against the item. This yields online estimates of both learner skill and item difficulty, with predictive performance close to IRT at far lower cost [6], [7]. Multivariate Elo variants track skill separately for several concepts [25]. Knowledge tracing models, from Bayesian to deep variants, model skill acquisition over time but require detailed sequences of item-level responses [26].

D. Cross-lingual Comparability of Items

When items are translated, differences in difficulty may arise from the translation itself or from the populations that answer each version. Such differences are studied as differential item functioning (DIF), and guidelines for test adaptation require empirical checks of equivalence [27]. Large bilingual item banks with response data are rare, especially for Kazakh and Russian.

III. DATA

A. Platform and Data Source

The data come from the production database of Aqyl Battle, a mobile quiz-based learning game in Kazakhstan, and were provided by Thousand IT Company LLP (MySQL, Laravel back-end) with Kazakh and Russian interfaces. Learners play asynchronous one-to-one duels of five rounds; in each round they answer three four-option multiple-choice questions from a chosen category (e.g., History of Kazakhstan, Geography, Physics & Mathematics, Literacy, English). They also take part in user-created mini-tournaments and in sponsored tournaments. At the time of export, the platform reported 818,386 registered users, 34.4 million games, and 47,035 active questions. The export contained 83 tables. Personally identifiable fields (names, phone numbers, e-mail addresses, tokens and message texts)

were removed by the data owner before export, and user identifiers are pseudonymous. Data from the same platform were analyzed in [2] at the level of users: that study estimated the effect of tournaments on activity, modeled retention and churn, and used a 3PL IRT model to derive learner ability as a predictor of churn. The present study works at the level of items and responses, uses different derived datasets (Section III-B), and addresses different questions: the prediction of item difficulty from text, the choice of an IRT model for calibrating tournament items, and rating models for predicting duel outcomes and response correctness.

1) *Ethics*: The study is a secondary analysis of operational data that were de-identified by the data owner before export. The data provider, Thousand IT Company LLP, approved their use for this research in writing. The study involved no intervention, no contact with users, and no collection of new data; it was not submitted to an institutional ethics committee, so no formal approval or exemption was obtained. To limit risks, no attempt was made to re-identify users; the analyses

used pseudonymous identifiers only, and, because the platform is also used by minors (17.5% of users with a stated age are under 18), only aggregated and item-level results are reported, and no learner-level records are published. One author is the chief executive of the platform (see Conflict of Interest).

Large tables were exported as the first 100,000 rows by primary key. Because primary keys increase over time, each truncated table contains the oldest records of that table rather than a random sample, and different tables cover different time windows: the duels come from the first months of the platform (April–July 2022), the round records from three weeks within that period (23 June–13 July 2022), and the item bank from questions created between June 2022 and May 2025, whose counters, however, accumulate attempts up to the export date. The tournament responses were not truncated. The duel results therefore describe an early cohort of players and may not generalize to the current player base, and the item-bank results describe older items. For this reason, three analyses were designed that each rely on an internally consistent subset rather than on joins across time windows (Fig. 1).

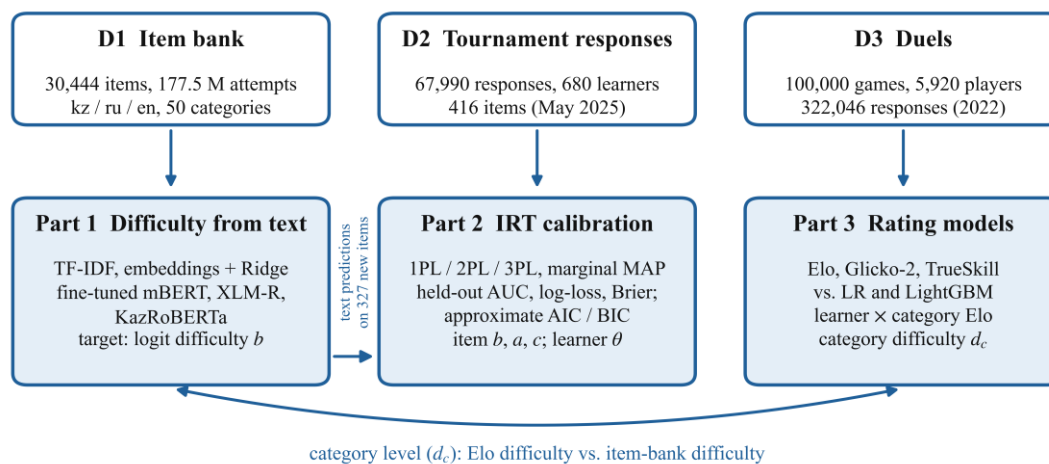


Fig. 1. Overview of the study. Three data subsets (top) feed three modeling parts (bottom). Text-based predictions from Part 1 are validated against IRT difficulties from Part 2 on cold-start items (arrow between Parts 1 and 2), and category difficulties from Part 3 are compared descriptively with item-bank difficulty (bottom arc).

B. Datasets

1) *D1 — Item bank (Part 1)*: Each question stores cumulative counts of correct and incorrect attempts across all game modes. Items with at least 30 attempts were retained: 30,444 items in 50 categories, with a median of 3,441 recorded attempts per item and 177.5 million attempts in total. The mean proportion correct was 0.532 (SD 0.176, range 0.00–0.99). The language of each item was detected from its script: items containing Kazakh-specific letters (ә, Ғ, қ, Һ, ө, ұ, Ү, һ, і) were labeled Kazakh, other Cyrillic items Russian, and Latin-alphabet items English (20,707, 8,051 and 1,686 items). Of the Russian items, 4,963 are translations linked to a Kazakh original. A native speaker labeled a stratified random sample of 200 items; the heuristic agreed for 195 of them (Kazakh 100%, Russian 97.1%, English 90.0%), that is, 97.5% of the unweighted sample. Weighting the per-stratum accuracies by

the stratum sizes in the item bank gives an estimated accuracy of 98.7% (95% interval 96.2–99.3%, from Beta posterior simulation within strata). The errors concerned short Kazakh items without Kazakh-specific letters and arithmetic items without natural-language text. Answer-option texts were available for 4,453 of the items.

2) *D2 — Tournament responses (Part 2)*: Response-level records from sponsored tournaments contain the learner, the item, the selected option, and a timestamp. Each response was scored against the answer key, the first attempt per learner–item pair was kept, and time-outs were treated as incorrect. The data were then filtered iteratively until every item had at least 30 responses and every learner at least 10. The result is 67,990 responses from 680 learners to 416 items (matrix density 0.24, overall accuracy 0.575), collected between 14 and 31 May 2025 in 46 rounds of three tournament series: Famous People (184 items, 6 rounds), History of Kazakhstan (182 items, 6 rounds)

and a regional cup (50 items, 34 regional qualifying rounds). Within a series, the rounds drew their questions from a common pool, so learners who played different rounds answered overlapping item sets, and 225 learners took part in two or three series. The learner–item graph is connected, and every pair of items was answered by at least one common learner (89% of pairs by at least 50). None of the 416 items appears in D1 by identifier, and none had accumulated attempts there. Of the 347 items with stored text, all in Kazakh, 20 have an exact or near-duplicate in D1 (character 3–5-gram TF-IDF cosine similarity ≥ 0.8 ; 10 are identical after normalization), and none is linked to a D1 item as a translation. These 20 items were excluded, leaving 327 items for cold-start validation (177 Famous People, 110 History of Kazakhstan and 40 cup items). The tournament categories are new identifiers, but they overlap in subject with item-bank categories (for example, Famous People and History of Kazakhstan). The validation therefore concerns new items within familiar subject areas, not new domains.

3) *D3 — Duels (Part 3)*: The duel data comprise 100,000 games among 5,920 players from April to July 2022, of which 7.2% ended in a draw. Round records (20,000 games, 23 June–13 July 2022) store, for each of the five rounds of a game, the chosen category and whether each player answered each of the three questions correctly, but not the identifier of each question. From them, 322,046 per-question responses (correct or incorrect) of 2,538 players in 18 categories were reconstructed; rounds without a category or without answers (games that ended early) contributed no responses. In 97.7% of the games, all round records share the creation time of the game, so the responses were ordered by the creation time of the game, then by round and by question position; a sensitivity analysis ordered them by the time at which each round record was last updated. In this order, the first 60% of the responses (193,228) were used only to update ratings, the next 20% (64,409) to tune parameters, and the final 20% (64,409) responses from 1,135 learners for evaluation (Section IV-C).

IV. METHODS

A. Part 1: Text-Based Item Difficulty Estimation

1) *Target*: For item i , let s_i denote the number of correct responses and n_i the total number of recorded attempts. Observed difficulty is defined as the smoothed negative logit of the proportion correct:

$$b_i = -\text{logit}\left(\frac{s_i+0.5}{n_i+1}\right) \quad (1)$$

Larger values indicate harder items, and results are also reported on the proportion-correct scale. This target is called observed difficulty to distinguish it from IRT difficulty. The counters are stored per item and include every attempt in every game mode, including repeated attempts by the same learner. Because they do not record who answered, first attempts cannot be separated from repeated ones, and the composition of learners behind each item is unknown; learners also choose the categories they play, so observed difficulty mixes item properties with the ability of the learners who select each

category. A target adjusted for learner ability, such as an IRT difficulty, would require response-level records, which the export contains for tournament items only (D2) and not for the item bank. Observed difficulty is therefore treated as a population-dependent proxy, and its relation to ability-adjusted difficulty is examined on the tournament items (Section VI-C). The 177.5 million attempts are not independent observations, and the target is not directly comparable with D2, in which only the first response of each learner to an item is used.

