Variabilidad en la evaluación de tribunales en oposiciones de Educación Física de Secundaria

Variability in the Assessment by Examination Boards in Secondary Physical Education Teacher Selection Examinations

Vicente Miñana-Signes[1*]
[1]IES Ferrer i Guàrdia, Valencia
*Autor para correspondencia: v.minanasignes@edu.gva.es
DOI: https://doi.org/10.55166/reefd.v440i3.4555

RESUMEN
El proceso de oposición docente es un mecanismo clave para el acceso a la función pública y requiere garantizar la equidad y coherencia en la evaluación realizada por los distintos tribunales calificadores. En la especialidad de Educación Física, la existencia de múltiples tribunales plantea interrogantes sobre la fiabilidad interevaluador y la consistencia del sistema de calificación. El objetivo de este estudio fue analizar la variabilidad en las calificaciones otorgadas por los tribunales evaluadores en el proceso selectivo para el cuerpo de profesorado de Educación Secundaria en Educación Física, convocado en 2025 en la Comunitat Valenciana. La muestra estuvo compuesta por 299 aspirantes (221 hombres y 78 mujeres), distribuidos en 22 tribunales de las provincias de Alicante, Castellón y València. Se analizaron las calificaciones oficiales correspondientes a la parte A (prueba teórica y práctica), la parte B (defensa oral) y la nota final. Se aplicaron análisis estadísticos descriptivos e inferenciales (t de Student, ANOVA unidireccional y factorial, y prueba post hoc de Tukey HSD), así como técnicas exploratorias de reducción de dimensionalidad y agrupamiento (análisis de componentes principales y análisis de clúster mediante k-means). Los resultados mostraron ausencia de diferencias significativas por sexo y provincia, pero sí diferencias estadísticamente significativas entre tribunales en la nota final ($p < .001$). El análisis de conglomerados identificó tres perfiles evaluativos diferenciados según la dispersión interna de las calificaciones. Estos hallazgos evidencian la existencia de variabilidad interevaluador y subrayan la necesidad de reforzar la formación, coordinación y supervisión de los tribunales para mejorar la equidad del proceso selectivo.
Palabras clave: Tribunales evaluadores; oposiciones docentes; educación física; fiabilidad interevaluador; variabilidad evaluativa

ABSTRACT
The teacher selection examination process is a key mechanism for access to the public teaching profession and requires ensuring equity and consistency in the assessment carried out by different examination boards. In the field of Physical Education, the existence of multiple examination boards raises questions about inter-rater reliability and the consistency of the grading system. The aim of this study was to analyze the variability in the grades awarded by examination boards in the 2025 selection process for Secondary Education Physical Education teachers in the Valencian Community. The sample consisted of 299 candidates (221 men and 78 women), distributed across 22 examination boards from the provinces of Alicante, Castellón, and València. The official grades corresponding to Part A (theoretical and practical examination), Part B (oral defense), and the final grade were analyzed. Descriptive and inferential statistical analyses were conducted, including Student's t-test, one-way and factorial ANOVA, and Tukey's HSD post hoc test. In addition, exploratory dimensionality reduction and clustering techniques were applied, including principal component analysis and k-means cluster analysis. The results showed no significant differences by sex or province, but statistically significant differences were found between examination boards in the final grade ($p < .001$). Cluster

analysis identified three differentiated assessment profiles according to the internal dispersion of grades. These findings provide evidence of inter-rater variability and highlight the need to strengthen the training, coordination, and supervision of examination boards in order to improve the equity of the selection process.
Keywords: Examination boards; teacher selection examinations; Physical Education; inter-rater reliability; assessment variability.

INTRODUCCIÓN

En España, el acceso a la profesión docente en la educación pública se articula mediante el sistema de concurso-oposición, concebido para garantizar los principios constitucionales de igualdad, mérito y capacidad (Constitución Española, 1978). En este marco, los aspirantes a profesor de Educación Secundaria deben superar diversas pruebas, cuya calificación recae en tribunales evaluadores designados por especialidad y provincia. En la Comunitat Valenciana, la convocatoria de 2025 para el cuerpo de profesorado de secundaria en la especialidad de Educación Física requirió la constitución de múltiples tribunales debido al elevado número de candidatos, distribuidos territorialmente entre València, Alicante y Castellón.

Este despliegue paralelo, aunque necesario desde el punto de vista organizativo, plantea un reto fundamental: preservar la homogeneidad y comparabilidad de la evaluación cuando diferentes tribunales aplican criterios comunes de forma autónoma. Incluso pequeñas divergencias interpretativas o variaciones en el nivel de exigencia pueden traducirse en diferencias relevantes en las calificaciones otorgadas a aspirantes con méritos similares (Aramendi & Aramendi, 2020; Bustos-Ceruelo & Callejo-Maudes, 2022; Glazerman et al., 2010).

