

EVALUACIÓN DEL INSTRUMENTO PARA EXAMINAR CON RECURSOS TECNOLÓGICOS A ESTUDIANTES DE ESTADÍSTICA CON TEORÍA DE RESPUESTA AL ÍTEM

Norma Patricia Caro, María Inés Ahumada, y Verónica Arias
Facultad de Ciencias Económicas Universidad Nacional de Córdoba, Argentina.
pacaro@unc.edu.ar

Con el objetivo de verificar el grado de dificultad y capacidad de discriminación de un instrumento de evaluación se examinó a un grupo de 124 estudiantes de un curso de Estadística de la Universidad Nacional de Córdoba, Argentina. Para ello se implementó un cuestionario en la plataforma Moodle, conformado por 21 preguntas sobre estadística inferencial y que fue respondido online. Se validó el instrumento a aplicando modelos de la Teoría de respuesta al ítem (TRI), con variables de respuesta binaria, detectando preguntas con diferente grado de dificultad y de discriminación. Los resultados indican que, aquellas preguntas que aportan al constructo (conocimientos en estadística), que poseen dificultad media y que son las que más segregan entre los estudiantes resultaron las adecuadas para evaluar.

INTRODUCCION

La evaluación del estudiantado es un tema complejo, contar con instrumentos y procedimientos idóneos ayuda a mejorar los procesos educativos y, por ende, influye en el rendimiento académico (Blanco y Ginovart, 2012). La necesidad de generar dispositivos de evaluación confiables es un tema recurrente en las instituciones de educación superior. En esta oportunidad se presenta una experiencia que incorpora recursos tecnológicos, cuya sistematización ayuda a su comprensión y validación.

En los formatos virtuales (plataformas) como Moodle, se ofrecen una serie de recursos, entre los que se encuentran los cuestionarios que pueden gestionarse en tiempo real, y ser evaluados de manera continua, para lograr su fiabilidad y validez (Iglesias Rodríguez et al., 2014; Llarena y Paparo, 2006; Llorente Cejudo, 2007; Martín Galán y Rodríguez Mateos, 2012; Santoveña Casal, 2010).

La relevancia de detenerse en la tipología de preguntas y cuestiones que integran los exámenes online, entre otros motivos es que los interrogantes que se formulan encaminan los aprendizajes que los estudiantes van construyendo (Vives et al., 2021).

En 2019 se evaluó a los estudiantes de un curso de Estadística en la Facultad de Ciencias Económicas de la Universidad Nacional de Córdoba, Argentina con un cuestionario, en el que se midieron los conocimientos adquiridos en estadística y las habilidades para identificar problemas (constructo latente). Para cuantificar dicho constructo se aplicó un modelo de ecuaciones estructurales generalizado, como es el modelo de respuesta al ítem (TRI). Asimismo, se analizó la relación de cada ítem con el concepto estadístico y cómo el cuestionario en su conjunto se relaciona con dicho rasgo latente. A fin de verificar el grado de dificultad y capacidad de discriminación de las preguntas que examinan al estudiantado en los conocimientos adquiridos en estadística, se plantean como objetivos, validar un cuestionario para una mejor evaluación del estudiantado e identificar fortalezas y debilidades en los instrumentos de evaluación.

Esta experiencia contribuye a contextos donde se puede evaluar de una manera totalmente virtual, permitiendo la mejora continua de los instrumentos. La contribución de este trabajo consiste en difundir estas estrategias de validación para que en la Facultad de Ciencias Económicas–Nacional de Córdoba (FCE– UNC) se trabaje en este sentido, logrando instrumentos de calidad.

MARCO CONCEPTUAL

En la búsqueda de la calidad en los instrumentos de evaluación, dentro de una serie de acciones de mejora, el uso de las plataformas como recurso educativo de aprendizaje (Tirado-Morueta y Hernando-Gómez, 2011) es utilizado conjugando no solo el uso de las tecnologías de la información y la comunicación, sino también las competencias digitales que traen los estudiantes para este contexto (Carrera Farran et al., 2011).

