Ardura, D. y Galán, A. (2020). Calibración del resultado de una prueba escrita en estudiantes de ciencias de secundaria: el efecto del sexo. *Revista de Investigación Educativa*, 38(2), 329-344.
DOI: http://dx.doi.org/10.6018/rie.384031

Calibración del resultado de una prueba escrita en estudiantes de ciencias de secundaria: el efecto del sexo

Secondary school science students' calibration of their grade in a written test: the effect of sex

Diego Ardura y Arturo Galán
Departamento de Métodos de Investigación y Diagnóstico en Educación I (MIDE I),
Universidad Nacional de Educación a Distancia (UNED), España

Resumen

Durante las últimas décadas se han encontrado importantes diferencias por sexo en la enseñanza y el aprendizaje de las disciplinas científicas. Por otro lado, la autoevaluación por parte de los estudiantes supone un aspecto fundamental en el ciclo de autorregulación del aprendizaje y, por tanto, en su rendimiento. El objetivo de este trabajo es analizar la metacognición de los estudiantes de secundaria y, en particular, el efecto del sexo en las mismas. Para ello se ha medido la calibración del resultado en una prueba escrita de 487 estudiantes. Nuestros análisis muestran que las chicas calibran mejor su nota que los chicos a pesar de que estos últimos muestran más seguridad en sus juicios. Se ha encontrado una tendencia de ambos sexos a la sobreestimación de sus calificaciones en una prueba escrita. Por otro lado, los estudiantes con rendimiento alto son más precisos y tienden a subestimar sus actuaciones. En cambio, los de rendimiento bajo son más imprecisos y tienden a sobreestimar sus calificaciones en la prueba. Aunque este efecto se observa en ambos sexos, su tamaño es superior en el caso de las chicas. En vista de los resultados, los estudiantes de rendimiento alto utilizan con más eficacia la retroalimentación que generan durante la prueba que los de rendimiento bajo. Las diferencias por sexo podrían tener su origen en las diferentes actitudes y motivaciones de los chicos y las chicas hacia la ciencia.

Palabras clave: educación secundaria; metacognición; calibración; sexo.

Correspondencia: Diego Ardura, dardura@edu.uned.es, Juan del Rosal, 14, D-2.48, 28040 Madrid (España).

Abstract

During the last decades, important sex differences have been found in the context of science education. Besides, self-assessment is crucial to the cycle of self-regulated learning and, consequently, to students' performance. The main goal of the present investigation is to analyze secondary school students' metacognition and, in particular, the effect of gender. To this aim, a sample of 487 students took part in our study. Our analyses show that girls are more accurate than boys in their predictions, despite the latter being more confident. A general tendency towards overestimations has been found for both sexes. Moreover, high-achieving students tend to be more precise and underestimate their performance, while low-achieving students tend to be less precise and overestimate their grade in the test. Although this effect was found in both sexes, the effect size was larger in the case of girls. In light of the results, high-achieving students make a better use of self-generated feedback than low-achievers. Sex differences in calibration could be explained by the different attitudes and motivations of boys and girls towards science.

Keywords: secondary education; metacognition; calibration; sex.

Introducción

En las últimas décadas ha habido un creciente interés por el estudio de las diferencias por sexo en el contexto de la enseñanza de las ciencias, tanto a nivel universitario (Pirmohamed, Debowska y Boduszek, 2017) como en enseñanza secundaria (Abraham y Barker, 2015). La principal razón para esta proliferación de estudios diferenciales por sexo en el ámbito de la educación científica se debe a que se ha observado un porcentaje bajo de mujeres especialmente en las áreas de las matemáticas, la física o la ingeniería (Vázquez y Mannasero, 2015). Por este motivo, diversos estudios recientes han abordado investigaciones por sexo sobre aspectos como las actitudes e interés hacia la ciencia (Potvin y Hasni, 2014), las diferencias en la motivación (Fischer, Schult y Hell, 2013), el abandono de las opciones científicas (Jacobs, 2005) o el rendimiento académico en el contexto de estas disciplinas (Eddy y Bronwell, 2016). Tal y como han demostrado estudios anteriores (Palmer, Burke y Aubusson, 2017), este último podría ser uno de los aspectos clave para equilibrar la presencia de ambos sexos en estudios de ciencia, tecnología, ingeniería y matemáticas (Science, Technology, Engineering and Mathematics, STEM).

El rendimiento académico está íntimamente relacionado con la autorregulación del aprendizaje de los estudiantes[1] (Dent y Koenka, 2016) y esta con sus capacidades de metacognición. Estas dos últimas facetas facilitan en gran medida la monitorización de las tareas que los aprendices abordan cuando estudian (Follmer y Sperling, 2016). A la precisión en los juicios que los estudiantes hacen sobre sus actuaciones en situaciones concretas se la denomina calibración (de Bruin, Kok, Lobbestael y de Grip, 2017). Así, un alumno calibrará mejor su desempeño en una tarea cuanto más se acerque su valoración personal al resultado real que alcanza en la misma. Por

1 Aquellas palabras que, por facilidad en la lectura, no se han podido sustituir por términos genéricos, se entenderá que se refieren indistintamente al sexo femenino o masculino.

tanto, la precisión en estos juicios será fundamental a la hora de tomar decisiones que le permitan autorregular su aprendizaje.

