Revista Española
de Educación Médica

Predictive validity of academy mock tests against MIR exam: a retrospective multicohort observational concordance study.

Validez predictiva de los simulacros de academia frente al examen MIR: un estudio observacional retrospectivo multicohorte de concordancia.

Pablo González-Castro[1], Jaime Campos-Pavón[2], Carlos Carazo-Casas[3].

[1] Virgen del Rocío University Hospital: Seville, Andalusia, ES, orcid.org/0009-0003-0077-126X
[2] Academia AMIR, Madrid, España
[3] Ramón y Cajal University Hospital: Madrid, ES, orcid.org/0000-0001-7568-7140

Received: 9/30/25: Accepted: 10/14/25; Published: 10/16/25

Abstract: Introduction: MIR examination is a requirement to access Spain's public specialist training system. This study evaluates, across five years, the concordance between mock exams administered by a MIR preparatory academy and the official MIR results. **Methods:** Candidate records were linked; on the matched intersection we computed Pearson's R and Spearman's ϱ (percentiles, net scores), bias (simulation−official), and RMSE. Annual key performance indicators (KPIs) were pair-weighted and analyses stratified by percentile cohorts (≤27, 28–73, >73); heatmap summarize percentile correlations. **Results:** Across years, weighted percentile correlations typically fell in the 0.71–0.76 range, with small average bias and moderate RMSE. Cohort analyses indicated larger errors at extreme percentiles in some years. Net-score bias weighted by pairs was near zero overall, with year-to-year variability. **Conclusions:** Academy mock tests provide a broadly consistent ranking signal relative to official MIR outcomes. Residual bias and dispersion suggest potential gains from calibration—particularly at cohort extremes—and motivate routine monitoring of concordance KPIs in future cohorts.

Keywords: MIR, mock examinations, predictive validity, concordance, Pearson correlation, Spearman correlation, RMSE, bias

Resumen: Introducción: El examen MIR es un requerimiento para acceder al sistema público de formación sanitaria especializada en España. Este estudio evalúa, a lo largo de cinco años, la concordancia entre los simulacros administrados por una academia preparatoria del MIR y los resultados oficiales del MIR. **Métodos:** Los registros de candidatos se vincularon; sobre la intersección emparejada se calcularon la R de Pearson y la ϱ de Spearman (percentiles y puntuaciones netas), el sesgo (simulacro−oficial) y el RMSE. Los indicadores clave anuales (KPI) se ponderaron por pares y los análisis se estratificaron por cohortes de percentil (≤27, 28–73, >73); los mapas de calor resumen las correlaciones en percentiles. **Resultados:** A lo largo de los años, las correlaciones ponderadas en percentiles se situaron habitualmente entre 0,71 y 0,76, con sesgo promedio pequeño y RMSE moderado. Los análisis por cohortes mostraron errores mayores en los percentiles extremos en algunos años. El sesgo en la puntuación neta ponderado por pares fue globalmente cercano a cero, con variabilidad interanual. **Conclusiones:** Los simulacros de academia proporcionan una señal de clasificación en gran medida consistente con respecto a los resultados oficiales del MIR. El sesgo residual y la dispersión sugieren un margen de mejora mediante calibración —especialmente en los extremos de las cohortes— y motivan la monitorización rutinaria de los KPI de concordancia en futuras promociones.

Palabras clave: MIR, simulacros, validez predictive, concordancia, correlación de Pearson, correlación de Spearman, RMSE, sesgo

1. Introduction

The commonly known "MIR examination" is a demanding assessment that all physicians—Spanish or foreign—must take to access Spain's public specialist training system. MIR usually consists on a 4.5-hour-long examination that includes 210 multiple choice questions with 4 options and only 1 correct answer. Preparation typically entails months of study, most often coordinated by experienced private academies. One of the three biggest academies in Spain is AMIR. A cornerstone of this preparation is the administration of full-length mock tests that aim to recreate the official exam's timing and format.