2) *Splits*: The data were split 80/10/10 into training, validation, and test sets at the level of item groups. A group joins a Kazakh original with its Russian translation and all items with an identical normalized stem, which prevents leakage through translations or duplicates. Because near duplicates with slightly different wording are not grouped, the partitions were audited after the experiments: for each test item, the maximum cosine similarity of character 3–5-gram TF-IDF vectors to any training or validation item was computed, and the test metrics were recomputed without the items whose similarity was at least 0.8.

3) *Models*: Five families of models were compared:

- Two naive baselines: the global mean and the category mean.
- TF-IDF features (word 1–2-grams and character 2–5-grams) with ridge regression, with and without metadata. The metadata were category, language, length in characters and words, and flags for digits, a question mark, an image, a translation, and a user-submitted item.
- LightGBM [28] on the metadata alone.
- Frozen multilingual sentence embeddings (LaBSE [29], multilingual-E5-large [30] and BGE-M3 [31]) with ridge regression.
- Four fine-tuned encoders: multilingual BERT [14], XLM-RoBERTa-base and -large [16], and the Kazakh RoBERTa model kaz-roberta-conversational [32].

Each encoder used mean pooling followed by a linear regression head. The models were trained on the standardized target using mean squared error and AdamW. The learning rate was 2×10^{-5} , except for XLM-RoBERTa-large, which used 1×10^{-5} . Training used a batch size of 32 and five epochs, with 10% linear warm-up. The maximum sequence length was 64 tokens; with this limit, 5 of the 30,444 inputs (0.02%; mBERT with the category prefix) were truncated, and none with the other tokenizers (median input length 12–22 tokens). Training used bfloat16 mixed precision, and the epoch with the lowest validation RMSE was kept. Two inputs were compared: the stem alone, and the category name followed by the stem. Every configuration was trained with three random seeds; the mean $\pm$ SD over seeds and a seed ensemble (the average of the three predictions) are reported. The best configuration was selected by mean validation RMSE only. Fine-tuning of large transformers is known to be unstable across seeds [33]. One of the six XLM-RoBERTa-large runs (text-only input, seed 3) collapsed to a near-constant predictor: its validation RMSE (0.924) was close to that of a constant training-mean predictor (0.926), and its

validation correlation was 0.03, whereas all other runs reached validation RMSEs of 0.842–0.887. No divergence criterion had been specified in advance; the collapse was noticed after training, and the run was replaced by seed 4 as a post hoc decision. Its effect is therefore reported: the selected configuration is not affected, and for the affected text-only configuration, using seed 3 instead of seed 4 lowers the mean single-seed correlation from 0.489 to 0.331 and the seed-ensemble correlation from 0.517 to 0.508.

4) *Evaluation*: On the held-out test set, RMSE on the logit and proportion-correct scales, Pearson r , and Spearman ρ are reported. Because translations and duplicate stems are grouped (see Splits), 95% confidence intervals were obtained by resampling item groups rather than individual items (2,000 resamples, 2,468 groups in the test set). A paired group bootstrap estimated the difference in Pearson r between the selected model and the strongest classical baseline, and absolute errors were compared with the Wilcoxon signed-rank test. Performance is also reported separately by language, with group-bootstrap intervals for the difference between models. To describe practical value, a screening task was defined: identifying the hardest and the easiest 10% of the test items, evaluated by the AUC of the predictions for membership in each decile and by the precision among the 10% of items predicted to be hardest or easiest.

5) *Ablation with answer options*: On the 4,453 items with option texts, 5-fold group cross-validation compared five inputs, adding one component at a time: the stem; the stem and the key; the stem, key and distractor texts; the stem and key with four surface features of the options (the length of the key relative to the distractors, a numeric key, the share of numeric distractors, and whether the key is the longest option); and all components together. The model was character TF-IDF with ridge regression. This design separates the contribution of the distractor texts from that of the surface features. Out-of-fold predictions were pooled, and 95% CIs for each input and for paired differences between inputs were obtained with a bootstrap over item groups (2,000 resamples).

6) *Cross-lingual difficulty*: For the 4,963 original–translation pairs, correlations between observed difficulties, the mean difference, and the share of pairs with $|\Delta b| > 0.5$ logits were computed. For each pair, Δb was defined as the difficulty of the Russian version minus that of the Kazakh original, so positive values indicate greater observed difficulty for the Russian version. Because pairs from the same category share content and learners, the uncertainty of these quantities was estimated with a bootstrap over the 13 categories (2,000 resamples). For comparison with common practice, a per-pair Wald test with the threshold $|z| > 3.29$ was also computed; its standard error was approximated as:

$$SE(\Delta b) = \sqrt{\frac{1}{n_{ru}p_{ru}(1-p_{ru})} + \frac{1}{n_{kz}p_{kz}(1-p_{kz})}} \quad (2)$$

Here, n_{ru} and n_{kz} denote the numbers of attempts, and p_{ru} and p_{kz} denote the proportions of correct responses for the Russian and Kazakh versions, respectively. Eq. (2) assumes

independent Bernoulli trials and independent samples for the two language versions. Repeated attempts by the same learners violate these assumptions, so the standard errors are too small by an unknown design effect. The share of significant pairs was therefore also computed with the sampling variance inflated by design effects from 2 to 50, and the Wald tests are reported descriptively only.

P -values were adjusted for 4,963 comparisons with the Holm [34] and Benjamini–Hochberg [35] procedures. The 0.5-logit threshold is a practical threshold chosen for this study, not a statistically derived one. This comparison does not control for learner ability and should therefore not be interpreted as a formal test of differential item functioning (DIF).

B. Part 2: IRT Calibration and Cold-start Validation

The 1PL, 2PL, and 3PL models were fitted by marginal maximum a posteriori estimation. Their general response probability is:

$$P(y_{pi} = 1 | \theta_p) = c_i + (1 - c_i)\sigma(a_i(\theta_p - b_i)) \quad (3)$$

Here, y_{pi} indicates whether the learner p answered item i correctly. The parameters θ_p, a_i, b_i and c_i represent learner ability, item discrimination, item difficulty, and pseudo-guessing, respectively. The logistic function is:

$$\sigma(x) = \frac{1}{1 + \exp(-x)} \quad (4)$$

The guessing parameter c_i is fixed at zero in the 1PL and 2PL models. The 1PL model estimates a common discrimination parameter, whereas the 2PL and 3PL models estimate item-specific discriminations.

Learner ability was assumed to follow a standard normal distribution $\theta_p \sim N(0, 1)$ and was integrated out using 41-point Gauss–Hermite quadrature [20]. Normal priors were used for item difficulty $b_i \sim N(0, 3^2)$ and log discrimination $\log a_i \sim N(0, 0.5^2)$, and a Beta prior for the guessing parameter in the 3PL model $c_i \sim \text{Beta}(5, 17)$. Optimization used L-BFGS in PyTorch.

The models were compared by prediction of held-out responses and by information criteria. *Held-out responses*. A random 10% of responses were held out. Item parameters were estimated from the remaining 90% only; the ability posterior of each learner used only that learner's training responses; the two baselines (the item mean, and an additive model of item and person logits) were computed from training responses. AUC, log-loss, and Brier score of posterior-predictive probabilities are reported, and differences between models were assessed with a bootstrap over learners (2,000 resamples), because responses of the same learner are dependent. Random response hold-out evaluates prediction for learners and items that were already observed. Approximately 20% of learners (135 of 680) were therefore also held out: item parameters were estimated from the other learners, and for each held-out learner, ability was estimated from a random half of their responses and used to predict the other half (6,610 responses).

1) *Information criteria*: The models were refitted to all responses in a separate step. AIC and BIC were computed from the marginal log-likelihood evaluated at the MAP estimates,

without the prior term. Because the estimates are penalized, these criteria are approximate and are interpreted together with held-out prediction. BIC used the number of learners ($N = 680$) as the sample size, because learners are the independent units of the marginal likelihood; values with N equal to the number of responses are reported in the text. Unpenalized maximum-likelihood estimation (L-BFGS on unconstrained parameters, without bounds) did not reach a stable solution for the 2PL model: the discriminations of 21 items drifted to extreme values (below 0.05 or above 20), which indicates an almost flat likelihood in these directions. This motivated the weakly informative priors, and the unpenalized fits are not used as a check for model comparison.

2) *Robustness checks*: Posterior standard errors of the 2PL discriminations were obtained from the curvature of the log-posterior at the MAP estimates. The sensitivity of the 3PL model to the prior on the pseudo-guessing parameter was assessed with Beta(5, 17), Beta(2.5, 7.5), Beta(1, 3), and Beta(2, 18) priors. Item fit was summarized by infit and outfit mean squares based on EAP abilities. Local dependence was assessed with Yen's Q3 residual correlations [36] for item pairs answered by at least 50 common learners; because the distribution of Q3 depends on the design [37], it was compared with a null distribution from 30 parametric-bootstrap replications, in which responses were simulated from the fitted 2PL with the observed response pattern and the model was refitted. To check whether a common ability scale is justified across the three tournament series, the unidimensional 2PL was compared with a confirmatory between-item three-dimensional 2PL [38], in which the items of each series measure their own trait, and the three traits are correlated (15 Gauss-Hermite nodes per dimension), and per-series EAP abilities were correlated for learners who took part in two series. Finally, the effect of the IRT model on the cold-start reference was examined by comparing the difficulties of the new items under the 1PL, 2PL, 3PL, and three-dimensional models.

