Simulación Clínica
Septiembre - Diciembre 2023
Vol. 5, núm. 3 / pp. 110-113

Diseños secuenciales para programas de evaluación más eficientes

Sequential designs for more efficient assessment programs

Jimmie Leppink*

Palabras clave:
diseños secuenciales, evaluación, conocimientos, habilidades, competencia.

Keywords:
sequential designs, assessment, knowledge, skills, competence.

RESUMEN

Es buena práctica de evaluación adquirir información suficiente para tomar decisiones sobre los conocimientos, habilidades o actitudes de los candidatos, utilizando no más recursos de los necesarios. Aunque la mayoría de los programas de evaluación tienden a utilizar los mismos recursos –por ejemplo, la misma cantidad de estaciones en un examen clínico estructurado objetivo (ECOE) o la misma cantidad de preguntas de opción múltiple– para todos los candidatos, se necesitan menos recursos para los candidatos que tienen niveles de competencia muy altos o muy bajos que para los candidatos que se encuentran en algún punto intermedio. Los diseños de evaluación secuencial pueden ayudar a reducir los recursos donde no se necesiten y, como tal, pueden ayudar a reducir el tiempo y otros costes. Este artículo presenta un ejemplo de un diseño de evaluación secuencial que se puede utilizar, independientemente del número de candidatos.

ABSTRACT

It is good practice of assessment to acquire sufficient information to make decisions regarding the knowledge, skills or attitudes of candidates using no more resources than needed. Although most assessment programs tend to use the same resources –for example the same number of objective structured clinical examination (OSCE) stations or same number of multiple-choice questions– for all candidates, we need fewer resources for candidates who have very high or very low levels competence than for candidates who are somewhere in between. Sequential assessment designs can help to reduce resources where those are not needed and can as such help to reduce time and other costs. This article presents an example of such a sequential assessment design which can be used, regardless of the number of candidates.

INTRODUCCIÓN

El objetivo central de la evaluación en la simulación y otras actividades en un programa de medicina u otra profesión de salud es adquirir la información necesaria para poder decidir si un candidato (estudiante, residente, profesional) tiene conocimientos, habilidades y/o actitudes al nivel deseado o mejor. Por un lado, una evaluación más larga suele resultar en más información sobre esta cuestión que una evaluación más corta. Por otro lado, evaluaciones más largas también requieren una inversión económica y de tiempo más alta. Si un candidato tiene un nivel de competencia al límite del nivel deseado, una inversión adicional puede ser justificada porque puede ser necesaria para llegar a una probabilidad alta de una decisión correcta (por ejemplo, en un examen: el candidato aprueba o no aprueba). A la vez, si el rendimiento de un candidato muestra un nivel de competencia claramente por encima o por debajo del nivel deseado, una evaluación más concisa puede ser suficiente para tomar la decisión correcta. El concepto de la evaluación secuencial puede ayudar a valorar dónde se necesita más recursos y dónde menos.[1,2] Aunque este concepto suele ser asociado con valoraciones cuantitativas en cohortes de muchos candidatos, se puede utilizar para otros tipos de valoraciones también y para cualquier número de candidatos. Por lo tanto, este artículo presenta un ejemplo de un diseño de evaluación secuencial para valoraciones dicótomas (por ejemplo, correcto o incorrecto, o acción hecha o no hecha), independientemente del número de candidatos (podría ser un participante solo).

* Hospital Virtual Valdecilla. España.

Recibido: 05/05/2023
Aceptado: 20/10/2023

doi: 10.35366/114033

Citar como: Leppink J. Diseños secuenciales para programas de evaluación más eficientes. Rev Latinoam Simul Clin. 2023; 5 (3): 110-113. https://dx.doi.org/10.35366/114033

MATERIAL Y MÉTODOS

En este ejemplo fictivo, se evaluaron a tres estudiantes en año 'X' de un programa de medicina mediante un examen clínico estructurado objetivo (ECOE) que consistió en una serie de estaciones de un nivel de dificultad parecido, pero sobre diferentes temas. Cada estación resultó en una valoración de '0' (incorrecto, no hecho) o '1' (correcto, hecho) en un total de 10 aspectos. Completaron cinco estaciones, resultando en un total de 50 valoraciones del tipo '0/1'. Se utilizó un modelo bayesiano binomial[3] para comparar cada candidato con el estándar mínimo de 50%. En este modelo bayesiano, se cuenta la frecuencia de éxitos (valoraciones de '1') y la frecuencia de fallos (valoraciones de '0') y añade una distribución a priori de un éxito y un fallo para llegar a la distribución a posteriori:

$$B(\text{éxitos}, \text{fallos}) + B(1, 1) = B(\text{éxitos} + 1, \text{fallos} + 1).$$