La evaluación es un proceso indisoluble de la enseñanza y el aprendizaje ya que se mira como están aprendiendo y como se enseña (Anijovich y Cappeletti, 2017), siendo en sí mismo un proceso en el que los estudiantes ponen en juego sus propias competencias y progreso (Ryan et al., 2013).

Tomatis et al. (2016) destacan que la evaluación es una preocupación constante en la práctica educativa dada sus consecuencias en el aprendizaje. La cultura de la evaluación, demanda información continua, de comprensión, de conocimiento y entendimiento de hechos o situaciones (Medina Moreno et al., 2014). Uno de los aspectos más importantes en la evaluación es la validez como el grado en que la evidencia apoya las inferencias que se hacen a partir de las puntuaciones obtenidas en dicha instancia de evaluación (Rojas, 2013).

Los cuestionarios en Moodle se clasifican en dos grupos: formativos y evaluativos. Los formativos tienen como objetivo, que los estudiantes aprendan a través de la reiteración y autoevaluación, mientras que los cuestionarios evaluativos o sumativos tienen como objetivo, valorar si se otorgan o no los créditos del curso. Todo instrumento de evaluación debe ser sujeto a una prueba empírica de validez que lo avale y justifique (Rodríguez Castro y Gómez López, 2010), lo que estriba en la posibilidad de contar con otros medios que permitan evaluar la efectividad de la misma, sus alcances y limitaciones. Organista-Sandoval et al. (2017) y Medina-Rebollo et al. (2019) desarrollan cuestionarios para estimar las habilidades digitales de estudiantes universitarios.

MATERIALES Y MÉTODO

Se realizó una investigación aplicada, cuantitativa y explicativa. Se trabajó con una muestra de 124 estudiantes de un total de 207 pertenecientes a un curso de Estadística de la Facultad de Ciencias Económicas de la Universidad Nacional de Córdoba, en 2019, quienes realizaron una evaluación sobre Inferencia Estadística. La evaluación estaba dirigida al total del estudiantado, pero con carácter optativo con el objetivo de autoevaluarse en una instancia previa al examen parcial que deberían realizar de manera obligatoria, con posterioridad. El 60% que realizó la evaluación constituye una muestra no probabilística, dado que los estudiantes eligen hacerla o no, pero representativa de la población, ya que mantiene sus características.

Se realizaron 21 preguntas (Tabla 1), consistentes en ejercicios a resolver y preguntas a contestar que fueron evaluados como correctas o incorrectas (respuesta binaria). El instrumento fue elaborado por el equipo de cátedra, es decir con juicio de expertos en lo que se quiere evaluar. Cada pregunta tiene dos resultados posibles: 1: aprobado y 0: no aprobado, ya sea por error o por ausencia de respuesta.

Se define como constructo latente aquel que representa los conocimientos adquiridos en inferencia estadística. Para la elaboración del instrumento de evaluación se cumplieron una serie de etapas:

- Diseño: se decidió evaluar en la plataforma Moodle para ser realizada en computadora, tanto de forma presencial como virtual.
- Preparación: implicó construir un banco de preguntas, estructurado en categorías conforme a los contenidos de la asignatura.
- Valoración: se definió la ponderación o peso de cada pregunta.
- Devolución: se incluyó una retroalimentación a cada ítem y/o al instrumento en su conjunto.