El interés por analizar la variabilidad intertribunal se justifica por sus implicaciones éticas, legales y educativas. Un proceso selectivo equitativo resulta decisivo para que las plazas sean adjudicadas a las personas candidatas más competentes, fortaleciendo la confianza social en el sistema de selección. Por el contrario, si el resultado de un aspirante depende en medida apreciable del tribunal asignado, se debilita el principio de igualdad de oportunidades y aumenta la percepción de arbitrariedad e indefensión. Además, discrepancias injustificadas en la calificación pueden generar conflictos y reclamaciones, desmotivación en el colectivo opositor y, en último término, pérdida de capital humano potencial para el sistema educativo. En el ámbito específico de la Educación Física, este proceso puede constituir también un factor de desgaste para el profesorado interino con trayectorias prolongadas, especialmente cuando la percepción de justicia, transparencia o coherencia del sistema se ve cuestionada (González-Calvo et al., 2023). Estas preocupaciones no se limitan al ámbito nacional: la literatura internacional ha advertido que las inconsistencias evaluativas en procesos de certificación y selección profesional erosionan la confianza institucional y pueden comprometer la legitimidad de las decisiones adoptadas (Darling-Hammond, 2014; Impara & Plake, 1998; OECD, 2013).

Desde el punto de vista normativo, la necesidad de objetividad y uniformidad está explícitamente reconocida. La normativa estatal (Real Decreto 276/2007) establece que los criterios de evaluación deben ser públicos, claros y aplicados de manera homogénea por los tribunales, de modo que la puntuación obtenida sea comparable con independencia del órgano evaluador o la ubicación. En coherencia con estos principios, se espera que baremos y rúbricas se definan con suficiente precisión para sustentar la motivación de cada calificación y minimizar sesgos o favoritismos. Idealmente, dos aspirantes con un desempeño equivalente deberían recibir puntuaciones similares con independencia del tribunal; sin embargo, la evidencia acumulada sugiere que la aplicación real de estándares comunes puede variar, especialmente cuando los criterios no son suficientemente operativos o cuando no existen mecanismos de calibración intertribunal (Koretz, 2008).

Conviene precisar que el presente estudio se centra exclusivamente en las calificaciones obtenidas en la fase de oposición, correspondientes a las pruebas evaluadas directamente por

los tribunales calificadores (parte A y parte B), y no incluye la fase de concurso de méritos. Esta delimitación responde a que la fase de oposición constituye el núcleo evaluativo del proceso selectivo y es donde la intervención directa de los tribunales puede introducir mayor variabilidad interevaluador. Al basarse en la valoración de desempeños complejos mediante juicios profesionales, esta fase resulta especialmente relevante para analizar la coherencia, fiabilidad y equidad de las decisiones evaluativas. En este sentido, la investigación reciente sobre evaluaciones de desempeño de alta escala ha mostrado que la variabilidad atribuible a los evaluadores puede afectar de manera sustantiva a la fiabilidad del sistema, lo que refuerza la necesidad de analizar no solo las calificaciones finales, sino también los posibles patrones de severidad, indulgencia o dispersión entre órganos evaluadores (Haviari et al., 2024).

En la investigación educativa española se han descrito problemas de subjetividad y desigualdad derivados de la falta de criterios suficientemente operativos en los procesos de evaluación. En particular, diversos estudios han señalado que la ausencia de pautas de corrección claramente definidas puede generar discrepancias entre tribunales y afectar a la comparabilidad de las calificaciones obtenidas por aspirantes con niveles de desempeño similares (Aramendi & Aramendi, 2020; Bustos-Ceruelo & Callejo-Maudes, 2022). Este tipo de situaciones ha reforzado la idea de que los procesos selectivos pueden incorporar, de forma no intencionada, sesgos procedimentales que comprometen la equidad del sistema.

Más recientemente, la especialidad de Educación Física ha comenzado a disponer de evidencia empírica contextualizada. En la Comunitat Valenciana, el estudio de Miñana-Signes (2025), centrado en la convocatoria de 2023, identificó diferencias significativas entre tribunales y patrones diferenciados de dispersión interna aún bajo criterios y normativa comunes. Dichos resultados apoyan la hipótesis de que el factor tribunal puede constituir una fuente relevante de variabilidad evaluativa y sugieren la pertinencia de profundizar en la coherencia y equidad del sistema en convocatorias posteriores (Miñana-Signes, 2025). Asimismo, el análisis de la percepción de los miembros de tribunales en procesos selectivos docentes ha puesto de relieve la importancia de clarificar los criterios de evaluación, reforzar la formación de los evaluadores y avanzar hacia procedimientos más coherentes con las competencias profesionales que se pretenden seleccionar (Miñana-Signes & Monfort-Pañego, 2025).