La distribución a posteriori da un intervalo creíble del 95%. Si este intervalo está completamente por encima de los 0.50, hay suficiente evidencia de que el nivel de competencia del candidato está suficientemente bien para no tener que completar la segunda mitad del examen que también consiste en cinco estaciones parecidas, resultando en un total de 50 valoraciones del tipo '0/1'. Si el intervalo está completamente por debajo de los 0.50, hay suficiente evidencia de que el nivel de competencia del candidato está por debajo del nivel deseado y tiene que prepararse para una segunda oportunidad del examen en tres meses. Y si el intervalo incluye 0.50, el candidato tiene que completar la segunda mitad del examen el próximo día para tener más información sobre su nivel de competencia y tomar una decisión. En el último caso, si el conjunto de las 10 estaciones resulta en un intervalo creíble del 95% totalmente por encima de los 0.50, aprueba el examen; y si el intervalo no está por encima de los 0.50, el candidato tiene que prepararse para la segunda oportunidad del examen en tres meses. En la segunda oportunidad del examen, la lógica es parecida: si la primera mitad de cinco estaciones resulta en un intervalo creíble del 95% por encima de los 0.50, el candidato aprueba; si este intervalo incluye 0.50, tiene que completar la segunda mitad del examen el próximo día para tomar una decisión (aprobar o volver el año que viene); y si el intervalo está por debajo de los 0.50, tiene que volver para el examen el año que viene.

Esto es un diseño secuencial, porque no todos los candidatos en la primera oportunidad tienen que completar 10 estaciones y tampoco todos los participantes en la segunda oportunidad tienen que completar 10 estaciones.

Las estadísticas se pueden calcular en varios programas, incluyendo los programas de fuente abierta (Open Source) JASP[4] y jamovi.[5]

RESULTADOS

La *Tabla 1* muestra los resultados de los tres candidatos.

Las primeras cinco estaciones resultaron en 27 puntos (es decir: 27 éxitos) para el candidato A, 36 puntos para el candidato B y 17 puntos para el candidato C. Esto significa unas distribuciones a posteriori de B(28, 24) para el candidato A, B(37, 15) para el candidato B y B(18, 34) para el candidato C. Los intervalos creíbles del 95% que correspondan con estos resultados son: 0.403-0.671 para el candidato A, 0.583-0.825 para el candidato B y 0.224-0.479 para el candidato C. Por lo tanto, el candidato A tuvo que volver para la segunda mitad del examen el día después, el candidato B aprobó sin necesidad de volver para la segunda mitad del examen, mientras el candidato C tuvo que volver para la segunda oportunidad del examen después de tres meses.

Tabla 1: Los resultados de los tres candidatos en el ejemplo.

Candidato	Oportunidad	Mitad	Intervalo	Resultado
A	1	1	[0.403; 0.671]	Volver para la segunda mitad
	1	2	[0.502; 0.691]	Aprobado
B	1	1	[0.583; 0.825]	Aprobado
C	1	1	[0.224; 0.479]	Volver en tres meses
	2	1	[0.501; 0.759]	Aprobado

La segunda mitad de la primera oportunidad resultó en 33 puntos para el candidato A, dando un total de 60 puntos (éxitos) para las 10 estaciones. Esto significa una distribución a posteriori de B(61, 41) para este candidato y un intervalo creíble del 95% de 0.502-0.691 y, por lo tanto, el candidato A aprobó.

El candidato C tuvo una nueva oportunidad de mostrar su competencia en la segunda oportunidad del examen. En la primera mitad de esta segunda oportunidad, su rendimiento fue mucho mejor que en la primera oportunidad del examen: con un total de 32 puntos, llegó a una distribución a posteriori de B(33, 19) y, por lo tanto, aprobó sin necesidad de hacer la segunda mitad porque el intervalo creíble del 95% fue 0.501-0.759.