En líneas generales, los estudiantes tienden a ser poco realistas y a sobreestimar, en promedio, su desempeño en las tareas escolares (de Bruin et al., 2017; Zamora y Ardura, 2014; Zamora, Suárez y Ardura, 2018). Por otro lado, en estudios anteriores se ha advertido un efecto diferencial del rendimiento académico sobre la calibración. Mientras que los estudiantes con rendimiento alto son más precisos y tienden a subestimar sus actuaciones, los de rendimiento bajo son más imprecisos y muestran una tendencia acusada a la sobreestimación (efecto Dunning-Kruger) (Kruger y Dunning, 1999). Este efecto se ha caracterizado en investigaciones previas, especialmente en el contexto universitario (Karatjas, 2013, 2014; Lindsey y Nagel, 2015). Los estudios en etapas educativas anteriores son más escasos, pero apuntan en la misma dirección (Bol, Hacker, Walck, Nunnery, 2012).

El momento en el que se solicita que los estudiantes estimen su nota puede ser relevante a la hora de investigar su precisión. La monitorización se puede realizar a priori, es decir, de manera prospectiva, o a posteriori (de manera retroactiva) (Baars, Vink, van Gog, de Bruin y Paas, 2014). Existen trabajos en la literatura sobre el tema que han abordado investigaciones sobre la comparación entre la calibración antes y después de llevar a cabo la tarea. La mayoría de estos estudios han demostrado que la precisión de los estudiantes mejora después de llevar a cabo la tarea (Erickson y Heit, 2015; Gutiérrez y Price, 2016). Sin embargo, otros encuentran que la retroalimentación que el propio estudiante genera para elaborar sus juicios durante el transcurso de la tarea no implica ninguna mejoría en la calibración (Hacker, Bol y Bahbahani, 2008; Schraw, Potenza y Nebelisick-Gullet, 1993). En esta línea, algunos trabajos recientes han mostrado que los estudiantes podrían estar utilizando dos mecanismos diferentes a la hora de elaborar sus juicios de metacognición: uno basado en la precisión y otro en los errores cometidos (Gutiérrez, Schraw, Kuch y Richmond, 2016).

Los estudios de los efectos del sexo sobre la calibración publicados hasta la fecha han revelado resultados muy dispares. Por un lado, algunos trabajos han encontrado que las mujeres muestran una menor confianza que los hombres en diversos ámbitos y tareas (Gutiérrez y Price, 2016). Por otro lado, Sharma y Bewes (2011), en un contexto universitario, y Nietfeld, Shores y Hoffman (2014) en enseñanza media, demostraron que las calibraciones de hombres y mujeres eran muy similares. Por su parte, Karatjas y Webb (2015) encontraron una interacción entre el sexo y el rendimiento académico sobre la calibración. En este estudio no se encontraron efectos del sexo en estudiantes con rendimiento alto ni con rendimiento bajo. Sin embargo, para los de rendimiento intermedio, los hombres sobreestimaron su actuación con respecto a las mujeres en la misma tarea. Estos autores explicaron las diferencias encontradas indicando que la autopercepción de los individuos podría ser clave para explicar las diferencias debidas al sexo. Por su parte, la madurez podría también jugar un papel importante, ya que se ha encontrado que las mujeres de más edad calibran mejor su actuación que los hombres (Volz-Sidiropoulou y Gauggel, 2012).

Método

Objetivo

Como se mencionó anteriormente, existe una escasez de trabajos de investigación en los que se estudie la calibración en el contexto de la enseñanza media. Por tanto, los objetivos de este trabajo son: (1) analizar los porcentajes de estudiantes que sobreestiman, aciertan y subestiman su calificación en la prueba escrita, (2) analizar las diferencias de medias en la calibración de los estudiantes y las relaciones entre las variables implicadas en el estudio y (3) analizar el efecto del rendimiento en la prueba sobre la calibración. En todos los casos se buscará aportar nuevas evidencias sobre el efecto del sexo en los aspectos investigados.

Población y Muestra

Para representar la población española de estudiantes de secundaria, se ha llevado a cabo un muestreo incidental basado en la accesibilidad de los dos centros participantes. Ambos están situados en el norte de España y se trata de centros concertados. El primer centro aportó un 41.7% de los estudiantes de la muestra y el segundo un 58.3%. La muestra se compuso de un total de 487 estudiantes de los que 241 fueron chicos (49.5%) y 246 chicas (50.5%). Estos estudiantes estaban matriculados en los cinco últimos cursos de la enseñanza secundaria en España: desde 2º de ESO a 1º de Bachillerato, ambos incluidos. La distribución de los estudiantes en los diferentes cursos fue la siguiente: 2º de ESO, 163 (33.5%); 3º de ESO, 140 (28.7%); 4º de ESO, 87 (17.9%); 1º de Bachillerato, 97 (19.9%). La edad media de los estudiantes en la muestra fue de 14.5 años. El estudio se llevó a cabo en el contexto de la asignatura de física y química en los cursos mencionados anteriormente y, en garantía de la ética de la investigación, los estudiantes y sus familias dieron su consentimiento informado para participar voluntaria y anónimamente en el estudio.