Estos diagnósticos encuentran respaldo en otros contextos. Aramendi y Aramendi (2020) han caracterizado el sistema de oposición como un procedimiento con componentes de subjetividad, destacando variaciones sustantivas entre tribunales y una insuficiente concreción pública de los criterios, y proponiendo modelos de selección más integrales y transparentes. De modo similar, Bustos-Ceruelo y Callejo-Maudes (2022) evidenciaron, en un estudio de caso en Castilla y León, que la probabilidad de éxito podía depender en gran medida del tribunal evaluador, subrayando el impacto de la asignación a una comisión concreta sobre la nota final. En la misma línea, distintas aportaciones han señalado que la ausencia de estándares de corrección suficientemente detallados facilita respuestas dispares para desempeños comparables, incrementando la percepción de arbitrariedad (Perrenoud, 2008).

En consecuencia, la literatura especializada ha recomendado diversas estrategias para mitigar la aleatoriedad y mejorar la fiabilidad interevaluador: definición de criterios más objetivos y operativos, uso de rúbricas detalladas, formación específica de los miembros de los tribunales en evaluación estandarizada y, en algunos casos, propuestas de profesionalización o independencia de los cuerpos evaluadores (Aramendi & Aramendi, 2020; Glazerman et al., 2010; Haviari et al., 2024; Miñana-Signes & Monfort-Pañego, 2025). Asimismo, algunas administraciones han incorporado mecanismos orientados a reducir la variabilidad entre tribunales, por ejemplo ajustando el número de aprobados por tribunal en proporción a los aspirantes asignados, de forma que la competencia se establezca dentro de cada comisión. Sin embargo,

continúa abierto el debate sobre la efectividad real de estas medidas y sobre cuál sería un modelo óptimo de selección docente que equilibre objetividad, viabilidad operativa y capacidad predictiva del desempeño profesional (Darling-Hammond et al., 2017).

En este marco, el presente estudio se centra en la convocatoria de 2025 del proceso selectivo para el cuerpo de profesorado de educación secundaria en la especialidad de Educación Física en la Comunitat Valenciana, con el objetivo de analizar la variabilidad en la evaluación entre los tribunales participantes. Se pretende comparar las calificaciones otorgadas en las distintas pruebas (parte A: teórica y práctica; parte B: defensa oral) y en la nota final, explorando posibles divergencias asociadas al tribunal, así como su relación con variables como el sexo y la provincia. Además, se busca clasificar a los tribunales según su comportamiento evaluativo mediante análisis exploratorios de reducción de dimensionalidad y agrupamiento, con el fin de identificar perfiles diferenciados de calificación. A partir de esta finalidad, se plantea como hipótesis que no existirán diferencias estadísticamente significativas en la nota final en función del sexo; y, en cambio, se anticipa que podrían observarse diferencias en las calificaciones entre tribunales, y potencialmente entre provincias, que evidenciarían variabilidad interevaluador. Asimismo, se espera que las puntuaciones de las distintas fases se relacionen de forma consistente con la nota final, lo que permitiría valorar la coherencia interna del proceso. Finalmente, se postula que el análisis de conglomerados permitirá identificar perfiles evaluativos diferenciados, lo que aportaría evidencia adicional sobre posibles divergencias prácticas en la aplicación de los criterios de corrección.

METODOLOGÍA

Diseño del estudio

Se realizó un estudio cuantitativo con un diseño no experimental, ex post facto y de carácter descriptivo-comparativo, orientado a analizar la equidad evaluativa y la homogeneidad intertribunales en el proceso selectivo de acceso al cuerpo de Profesorado de Educación Secundaria en la especialidad de Educación Física, convocado en 2025 por la Conselleria de Educación, Cultura, Universidades y Empleo de la Generalitat Valenciana.

El estudio se basa en el análisis de resultados ya producidos tras la finalización del proceso selectivo, sin manipulación de variables, a partir de fuentes documentales oficiales. El análisis se circunscribe exclusivamente a la fase de oposición (parte A y parte B), evaluada directamente por los tribunales calificadores, quedando excluida la fase de concurso de méritos por no formar parte del objeto de estudio.

Participantes

La población estuvo formada por las 1.070 personas presentadas al procedimiento selectivo. La muestra final incluyó a 299 aspirantes (27,9%) que completaron ambas fases de la oposición y disponían de calificación final registrada. De ellos, 221 fueron hombres (73,9%) y 78 mujeres (26,1%). Los aspirantes fueron evaluados por 22 tribunales distribuidos en las tres provincias de la Comunitat Valenciana: València (53,8%, n = 161), Alicante (25,8%, n = 77) y Castellón (20,4%, n = 61), conforme a la organización administrativa establecida en la convocatoria oficial.

Instrumentos y procedimiento

Los datos proceden de las actas oficiales de calificaciones publicadas por los tribunales en la página web institucional de la Conselleria de Educación, Cultura, Universidades y Empleo (2026). Se registraron las puntuaciones obtenidas en la parte A (prueba teórica y práctica), la parte B (defensa oral) y la nota final de la fase de oposición, así como el tribunal evaluador, la provincia y el sexo del aspirante.

