

El History Class Observation Tool (HCOT) está compuesto por tres dimensiones: Discurso Docente (Verbal), Materiales Didácticos y Actividad de los estudiantes. La primera dimensión, Discurso Docente, se subdivide en cinco categorías: Exploración y activación de Conocimientos Previos, Contextualización, Interpretación, Pensamiento Histórico y Métodos, estrategias y técnicas de enseñanza. Esto da un total de 38 ítems. Por su parte, los grupos focales constan de entrevistas con alumnado, profesorado en formación y tutores de aulas de secundaria, con un total de dieciséis, nueve y dieciséis preguntas respectivamente.

La validez a nivel de contenido y del constructo de estos instrumentos se realizó entre los meses de enero y marzo de 2023.

Se calcularon los coeficientes de concordancia estrictos (B_N) y ponderados o de acuerdo parcial (B_N^w) de Bangdiwala (Bangdiwala, 1987; Friendly & Meyer, 2016) obtenidos tanto para el instrumento de observación como para los grupos focales. Los coeficientes ponderados de concordancia de Bangdiwala (Bangdiwala, 1987) obtenidos fueron: $B_N^w = .94$ (idoneidad); $B_N^w = .94$ (importancia); $B_N^w = .92$ (claridad de la redacción); y $B_N^w = .88$ (observabilidad). Los resultados obtenidos, de acuerdo con esta interpretación, sugieren un nivel de acuerdo excelente según los coeficientes de concordancia establecidos por Muñoz y Bangdiwala (1997), de modo que los ítems, globalmente, pueden considerarse idóneos, importantes y observables. Además, la claridad de la redacción, a juicio de los expertos, es adecuada para observar la eficacia de un programa formativo para la enseñanza de la historia en bachillerato basado en cambios epistemológicos y metodológicos.

Para completar el análisis se calcularon también los coeficientes de acuerdo AC_2 propuesto por Gwet (2008) para superar las limitaciones y paradojas propias del coeficiente Kappa de Cohen (Cicchetti & Feinstein, 1990; Feinstein & Cicchetti, 1990) especialmente en aquellas situaciones en las que el grado de acuerdo es alto. En este caso se empleó el AC_2 , por tratarse de datos ordinales, ya que permite el acuerdo parcial. Los índices de acuerdo AC_2 obtenidos pueden considerarse excelentes para todas las variables examinadas atendiendo a los criterios propuestos por Muñoz y Bangdiwala (1997), o bien tomando en consideración la propuesta de Gwet (2021) para interpretar los valores obtenidos calculando la probabilidad de pertenencia a cada uno de los intervalos (Interval Membership Probability), las probabilidades acumuladas para el coeficiente de concordancia AC_2 utilizando la escala de referencia propuesta por Muñoz y Bangdiwala (1997). Atendiendo a estos resultados, el nivel de acuerdo obtenido respecto a las conductas contempladas en el instrumento de observación se puede considerar excelente con una confianza superior al 95 % en todas las variables analizadas. En lo que se refiere a los grupos focales, el acuerdo se puede considerar excelente en cuanto a la idoneidad e importancia de los ítems planteados, con una confianza superior al 95 % y se puede considerar bueno o superior con una confianza del 100 % en cuanto a la claridad de la redacción de los ítems.

El análisis de las respuestas del mismo conjunto de jueces sobre su valoración de las preguntas planteadas para los grupos focales sugieren que los ítems, globalmente, pueden considerarse idóneos e importantes y, además, la claridad de la redacción, a juicio de los expertos, es adecuada para plantear las preguntas propuestas en los grupos focales con el objetivo de identificar cambios y permanencias en las prácticas docentes, en el papel del alumnado y en el aprendizaje de competencias históricas dentro del grupo, e identificar el papel que tiene el contexto del centro, el alumnado y el profesorado en los resultados de la experimentación.

Posteriormente, se calcularon las razones de validez de contenido (*Content Validity Ratio*, CVR) (Lawshe, 1975) para cada una de las conductas incorporadas en el instrumento de observación y para cada una de las preguntas de los grupos focales, y en cada una de las variables analizadas y el índice de validez de contenido (*Content Validity Index*, CVI) (Lawshe, 1975) para el conjunto de ítems.



Para la interpretación de los resultados con 9 jueces, Ayre y Scally (2014) proponen el valor crítico de CVR = .778 ($p = .020$), lo que supone que al menos 8 de los 9 jueces estén de acuerdo en la valoración del ítem y supera la probabilidad de acuerdo por efecto del azar a un nivel de confianza del 95 % ($\alpha = .05$), en tanto que para 8 jueces el valor crítico de CVR propuesto por estos autores es .750 ($p = .035$). Para el cálculo de los valores CVR, se agruparon las valoraciones de los jueces que seleccionaron las opciones 2 (bastante) y 3 (mucho) y se calcularon las razones aplicando la fórmula propuesta por Lawshe (1975):

$$RVC = \frac{n_e - \left(\frac{N}{2}\right)}{\left(\frac{N}{2}\right)}$$

En lo que se refiere a las conductas presentes en el instrumento de observación, los valores RVC oscilaron entre .78 y 1 en las cuatro variables consideradas, de modo que todas superaron el valor RVC crítico propuesto por Ayre y Scally (2014) de .778 para 9 jueces. El CVI global para el conjunto de conductas fue .96 (idoneidad), .97 (importancia), .93 (claridad) y .93 (observabilidad).

En cuanto a las preguntas planteadas para los grupos focales, los valores RVC de todos los ítems superaron el valor crítico de .750 ($p = .035$); excepto el ítem 14 del grupo focal del profesorado en activo en idoneidad ($RVC = .5$) y claridad ($RVC = .5$); el ítem 16 del grupo focal dirigido al alumnado en idoneidad ($RVC = .5$); el ítem 6 del grupo focal de profesorado en activo en claridad ($RVC = .5$); y el ítem 8 del grupo focal de profesorado en activo en claridad ($RVC = .25$). Los valores CVI para el conjunto de ítems propuestos para los grupos focales fueron .93 (idoneidad), .96 (importancia) y .84 (claridad). A la luz de estos resultados se revisaron las valoraciones cualitativas de los jueces sobre los ítems señalados que no superaron el umbral de razón de validez de contenido.