Instrumento

Los estudiantes emitieron los juicios sobre su desempeño en una prueba escrita que se diseñó *ad hoc* para esta investigación para cada uno de los niveles educativos implicados en la muestra y que se valoró en una escala de 0 a 10. Para ganar validez ecológica, estas pruebas incluyeron diferentes tipos de preguntas que se usan comúnmente en los exámenes de esta asignatura: dos preguntas de verdadero y falso en las que el estudiante debía razonar su respuesta, una pregunta corta abierta como por ejemplo el enunciado de una ley o la definición de un concepto, tres preguntas de elección múltiple tanto teóricas como aplicadas y un problema de aplicación numérica (ver Anexo I). Para realizar la corrección de la prueba se planteó un guion de evaluación que aplicaron, de manera independiente, dos expertos: un profesor y un investigador (ver Anexo I). Después de la corrección individual, los correctores discutieron las discrepancias en la calificación otorgada basándose en el guion de evaluación hasta llegar a un acuerdo.

Durante la recogida de datos se utilizó además un formulario diseñado *ad hoc* para esta investigación, en el que los estudiantes pudieron registrar las estimaciones

de sus notas en cada uno de los dos momentos investigados y la confianza que en esos momentos tenían en sus propios juicios de calibración (juicios de calibración de segundo orden).

Procedimiento de recogida y análisis de datos

El día de la prueba escrita se facilitó a los estudiantes el formulario diseñado para el registro de sus estimaciones. A continuación, se suministró el enunciado del examen a los estudiantes y se solicitó que leyesen las preguntas sin empezar a contestarlas. En ese momento se les pidió que, en vista de las preguntas, hiciesen su primera estimación sobre la nota que creían que iban a obtener en el examen. Además, se les pidió que registrasen en el formulario la seguridad que tenían en esa predicción en una escala tipo Likert de 5 puntos. Al final de la investigación, los centros implicados facilitaron la calificación obtenida por los estudiantes en la asignatura de física y química.

Una vez concluida la recogida de datos y hecha la doble corrección de las pruebas, se calcularon los porcentajes de error en las estimaciones de los estudiantes (error de calibración, EC) empleando el enfoque propuesto por Dulonsky y Metcalfe (2008):

$$EC = \frac{Estimación\ del\ estudiante - Nota\ del\ profesor}{10} \times 100$$

Por tanto, a partir de esta fórmula se definen los errores de calibración para las estimaciones que realiza el estudiante de su nota antes (EC_1) y después de resolver el examen (EC_2). Los estudiantes, como se indicó anteriormente, valoraron la confianza en la certeza de sus estimaciones (juicios de calibración de segundo orden, CSO) en ambos momentos (CSO_1 y CSO_2). Finalmente, a partir de las diferencias entre los errores de calibración antes y después de resolver el examen se define una variable para comparar ambos errores (ΔEC). Conviene tener en cuenta que un signo menos en los errores de calibración implica una subestimación y un signo más se correspondería con una sobreestimación.

Una primera exploración de los datos ha permitido observar la dificultad de que los estudiantes calibren su nota de manera totalmente precisa. Así, en cualquiera de los dos momentos investigados, el porcentaje de estudiantes cuya estimación coincide con la nota otorgada por el docente no ha superado el 4% de los sujetos. Por esta razón, se consideró que los sujetos con un error de calibración menor que un 5% en valor absoluto, fueron lo suficientemente precisos como para considerar correcta la estimación. Por tanto, un error de más del 5% y de menos del -5% se consideraron, respectivamente, una sobreestimación y una subestimación.

Se han llevado a cabo análisis descriptivos y correlacionales. Para los análisis de diferencia de medias entre dos grupos se ha utilizado la prueba t de Student para muestras independientes o relacionadas en función de su objetivo. Para las comparaciones entre más de dos grupos se ha utilizado el análisis de la varianza (ANOVA) y de la covarianza (ANCOVA). Los tamaños del efecto se miden mediante el estadístico

d de Cohen en el caso de las primeras y omega cuadrado en el caso de las segundas. Para todos los análisis estadísticos se ha utilizado el programa SPSS (v.25).