La información fue organizada en una base de datos anonimizada y depurada para garantizar la consistencia de los registros. Al tratarse de datos administrativos públicos y anonimizados, no fue necesario consentimiento informado individual.

Análisis estadístico

Se calcularon estadísticos descriptivos (media y desviación típica) y se aplicaron análisis inferenciales mediante pruebas t de Student y análisis de varianza (ANOVA) unidireccional y factorial, comprobándose la homogeneidad de varianzas mediante la prueba de Levene. Las comparaciones post hoc se realizaron con la prueba de Tukey HSD.

Asimismo, se analizaron las correlaciones de Pearson entre las distintas dimensiones evaluativas. Para identificar perfiles de comportamiento evaluador entre tribunales se aplicaron técnicas exploratorias de reducción de dimensionalidad y agrupamiento, combinando un análisis de componentes principales (PCA) y un análisis de conglomerados mediante k-means. Se consideraron estadísticamente significativos los valores de $p < .05$.

RESULTADOS

Resultados descriptivos de las calificaciones

El análisis descriptivo de las calificaciones de la fase de oposición muestra diferencias claras entre las dimensiones evaluadas. La parte B (defensa oral) presentó la media más elevada ($M = 7.81$; $DT = 1.44$), mientras que la parte práctica ($M = 3.09$; $DT = 0.85$) y la parte teórica ($M = 3.41$; $DT = 0.83$) registraron puntuaciones inferiores. La media global de la parte A se situó en 6.50 puntos ($DT = 1.16$), y la nota final alcanzó una media de 7.15 ($DT = 1.07$), lo que indica una dispersión moderada en el rendimiento global de los aspirantes.

Comparación de las calificaciones en función del sexo

No se observaron diferencias estadísticamente significativas en la nota final en función del sexo del aspirante. La media obtenida por los hombres ($M = 7.17$; $DT = 1.04$) fue similar a la de las mujeres ($M = 7.11$; $DT = 1.16$), sin diferencias significativas en la prueba t ($t = 0.38$; $p = .704$), resultado confirmado mediante ANOVA ($F = 0.16$; $p = .689$).

Comparación de las calificaciones en función de la provincia

Tampoco se identificaron diferencias significativas según la provincia del tribunal evaluador. Aunque descriptivamente Alicante presentó una media ligeramente inferior ($M = 7.00$) respecto a Castellón ($M = 7.23$) y València ($M = 7.20$), el análisis de varianza no mostró efectos significativos ($F = 1.62$; $p = .200$), y las comparaciones post hoc de Tukey confirmaron la ausencia de diferencias relevantes entre provincias.

Comparación de las calificaciones en función del tribunal

En contraste con los resultados anteriores, el tribunal evaluador emergió como una fuente relevante de variabilidad. Las medias de la nota final oscilaron entre 6.58 puntos (tribunales V10 y C2) y 8.04 puntos (V9), con diferencias notables también en la dispersión interna de las calificaciones.

El ANOVA unidireccional mostró un efecto significativo del tribunal sobre la nota final ($F = 2.57$; $p < .001$), confirmando que el tribunal asignado influyó de forma estadísticamente significativa en el resultado obtenido por los aspirantes. Las comparaciones post hoc mediante Tukey HSD identificaron cinco pares de tribunales con diferencias significativas, con discrepancias superiores a un punto en algunos casos (ver Tabla 1), lo que evidencia patrones evaluativos diferenciados.

Tabla 1

Comparaciones post hoc significativas entre tribunales en la nota final de la fase de oposición.