En este trabajo se aplica la TRI en el campo de la educación, sobre los conocimientos adquiridos en estadística por estudiantes que cursan en una institución de educación superior. Para ello, se utiliza un modelo de ecuaciones estructurales, en particular un modelo logístico de dos parámetros para datos dicotómicos, que tiene como variables respuesta, variables binarias que siguen una distribución logística. Nazario et al. (2021) presentan el análisis TRI utilizando el modelo logístico de dos parámetros, para variables dicotómicas, propuesto por Birnbaum (1968), que es el que se aplica en este trabajo, según la ecuación:

$$P_{i(\theta)} = \frac{e^{Da(\theta-b)}}{1+e^{Da(\theta-b)}} \quad \text{Para } i=1,2,3,\dots,n \quad (1)$$

donde:

- $P_{i(\theta)}$: es la probabilidad de que un o una examinado/a elegido/a al azar con aptitud θ conteste correctamente el ítem i .
- n : número de ítems del test.
- b : parámetro de posición o de la dificultad del ítem.
- D : factor igual a 1,7, es un valor arbitrario introducido para que la función logística sea ajustada a la ojiva normal con una exactitud de 0,01.
- a : parámetro de discriminación, es la pendiente de la curva característica de cada ítem en el punto b .

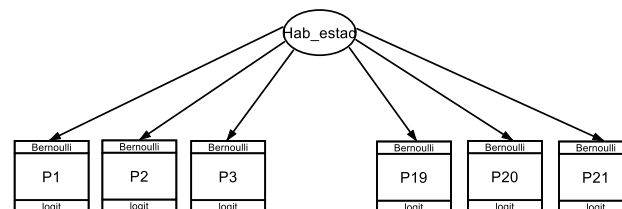
Tabla 1. Contenidos incluidos en el cuestionario de evaluación

Tema principal	Sub temas	Id
Prueba de hipótesis para la media poblacional con varianza conocida	Planteo de hipótesis nula y alternativa	P1
	Estadístico y distribución de probabilidad	P2
	Regla de decisión	P3
	Decisión en el marco del problema	P4
Estimación por intervalos para la media poblacional	Interpretación del intervalo de confianza	P5
	Cálculo e interpretación de la precisión de la estimación.	P6
Prueba de hipótesis para la media poblacional con varianza conocida	Interpretación de salida de software	P7
Determinación del tamaño de muestra para estimar la proporción	Cálculo del tamaño de muestra requerido	P8
Muestreo y tipos de muestreo	Similitudes y diferencias entre métodos de muestreo	P9
Prueba de hipótesis para la proporción poblacional	Planteo de hipótesis y selección de estadístico con su distribución de probabilidad	P10
	Regla de decisión	P11
	Cálculo del valor des estadístico	P12
	Conclusión en términos del problema	P13
Potencia de una prueba de hipótesis	Calcular la probabilidad de cometer el error tipo II	P15
Prueba de hipótesis para la diferencia de proporciones	Identificación de la prueba de hipótesis	P16
	Planteo de las hipótesis y supuestos requeridos	P17
	Valor del estadístico, regla de decisión y decisión	P18
		P19
	Conclusión en el marco del problema	P20
		P21

Fuente: elaboración propia

Los ítems con pendiente mayor son más útiles para separar al estudiantado en distintos niveles de aptitud, que aquellos que tienen menor pendiente.

El modelo de ecuaciones estructurales utilizado posee una notación gráfica que se presenta en la Figura 1. El rasgo latente constituido por los conocimientos en estadística y cada uno de los resultados de las preguntas son las variables observadas.



Fuente: elaboración propia

Figura 1. Modelo de ecuaciones estructurales generalizado (teoría de respuesta al ítem)

RESULTADOS

El estudiantado que realizó la evaluación estaba distribuido uniformemente entre varones y mujeres, sus edades eran en promedio de 20 años, el 80% eran cursantes por primera vez y estaban distribuidos proporcionalmente en las tres carreras de la Facultad de Ciencias Económicas.