Resultados

Análisis de los porcentajes de error en las calibraciones

En la Tabla 1 se recogen los resultados del análisis del porcentaje de estudiantes que sobreestiman, aciertan y subestiman su calificación en la prueba. La mayoría de los estudiantes, independientemente de su sexo, tiende a sobreestimar su calificación, si bien las chicas lo hacen en menor porcentaje que los chicos. Es interesante advertir que en la estimación que se lleva a cabo después completar la prueba, el porcentaje de chicas que calibra bien su nota aumenta, mientras que el de chicos disminuye. Simultáneamente, para ambos sexos el porcentaje de sobreestimaciones aumenta una vez contestadas las preguntas (ver Tabla 1).

Tabla 1

Comparación de las sobreestimaciones, aciertos y subestimaciones en la calibración en función del sexo

Momento	Calibración	Chicas				Chicos			
		N	%	M	DT	N	%	M	DT
Antes	EC_1 > 5	110	45.1	22.2	13.7	123	51.0	24.9	13.4
	-5 ≤ EC_1 ≤ 5	57	23.2	.5	3.1	58	24.1	1.1	3.0
	EC_1 < -5	79	31.7	-16.0	6.8	60	24.9	-16.5	9.1
Después	EC_2 > 5	128	52.0	21.3	11.5	153	63.5	21.5	13.1
	-5 ≤ EC_2 ≤ 5	75	30.5	-.2	3.3	46	19.1	.11	3.4
	EC_2 < -5	43	17.5	-13.6	6.0	42	17.4	-12.1	5.9

EC_1: Error de calibración antes de contestar; EC_2: Error de calibración después de contestar.

Análisis de las diferencias de medias de calibración en función del sexo

La comparación por sexo del rendimiento en la asignatura arroja diferencias significativas en favor de las chicas (6.51) que promedian medio punto por encima de los chicos (6.13). Es interesante observar que, cuando se consideran todos los estudiantes de cada sexo, se detecta, en primer lugar, una tendencia general a la sobreestimación, que resulta más acusada en el caso de los chicos (ver Tabla 2). Las diferencias entre ambos sexos en cada instante fueron estadísticamente significativas, aunque el tamaño del efecto se puede considerar bajo (ver Tabla 2). En cuanto a la evolución temporal del error de calibración, se observan diferencias para ambos sexos. Por un lado, antes de resolver el examen, el error de calibración de las chicas fue de un 4.91% y después de un 8.66%. Estas diferencias fueron estadísticamente significativas ($t=4.62$, $p < .001$, $d=.21$). Por su parte, en el caso de los chicos, las diferencias también fueron significativas ($t=3.29$, $p =$

.001, d=.15), incrementándose el error de calibración desde un 8.89% hasta un 11.75%. Dada la falta de equivalencia de los sexos en el error de calibración antes de resolver las preguntas del examen, para comparar los errores de calibración según el sexo después de ver las preguntas, se llevó a cabo un análisis de la covarianza (ANCOVA) que indicó que, una vez controlado el error de calibración antes de contestar las preguntas de la prueba, no hubo diferencias por sexo significativas en el error de calibración después de resolver el examen, $F(1,484)=.750$, $p=.450$, $\omega^2 < .01$.

Tabla 2

Descriptivos y diferencias de medias en función del sexo

	Chicas		Chicos		p	d[1]
	M	DT	M	DT		
Rend_FQ	6.51	2.11	6.13	2.22	.039	.18
EC_1	4.91	19.45	8.89	20.53	.028	.20
CSO_1	3.42	.61	3.60	.60	.002	.30
EC_2	8.66	16.49	11.75	15.53	.045	.19
CSO_2	3.53	.70	3.77	.63	<.001	.36
ΔEC	3.75	12.73	2.86	12.73	.441	.07

Rend_FQ: rendimiento en física y química; EC_1: Error de calibración antes de contestar; EC_2: Error de calibración después de contestar; CSO_1: juicio de calibración de segundo orden antes de contestar; CSO_2: juicio de calibración de segundo orden después de contestar; ΔEC: variación en el error de calibración entre la primera y la segunda estimación.
[1] El tamaño del efecto se reporta mediante el estadístico d de Cohen

Finalmente no se observó un efecto significativo del curso ni en la calibración antes de la prueba, $F(3,486)=0.851$, $p=.466$, $\omega^2 < .01$, ni después de la prueba, $F(3,486)=0.682$, $p=.682$, $\omega^2 < .01$. Por otro lado, tampoco se apreció ningún efecto de interacción entre el curso y el género, $F(3,479)=1.228$, $p=.299$, $\omega^2 < .01$. Por tanto, la supuesta mayor madurez que tendrían los alumnos a medida que pasan de curso no implica una mayor precisión en la calibración de resultados.