Although dozens of training centers for specialized healthcare examinations exist internationally, all of them provide mock exams as part of their test preparation, with methodological differences between institutions. However, none of them has published a study similar to the present one. Overseas, in the USMLE, we found several studies comparing performance predictors for Step 1 (1)(2) or Step 2 CK (3)(4). Some of the metrics have been used to predict better performance during residence (5).

In Spain, studies have been conducted on the MIR exam; however, they have not included the most decisive factor in preparation: mock exams. Some authors have emphasized university grades as a predictor (6), or even the students' educational background (7). This study aims to fill a gap in the literature, given that mock examinations are ubiquitous in MIR preparation. We did not find any similar study in Spanish literature, maybe because of business implications related to publication bias.

This study evaluates, across five years, the concordance between mock exams administered by a MIR preparatory academy (AMIR) and the official MIR results. We estimated concordance between academy mock results and official MIR outcomes for 2021–2025, focusing on rank agreement (Pearson and Spearman), systematic bias, error magnitude (RMSE), and subgroup performance by official percentile cohorts. The conclusions drawn from this type of study—whether positive or negative—may be of interest to all those involved in the learning process: Students receive statistical feedback on the validity of the method they have chosen; the departments responsible for preparing mock exams gain insight into their performance; and academy directors can either reinforce the classical model they are currently developing or, in case of adverse results, consider shifting toward new learning models.

2. Methods

Study design and data linkage

Retrospective multi-cohort observational study (2021–2025). As a part of the teaching department at AMIR academy, the researchers conducted the study using mock exams from the academy itself, after a formal request to the responsible department. No data from students or mock exams from other training centers were included. We linked last 5 original academyś mock-exam results to official MIR outcomes, anonymizing data. Mock-exams number 36, 37 and 38 were not analyzed to focus the results on original academyś mock exams, due to their correspondence to passed MIR exams. Exclusion criteria were duplicated registers, candidates with missing values in key variables and candidates who were not in the intersection of both sources, mock examinations and official MIR exam. As a retrospective analysis using the entire available

sample, no prior sample size calculation was performed. Mock-exam data were obtained from internal AMIR registers. Official MIR exam results were obtained from EstimAMIR, a platform where students communicate their exam choices after MIR realization, getting corrected when Spanish Health Ministry assigns correct answers.

MIR exam

The official MIR examination is a national, high-stakes multiple-choice test used to allocate residency positions in Spain. The assessment consists of a single session of 4.5 hours comprising 210 questions with four options and one correct answer, designed to sample core clinical knowledge and reasoning across specialties. The scoring follows a net scheme whereby correct responses add one point, and penalization applies to incorrect responses (three wrong answers subtract one point), yielding a net score that is subsequently normalized and combined with academic record according to Ministry procedures. Candidates sit the exam simultaneously under standardized conditions and the test blueprint is broadly stable year-to-year, which enables meaningful multi-cohort comparisons. In this study we use official percentiles as an outcome that is robust to annual difficulty shifts and net scores to evaluate absolute-level calibration. The number of applicants admitted to the MIR exam and the proportion of AMIR students were as follows:

- **2021**: 14,425 admitted: 2,687 trained with AMIR: **18.6%**
- **2022**: 13,059 admitted: 2,695 trained with AMIR: **20.6%**
- **2023**: 12,251 admitted: 2,090 trained with AMIR: **17.1%**
- **2024**: 13,966 admitted: 2,148 trained with AMIR: **15.4%**
- **2025**: 15,000 admitted: 1,990 trained with AMIR: **13.3%**

The remaining candidates were distributed among those who prepared with other academies, those who did not subscribe to any preparation program, and a small proportion who ultimately did not attend the exam.

Outcomes and metrics. Endpoints:

Primary: annual/global concordance (Pearson's R and Spearman's ϱ on percentiles and net scores, RMSE, and bias for both), summarized by year.

Secondary: cohort-stratified concordance by official percentile (≤27, 28–73, >73) using the same metrics.

For each mock we computed Pearson's R and Spearman's ϱ on percentiles and on net scores (correct answer in both, MIR and mock test, adds one point, three incorrect answers subtract one point). Bias was defined as the mean difference (simulation minus official), and RMSE as the square root of the mean squared error. Annual KPIs were pair-weighted. We also stratified analyses by official percentile cohorts: ≤27 (weak group), 28–73 (standard group), and >73 (strong group).