DISCUSIÓN

En un examen no secuencial, los tres candidatos habrían completado la primera y segunda mitad en la primera oportunidad del examen, y el candidato C habría completado la primera y segunda mitad en la segunda oportunidad del examen, requiriendo un total de recursos de ocho series de cinco estaciones para los tres candidatos. Con el diseño secuencial utilizado en este ejemplo, cinco series de cinco estaciones para los tres candidatos ha sido suficiente, que significa una reducción de recursos utilizados de un 37.5%, una reducción que además es fácil de justificar. Un candidato que en la primera mitad llega a sólo 17 puntos necesitaría 43 puntos más para llegar al mínimo para aprobar después de las dos mitades (60 puntos), que es un evento de una probabilidad de casi 0 (obtener al menos 29 puntos en esta segunda mitad ya tendría una probabilidad por debajo de 0.001 o 1/1,000) y, por lo tanto, utilizar recursos para esta segunda mitad probablemente se puede considerar innecesario. En la misma línea, un candidato que en la primera mitad llega a 36 puntos necesitaría solo 24 puntos más para llegar a 60 en dos mitades, que es muy probable de ocurrir (más de 99.9%). A la vez, si entre dos oportunidades de un examen hay un periodo de unos meses, un candidato que tiene que volver para esta segunda oportunidad tiene tiempo para mejorar donde tenga que mejorar y, por lo tanto, no utilizar el rendimiento malo de la primera oportunidad es justificable.

Si el estándar debe ser 50% (0.50) u otro número depende del contexto –el programa, tipo de examen, tipo de riesgo– y es algo que puede establecer la junta de evaluación que se encarga del examen (y/o del programa del que este examen forma parte), pero la lógica sigue la misma. Si, por ejemplo, en una evaluación de bajo riesgo se prefiere utilizar 40% como estándar, se compara los intervalos creíbles del 95% no con 0.50 sino con 0.40, y si en un examen de alto riesgo se considera la necesidad de utilizar 60% como estándar, se compara los intervalos con 0.60.

En cuanto al número de estaciones (o preguntas), también depende del tipo de examen y tipo de riesgo. Evaluaciones más largas resultan en intervalos creíbles del 95% más estrechos que evaluaciones más cortas, es decir, que un examen de un riesgo más alto necesitará más estaciones (o preguntas) que un examen de un riesgo más bajo. Por ejemplo, en un examen de 200 valoraciones (por ejemplo, 200 preguntas de opción múltiple) se necesita sólo 114 puntos (57% en vez de 60% como en el ejemplo principal en este artículo) para aprobar si el estándar es 50% y unos 134 puntos (67%) si el estándar es 60%.

El modelo presentado en este artículo es relativamente intuitivo y fácil de utilizar para evaluaciones secuenciales como en el ejemplo en este artículo y también para definir un estándar empírico razonable donde las mismas estaciones o preguntas están reutilizadas en cohortes distintos.[3] Además, también existen otros tipos de modelos donde las valoraciones no son dicótomas sino cuantitativas.[1,2] Los cálculos se pueden hacer relativamente fácil en programas estadísticos accesibles de modo gratuito[4,5] y el uso de un diseño secuencial puede ayudar a reducir considerablemente los recursos necesarios para una evaluación sin tener que preocuparse de aumentar sustancialmente la probabilidad de decisiones erróneas. Menos tiempo en exámenes significa más tiempo para un candidato de aprender en otros sitios (incluyendo en la práctica clínica), una reducción de la inversión económica (y tiempo) en la evaluación de parte de la institución, y para los examinadores más tiempo para atender pacientes y otras necesidades. El uso de diseños secuenciales significa un esfuerzo potencialmente mínimo con un potencial importante para todos involucrados en la evaluación en un rol u otro. Por lo tanto, la evaluación secuencial se merece más consideración en los programas de medicina y otras profesiones de salud.

REFERENCIAS

1. Mancuso G, Strachan S, Capey S. Sequential testing in high stakes OSCE: a stratified cross-validation

approach. MedEdPublish [Internet]. 2019. Available in: https://doi.org/10.15694/mep.2019.000132.1
2. Leppink J. Assessment of individual competence: a sequential mixed model. Sci Med [Internet]. 2021; 31: e40128. Available in: https://doi.org/10.15448/1980-6108.2021.1.40128
3. Leppink J. In god we trust, all others bring data: a Bayesian approach to standard setting. Health Prof Educ [Internet]. 2020; 6 (2): 291-299. Available in: https://doi.org/10.1016/j.hpe.2020.01.003
4. JASP Team. JASP (version 0.17) [Computer software]. Retrieved (May 5, 2023). Available in: https://jasp-stats.org
5. The jamovi project. jamovi (version 2.3.21) [Computer software]. Retrieved (May 5, 2023). Available in: https://www.jamovi.org

Correspondencia:
Dr. Jimmie Leppink
E-mail: j.leppink@gmail.com