Los juicios de calibración de segundo orden en los dos instantes también muestran diferencias significativas en función del sexo. En ambos casos, los chicos se muestran más seguros de su predicción que las chicas a pesar de que, como se ha comentado anteriormente, son más imprecisos. En concreto, antes de resolver las preguntas, los chicos muestran una seguridad en la estimación que hacen en ese momento de 3.60, mientras que la seguridad de las chicas es de 3.42. En la segunda estimación los chicos vuelven a sentirse más seguros (3.77) que las chicas (3.53). En este último caso, la diferencia de medias presenta un tamaño del efecto medio. La comparación para cada sexo en cada instante arrojó diferencias estadísticamente significativas tanto para los chicos ($t=4.79$, $p<.001$, $d=.31$) como para las chicas ($t=2.87$, $p = .005$, $d=.17$), mostrándose, por tanto, más seguros de su predicción después de completar el examen que una vez que examinan las preguntas sin todavía haber tenido la ocasión de responder.

Tabla 3

Resultados del análisis de correlación por sexo

	Rend_FQ	EC_1	CSO_1	EC_2	CSO_2	EC
Rend_FQ	1	-.651**	.236**	-.664**	.279**	.134*
EC_1	-.751**	1	.093	.761**	-.052	-.542**
CSO_1	.227**	.055	1	.096	.608**	-.018
EC_2	-.623**	.787**	.044	1	.051	.132*
CSO_2	.251**	.010	.536**	.082	1	.146*
ΔEC	.295**	-.529**	-.028	.107	.132	1

*Rend_FQ: rendimiento en física y química; EC_1: Error de calibración antes de contestar; EC_2: Error de calibración después de contestar; CSO_1: juicio de calibración de segundo orden antes de contestar; CSO_2: juicio de calibración de segundo orden después de contestar; ΔEC: variación en el error de calibración entre la primera y la segunda estimación. *p<.05 **p<.01. NOTA: En la parte superior de la matriz de correlaciones se muestran los coeficientes correspondientes a las chicas y en la inferior las de los chicos.*

Análisis de correlación

En la Tabla 3 se recogen los coeficientes de correlación para chicos (parte inferior de la matriz) y chicas (parte superior de la matriz) entre las parejas de variables seleccionadas para la investigación. Como se puede apreciar, los errores de calibración tanto antes de resolver la prueba como después, presentan una correlación significativa y negativa con el rendimiento en la asignatura. Sin embargo, en el caso de los juicios de segundo orden, las correlaciones con el rendimiento son positivas. Por su parte, los errores de calibración antes y después de resolver el examen presentan una correlación alta tanto en el grupo de las chicas (.761**) como en el de chicos (-.787**). Conviene destacar la ausencia de correlación entre los errores de calibración y los juicios de segundo orden sobre las estimaciones. La variación en los errores de calibración correlaciona, en ambos sexos, con el error de calibración antes de resolver la prueba de manera significativa y negativa. Las únicas parejas de coeficientes estadísticamente diferentes, según el sexo, de todas las investigadas, son las que relacionan la variación en el error de calibración antes y después de la prueba y el error de calibración y la calibración de segundo orden después de resolver la prueba cuyos coeficientes son significativos únicamente en el caso de las chicas (ver Tabla 3).

Tabla 4

Resultados de las variables de calibración en función del rendimiento y el sexo

	Rendimiento	Chicas		Chicos	
		M	DT	M	DT
EC_1	Alto	-10.41	10.15	-10.02	12.15
	Medio	1.86	16.61	7.19	13.84
	Bajo	22.86	14.50	24.97	18.93

	Rendimiento	Chicas		Chicos	
		M	**DT**	**M**	**DT**
CSO_1	Alto	3.62	.64	3.74	.52
	Medio	3.32	.61	3.66	.60
	Bajo	3.36	.56	3.41	.63
EC_2	Alto	-4.26	9.94	-3.35	11.01
	Medio	6.47	12.34	11.10	12.90
	Bajo	23.15	14.59	23.80	17.05
CSO_2	Alto	3.77	.64	3.93	.73
	Medio	3.50	.65	3.86	.60
	Bajo	3.34	.74	3.56	.51
ΔEC	Alto	6.15	8.70	6.67	10.07
	Medio	4.79	14.72	3.90	9.96
	Bajo	.28	12.56	-1.16	15.93

Rend_FQ: rendimiento en física y química; EC_1: Error de calibración antes de contestar; EC_2: Error de calibración después de contestar; CSO_1: juicio de calibración de segundo orden antes de contestar; CSO_2: juicio de calibración de segundo orden después de contestar; ΔEC: variación en el error de calibración entre la primera y la segunda estimación.