Statistical analysis

For each mock, we computed Pearson's correlation (R) and Spearman's rank correlation (ϱ) between the mock and official outcomes on two variables: (i) official percentile and (ii) net score. Pearson quantifies linear agreement, whereas Spearman captures monotonic rank agreement and is less sensitive to non-linearities and outliers. Correlations were computed on the intersection of candidates with non-missing values. All analyses were conducted on matched candidate pairs present in both the academy mock dataset and the official MIR dataset for the same year. Names were standardized (lower-casing and whitespace normalization), anonymized and duplicate records within a source were erased. Visualization relied on correlation heatmap among mocks and the official percentile, using a unified color scale (0.60–0.90; center 0.75) to enhance discrimination in the observed range. Each cell is annotated as "R / ϱ". All computations were performed in Python (pandas/numpy/matplotlib). No null-hypothesis significance testing was pre-specified; emphasis is

on effect sizes and their stability across cohorts. Annual summaries were obtained via pair-weighted aggregation across all mocks in the year: R_weighted = $\Sigma(R_i \cdot n_i)/\Sigma n_i$; bias_weighted = $\Sigma(bias_i \cdot n_i)/\Sigma n_i$; RMSE_aggregated = sqrt[$\Sigma(RMSE_i^2 \cdot n_i)/\Sigma n_i$], where n_i is the number of matched candidates for mock i. We report these KPIs globally and by official percentile cohorts (≤27, 28–73, >73). Calibration and dispersion were summarized with bias and RMSE. Bias was defined as the mean difference (mock − official) in percentiles and in net scores. RMSE was the square root of the mean squared error on the same differences, thus reflecting typical magnitude of the deviation irrespective of sign.

Ethics

This study was conducted using anonymized secondary data, and therefore, it did not require approval from an ethics committee, in accordance with national regulations. The research complied with the principles of the Declaration of Helsinki and applicable national regulations regarding the use of anonymous data.

3. Results

Year-weighted correlations between mock ("Sim" in Figure 1) and official percentiles typically ranged 0.71–0.76, with small mean bias and moderate RMSE for both percentiles and net scores (Table 1). Overall stability was high across the five cohorts, with only modest year-to-year variation. By mock, a relatively stable hierarchy of concordance emerged: certain mocks showed consistently higher R/ϱ values (e.g., Sim 40 in multiple years), whereas others remained in the mid-to-upper range (Figure 1). In 2025, a slight reduction in correlation was observed for some mocks compared with prior years, though values remained within acceptable ranges for ranking concordance (figure 1; table 1).

In the percentile-cohort stratification (≤27; 28–73; >73), larger errors at distribution tails were noted in some years, consistent with floor/ceiling effects and reduced separation among candidates in those regions (table 3). This variability did not materially affect relative ranking (table 1).

Figure 1 presents a single consolidated panel summarizing the mock–official relationship by year and mock: color encodes Pearson's R on a unified 0.60–0.90 scale, while cell text shows "R / ϱ" (Spearman). This visualization enables quick comparison of annual patterns and differences across mocks, complementing the detailed values in tables 1 and 2.

Table 1. Annual weighted KPIs (percentile-based).

Year	Total pairs	R weighted percentil	Bias weighted percentil	RMSE aggregated percentil
2021	6512	0.739	-3.962	19.772
2022	6775	0.752	-7.776	20.151
2023	4932	0.763	-8.914	20.359
2024	5286	0.760	-8.741	19.985
2025	4210	0.714	-7.119	20.938

Table 2. Annual weighted KPIs (net-based).