3) *Cold-start validation*: The text-only models from Part 1, trained on D1 alone, predicted the difficulty of the 327 tournament items without duplicates in D1 (Section III-B). Their predictions were correlated with the IRT difficulties b . This tests whether difficulty learned from aggregate platform statistics transfers to IRT difficulty estimated from a different set of responses, for items that were not part of the training data. Because the subject areas overlap, it is a test of transfer to new items rather than to new domains. Confidence intervals for each correlation and for the difference between two models were obtained with a paired bootstrap over items (2,000 resamples). Because the intended use of such predictions is to flag items before they are used, a screening task was added: the AUC of the predictions for membership in the hardest and the easiest 10% of the 2PL difficulties, and the precision among the 33 items predicted to be hardest or easiest, with a bootstrap over items. Chance levels are an AUC of 0.5 and a precision of 0.10.

4) *Calibration experiment*: To test whether text-based predictions are useful as prior information for calibrating a new

item, the estimation of the difficulty of each of the 327 items from $k \in \{5, 10, 20, 40\}$ of its responses was simulated with item-level cross-fitting. The 327 items were divided into five folds. For each fold, the 2PL was refitted to all responses except those to the held-out items, so that the parameters of the anchor items and the learner abilities (EAP estimates from the anchor items only) did not depend on any response to a held-out item. The responses to each held-out item were then split at random into two halves: a reference half, from which the reference difficulty was estimated under a weak prior $N(0, 3^2)$, and a calibration half (at least 58 responses per item), from which 200 subsets of k responses were drawn without replacement. In each subset, the difficulty was estimated by MAP under three priors: the weak prior; a population prior with the mean and variance of the reference difficulties of the items in the other folds; and a text prior $N(\alpha + \beta \cdot \hat{y}, \sigma^2)$, where $\hat{y}$ is the text-based prediction and α, β and σ were estimated on the items of the other folds. The text model was fixed in advance as the validation-selected architecture with text-only input (XLM-R large); KazRoBERTa and TF-IDF were used as sensitivity analyses. The main analysis fixed the discrimination at 1. Sensitivity analyses fixed it at the geometric mean of the anchor discriminations, or estimated discrimination and difficulty jointly (with a log-normal prior on the discrimination), and repeated every design with the reference estimated from all responses to the item, so that the drawn responses overlap with the reference, as in a non-cross-fitted design. RMSE against the reference is reported with bootstrap confidence intervals over items, including $k = 0$, the prior alone. Because the disjoint reference is itself estimated with error, these RMSEs include reference noise; the noise adds the same variance to the mean squared error of every prior, so the comparison between priors is not biased, although differences in RMSE are slightly compressed.

C. Part 3: Rating Models of Learner Performance

1) *Duel outcomes*: Games were processed in chronological order and divided by time into 60% training, 20% validation, and 20% test. Online Elo [22], Glicko-2 [23], and TrueSkill [24] each produced a pre-game probability that the initiator wins. Elo used the logistic curve with scale 400; its K-factor and an initiator advantage were tuned on validation log-loss. In the main configuration, Glicko-2 treated each game as a rating period, with τ tuned on validation, and TrueSkill used its default parameters with a draw probability of 0.07. Because Elo was tuned more extensively, sensitivity analyses tuned TrueSkill on validation log-loss over 36 settings (performance variability β , dynamics factor, initial uncertainty σ_0 and draw probability) and ran Glicko-2 with rating periods of one game, one hour, one day and one week, re-tuning its volatility parameter for each. Draws updated the ratings with a score of 0.5 and were excluded from evaluation. Elo and Glicko-2 return an expected score, $P(\text{win}) + 0.5 \cdot P(\text{draw})$, which was used as the probability that the initiator wins a decisive game; for TrueSkill, the probability that the initiator's performance

exceeds the opponent's was used and, as a sensitivity analysis, the probability of a win conditional on a decisive outcome under the TrueSkill draw margin. The rating models were compared with logistic regression on smoothed historical win rates and game counts, and with LightGBM on the same features plus the Elo logit. AUC, log-loss, Brier score, and accuracy are reported on all decisive test games and on games between experienced players (both with at least 10 previous games). Because the same players take part in many games, confidence intervals for differences in AUC were obtained with a two-way player bootstrap [39], in which every player receives a Poisson(1) weight and every game the product of the weights of its two players, and with a bootstrap over calendar days; the two-way intervals were the wider ones and are reported.

2) *Response correctness by category*: Following the educational adaptation of Elo [6], the probability of a correct response was modeled as:

$$P(\text{correct}) = \sigma(\theta_u - d_c) \quad (5)$$

Here, θ_u denotes learner skill and d_c denotes category difficulty. The logistic function σ is defined in (4). Both estimates were updated online using the following update-factor function:

$$K(n) = \frac{\alpha}{1 + \beta n} \quad (6)$$

The parameters α and β were tuned on the validation data. An extended model included an additional learner-by-category skill term, θ_{uc} .

Here, n is the number of previous updates to the corresponding learner or category parameter. The baselines were running means per category and per learner. The final category difficulties d_c were compared with the mean item-bank difficulty of the same categories (D1) and with the mean difficulty predicted by the Part 1 model. This is the category-level link between the item-content and learner-interaction views of difficulty.

In the extended model, $P(\text{correct}) = \sigma(\theta_u + \theta_{uc} - d_c)$, and after each response y the three terms are updated as $\theta_u \leftarrow \theta_u + K(n_u)(y - P)$, $\theta_{uc} \leftarrow \theta_{uc} + K(n_{uc})(y - P)$ and $d_c \leftarrow d_c - K(n_c)(y - P)$. Responses were processed in chronological order, and each prediction was made before the response was observed. The first 60% of responses (193,228) were used only to update ratings, the next 20% (64,409) to tune α and β on log-loss and the final 20% (64,409 responses from 1,135 learners) for evaluation. Validation selected $\alpha = 0.1$ and $\beta = 0.001$ for the base model, $\alpha = 0.2$ and $\beta = 0.02$ for the extended model, an Elo K-factor of 96 and a Glicko-2 τ of 0.2; TrueSkill used default parameters with a draw probability of 0.07.

D. Implementation

The experiments used Python 3.14, PyTorch 2.11 with CUDA 12.8 on an NVIDIA RTX 5070 Ti laptop GPU, Hugging Face Transformers [40], sentence-transformers, scikit-learn, LightGBM [28], and the TrueSkill package. Hyperparameter grids, baseline settings, and random seeds are documented with the source code, which is available as Supplementary Material

S1 on Zenodo (<https://doi.org/10.5281/zenodo.23268847>) and on GitHub (<https://github.com/sakenski-afk/kazakh-item-difficulty-irt-elo>, release v1.1.0); the analyses added during the revision are provided as separate scripts in the same release. The item-level data are available as Supplementary Material S2 (<https://doi.org/10.5281/zenodo.22960055>); see the Supplementary Materials and Data Availability section.

V. RESULTS

A. Part 1: Text-Based Estimation of Item Difficulty

1) *Overall comparison*: Table I summarizes performance on the 3,045 held-out test items. Predicting the training mean gives RMSE = 0.822 logits, and the category mean already explains part of the variance ($r = 0.349$). The strongest classical model, TF-IDF with metadata and ridge regression, reaches $r = 0.461$ (95% CI 0.420–0.498) and RMSE = 0.733. Frozen sentence embeddings with ridge regression achieved correlations of 0.396–0.410 with text alone and 0.452–0.461 with metadata. These correlations were comparable to those obtained with TF-IDF features.

All four fine-tuned encoders achieved higher mean Pearson correlations than the classical baselines, although they did not all achieve lower RMSE. The configuration selected on validation data, XLM-RoBERTa-large with the category name prepended to the stem, reaches mean $r = 0.510 \pm 0.013$ over seeds (RMSE = 0.717 ± 0.016). Its seed ensemble (last row of Table I) reaches $r = 0.527$ (95% CI 0.475–0.574), RMSE = 0.702 logits and RMSE = 0.149 on the proportion-correct scale. The paired group bootstrap gives $\Delta r = +0.066$ (95% CI 0.032–0.099) over the TF-IDF + metadata baseline and an RMSE reduction of 0.031 logits (95% CI 0.010–0.051); absolute errors are significantly lower (Wilcoxon $p = 2.4 \times 10^{-8}$). Differences between the four encoders are small relative to their seed variability: the seed ensembles of all eight configurations range from $r = 0.494$ to 0.535, with overlapping confidence intervals. The mBERT ensemble achieved the highest test correlation ($r = 0.535$). However, XLM-RoBERTa-large remained the selected configuration because model selection was based on validation RMSE.