Efecto del rendimiento sobre la calibración y los juicios de segundo orden

El análisis de la varianza permitió caracterizar un efecto estadísticamente significativo del rendimiento en la asignatura sobre los errores de calibración antes de resolver el examen tanto en el caso de las chicas, $F(2, 243)=104.28$, $p<.001$, $\omega^2 = .46$, como en el de los chicos, $F(2, 238)=93.73$, $p<.001$, $\omega^2 = .43$. Los análisis *post hoc* revelaron que los estudiantes con rendimiento alto son los que mejor calibran su actuación seguidos de los de rendimiento medio y bajo (ver Tabla 4). Es interesante destacar la tendencia, presente en ambos sexos, a que los estudiantes de rendimiento bajo sobreestimen en gran medida su calificación en la prueba en comparación con los otros dos grupos formados para este análisis. Los resultados del ANOVA para el error de calibración cometido por los estudiantes una vez que han resuelto el examen siguen la misma tendencia: las chicas con rendimiento alto son las que mejor calibran su actuación, $F(2, 243)=91.87$, $p<.001$, $\omega^2 = .42$ y lo mismo pasa con los chicos $F(2, 238)=67.18$, $p<.001$, $\omega^2 = .35$. Finalmente, en ninguno de los dos momentos investigados se observaron efectos de interacción entre el rendimiento y el sexo sobre los errores de calibración: $F(2, 481)=1.267$, $p=.283$, $\omega^2 < .01$ (antes de resolver la prueba) y $F(2, 481)=1.224$, $p=.295$, $\omega^2 < .01$ (después de resolver la prueba).

Los juicios de calibración de segundo orden también se ven afectados significativamente por el rendimiento. Cuando los estudiantes elaboran el juicio antes de responder a las preguntas, los que tienen un rendimiento más alto muestran una mayor seguridad en su estimación que los de rendimiento medio y bajo (ver Tabla 4), $F(2, 243)=5.274$, $p=.006$, $\omega^2 =.03$, para las chicas y $F(2, 238)=6.744$, $p=.001$, $\omega^2 =.05$, para los chicos. Una vez que han resuelto la prueba, este efecto del rendimiento en la seguridad del juicio siguió presente en ambos sexos: $F(2, 243)=7.005$, $p=.001$, $\omega^2 = .05$ (chicas) y $F(2, 238)=10.46$, $p < .001$, ω^2

= .05, (chicos). Como en el caso de los errores de calibración, no se detectaron efectos de interacción entre el sexo y el rendimiento ni antes de responder a las preguntas del examen, F(2, 481)=2.951, p=.053, ω^2 = .01, ni después, F(2, 481)=1.038, p=.355, ω^2 < .01.

En cuanto a los errores de calibración antes y después de resolver la prueba, se observaron diferencias estadísticamente significativas en función del rendimiento en la asignatura en el caso de los chicos, F(2, 243)=4.607, p = .011, ω^2 = .03, y en el de las chicas, F(2, 238)=7.723, p = .001, ω^2 = .05 (ver Tabla 4). Como en los casos anteriores no se detectó ningún efecto de interacción entre el sexo y el rendimiento académico, F(2, 481)=.236, p=.790, ω^2 < .01.

Discusión y conclusiones

El objetivo de este trabajo fue investigar las diferencias por sexo en la capacidad de los estudiantes de la asignatura de física y química de secundaria para predecir sus calificaciones en una prueba escrita. Dado que el momento en el que se solicitan las estimaciones puede resultar importante, se recogieron datos antes y después de que los estudiantes realizasen la prueba.

El análisis de las calificaciones de los estudiantes en la asignatura confirmó la presencia de diferencias por sexo en el rendimiento en favor de las chicas. Este resultado apunta en la línea de algunas investigaciones recientes que demuestran que a nivel de enseñanza secundaria las chicas rinden mejor que los chicos (Fischer et al., 2013). Las razones que se han esgrimido para racionalizar estas diferencias pasan por una multitud de factores como, por ejemplo, una mayor motivación hacia el logro en el caso de las chicas (Fischer et al., 2013) o su mayor percepción de la utilidad de la ciencia (Acar, Türkmen y Bilgin, 2015). Es interesante destacar que las investigaciones anteriores con estudiantes universitarios plantean la tendencia opuesta, siendo generalmente los hombres los que muestran un mejor rendimiento (Eddy y Bronwell, 2016).

En lo referente a la calibración de los estudiantes, una primera cuestión a subrayar es la dificultad que entraña la tarea que se ha planteado a los estudiantes. En efecto, el porcentaje de los mismos con una buena calibración es bajo. En el caso de las chicas se incrementa ligeramente después de resolver la prueba y en el de los chicos, disminuye. En comparación con estudios anteriores, se observa una peor calibración. Las mayores desviaciones que hemos encontrado en este trabajo eran previsibles dado que, en primer lugar, nuestra investigación se he llevado a cabo en un aula y, por tanto, no se trata de una situación tan controlada como la mayoría de los estudios anteriores que se desarrollaron en condiciones de laboratorio (Hacker, Bol, Horgan y Rakow, 2000). En segundo lugar, el tipo de prueba que hemos utilizado, precisamente por aumentar la validez ecológica de nuestra investigación, contiene preguntas abiertas que posiblemente sean más difíciles de calibrar por parte de los estudiantes que las utilizadas habitualmente en este campo que suelen ser preguntas de elección múltiple (Hacker, et al., 2008).