Se midió el aporte de cada pregunta (ítem) sobre el constructo (conocimientos en estadística). La mayoría de las preguntas aportan significativamente al constructo significando que reflejan los conocimientos y habilidades adquiridos en estadística ya que presentan una relación significativa entre cada ítem y la variable latente en casi todas las preguntas. Las excepciones fueron las preguntas P9 y P15 que no son significativas, a un nivel del 5%, dado que sus valores de probabilidad (p-valor) fueron de 0,058 y 0,123, respectivamente. Estas preguntas versan sobre:

- *P9: Similitudes y diferencias entre los modelos de muestreo estadístico, Conceptualmente, suele ser un tema no complejo y por ello es una pregunta que puede ser catalogada como fácil.*
- *P15: Calcular la probabilidad de error tipo II para un valor particular de la hipótesis alternativa, que constituye un tema complejo, y que fue respondido correctamente sólo por el 15% de los estudiantes.*

El índice de discriminación de un ítem indica en qué medida el ítem diferencia a los y las estudiantes, ya que cada uno aporta información en diferentes niveles de la escala de conocimientos de estadística. Por tanto, se pueden elegir del banco de preguntas los ítems que presenten una discriminación máxima en el nivel del rasgo asociado. Si bien se puede determinar un punto de corte, que depende del área de interés en la que se aplican estos modelos, indicados por expertos, también pueden, estos índices de discriminación ser considerados en términos comparativos, seleccionando a aquellos que poseen mayores valores.

En cuanto al análisis de los coeficientes que miden discriminación, la Tabla 2 presenta de manera ordenada que preguntas discriminan menos y que preguntas discriminan más. Dejando de lado las preguntas que resultaron ser no significativas, se observa que P11 es la que más discrimina, es decir que es la que más aporta al constructo (sobre el tema de selección del estadístico y su distribución de probabilidad para una prueba de hipótesis de la proporción poblacional), mientras que P8 es la que menos aporta (sobre el tema determinación del tamaño de muestra para estimar proporción).

De esta manera, en función de los objetivos planteados, se validó el cuestionario en cuanto al aporte que cada pregunta tiene sobre esos conocimientos, donde se detectaron preguntas que no aportan significativamente al constructo, por un lado, y preguntas que sí aportan al conocimiento estadístico con mayores o menores grados de dificultad y de discriminación.

Por otro lado, analizando la dificultad de cada ítem, dentro del conjunto de preguntas que resultaron significativas, las preguntas 11 y 16 son las más fáciles.

- *P11: Determinar el estadístico y su distribución de probabilidad en una prueba de hipótesis para la proporción.*
- *P16: Identificar la prueba de hipótesis correspondiente a una diferencia de proporciones.*

Mientras que las preguntas más difíciles son la 8 y 1.

- *P8: Calcular el tamaño de muestra requerido*
- *P17: Plantear las hipótesis de una prueba de diferencia de proporciones y explicar los supuestos requeridos.*

Esto permitió identificar fortalezas y debilidades en el instrumento de evaluación, con la finalidad de mejorar el cuestionario, o elaborar uno diferente. Aquellas preguntas que aportan al constructo, que poseen dificultad media y que son las que más discriminan entre el estudiantado se retienen en el cuestionario. Estas permiten no solo obtener los conocimientos sino desarrollar otras capacidades que también deben ser aprendidas, mientras que las preguntas no significativas o muy fáciles y que no discriminan, pueden ser reemplazadas con la finalidad de obtener mejores instrumentos de evaluación.

DISCUSIÓN

Enfatizar en la importancia de la evaluación en entornos virtuales con instrumentos idóneos es un asunto muy importante. Del análisis exhaustivo del instrumento aplicado se observan algunas falencias importantes en el diseño de los mismos, como son preguntas no significativas para medir el aprendizaje y el nivel (ranking) en que cada una discrimina en el conocimiento.

De esta manera se presenta una herramienta, entre otras, que permite mejorar los cuestionarios con los que se examina al estudiantado y que refleje el aprendizaje adquirido.