2) *Near duplicates and practical value*: In the audit, 229 test items (7.5%) had a training or validation item with a cosine similarity of at least 0.8 (71 items at least 0.9). Without them ($n = 2,816$), the selected ensemble reached $r = 0.514$ (95% CI 0.465–0.562) against 0.449 for TF-IDF with metadata ($\Delta r = 0.065$, 95% CI 0.032–0.098), so the improvement does not depend on near duplicates. In the screening task, the selected ensemble identified the hardest 10% of the test items with AUC = 0.736 (95% CI 0.706–0.767) and the easiest 10% with AUC = 0.800 (0.753–0.838), against 0.689 and 0.760 for TF-IDF with metadata (difference for the hardest decile 0.047, 95% CI 0.019–0.077). Of the items it placed in the predicted hardest decile, 27.0% were in the observed hardest decile (chance 10%) and 16.8% in the easier half. On the proportion-correct scale, 51.2% of its predictions were within 0.10 of the observed proportion correct and 83.3% within 0.20, against 44.3% and 77.6% for the category mean. The gains are thus modest in

absolute terms but improve coarse screening; they concern observed difficulty, not IRT difficulty.

TABLE I. PREDICTION OF ITEM DIFFICULTY ON THE TEST SET (N = 3,045). FINE-TUNED MODELS: MEAN $\pm$ SD OVER THREE SEEDS; THE LAST ROW IS THE SEED ENSEMBLE OF THE CONFIGURATION SELECTED ON VALIDATION DATA. BOLD MARKS THE SELECTED CONFIGURATION

Family	Model	Input	RMSE (logit)	RMSE (p)	Pearson r	Spearman ρ
Baseline	Global mean	–	0.822	0.174	–	–
Baseline	Category mean	–	0.770	0.163	0.349	0.297
Baseline	LightGBM	metadata	0.765	0.161	0.387	0.355
Baseline	TF-IDF + Ridge	text	0.750	0.158	0.411	0.388
Baseline	TF-IDF + Ridge	text + metadata	0.733	0.157	0.461	0.425
Embeddings	LaBSE + Ridge	text + metadata	0.734	0.156	0.452	0.419
Embeddings	mE5-large + Ridge	text + metadata	0.732	0.155	0.453	0.425
Embeddings	BGE-M3 + Ridge	text + metadata	0.729	0.155	0.461	0.438
Fine-tuned	mBERT	category + text	0.736 $\pm$ 0.014	0.158 $\pm$ 0.003	0.515 $\pm$ 0.006	0.484 $\pm$ 0.003
Fine-tuned	KazRoBERTa	category + text	0.720 $\pm$ 0.005	0.155 $\pm$ 0.001	0.497 $\pm$ 0.020	0.453 $\pm$ 0.022
Fine-tuned	XLM-R base	category + text	0.714 $\pm$ 0.003	0.152 $\pm$ 0.001	0.505 $\pm$ 0.006	0.479 $\pm$ 0.008
Fine-tuned	XLM-R large	category + text	0.717 $\pm$ 0.016	0.152 $\pm$ 0.003	0.510 $\pm$ 0.013	0.494 $\pm$ 0.003
Fine-tuned	XLM-R large (ensemble)	category + text	0.702	0.149	0.527	0.508

3) *Effect of the category cue*: Adding the category name increased the mean test correlation for all four encoders. For mBERT, r increased from 0.470 to 0.515; for KazRoBERTa, it increased from 0.476 to 0.497. XLM-R base improved from 0.479 to 0.505, and XLM-R large from 0.489 to 0.510. The short stems (median 44 characters) often omit the subject area, so category names provide additional context.

4) *Language*: The selected model had higher correlations than the classical baseline in all three language groups (Table II). The group-bootstrap interval of the difference excluded zero for Kazakh items (r from 0.453 to 0.503; Δr 95% CI 0.020–0.079) and Russian items (0.425 to 0.530; 0.043–0.173), but not for the 168 English items (0.558 to 0.622; –0.018 to 0.146). In absolute terms, Kazakh items had the lowest error for both models (RMSE 0.684 for the fine-tuned model).

TABLE II. TEST PERFORMANCE BY LANGUAGE: TF-IDF + METADATA BASELINE VERSUS THE SELECTED FINE-TUNED MODEL (SEED ENSEMBLE)

Language	n	r base	r FT	RMSE base	RMSE FT
Kazakh	2,081	0.453	0.503	0.706	0.684
Russian	796	0.425	0.530	0.793	0.742
English	168	0.558	0.622	0.760	0.718

5) *Answer options*: On the 4,453 items with option texts, adding the correct answer increased r from 0.470 to 0.487 (Table III; $\Delta r = 0.017$, 95% CI 0.003–0.031). Adding the distractor texts increased it only to 0.493 ($\Delta r = 0.005$, 95% CI –0.003 to 0.014), whereas the surface features of the options increased it to 0.501 ($\Delta r = 0.014$, 0.007–0.020); with all components, r reached 0.509 ($\Delta r = 0.038$ over the stem alone, 0.020–0.056) and RMSE fell from 0.720 to 0.699. The options therefore carry information beyond the stem, but the gain came

from the key and from simple surface properties of the options; the contribution of the distractor texts was not distinguishable from zero.

TABLE III. ABLATION WITH ANSWER OPTIONS (CHARACTER TF-IDF + RIDGE, 5-FOLD GROUP CROSS-VALIDATION, N = 4,453). S = STEM, K = KEY, D = DISTRATOR TEXTS, F = SURFACE FEATURES OF THE OPTIONS; 95% CI OF R FROM A BOOTSTRAP OVER ITEM GROUPS

Input	RMSE	r	95% CI	ρ
S	0.720	0.470	0.443–0.496	0.446
S + K	0.709	0.487	0.463–0.512	0.472
S + K + D	0.707	0.493	0.467–0.517	0.472
S + K + F	0.703	0.501	0.476–0.525	0.490
S + K + D + F	0.699	0.509	0.484–0.533	0.493

6) *Cross-lingual difficulty*: Across the 4,963 Kazakh–Russian pairs, observed difficulties were highly consistent (Pearson $r = 0.941$, 95% CI 0.919–0.959 from the bootstrap over categories; Spearman $\rho = 0.932$; Fig. 2). Russian versions had lower observed proportions of correct responses: the mean proportion correct was 0.539 versus 0.583, and the mean shift was $\Delta b = +0.170$ logits (95% CI 0.086–0.266); 70.0% of the Russian versions were harder than their originals, and the mean shift was positive in 11 of the 13 categories. For 21.8% of pairs, $|\Delta b|$ exceeded 0.5 logits (95% CI 15.4–29.0%). Under the independence assumption, the per-pair Wald test flagged 62.7% of pairs at $|z| > 3.29$, and 53.0% remained significant after Holm adjustment (75.4% after Benjamini–Hochberg adjustment; median SE 0.065 logits). These shares depend strongly on the unknown design effect of repeated attempts: with the sampling variance inflated by factors of 2, 5 and 10, the Holm-significant share fell to 36.1%, 11.8% and

2.4%. The per-pair tests are therefore not used for inference, and the comparison is reported descriptively through the size of the differences. The shift varied by category. It was negligible for Logos (−0.02) and Transport (−0.02) and largest for Independent Kazakhstan (+0.44), Fashion (+0.34), Famous People (+0.33) and Currencies (+0.30), which are categories tied to local and cultural knowledge. These differences reflect both potential translation-related effects and differences between the learner populations answering the two language versions; the present analysis cannot separate these contributions.

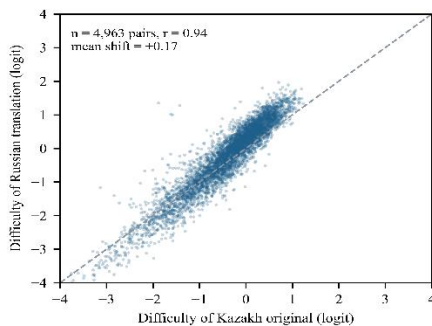


Fig. 2. Difficulty of Kazakh originals and their Russian translations (4,963 pairs). The dashed line marks equal difficulty.

B. Part 2: IRT Calibration and Cold-Start Validation

1) *Model comparison*: All three IRT models predicted held-out responses better than the item-mean baseline (AUC = 0.807) and the additive person-plus-item baseline (AUC = 0.820) (Table IV). Predictive accuracy increased slightly with model complexity: AUC was 0.831 for the 1PL, 0.833 for the 2PL, and 0.834 for the 3PL model, with the same ordering in log-loss and Brier score. These differences are small. With the bootstrap over learners, the AUC difference was 0.0018 between the 2PL and 1PL (95% CI −0.0003 to 0.0035) and 0.0017 between the 3PL and 2PL (−0.0000 to 0.0034), whereas the log-loss differences favored the more complex models (2PL − 1PL: −0.004, 95% CI −0.007 to −0.001; 3PL − 2PL: −0.003, −0.006 to −0.0005). The ordering was the same for the 135 held-out learners (AUC 0.817, 0.821 and 0.823 for the 1PL, 2PL and 3PL models, against 0.806 for the additive baseline). The information criteria disagreed. AIC favored the 2PL (68,564.8, against 69,191.6 for the 1PL and 68,837.1 for the 3PL), whereas BIC favored the 1PL (71,077.3). BIC also favored the 1PL when the number of responses was used as the sample size. Because the criteria are computed at penalized estimates and disagree, the choice between the 1PL and 2PL models remains tentative. It has little effect on the cold-start reference: the difficulties of the 327 new items correlated at 0.94 (Spearman 0.96) between the 1PL and 2PL and at 0.97 between the 2PL and 3PL, and the correlations of the text predictions with difficulty differed by at most 0.03 between the three models (Table V).