Nuestros análisis apuntan a una tendencia de ambos sexos a la sobreestimación de sus calificaciones en comparación con la otorgada con el profesor a nivel de enseñanza secundaria. Estos resultados confirman los anteriormente publicados, principalmente en el contexto de la enseñanza de las disciplinas científicas en la universidad (Dunning,

2005; Hacker, et al., 2008). Esta tendencia a la sobreestimación se debe, no solo a que el porcentaje de estudiantes que se muestran excesivamente optimistas sobre su calificación es mayor que los que aciertan y los que son pesimistas, sino también a que los errores de calibración en el caso de los estudiantes que sobreestiman son de mayor magnitud que los de los que subestiman su nota. Nuestra investigación ha revelado que las chicas son más precisas que los chicos en los dos momentos investigados. Diversos estudios por sexo anteriores en el contexto de la enseñanza de la ciencia han demostrado diferencias de actitud y motivación hacia la ciencia entre hombres y mujeres que podrían explicar este hecho (Abraham y Barker, 2015; Mutjaba y Reiss, 2013). Otros trabajos han encontrado niveles de autoeficacia más altos en hombres que en mujeres (Glynn, Brickman, Armstrong y Taasoobshirazi, 2013; Schumm y Bogner, 2016). En vista de los resultados que se presentan en este trabajo, esta tendencia podría llevar a los chicos a juicios excesivamente optimistas sobre su desempeño.

De acuerdo con el análisis de correlaciones y los ANOVA que se presentan en este trabajo, existe una relación entre el rendimiento académico y la precisión de los juicios de los estudiantes. Así, aquellos alumnos con un rendimiento académico alto calibran mejor su actuación que los de rendimiento medio y bajo; además, estos últimos tienden a cometer un error de calibración mayor que los primeros. Estos resultados respaldan la presencia, en nuestra muestra de estudiantes, del efecto Dunning-Kruger presentado en la introducción. Este efecto se ha observado previamente tanto en el dominio de la enseñanza de la ciencia (Karatjas, 2013; Linsey y Nagel, 2015) como en general (Hacker, et al., 2008; Kruger y Dunning, 1999). Además, se ha manifestado para ambos sexos, si bien cabe destacar que el tamaño del efecto es el doble para los chicos que para las chicas.

Como se indicó en la introducción, el momento en el que se solicita la estimación al estudiante podría ser relevante en su precisión. En la presente investigación se ha encontrado que los estudiantes de ambos sexos son más optimistas en sus juicios una vez concluida la prueba que al principio de la misma, lo que provoca una calibración peor. Esto parece indicar que la retroalimentación autogenerada por los estudiantes durante el transcurso de la prueba les lleva a empeorar el juicio sobre el desempeño en la prueba. Este hecho contrasta con estudios previos que demuestran que la retroalimentación interna mejora la calibración (Brannick, Miles y Kisamor, 2005) o que no influye en ella (Guitérrez y Price, 2016; Schraw, Potenza y Nebelsick-Gullet, 1993). Una posible explicación del comportamiento observado en este estudio, podría ser que en la primera estimación los estudiantes se mostraran más conservadores ante la posibilidad de que, una vez abordada la tarea, se pudieran presentar complicaciones. Es interesante destacar, en esta línea de razonamiento, que las chicas muestran una menor seguridad en sus predicciones que los chicos, pero son estos últimos los que peor calibran su desempeño en la prueba. Alternativamente, conviene tener en cuenta que, en algunos estudios anteriores, se ha comprobado la dificultad de los estudiantes de secundaria para localizar y caracterizar sus propios errores (Zamora y Ardura, 2014; Zamora, Suárez y Ardura, 2018). Por lo tanto, el hecho de que los estudiantes lleven a cabo la prueba no parece una garantía para generar una información de suficiente calidad como para que la estimación que hacen de su nota al finalizar la prueba mejore significativamente.

Los resultados encontrados en este trabajo constituyen una primera aproximación a una situación de estudio en el que el estudiante monitoriza su aprendizaje. Este mecanismo le permitiría para tomar decisiones estratégicas tan importantes como cuándo terminar su sesión de estudio o cómo abordar aquellos aspectos en los que percibe que debe mejorar. Nuestros resultados apuntan a que la retroalimentación generada por los propios estudiantes puede no ser suficiente para favorecer su autorregulación, por lo que el profesorado debe facilitarles estrategias para mejorar su capacidad de autoevaluación. En esta línea se ha demostrado en una investigación anterior que la precisión en los juicios de los estudiantes se puede mejorar mediante la instrucción adecuada (Baars et al, 2014).

Además de las limitaciones habituales de la investigación en educación, se deben tener en cuenta algunas limitaciones específicas a la hora de interpretar los resultados que se presentan y que deberían conducir al planteamiento de nuevos estudios. En primer lugar, los juicios de los estudiantes pueden estar influenciados por el hecho de que se recogen durante un examen con un cierto peso en su evaluación final. La ansiedad derivada de la situación podría provocar imprecisiones añadidas a las estimaciones de los alumnos. Por tanto, convendría extender la investigación utilizando tareas que no tengan una repercusión en la calificación de los estudiantes y que sean más cercanas a las condiciones de estudio. Debido a que se buscaba ganar en validez ecológica, la prueba que se ha utilizado para medir la calibración contenía los tipos de preguntas más comúnmente utilizados en exámenes de física y química en secundaria. A partir de los resultados generales obtenidos en el presente estudio, se podría abordar una investigación para analizar el efecto del tipo de pregunta en la calidad de las estimaciones de los estudiantes. En vista de los resultados, parece conveniente plantear investigaciones que permitan estudiar el efecto en la calibración del uso de retroalimentación externa a los estudiantes que pudiera ser suministrada por los docentes como pueden ser las rúbricas o los guiones de evaluación.