Tabla 2. Coeficientes de discriminación

Pregunta	Ranking	Coef. Discriminación	p-value	Coef. Dificultad
P9	1	0,3859	0,058	0,365
P15	2	0,4691	0,123	-2,052
P8	3	0,5355	0,018	-0,377
P7	4	0,621	0,011	0,566
P1	5	0,6337	0,029	1,706
P20	6	0,6775	0,005	0,247
P17	7	0,8128	0,002	-0,259
P18	8	0,8641	0,005	0,947
P2	9	0,8644	0,033	1,766
P3	10	0,9201	0,023	1,706
P21	11	0,9603	0,003	0,606
P19	12	0,9773	0,001	-0,064
P13	13	10,144	0,002	0,566
P5	14	10,365	0	0,484
P6	15	10,967	0,002	1,274
P16	16	12,635	0,001	1,892
P10	17	13,312	0,002	1,177
P14	18	14,014	0,001	0,208
P12	19	14,899	0	0,947
P4	20	21,243	0,009	1,647
P11	21	137,242	0	2,1

Fuente: elaboración propia

El hecho de que los cuestionarios no evalúen los aspectos que se pretenden evaluar tiene múltiples causas, algunas de las cuales pueden ser detectadas a través del análisis de estos instrumentos y así mejorarlos. Así, se torna necesario dar a conocer estos métodos y transmitir este tipo de herramientas de análisis al cuerpo docente de la institución para revisar una de las prácticas de la enseñanza más controvertidas y traducirlas en evaluaciones de calidad.

En una segunda etapa se realizaron ajustes sobre la estructura del cuestionario y sobre los ítems que lo conforman, con la finalidad de lograr una mejora significativa en tal instrumento. Luego, esa tarea se traduce en aportar calidad al proceso de evaluación, es decir logrando evaluar mejor los aprendizajes, mediante el perfeccionamiento de los exámenes. Así, mediante sucesivos ajustes y modificaciones basados en indicadores objetivos, se obtienen pruebas con atributos singulares, a fin de inferir en cada estudiante, el verdadero valor del rasgo latente de manera más justa.

La experiencia realizada en el curso de Estadística repercutió positivamente, ya que ha permitido, a través de la validación de cuestionarios, generar instrumentos de evaluación que contribuyen a la mejora del proceso evaluativo.

Debido a la cantidad de estudiantes y a los diferentes instrumentos que se utilizan en cada prueba, estas validaciones se realizan en grupos que van entre los 100 y 150 estudiantes obteniéndose resultados satisfactorios. Se han realizado validaciones de cuestionarios en otros grupos, con algunas adaptaciones y los resultados se confirman.

REFERENCIAS

- Anijovich, R., y Capelletti, G. (2017). *La evaluación como oportunidad*. Paidós.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.
- Blanco, M., y Ginovart, M. (2012). Los cuestionarios del entorno Moodle su contribución a la evaluación virtual formativa de los alumnos de matemáticas de primer año de las titulaciones de