Year	Total pairs	R weighted net	Bias weighted net	RMSE aggregated net
2021	6512	0.709	-1.706	18.547
2022	6775	0.707	-2.199	21.527
2023	4932	0.734	-12.518	24.728
2024	5286	0.731	-15.026	24.880
2025	4210	0.689	8.401	22.667

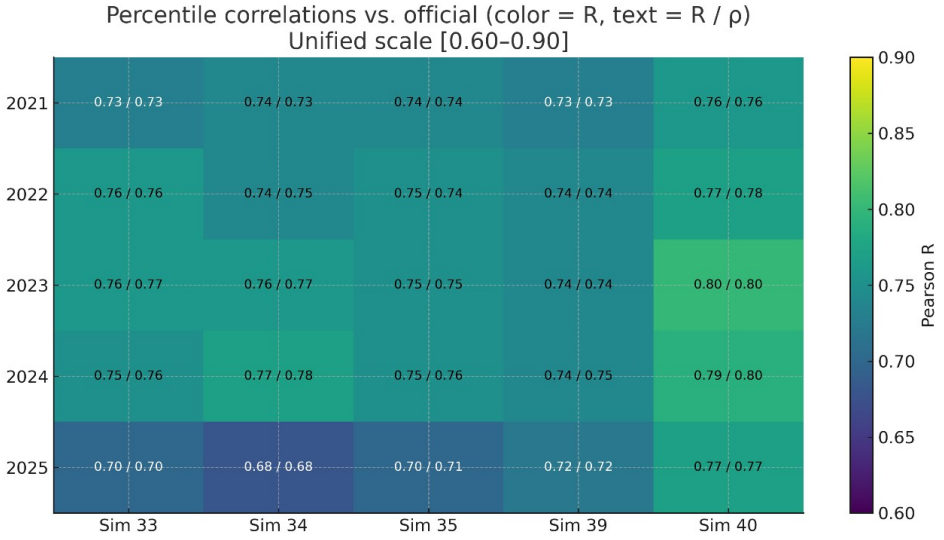

Figure 1. Percentile correlation heatmap (color=Pearson; annotation=R/ϱ).

4. Discussion

The findings of this study confirm that academy mock examinations represent a valid instrument to approximate expected performance on the official MIR examination. The observed correlations (R and ϱ around 0.70–0.80) fall within the upper range and remain consistent across five cohorts, reinforcing the robustness of the results. In medical education research, correlations of this magnitude are generally regarded as acceptable evidence of predictive validity, particularly for high-stakes, wide-scope assessments such as the MIR.

The cohort-stratified analysis adds important nuance. The greater dispersion observed among candidates at the lowest (≤27) and highest (>73) percentiles suggests that mock exams are more reliable in distinguishing performance within the mid-range. This pattern is consistent with well-known floor and ceiling effects: at the lower end, incomplete preparation or random guessing may exert stronger influence, whereas at the upper end, knowledge saturation and small score fluctuations around the maximum limit hinder stable rank differentiation.

Methodologically, the use of official percentiles as the outcome variable neutralizes annual variability in exam difficulty and constitutes a strength compared to approaches relying solely on raw scores. Nevertheless, percentile-based scaling tends to compress variation at the extremes, which partly explains the subgroup limitations observed. Complementary analyses based on net scores provide an additional perspective on absolute calibration, particularly useful to detect systematic bias.

The practical implications are significant. For the academy, the consistent predictive value of mock exams supports their role as a longitudinal monitoring tool and a basis for tailoring individualized study plans. The small annual bias identified could be corrected with simple linear adjustments, preserving the integrity of rank ordering while improving communication of absolute expectations. Cohort-level findings also highlight potential benefits of targeted interventions, such as adaptive item banks or subgroup-focused formative feedback, to enhance predictive accuracy among candidates at distribution tails.

From a comparative perspective, the scarcity of similar analyses in the MIR context underscores the pioneering contribution of this study. While predictive validity of mock or practice tests has been explored in other residency examinations—such as the USMLE in the United States—published evidence remains limited and highly context-dependent.

Studies conducted on the USMLE Step 1 did not find strong predictors in university academic performance (1), although they focused on an earlier stage of education than our study. One study found a correlation—with much smaller sample size—with the percentage of correct answers in a question bank; however, unlike our study, it did not expose students to real standardized exam conditions, but only to individual questions (2). With the same limitation in sample size, other studies also failed to find strong correlations when comparing standardized tests with Step 2 performance (3).