The estimated discriminations varied widely (mean $a = 1.08$, range 0.29–3.39). The standard deviation of $\log a$ (0.48) was larger than its typical posterior standard error (0.27), and for 164 items (39.4%) the 95% posterior interval excluded the common 1PL discrimination (0.75). This indicates possible heterogeneity of discriminations beyond estimation noise. Forty-four items (10.6%) had $a < 0.5$, but for none of them did the 95% interval lie entirely below 0.5; they are candidates for review rather than confirmed weak items. The pseudo-guessing parameter of the 3PL model was not well identified: its mean was 0.21 with the Beta(5, 17) prior (prior mean 0.23) and with Beta(2.5, 7.5), but 0.12 with Beta(1, 3) and 0.07 with Beta(2, 18), while held-out AUC changed by at most 0.001. Infit mean squares lay within the conventional range of 0.7–1.3 for 99.8% of the items. Outfit mean squares, which are more sensitive to unexpected responses of learners whose ability is far from the item's difficulty, lay within this range for 88.0% of the items, so about one item in eight showed outlier-sensitive misfit. Yen's Q3 residual correlations averaged −0.004 over 77,003 item pairs, and 12.6% of the pairs had $|Q3^*| > 0.2$ ($Q3^*$ is Q3 centered at its mean). In the parametric-bootstrap null distribution, this share was 10.4% (range 10.1–10.8% over 30 replications), and 99 observed pairs exceeded the 99.9th percentile of the null distribution, against 77 expected. This indicates a small excess of local dependence concentrated in a limited number of item pairs rather than a general violation; pairs from the same series had a mean $Q3^*$ of 0.006, against −0.005 for pairs from different series. Seventy-seven items (18.5%) were answered correctly by more than 90% of participants, so the tournament pool was easy for this population. The marginal reliability of the EAP ability estimates was 0.77.

2) *Common scale and dimensionality*: The between-item three-dimensional 2PL fitted better than the unidimensional 2PL (log-likelihood −33,332.1 vs. −33,450.4 with three additional parameters; AIC 68,334.2 vs. 68,564.8; BIC 72,110.2 vs. 72,327.2) and predicted held-out responses slightly better (AUC 0.835 vs. 0.833). The latent correlation between the Famous People and History of Kazakhstan series was 0.87, but the regional-cup trait correlated only 0.47 and 0.50 with the other two; per-series EAP abilities of learners who took part in two series gave the same picture (disattenuated correlations 0.87, 0.51 and 0.40). A single ability scale is thus a reasonable approximation for the two thematic series but not for the regional cup, which was held one to two weeks later as a regional qualifier with a larger set of participants (549 learners). The difficulties of the 327 new items were nevertheless almost identical under the two models ($r = 0.996$; 0.990–0.999 within series), and the cold-start correlations changed by less than 0.005 (Table V uses the unidimensional 2PL). Without the 40 cup items, the correlations of the fine-tuned encoders and TF-IDF with 2PL difficulty were 0.24–0.30, and within the Famous People ($n = 177$) and History of Kazakhstan ($n = 110$) series they were 0.17–0.34.

TABLE IV. IRT MODELS: PREDICTION OF 10% HELD-OUT RESPONSES AND INFORMATION CRITERIA ON THE FULL DATA (67,990 RESPONSES, 680 LEARNERS, 416 ITEMS). AIC AND BIC USE THE MARGINAL LOG-LIKELIHOOD AT THE MAP ESTIMATES; BIC USES $N = 680$ LEARNERS

Model	AUC	Log-loss	Brier	Parameters	AIC	BIC
Item mean	0.807	0.528	0.176	–	–	–
Item + person (additive)	0.820	0.525	0.174	–	–	–
1PL	0.831	0.499	0.165	417	69,191.6	71,077.3
2PL	0.833	0.495	0.164	832	68,564.8	72,327.2
3PL	0.834	0.491	0.163	1,248	68,837.1	74,480.7

TABLE V. COLD-START VALIDATION: CORRELATION OF TEXT-ONLY PREDICTIONS WITH IRT DIFFICULTY ON 327 TOURNAMENT ITEMS WITHOUT DUPLICATES IN THE ITEM BANK (FINE-TUNED MODELS: SEED ENSEMBLES). THE LAST COLUMN GIVES THE PAIRED-BOOTSTRAP 95% CI OF THE CORRELATION WITH 2PL DIFFICULTY

Model	r (b, 1PL)	r (b, 2PL)	r (b, 3PL)	ρ (b, 2PL)	95% CI of r (b, 2PL)
KazRoBERTa (fine-tuned)	0.340	0.331	0.311	0.346	0.228–0.433
XLM-R base (fine-tuned)	0.311	0.311	0.292	0.335	0.210–0.410
XLM-R large (fine-tuned)	0.260	0.269	0.238	0.267	0.168–0.367
BGE-M3 + Ridge	0.281	0.253	0.255	0.257	0.148–0.357
mBERT (fine-tuned)	0.245	0.251	0.227	0.295	0.151–0.355
TF-IDF + Ridge	0.236	0.240	0.233	0.233	0.137–0.342
LaBSE + Ridge	0.162	0.178	0.158	0.183	0.071–0.287
mE5-large + Ridge	0.110	0.111	0.098	0.109	0.008–0.214

3) *Cold-start validation*: After the exclusion of 20 items that duplicated item-bank content, 327 tournament items remained. All text-only models produced predictions that correlated positively with the 2PL difficulties (Table V, Fig. 3). KazRoBERTa had the highest observed correlation ($r = 0.331$, 95% CI 0.228–0.433; $\rho = 0.346$), followed by XLM-R base ($r = 0.311$). The differences between KazRoBERTa and the other fine-tuned encoders ($\Delta r = 0.02$ –0.08) and between KazRoBERTa and TF-IDF ($\Delta r = 0.091$, 95% CI -0.005 to 0.190) were not statistically significant; KazRoBERTa differed significantly only from LaBSE and mE5-large. On average, the four fine-tuned encoders exceeded TF-IDF by 0.051 (95% CI -0.033 to 0.136). When the 20 duplicated items were included, all correlations were higher (KazRoBERTa $r = 0.363$, TF-IDF $r = 0.271$), and the ordering of the models was similar.

4) *Screening of new items*: Table VI shows how well the text predictions identified the hardest and the easiest 10% of the 327 new items (33 items each). All fine-tuned encoders identified the hardest decile better than chance (AUC 0.66–0.69, lower 95% limits 0.55–0.57); for the easiest decile, KazRoBERTa, XLM-R base and mBERT reached AUC 0.66–0.70, whereas XLM-R large reached only 0.60 (95% CI 0.51–0.70). Among the 33 items predicted to be hardest, 24–36% belonged to the hardest decile (chance 10%) and 9–21% to the easier half. Against the cross-fitted reference of the calibration experiment, the hardest-decile AUCs of the fine-tuned encoders were similar (0.66–0.71). With correlations of about 0.3, the predictions can thus flag candidates for review or help to balance tournament rounds, but they misclassify a substantial share of items and cannot replace response data.

TABLE VI. SCREENING OF NEW ITEMS: IDENTIFICATION OF THE HARDEST AND EASIEST 10% OF THE 327 TOURNAMENT ITEMS (2PL DIFFICULTY) FROM TEXT-ONLY PREDICTIONS; AUC WITH BOOTSTRAP 95% CI, AND PRECISION AMONG THE 33 ITEMS PREDICTED TO BE HARDEST OR EASIEST (CHANCE: AUC 0.5, PRECISION 0.10)

Model	AUC hardest 10% (95% CI)	AUC easiest 10% (95% CI)	Precision, hardest	Precision, easiest
KazRoBERTa (fine-tuned)	0.673 (0.556–0.765)	0.698 (0.619–0.793)	0.30	0.27
XLM-R base (fine-tuned)	0.686 (0.570–0.780)	0.667 (0.578–0.761)	0.30	0.18
XLM-R large (fine-tuned)	0.662 (0.553–0.763)	0.597 (0.512–0.698)	0.36	0.09
mBERT (fine-tuned)	0.665 (0.554–0.762)	0.664 (0.579–0.753)	0.24	0.12
TF-IDF + Ridge	0.581 (0.475–0.706)	0.693 (0.590–0.780)	0.21	0.27

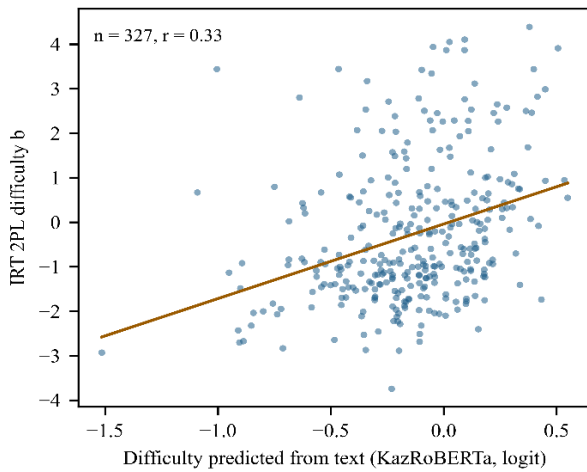


Fig. 3. Cold-start validation: difficulty predicted from text by KazRoBERTa versus 2PL difficulty for 327 tournament items without duplicates in the item bank. The line is the least-squares fit.