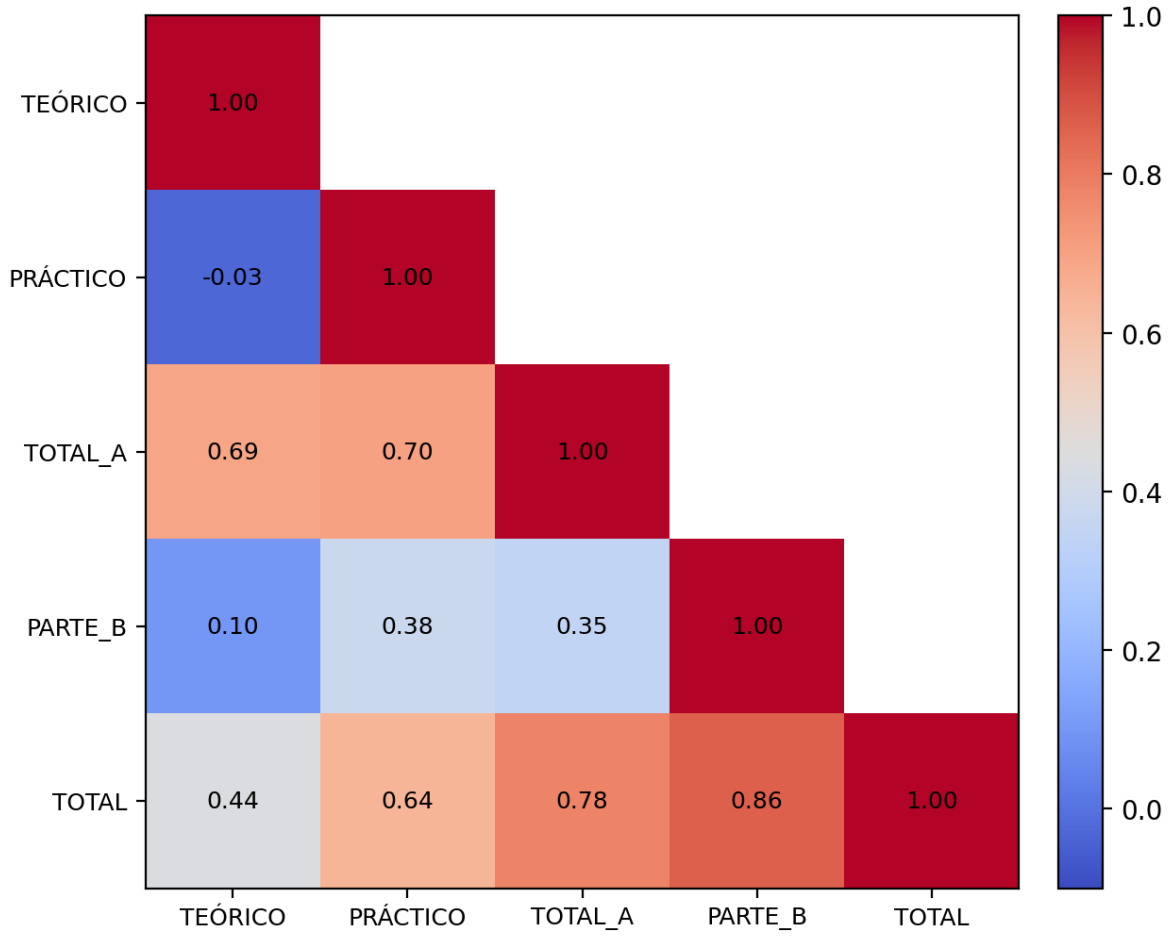
Tribunal 1	Tribunal 2	Diferencia de medias	P
C1	V8	-1.1895	.038
C2	V9	1.4631	.014
C3	V9	1.1122	.040
V10	V9	1.4642	.021
V8	V9	1.3986	.005

Correlaciones entre las diferentes calificaciones del proceso selectivo

El análisis de correlaciones de Pearson mostró una asociación muy elevada entre la parte B y la nota final ($r = .861$; $p < .001$), indicando que la defensa oral fue el principal componente explicativo del resultado global. Asimismo, la parte A se correlacionó de forma fuerte con la nota final ($r = .776$; $p < .001$), mientras que las correlaciones entre la parte teórica y práctica fueron de magnitud moderada. Estos resultados evidencian un peso predominante de la fase oral en la calificación final, aspecto que se visualiza de forma sintética en el mapa de calor de correlaciones (ver Figura 1).

Figura 1

Mapa de calor de correlaciones entre dimensiones evaluativas



Efecto conjunto de sexo, provincia y tribunal sobre la nota final

El análisis de varianza factorial confirmó que únicamente el tribunal evaluador ejerció un efecto significativo sobre la nota final ($F = 2.54$; $p < .001$). Ni el sexo ($F = 0.09$; $p = .767$) ni la

provincia (F = 0.76; p = .471) mostraron efectos significativos, reforzando la idea de que la principal fuente de variabilidad del sistema se localiza en los órganos evaluadores y no en variables individuales o territoriales.

Variabilidad interna de las calificaciones por tribunal

El análisis de la dispersión interna reveló una heterogeneidad considerable entre tribunales. Mientras algunos presentaron una alta consistencia interna (por ejemplo, A3: DT = 0.71; CV = 9.82%), otros mostraron una dispersión elevada (A2 y V5: CV > 21%), con rangos superiores a cuatro puntos.

Más del 40% de los tribunales presentaron coeficientes de variación superiores al 17%, lo que sugiere diferencias sustanciales en la aplicación de los criterios de evaluación. Aunque cierta variabilidad puede atribuirse a diferencias reales en el desempeño del alumnado, niveles elevados de dispersión podrían comprometer la equidad y la coherencia del proceso evaluativo (Tabla 2).

Tabla 2

Estadísticos descriptivos y de dispersión de la nota final por tribunal.