This study therefore helps fill a knowledge gap in Spain and establishes a foundation for future educational research in this domain. We acknowledge several limitations, such as the possibility of data entry errors by students (despite strict exclusion criteria), the restriction to data from a single academy, and unmeasured factors such as curricular variability, which may affect the generalizability of the results. Analysis by percentiles, while robust, compress extremes (ties), whereas net scores are more sensitive to scale.

5. Conclusions

- Academy mock tests provide a stable proxy for official MIR performance in terms of relative ranking.
- Targeted calibration (bias) and attention to cohorts with larger errors could further improve predictive usefulness.

Funding: There was no funding.

Declaration of conflict of interest: Authors work for Healthcademia.

Author contributions: (PGC) Conceptualization, Data curation, Formal analysis, Research, Methodology, Writing, original draft (CCC) Validation, Visualization, Writing–review and editing (JCP) Data acquisition, Project administration

6. References.

1. Puri N, McCarthy M, Miller B. Validity and reliability of pre-matriculation and institutional assessments in predicting USMLE STEP 1 success: lessons from a traditional 2 × 2 curricular model. *Front Med* (Lausanne). **2021**, 8, 798876. https://doi.org/10.3389/fmed.2021.798876

2. Tiffin PA, Paton LW. A predictive model for USMLE Step 1 scores. *Cureus.* **2016**, 8(9), e769. https://doi.org/10.7759/cureus.769

3. Aharonian K, Sanders M, Schlesinger T, Winter V, Simanton E. Predictive validity of preclerkship performance metrics on USMLE Step 2 CK outcomes in the Step 1 pass/fail era. *Adv Med Educ Pract*. **2025**, 16, 43–50. https://doi.org/10.2147/AMEP.S505612

4. Bird JB, Olvet DM, Willey JM, Brenner JM. A generalizable approach to predicting performance on USMLE Step 2 CK. *Adv Med Educ Pract.* **2022**, 13, 939–944. https://doi.org/10.2147/AMEP.S373300

5. Jones A, Benns M, Farmer R. Using resident performance on Step 2 to predict surgical residency success. *Surgery.* **2025**, 179, 108801. https://doi.org/10.1016/j.surg.2024.07.058

6. Baladrón Romero J, López Criado MS, Escudero Carretero MJ, Gil Navarro MV, et al. Resultados obtenidos en la prueba MIR según baremo académico. Convocatorias de 2019 y 2020. *Investig Educ Méd.* **2022**, 11(43), 51–62. https://doi.org/10.22201/fm.20075057e.2022.43.22420

7. Baladrón-Romero J, López Criado MS, Escudero Carretero MJ, Gil Navarro MV, et al. Resultados obtenidos en la prueba MIR de 2021, según nacionalidad y baremo académico. *FEM.* **2022**, 25(5), 205–213. https://doi.org/10.33588/fem.255.1230

Table 3. Weighted KPIs by official percentile cohort and year.

Year	Cohort	Total pairs	R weighted percentil	Bias weighted percentil	RMSE aggregated percentil	Bias weighted net
2021	28–73	3514	0.524	-3.533	19.188	-1.604
2022	28–73	3744	0.565	-7.671	20.086	-2.898
2023	28–73	2464	0.540	-8.618	20.183	-13.097
2024	28–73	2757	0.488	-9.280	20.626	-15.981
2025	28–73	2185	0.456	-6.569	20.983	8.366
2021	>73	2032	0.627	-8.943	17.302	-3.801
2022	>73	2371	0.557	-10.453	19.365	-3.746
2023	>73	1939	0.609	-11.725	19.461	-14.705
2024	>73	1998	0.684	-10.813	17.340	-17.040
2025	>73	1563	0.591	-11.615	18.754	6.154
2021	≤27	966	0.141	4.958	25.812	6.578
2022	≤27	660	-0.096	1.250	23.077	7.327
2023	≤27	529	-0.085	0.016	24.058	-1.802
2024	≤27	531	-0.036	1.847	25.215	-2.490
2025	≤27	462	0.080	5.489	26.885	16.172