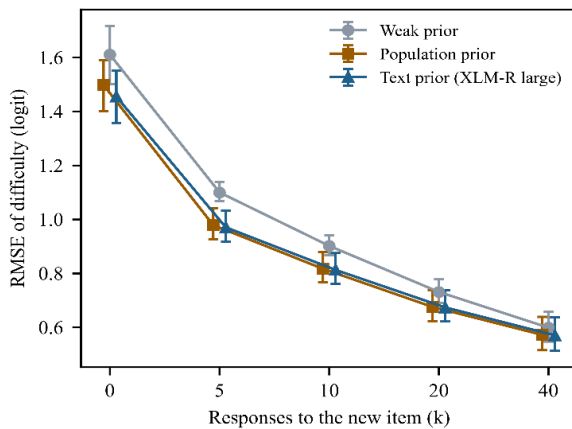


Fig. 4. Calibration experiment (cross-fitted, disjoint reference): RMSE of the difficulty of a new item estimated from k of its responses under three priors; $k = 0$ is the prior alone (327 items, 200 draws per item and k ; error bars are bootstrap 95% CIs over items).

5) *Calibration experiment*: In the cross-fitted design with a disjoint reference (Fig. 4), all informative priors reduced the error of a new item's difficulty compared with the weak prior. With 5 responses, the RMSE was 1.100 logits under the weak prior, 0.979 under the population prior and 0.970 under the text prior; with 20 responses it was 0.731, 0.676 and 0.674. Almost

all of this gain came from the population prior. The text prior was more accurate than the population prior before any response was observed ($k = 0$; RMSE 1.456 vs. 1.498, 95% CI of the difference -0.079 to -0.007), but for $k \geq 5$ the difference was at most 0.009 logits, and its confidence interval included zero at every k . The same pattern held with the discrimination fixed at the mean anchor value (0.96–0.99), with discrimination and difficulty estimated jointly, and with the reference estimated from all responses; the original non-cross-fitted design had given nearly the same values (RMSE 0.868 vs. 0.876 at $k = 5$). KazRoBERTa, the encoder with the highest cold-start correlation, gave the largest gains ($k = 0$: RMSE 1.419; $k = 5$: 0.960 vs. 0.979, 95% CI of the difference -0.033 to -0.005), but these, too, vanished by $k = 10$. In the simulated conditions, text predictions therefore add information beyond the typical difficulty only before responses are available.

C. Part 3: Rating Models of Learner Performance

1) *Duel outcomes*: Results for the 18,268 decisive games in the final 20% of the timeline are shown in Table VII. Elo, with $K = 96$ selected on validation, reached AUC = 0.828 and an accuracy of 74.9%, above logistic regression on win-rate history (AUC = 0.804; Δ AUC = 0.024, 95% CI 0.016–0.033). In the main configurations, Glicko-2 (0.813) and TrueSkill with default parameters (0.799) were weaker than Elo (Δ AUC = 0.016, 95% CI 0.011–0.022, and 0.029, 0.022–0.037). Tuning TrueSkill on validation data (β and σ_0 halved and the dynamics factor quadrupled relative to the defaults) raised its AUC to 0.823, still slightly below Elo (Δ AUC = 0.005, 95% CI 0.002–0.008); conditioning on a decisive outcome through the draw margin changed its AUC by less than 0.001. Longer Glicko-2 rating periods reduced accuracy (AUC 0.799 for one hour, 0.762 for one day and 0.721 for one week), because ratings are then updated only at the end of each period. All intervals come from the two-way player bootstrap. LightGBM on history features plus the Elo logit gave a small overall gain that was not significant (AUC 0.830 vs. 0.828; Δ AUC = 0.002, 95% CI -0.002 to 0.006). For games between experienced players ($n = 13,915$) plain Elo and tuned TrueSkill had the highest AUC (both 0.827). A term for an initiator advantage did not improve validation log-loss within the tested range (-25 to 25 rating points) and was set to zero. The advantage of Elo therefore holds for the configurations tested here and is small relative to a tuned TrueSkill.

TABLE VII. PRE-GAME PREDICTION OF DUEL OUTCOMES IN THE TEST PERIOD (18,268 DECISIVE GAMES); THE LAST COLUMN REFERS TO GAMES BETWEEN EXPERIENCED PLAYERS ($N = 13,915$). TRUESKILL IS SHOWN WITH DEFAULT AND WITH VALIDATION-TUNED PARAMETERS

Model	AUC	Log-loss	Brier	Accuracy	AUC (experienced)
Base rate	0.500	0.693	0.250	0.503	–
Logistic regression (win-rate history)	0.804	0.555	0.187	0.722	0.793
TrueSkill (default)	0.799	0.543	0.183	0.721	0.806
TrueSkill (tuned)	0.823	0.517	0.172	0.743	0.827
Glicko-2	0.813	0.528	0.177	0.732	0.818
Elo	0.828	0.510	0.169	0.749	0.827
LightGBM (history + Elo)	0.830	0.509	0.168	0.751	0.823

2) *Response correctness*: On the 64,409 test responses (1,135 learners), the Elo model with learner skill and category difficulty reached $AUC = 0.684$, and adding a learner-by-category skill raised it to 0.697 (log-loss 0.597; Table VIII). The unrounded AUCs are 0.6836 and 0.6974, so the difference is 0.014 (95% CI 0.011–0.017, bootstrap over learners), although the rounded values differ by 0.013. Running means per category and per learner reached $AUC = 0.607$ and 0.635. Ordering the responses by the time at which each round was completed, instead of the creation time of the game, gave nearly the same results (AUC 0.682 and 0.697; $\Delta AUC = 0.014$, 95% CI 0.012–0.017). The improvement suggests that learner-by-category information contributes to response prediction beyond a single overall learner rating.

TABLE VIII. ONLINE PREDICTION OF PER-QUESTION CORRECTNESS ON THE TEST PERIOD: 64,409 RESPONSES FROM 1,135 LEARNERS (322,046 RESPONSES FROM 2,538 PLAYERS IN 18 CATEGORIES PROCESSED IN TOTAL; PREDICTIONS MADE BEFORE EACH RESPONSE)

Model	AUC	Log-loss	Brier	Acc.
Category mean	0.607	0.639	0.224	0.639
Learner mean	0.635	0.632	0.221	0.644
Elo: $\theta_u - d_c$	0.684	0.605	0.210	0.661
Elo + θ_{uc}	0.697	0.597	0.207	0.663

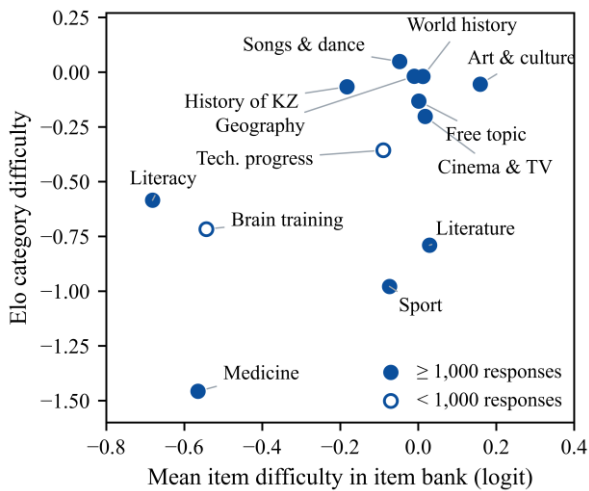


Fig. 5. Category level: Elo category difficulty in duels (Part 3) versus mean item difficulty in the item bank (Part 1) for the 13 categories present in both data sets. Higher values mean harder categories; open circles mark categories with fewer than 1,000 duel responses.

3) *Linking parts 1 and 3*: Thirteen categories appear both in the duel data (2022) and in the calibrated item bank. For them, Elo category difficulty correlated with mean item-bank difficulty at $r = 0.59$ ($p = 0.033$), and with the mean difficulty predicted by the Part 1 model at $r = 0.60$ ($p = 0.031$). The rank correlation was not significant ($\rho = 0.40$, $p = 0.18$). Fig. 5 shows the 13 categories. In both sources, Medicine and Literature were among the easiest categories, and the other categories with at least 1,000 duel responses, except Literature and Sport, lay close together (item-bank difficulty -0.18 to 0.16 logits, Elo difficulty -0.20 to

0.05). Literature and Sport were among the easiest categories in duels but of average difficulty in the item bank. A possible reason is that learners choose the category of each duel round, so these categories may be answered mainly by players who prefer them, whereas the item bank pools all game modes and years. The correlation was sensitive to single categories: leaving out one category at a time gave $r = 0.43 - 0.68$ ($p = 0.015 - 0.17$), with the lowest value when Medicine was excluded. Without the two categories with few responses (Brain training, 33 responses; Technological progress, 288 responses), r was 0.57 ($p = 0.064$) and ρ was 0.28 ($p = 0.40$). With only 13 categories, this link has little statistical power and is therefore treated as an exploratory, descriptive observation rather than a finding.