Tribunal	M	DT	R	CV (%)
A1	6.73	1.17	5.48	16.51
A2	6.80	1.47	5.31	22.06
A3	6.70	0.71	2.37	9.82
A4	7.11	1.18	4.29	16.55
A5	6.68	1.01	4.06	13.65
A6	7.35	1.20	4.35	17.91
A7	7.33	1.16	5.40	17.00
C1	7.83	1.38	5.36	19.55
C2	6.58	0.97	3.16	13.92
C3	6.93	1.26	4.99	18.57
V1	7.34	1.40	5.63	21.63
V2	7.04	1.11	4.03	16.01
V3	7.20	1.16	4.54	16.51
V4	7.10	1.24	4.57	18.33
V5	6.69	1.44	4.83	21.77
V6	7.57	1.18	4.01	17.70
V7	7.13	1.31	5.20	20.51
V8	6.65	1.25	4.70	17.41
V9	8.04	1.05	3.73	14.58
V10	6.58	1.12	3.77	17.00
V11	7.41	1.27	4.84	19.55
V12	7.10	1.22	4.61	17.68

Nota. M = media; DT = desviación típica; R = rango; CV = coeficiente de variación. Las letras A, C y V identifican tribunales de Alicante, Castellón y València, respectivamente.

Perfiles evaluativos de los tribunales mediante análisis de conglomerados

El análisis de conglomerados (k-means), precedido de un análisis de componentes principales, permitió sintetizar la variabilidad observada en tres perfiles evaluativos diferenciados:

- Clúster 0 (perfil exigente y homogéneo): tribunales con medias moderadamente bajas y baja dispersión interna (CV ≈ 13.3%), caracterizados por una aplicación coherente y restrictiva de los criterios.

• Clúster 1 (perfil heterogéneo): tribunales con mayor dispersión interna (CV ≈ 18.8%), reflejo de una aplicación menos uniforme de los criterios evaluativos.

• Clúster 2 (perfil más benevolente): tribunales con medias más elevadas y dispersión moderada (CV ≈ 15.8%), asociados a patrones de calificación más permisivos.

DISCUSIÓN

Este estudio aporta evidencia sobre la equidad y la consistencia evaluativa en las oposiciones de Educación Física de Secundaria en la Comunitat Valenciana (2025). Los resultados muestran que la variabilidad de las calificaciones no se explica por el sexo ni por la provincia, sino principalmente por el tribunal evaluador, lo que confirma la relevancia del denominado efecto tribunal y refuerza la necesidad de analizar la homogeneidad en la aplicación de los criterios de evaluación. Estos hallazgos son coherentes con trabajos previos sobre procesos selectivos docentes y con la evidencia reciente disponible en la misma especialidad y contexto autonómico (Aguilar-Parra et al., 2015; Miñana-Signes, 2025).

Los análisis no mostraron diferencias significativas en la nota final por sexo, lo que sugiere que, en esta convocatoria, la evaluación no se asoció a sesgos sistemáticos de género y se ajustó al principio de igualdad de trato. Este patrón es consistente con estudios en otros ámbitos de oposiciones donde la diferencia media por sexo no implica necesariamente discriminación, sino cumplimiento de estándares objetivos (Salgado, 2023).

Del mismo modo, no se observaron diferencias significativas por provincia. Aunque la literatura ha descrito disparidades territoriales en otros contextos autonómicos (Aguilar-Parra et al., 2015), en este caso la homogeneidad provincial sugiere que la gestión autonómica unificada (pruebas, temarios y baremos) puede amortiguar variaciones geográficas. En conjunto, sexo y provincia no explicaron la variabilidad de las calificaciones finales.

El hallazgo más relevante fue la existencia de diferencias significativas entre tribunales, confirmando el "efecto tribunal" descrito en investigaciones previas: el resultado de un aspirante puede depender en parte del órgano evaluador asignado, aun bajo criterios formalmente comunes (Aguilar-Parra et al., 2015; Bustos-Ceruelo & Callejo-Maudes, 2022). Este patrón reproduce lo observado en la Comunitat Valenciana en convocatorias anteriores de Educación Física, donde se subrayó la necesidad de reforzar la coordinación evaluativa (Miñana-Signes, 2025). La implicación principal es de equidad: si la severidad o el estilo de calificación difiere sistemáticamente, el principio de igualdad de oportunidades queda comprometido.

Desde una perspectiva psicométrica, estas diferencias remiten a la fiabilidad interevaluador. La literatura internacional sobre evaluación del desempeño (p. ej., OSCEs) muestra que la variabilidad atribuible a los examinadores puede afectar de forma sustantiva la fiabilidad del sistema (Haviari et al., 2024; Sebok et al., 2015; Yeates et al., 2021). La literatura apunta como fuentes probables interpretaciones divergentes de los criterios, ponderación desigual de dimensiones, uso de juicios holísticos frente a rúbricas analíticas y factores organizativos (p. ej., carga de corrección, fatiga o experiencia del evaluador) (Haviari et al., 2024; Yeates et al., 2013).