- Ingeniería. *RUSC, Universities and Knowledge Society Journal*, 9(1), 354–370. <http://doi.org/10.7238/rusc.v9i1.1277>
- Carrera Farran, F. X., Vaquero Tió, E., y Balsells Bailón, M. (2011). Instrumento de evaluación de competencias digitales para adolescentes en riesgo social. *EDUTEC. Revista Electrónica de Tecnología Educativa*, (35), Article 154. <https://doi.org/10.21556/edutec.2011.35.410>
- Iglesias Rodríguez, A., Olmos Migueláñez, S., Torrecilla Sánchez, E. M., y Mena Marcos, J. J. (2014). Evaluar para optimizar el uso de la plataforma Moodle (studium) en el departamento de didáctica, organización y métodos de investigación. *Tendencias Pedagógicas*, 23, 155–170. <https://revistas.uam.es/tendenciaspedagogicas/article/view/2077>
- Llarena, M., y Paparo, M. (2006). Propuesta de una metodología de seguimiento y evaluación de cursos a distancia. *Revista Iberoamericana de Educación*, 37(4). <https://doi.org/10.35362/rie3742702>
- Llorente Cejudo, M. (2007). Moodle como entorno virtual de formación al alcance de todos. *Comunicar*, 15(28), 197–202. <https://www.redalyc.org/pdf/158/15802827.pdf>
- Martín Galán, B., y Rodríguez Mateos, D. (2012). La evaluación de la formación universitaria semipresencial y en línea en el contexto del EEES mediante el uso de los informes de actividad de la plataforma Moodle. *RIED. Revista Iberoamericana de Educación a Distancia*, 15(1), 159–178. <https://doi.org/10.5944/ried.1.15.782>
- Medina Moreno, P., Aguirre, M., y Luengo, M. L. (2014). Predictibilidad de las notas de enseñanza media según establecimiento de origen sobre el rendimiento académico en estudiantes odontología. *Educación Médica Superior*, 28(1), 65–73. http://scielo.sld.cu/scielo.php?script=sci_abstract&pid=S0864-21412014000100008
- Medina-Rebollo, D., Grimaldi-Puyana, M., Del Río-Rama, M., y Sánchez-Oliver, A. (2019). Elaboración y validación de un cuestionario para evaluar el uso didáctico de la plataforma de teleformación Moodle en el contexto universitario. *Revista ESPACIOS*, 40(9). <https://www.revistaespacios.com/a19v40n09/a19v40n09p04.pdf>
- Nazario, P. F., Ferreira, L., Both, J., y Lopes Vieira, J. L. (2021). Qualidade avaliativa do test of gross motor development 2: Uma investigação com base na teoria de resposta ao item [Calidad evaluativa de la prueba de desarrollo motor grueso 2: Una investigación basada en la teoría de respuesta al ítem]. *Retos*, 42, 696–704. <https://doi.org/10.47197/retos.v42i0.86427>
- Organista-Sandoval, J., Lavigne, G., Serrano-Santoyo, A., y Sandoval-Silva, M. (2017). Desarrollo de un cuestionario para estimar las habilidades digitales de estudiantes universitarios. *Revista Complutense de Educación*, 28(1), 325–343. https://doi.org/10.5209/rev_RCED.2017.v28.n1.49802
- Rodríguez Castro, M., y Gómez López, V. (2010). Indicadores al ingreso en la carrera de medicina y su relación con el rendimiento. *Revista de la Educación Superior*. 39(153), 43–50. <https://www.redalyc.org/pdf/604/60416813003.pdf>
- Rojas, T. L. (2013). Validez predictiva de los componentes del promedio de admisión a la Universidad de Costa Rica utilizando el género y el tipo de colegio como variables control. *Revista Electrónica Actualidades Investigativas en Educación*, 13(1), 1–24. <http://doi.org/10.15517/aie.v13i1.8562>
- Ryan, S., Scott, B., Freeman, H., y Patel, D. (2013). *The virtual university: The internet and resource-based learning*. Routledge. <https://doi.org/10.4324/9781315042022>
- Santoveña Casal, S. M. (2010). Cuestionario de evaluación de la calidad de los cursos virtuales de la UNED. *Revista de Educación a Distancia (RED)*, (25). <https://revistas.um.es/red/article/view/125311>
- Tirado-Morueta, R., y Hernando-Gómez, Á. (2011). Comunidades de aprendizaje a través de plataformas de teleformación. *RIED. Revista Iberoamericana de Educación a Distancia*, 14(2), 99–120. <https://doi.org/10.5944/ried.2.14.791>
- Tomatis, M. C., Burrone, M. S., Romero, D., Novella, M. L., Olivero, M., Antuña, A. M., Lucchese, M., Enders, J., y Fernández, A. (2016). Validez predictiva del examen de ingreso a la carrera de medicina de la Facultad de Ciencias Médicas (UNC). *Revista de Educación*, 7(9), 357–367. https://fh.mdp.edu.ar/revistas/index.php/r_educ/article/view/1917
- Vives, M., Orte, C., Amer, J., y Quesada, V. (2021). Evaluación del cambio en los participantes del programa intergeneracional de educación primaria “compartir la infancia.” *Revista Iberoamericana de Evaluación Educativa*, 14(2), 5–19. <https://doi.org/10.15366/rie2021.14.2.001>