VI. DISCUSSION

A. Text Carries a Moderate Signal About Observed Difficulty

The selected XLM-RoBERTa-large ensemble improved on TF-IDF with metadata, also after near duplicates were removed, but prediction remained imperfect: item text alone does not capture all sources of observed difficulty. Answer options and category names both added information (Table III), consistent with earlier work on multiple-choice questions [11]; in the ablation, the gain came from the key and from surface properties of the options, whereas the distractor texts added no measurable information. In absolute terms, the gain over TF-IDF with metadata is modest ($\Delta r = 0.066$, RMSE reduction 0.031 logits), but it improves the identification of the hardest and easiest items (AUC 0.736 vs. 0.689 and 0.800 vs. 0.760). Because the target is observed difficulty, these figures describe how well text predicts difficulty for the platform's population of players, not IRT difficulty.

B. Transfer to Tournament Items was Moderate and Did Not Separate the Encoders

On new tournament items, KazRoBERTa had the highest observed correlation with IRT difficulty, but its differences from the other encoders and from TF-IDF were not significant. The results are consistent with, but do not demonstrate, an advantage of language-specific pretraining for new Kazakh content of the kind reported for other Kazakh NLP tasks [17]. Model selection for a production item bank should include an evaluation on new items, because the ranking on a random split need not carry over to new content.

C. Text-Based Predictions Support Only Coarse Screening of New Items

Text predictions trained on cumulative platform counters correlated with IRT difficulties estimated from tournament responses, which indicates transfer between the two data sources. On the new tournament items, observed logit difficulty correlated at 0.94 with the 2PL difficulty and at 0.999 with the 1PL difficulty, and the text predictions correlated almost equally with observed and IRT difficulty (for KazRoBERTa, 0.341 and 0.331). Where all learners in a round answer the same pool of items, population-level and ability-adjusted difficulty therefore rank items similarly. In the item bank, however, learners choose categories, and part of what the text models learn may be the ability of the learners who play each category rather than an item

property; the moderate cold-start correlations are consistent with this. The screening task and the calibration experiment show the practical limits. The predictions identified the hardest new items better than chance, so they can help to flag likely very easy or very hard items, to balance tournament rounds and to give authors early feedback. Once a few responses were available, however, the text prior brought no gain over the population prior in the cross-fitted simulation; with correlations of about 0.3, most of the benefit of an informative prior comes from shrinkage towards the typical difficulty. Higher predictive accuracy, for example from an ability-adjusted training target once response-level data are available for the item bank, would be needed for text predictions to reduce calibration effort.

D. Observed Difficulty Differs Between Language Versions

Observed difficulties were strongly correlated across Kazakh originals and Russian translations ($r = 0.94$). Russian versions nevertheless had higher observed difficulty. The largest positive shifts occurred in Independent Kazakhstan, Fashion, Famous People, and Currencies. Two explanations fit the data, and they cannot be separated without learner-level information. The first is translation effects, such as loss of cues or less familiar terminology. The second is population effects: learners using the Russian interface may have less exposure to Kazakh-specific content. Either way, bilingual item banks should not assume that a translated item keeps the difficulty of its original. About one pair in five differs by more than 0.5 logits, which is enough to move an item between difficulty bands. Because repeated attempts make the per-pair significance tests uninterpretable, these findings are descriptive. They support screening language versions for observed difficulty differences, followed by ability-controlled DIF analysis to determine whether the differences persist among learners with comparable proficiency [27].

E. Choice of IRT Model for Tournament Calibration

Held-out prediction improved slightly from the 1PL to the 3PL model, for held-out responses and for held-out learners, but the AUC differences were within bootstrap uncertainty, and only the log-loss differences were distinguishable from zero; AIC preferred the 2PL and BIC the 1PL, and because the criteria are computed at penalized estimates, the comparison is approximate. The posterior intervals indicate possible heterogeneity of discriminations, but a formal test of the Rasch model was not possible because unpenalized estimation did not yield a stable solution for all item parameters. The pseudo-guessing parameter followed its prior, so the 3PL adds little for these data. Item fit was acceptable for most items, and local dependence was only slightly above its null distribution. The common-scale assumption held approximately for the two thematic series but not for the regional cup, which in practice should be calibrated separately or with a multidimensional model. For this study, these choices hardly affected the difficulties of new items (correlations of 0.94–0.996 between models), so the cold-start conclusions do not depend on them. The 2PL is therefore a reasonable working model, but the choice between the 1PL and 2PL remains tentative. The share of very easy items (18.5% with $p > 0.9$) indicates that tournament content could be targeted better; without these items, the cold-start correlations were lower (0.18–0.25), so part of the agreement comes from recognizing very easy items.

F. Simple Rating Models are Strong Baselines for Predicting Learner Performance

Plain Elo predicted duel outcomes better than Glicko-2 and TrueSkill in all configurations tested, consistent with evidence that Elo is an efficient learner model in adaptive educational systems [6], [7]. The advantage over TrueSkill, however, shrank from 0.029 to 0.005 AUC once TrueSkill was tuned, and Glicko-2 performed best with single-game rating periods; the comparison therefore supports Elo as a strong and simple baseline rather than a general superiority over Bayesian rating systems. The learner-by-category term improved response prediction, which supports subject-specific ratings. These results concern predictive performance only. A high AUC for duel outcomes does not show that the ratings measure educational proficiency, because outcomes also depend on category choice, game experience, and the matching of opponents; establishing them as measures of proficiency would require a comparison with an external assessment, such as national test scores. The category-level link between Elo and item-bank difficulty depended on single categories and remains exploratory.

G. Limitations

The study has the following limitations:

- The export contained the oldest 100,000 rows of large tables, so the three analyses rest on different time windows and cohorts (the duels describe the first months of the platform in 2022) and could not be linked at the learner level; round records do not store item identifiers, so the category-level link rests on 13 points.
- Item-bank difficulty is observed difficulty based on cumulative counters across game modes that include repeated attempts and self-selected categories; it cannot be restricted to first attempts or adjusted for learner ability and is not directly comparable with the tournament data.
- Cold-start validation used 327 items from three tournament series, too few to resolve differences between encoders; the regional-cup items did not share a common ability scale with the other series; and the calibration experiment was a simulation on existing responses rather than a prospective study.
- Cross-lingual differences combine translation and population effects, and their per-pair significance cannot be assessed because repeated attempts make the attempts dependent.
- The rating models were evaluated by predictive accuracy only; their validity as measures of proficiency was not tested against an external assessment, and the platform is a gamified quiz rather than a formal assessment.
- Answer-option texts were available for only 4,453 items, and item language was assigned by a heuristic (estimated accuracy 98.7%).

VII. CONCLUSION

This study evaluated text-based item difficulty prediction, IRT calibration, and online learner ratings on a Kazakh–Russian quiz platform in three related analyses. RQ1: fine-tuned transformer encoders predicted observed item difficulty from text better than lexical and embedding baselines ($r = 0.527$ vs. 0.461 for the selected model, confirmed by a bootstrap over item groups and after the removal of near duplicates); category names and the key and surface properties of the answer options added information, but the correlation remained moderate (RMSE 0.70 logits). RQ2: text-based predictions agreed weakly to moderately with IRT difficulty on 327 new tournament items ($r = 0.24$ – 0.33 for the main models) and identified the hardest items better than chance, but the encoders could not be distinguished statistically, and in a cross-fitted calibration simulation, text-based priors improved difficulty estimates only before any responses were observed. RQ3: plain Elo predicted duel outcomes slightly better than Glicko-2 and TrueSkill in the configurations tested (AUC 0.828), a gradient-boosted model added no significant gain, and a learner-by-category term improved the prediction of response correctness (AUC 0.697 vs. 0.684); these results concern predictive performance, not the validity of the ratings as measures of proficiency. In addition, Russian item versions had greater observed difficulty than their Kazakh originals, but translation effects could not be separated from learner-population differences; among the IRT models, AIC favored the 2PL and BIC the 1PL, and the choice hardly affected the difficulties of new items.

Potential applications include coarse screening of new items by predicted difficulty, screening language versions for observed difficulty differences, and category-specific learner ratings. Future work will use a complete export covering consistent time periods and link dual responses to individual items. This linkage, together with response-level data for the item bank, will support joint estimation of learner ability and item difficulty and an ability-adjusted training target. Further experiments will incorporate the full answer-option data, which may raise predictive accuracy to the level at which text-based priors reduce calibration effort. The relation of learners' category choices to engagement and learning will also be examined.

ACKNOWLEDGMENT

The authors thank Thousand IT Company LLP for providing anonymized data. This research received no external funding.

CONFLICT OF INTERESTS

A. Kulatay is the CEO of Aqyl Battle. The data were provided by Thousand IT Company LLP. The study evaluates measurement models and does not assess the platform's commercial or educational performance. The other authors declare no conflict of interest.

Supplementary Materials and Data Availability

The supplementary materials are deposited in Zenodo, a public long-term repository, and are openly accessible without registration:

Supplementary Material S1 (source code, result tables and figures): <https://doi.org/10.5281/zenodo.23268847> (release v1.1.0, MIT license). The same release is available on GitHub at

<https://github.com/sakenski-afk/kazakh-item-difficulty-irt-elo/releases/tag/v1.1.0>.