Además de la variación en medias, se observó heterogeneidad en la dispersión interna de las calificaciones por tribunal (rango, desviación típica y coeficiente de variación). Este patrón sugiere diferencias en el grado de consistencia interna con que se aplican los criterios: algunos tribunales presentan calificaciones más concentradas, mientras otros muestran mayor dispersión, lo que puede reflejar desde mayor capacidad discriminativa hasta aplicación menos uniforme de estándares. En procesos de alto impacto, la variabilidad interna elevada puede afectar a la percepción de justicia y a la comparabilidad entre aspirantes, especialmente cuando existen efectos sistemáticos de severidad o indulgencia entre evaluadores (Engelhard, 1994; Wolfe & McVay, 2012).

El análisis de clúster aportó un valor añadido al identificar tres perfiles evaluativos (más estrictos, intermedios y más benevolentes), coherentes con la literatura sobre sesgos de severidad/indulgencia en sistemas con múltiples evaluadores (Engelhard, 1994; Wolfe & McVay, 2012). Esta clasificación ayuda a interpretar la variabilidad no como casos aislados, sino como patrones relativamente estables, lo que refuerza la necesidad de calibración intertribunal para reducir diferencias sistemáticas.

IMPLICACIONES PRÁCTICAS

Los resultados apuntan a que la mejora del sistema debe centrarse prioritariamente en la homogeneización intertribunal, ya que las diferencias significativas observadas entre tribunales y los distintos niveles de dispersión interna sugieren posibles variaciones en la aplicación de los criterios de evaluación. En este sentido, las rúbricas deberían formularse de manera más operativa, compartida y vinculada a indicadores observables de desempeño, con el fin de reducir márgenes interpretativos excesivamente amplios y favorecer una calificación más comparable entre aspirantes.

Asimismo, la identificación de patrones diferenciados de severidad, heterogeneidad y benevolencia entre tribunales refuerza la conveniencia de realizar sesiones previas de calibración con ejemplos ancla, en las que los miembros de los tribunales puedan contrastar criterios, discutir casos límite y ajustar expectativas de calificación antes del inicio del proceso. Esta medida debería complementarse con una formación específica en evaluación, fiabilidad interevaluador y sesgos, especialmente orientada a mejorar la consistencia en la valoración de pruebas complejas y de alto impacto (Miñana-Signes, 2025; Miñana-Signes & Monfort-Pañego, 2025).

En las pruebas escritas, la doble corrección anónima podría contribuir a reducir la varianza atribuible al evaluador y aumentar la fiabilidad de las decisiones, especialmente cuando se detecten discrepancias relevantes entre correctores o tribunales (Haviari et al., 2024). En la fase oral, donde la replicación evaluativa resulta más compleja y el peso sobre la calificación final puede ser elevado, la grabación de las defensas para su eventual revisión en casos de reclamación o discrepancia podría mejorar la transparencia, la trazabilidad y la confianza en el procedimiento. Finalmente, una mayor desagregación de las calificaciones y su justificación vinculada a criterios concretos permitiría reforzar la percepción de justicia, disminuir la conflictividad y avanzar hacia un sistema más consistente, equitativo y sostenible, especialmente para colectivos con trayectorias prolongadas de interinidad (González-Calvo et al., 2023).

CONCLUSIONES

El presente estudio aporta evidencia empírica sobre la equidad y consistencia del proceso de evaluación en las oposiciones al cuerpo de Profesorado de Educación Secundaria, especialidad Educación Física, en la Comunitat Valenciana, convocatoria 2025. Los resultados no muestran diferencias significativas en la nota final por sexo ni por provincia, lo que sugiere una ausencia de sesgos sistemáticos asociados a estas variables y una aplicación territorial globalmente homogénea de los criterios.

Sin embargo, se identificó una variabilidad significativa entre tribunales, tanto en las medias de calificación como en los indicadores de dispersión interna. Además, el análisis de clúster permitió distinguir tres perfiles evaluativos diferenciados —estrictos, intermedios y benevolentes—, lo que apunta a patrones de calificación que pueden afectar a la comparabilidad de los resultados y a la percepción de equidad del proceso selectivo.

Estos hallazgos respaldan la necesidad de reforzar la estandarización y la fiabilidad interevaluador mediante formación específica de tribunales, rúbricas operativas comunes, sesiones previas de calibración y, cuando sea viable, procedimientos de doble corrección o supervisión externa. Avanzar hacia modelos de evaluación más objetivos, transparentes y trazables

contribuiría a mejorar la justicia procedimental y la confianza en el sistema, en coherencia con los principios de mérito, capacidad e igualdad.

Pese a las limitaciones del diseño y del contexto analizado, los resultados ofrecen una base útil para futuras investigaciones y para la toma de decisiones orientadas a consolidar un proceso selectivo más consistente, equitativo y formativo.

REFERENCIAS BIBLIOGRÁFICAS

Aguilar-Parra, J. M., Álvarez, J., & Lorenzo Torrecillas, J. J. (2015). Estudio sobre las pruebas de la oposición de acceso a la función pública docente: Variables influyentes en cada fase de oposición. *Educación XX1, 19*(1), 353–376. https://doi.org/10.5944/educxx1.15590

Aramendi Lekuona, G., & Aramendi Jauregi, P. (2020). El acceso a la función pública docente en Educación Infantil, Primaria y Secundaria. *Ikastorratza. e-Revista de Didáctica, 24*, 1–22.

Bustos-Ceruelo, D., & Callejo-Maudes, J. (2022). El sistema de selección de los maestros de Educación Primaria a debate. Profesorado. *Revista de Currículum y Formación del Profesorado, 26*(3), 483–504. https://doi.org/10.30827/profesorado.v26i3.21511

Conselleria de Educación, Cultura, Universidades y Empleo. (2026). Actas oficiales de calificaciones del procedimiento selectivo para el cuerpo de Profesorado de Educación Secundaria, especialidad Educación Física, convocatoria 2025. Generalitat Valenciana.

Constitución Española. (1978). Boletín Oficial del Estado, 311. https://www.boe.es/eli/es/c/1978/12/27/(1)

Darling-Hammond, L. (2014). Want to close the achievement gap? Close the teaching gap. *American Educator, 38*(4), 14–18.

Darling-Hammond, L., Hyler, M. E., & Gardner, M. (2017). Effective teacher professional development. Learning Policy Institute. https://doi.org/10.54300/122.311

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-facet Rasch model. *Journal of Educational Measurement, 31*(2), 93–112. https://doi.org/10.1111/j.1745-3984.1994.tb00436.x

Glazerman, S., Loeb, S., Goldhaber, D. D., Raudenbush, S., & Whitehurst, G. J. (2010). *Evaluating teachers: The important role of value-added*. Brookings Institution.

González-Calvo, G., Hortigüela-Alcalá, D., & Bores-García, D. (2023). Physical education public examination as an obstacle for long-term contract teachers. *Apunts. Educación Física y Deportes, 151*, 27–35. https://doi.org/10.5672/apunts.2014-0983.es.(2023/1).151.03

Haviari, S., de Tymowski, C., Burnichon, N., Lemogne, C., Flamant, M., Ruszniewski, P., Bensaadi, S., Mercier, G., Hamaoui, H., … Bouzid, D. (2024). Measuring and correcting staff variability in large-scale OSCEs. *BMC Medical Education, 24*, Article 817. https://doi.org/10.1186/s12909-024-05803-6

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35*(1), 69–81. https://doi.org/10.1111/j.1745-3984.1998.tb00528.x

Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Harvard University Press.

Miñana-Signes, V. (2025). Equidad y coherencia en la evaluación de las oposiciones de Educación Física (2023) en la Comunitat. *Retos, 73*, 239–248. https://doi.org/10.47197/retos.v73.117167

Miñana-Signes, V., & Monfort-Pañego, M. (2025). El opositor/a ideal desde la percepción de los miembros de los tribunales: Estudio de casos múltiple sobre el cuerpo de maestros/as de educación infantil. *Actividad Física y Deporte: Ciencia y Profesión, 41*, 9–23.

OECD. (2013). Teachers for the 21st century: Using evaluation to improve teaching. OECD Publishing. https://doi.org/10.1787/9789264193864-en

Perrenoud, P. (2008). *Diez nuevas competencias para enseñar*. Graó.

Real Decreto 276/2007, de 23 de febrero. (2007). Boletín Oficial del Estado, 53. https://www.boe.es/eli/es/rd/2007/02/23/276

Salgado, J. F. (2023). Validez predictiva e igualdad de trato de la oposición como método de selección para inspectores de Hacienda y técnicos de Hacienda del Estado. *Revista de Psicología del Trabajo y de las Organizaciones, 39*(2), 109–119. https://doi.org/10.5093/jwop2023a13

Sebok, S. S., Roy, M., Klinger, D. A., & de Champlain, A. F. (2015). Examiners and content and site: Oh my! A national organization's investigation of score variation in large-scale performance assessments. *Advances in Health Sciences Education, 20*(3), 581–594. https://doi.org/10.1007/s10459-014-9547-z

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identify rater effects in performance assessments. *Educational Measurement: Issues and Practice, 31*(2), 31–38. https://doi.org/10.1111/j.1745-3992.2012.00241.x

Yeates, P., Moult, A., Cope, N., McCray, G., Xilas, E., Lovelock, T., et al. (2021). Measuring the effect of examiner variability in a multiple-circuit objective structured clinical examination (OSCE). *Academic Medicine, 96*(8), 1189–1196. https://doi.org/10.1097/ACM.0000000000004028

Yeates, P., O'Neill, P., Mann, K., & Eva, K. (2013). Seeing the same thing differently: Mechanisms that contribute to assessor differences in directly observed performance assessments. *Advances in Health Sciences Education, 18*(3), 325–341. https://doi.org/10.1007/s10459-012-9372-1