Supplementary Material S2 (item-level data): <https://doi.org/10.5281/zenodo.22960055> (item-level difficulty statistics without item texts, with platform identifiers replaced by random identifiers; CC BY 4.0).

Question texts and learner-level records belong to Thousand IT Company LLP; all records are de-identified. They are available from the corresponding author on reasonable request, subject to the data provider's approval and a data-use agreement.

REFERENCES

- [1] J. Zeng, D. Sun, C.-K. Looi, and A. C. W. Fan, "Exploring the impact of gamification on students' academic performance: A comprehensive meta-analysis of studies from the year 2008 to 2023," *British Journal of Educational Technology*, vol. 55, no. 6, pp. 2478–2502, 2024, doi: 10.1111/bjet.13471.
- [2] A. A. Kulatay, D. S. Zhaisanova, N. A. Daurenbayeva, S. Y. Mamanova, and M. Tolegen, "Machine learning for personalized learning in gamified EdTech platforms: Aqyl Battle case," *Academic Scientific Journal of Computer Science*, no. 2, pp. 248–262, 2026, doi: 10.32014/2026.2518-1726.437.
- [3] Y. Zhuang et al., "Survey of computerized adaptive testing: A machine learning perspective," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 48, no. 8, pp. 8744–8763, 2026, doi: 10.1109/TPAMI.2026.3672850.
- [4] L. Benedetto, P. Cremonesi, A. Caines, P. Buttery, A. Cappelli, A. Giussani, and R. Turrin, "A survey on recent approaches to question difficulty estimation from text," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–37, 2023, doi: 10.1145/3556538.
- [5] S. Alkhuzaey, F. Grasso, T. R. Payne, and V. Tamma, "Text-based question difficulty prediction: A systematic review of automatic approaches," *Int. J. Artificial Intelligence in Education*, vol. 34, no. 3, pp. 862–914, 2024, doi: 10.1007/s40593-023-00362-1.
- [6] R. Pelánek, "Applications of the Elo rating system in adaptive educational systems," *Computers & Education*, vol. 98, pp. 169–179, 2016, doi: 10.1016/j.compedu.2016.03.017.
- [7] S. Klinkenberg, M. Straatemeier, and H. L. J. van der Maas, "Computer adaptive practice of Maths ability using a new item response model for on-the-fly ability and difficulty estimation," *Computers & Education*, vol. 57, no. 2, pp. 1813–1824, 2011, doi: 10.1016/j.compedu.2011.02.003.
- [8] L. Baitenova, G. Mukhamejanova, G. Munaitbas, S. Mambetov, and Z. Mukanova, "A multi-branch transformer-enhanced neural framework for joint morphological representation learning," *Computers, Materials & Continua*, vol. 88, no. 3, 2026, doi: 10.32604/cmc.2026.081004.
- [9] L. Štěpánek, J. Dlouhá, and P. Martinková, "Item difficulty prediction using item text features: Comparison of predictive performance across machine-learning algorithms," *Mathematics*, vol. 11, no. 19, Art. no. 4104, 2023, doi: 10.3390/math11194104.
- [10] L. Benedetto, A. Cappelli, R. Turrin, and P. Cremonesi, "R2DE: A NLP approach to estimating IRT parameters of newly generated questions," in *Proc. 10th Int. Conf. Learning Analytics & Knowledge (LAK)*, 2020, pp. 412–421, doi: 10.1145/3375462.3375517.
- [11] L. Benedetto, G. Aradelli, P. Cremonesi, A. Cappelli, A. Giussani, and R. Turrin, "On the application of Transformers for estimating the difficulty of multiple-choice questions from text," in *Proc. 16th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, 2021, pp. 147–157. [Online]. Available: <https://aclanthology.org/2021.bea-1.16/>
- [12] B. Settles, G. T. LaFlair, and M. Hagiwara, "Machine learning-driven language assessment," *Trans. Association for Computational Linguistics*, vol. 8, pp. 247–263, 2020, doi: 10.1162/tacl_a_00310.
- [13] A. D. McCarthy, K. P. Yancey, G. T. LaFlair, J. Egbert, M. Liao, and B. Settles, "Jump-starting item parameters for adaptive language tests," in *Proc. 2021 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2021, pp. 883–899, doi: 10.18653/v1/2021.emnlp-main.67.

- [14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
- [15] G. Munaitbas, L. Baitenova, S. Mambetov, and S. Tussupova, "Development of a framework for analyzing Kazakh language errors in semi-structured data," *Procedia Computer Science*, vol. 269, pp. 69–77, 2025, doi: 10.1016/j.procs.2025.08.260.
- [16] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," in Proc. 58th Annu. Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 8440–8451, doi: 10.18653/v1/2020.acl-main.747.
- [17] L. Baitenova, S. Tussupova, S. Mambetov, G. Munaitbas, and G. Mukhamejanova, "Hybrid artificial intelligence architectures for automatic text correction in the Kazakh language," *Frontiers in Artificial Intelligence*, vol. 8, Art. no. 1708566, 2025, doi: 10.3389/frai.2025.1708566.
- [18] G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen, Denmark: Danish Institute for Educational Research, 1960.
- [19] A. Birnbaum, "Some latent trait models and their use in inferring an examinee's ability," in *Statistical Theories of Mental Test Scores*, F. M. Lord and M. R. Novick, Eds. Reading, MA, USA: Addison-Wesley, 1968, pp. 397–479.
- [20] R. D. Bock and M. Aitkin, "Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm," *Psychometrika*, vol. 46, no. 4, pp. 443–459, 1981, doi: 10.1007/BF02293801.
- [21] A. Ihichr, O. Oustous, Y. E. B. El Idrissi, and A. Ait Lahcen, "A systematic review on assessment in adaptive learning: Theories, algorithms and techniques," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 7, 2024, doi: 10.14569/IJACSA.2024.0150785.
- [22] A. E. Elo, *The Rating of Chessplayers, Past and Present*. New York, NY, USA: Arco, 1978.
- [23] M. E. Glickman, "Example of the Glicko-2 system," Boston University, 2012. [Online]. Available: <http://www.glicko.net/glicko/glicko2.pdf>
- [24] R. Herbrich, T. Minka, and T. Graepel, "TrueSkill™: A Bayesian skill rating system," in *Advances in Neural Information Processing Systems* 19, 2007, pp. 569–576, doi: 10.7551/mitpress/7503.003.0076.
- [25] E. N. Kandemir, J.-J. Vie, A. Sanchez-Ayte, O. Palombi, and F. Ramus, "Adaptation of the multi-concept multivariate Elo rating system to medical students' training data," in Proc. 14th Learning Analytics and Knowledge Conf. (LAK), 2024, pp. 123–133, doi: 10.1145/3636555.3636858.
- [26] S. Shen et al., "A survey of knowledge tracing: Models, variants, and applications," *IEEE Transactions on Learning Technologies*, vol. 17, pp. 1858–1879, 2024, doi: 10.1109/TLT.2024.3383325.
- [27] International Test Commission, "ITC guidelines for translating and adapting tests (second edition)," *International Journal of Testing*, vol. 18, no. 2, pp. 101–134, 2018, doi: 10.1080/15305058.2017.1398166.
- [28] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems* 30, 2017, pp. 3146–3154.
- [29] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT sentence embedding," in Proc. 60th Annu. Meeting of the Association for Computational Linguistics (ACL), 2022, pp. 878–891, doi: 10.18653/v1/2022.acl-long.62.
- [30] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Multilingual E5 text embeddings: A technical report," arXiv:2402.05672, 2024, doi: 10.48550/arXiv.2402.05672.
- [31] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 2318–2335, doi: 10.18653/v1/2024.findings-acl.137.
- [32] B. Sagyndyk, S. Murzakhmetov, and K. Yakunin, "Kaz-RoBERTa conversational technical report," TechRxiv, 2025, doi: 10.36227/techrxiv.175942902.25827042/v1.
- [33] M. Mosbach, M. Andriushchenko, and D. Klakow, "On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines," in Proc. Int. Conf. Learning Representations (ICLR), 2021.
- [34] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.
- [35] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995, doi: 10.1111/j.2517-6161.1995.tb02031.x.
- [36] W. M. Yen, "Effects of local item dependence on the fit and equating performance of the three-parameter logistic model," *Applied Psychological Measurement*, vol. 8, no. 2, pp. 125–145, 1984, doi: 10.1177/014662168400800201.
- [37] K. B. Christensen, G. Makransky, and M. Horton, "Critical values for Yen's Q3: Identification of local dependence in the Rasch model using residual correlations," *Applied Psychological Measurement*, vol. 41, no. 3, pp. 178–194, 2017, doi: 10.1177/0146621616677520.
- [38] M. D. Reckase, *Multidimensional Item Response Theory*. New York, NY, USA: Springer, 2009, doi: 10.1007/978-0-387-89976-3.
- [39] A. B. Owen and D. Eckles, "Bootstrapping data arrays of arbitrary order," *The Annals of Applied Statistics*, vol. 6, no. 3, pp. 895–927, 2012, doi: 10.1214/12-AOAS547.
- [40] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in Proc. EMNLP 2020: System Demonstrations, 2020, pp. 38–45, doi: 10.18653/v1/2020.emnlp-demos.6.