Finalmente, otra potencial aplicación, a juicio de los autores, se encuentra en las asignaturas metodológicas en los títulos de Educación, donde los estudiantes suelen anticipar un mayor fracaso en el rendimiento académico. Acciones de evaluación y calibración intermedias durante el curso podrían ayudar a vencer las creencias erróneas sobre el aprendizaje.

Referencias

Abraham, J. y Barker, K. (2015). Exploring gender difference in motivation, engagement and enrolment behavior of senior secondary physics students in New South Wales. *Research in Science Education, 45*(1), 59-73, doi: 10.1007/s11165-014-9413-2.

Acar Ö., Türkmen L., y Bilgin A., (2015). Examination of Gender Differences on Cognitive and Motivational Factors that Influence 8th Graders' Science Achievement in Turkey. *Eurasia Journal of Mathematics Science Technology Education*, *11*(5), 1027–1040, doi: 10.12973/eurasia.2015.1372a

Baars, M., Vink, S., van Gog, T., de Bruin, A., y Paas, F. (2014). Effects of training self-assessment and using assessment standards on retrospective and prospective

monitoring of problem solving. *Learning and Instruction*, *33*, 92–107, doi: 10.1016/j.learninstruc.2014.04.004

Bol, L., Hacker, D. J., Walck, C. C., y Nunnery, J. A. (2012). The effects of individual or group guidelines on the calibration accuracy and achievement of high school biology students. *Contemporary Educational Psychology*, *37*(4), 280–287, doi: 10.1016/j.cedpsych.2012.02.004

Brannick, M. T., Miles, D. E., y Kisamore, J. L. (2005). Calibration between student mastery and self-efficacy. *Studies in Higher Education*, *30*(4), 473–483, doi: 10.1080/03075070500160244

Brown, G. T. L., Andrade, H. L., y Chen, F. (2015). Accuracy in student self-assessment: directions and cautions for research. *Assessment in Education: Principles, Policy & Practice*, *22*(4), 444–457, doi: 10.1080/0969594X.2014.996523

Chiu, M. M., y Klassen, R. M. (2010). Relations of mathematics self-concept and its calibration with mathematics achievement: Cultural differences among fifteen-year-olds in 34 countries. *Learning and Instruction*, *20*(1), 2–17, doi: 10.1016/j.learninstruc.2008.11.002

de Bruin, A. B. H., Kok, E. M., Lobbestael, J. y de Grip, A. (2017). The impact of an online tool for monitoring and regulating learning at university: overconfidence, learning strategy, and personality. *Metacognition and Learning*, *12*(1), 21–43, doi: 10.1007/s11409-016-9159-5

Dent, A. L., y Koenka, A. C. (2016). The Relation Between Self-Regulated Learning and Academic Achievement Across Childhood and Adolescence: A Meta-Analysis. *Educational Psychology Review*, *28*(3), 425–474, doi: 10.1007/s10648-015-9320-8

Dunlosky, J., y Metcalfe, J. (2008). *Metacognition*. Los Angeles, CA: SAGE Publications

Dunning, D. (2005). *Self-insights: Roadblocks and detours on the path of knowing thyself*. New York: Psychology Press.

Eddy, S. L., y Brownell, S. E. (2016). Beneath the numbers: A review of gender disparities in undergraduate education across science, technology, engineering, and math disciplines. *Physical Review Physics Education Research*, *12*(2), 020106, doi: 10.1103/PhysRevPhysEducRes.12.020106

Erickson, S., y Heit, E. (2015). Metacognition and confidence: comparing math to other academic subjects. *Frontiers in Psychology*, *6*, 742, doi: 10.3389/fpsyg.2015.00742

Fischer, F., Schult, J., y Hell, B. (2013). Sex differences in secondary school success: Why female students perform better. *European journal of psychology of education*, *28*(2), 529-543, doi: 10.1007/s10212-012-0127-4

Follmer, D. J., y Sperling, R. A. (2016). The mediating role of metacognition in the relationship between executive function and self-regulated learning. *British Journal of Educational Psychology*, *86*(4), 559–575, doi: 10.1111/bjep.12123

Glynn, S. M., Brickman, P., Armstrong, N., y Taasoobshirazi, G. (2011). Science motivation questionnaire II: Validation with science majors and nonscience majors. *Journal of Research in Science Teaching*, *48*(10), 1159–1176, doi: 10.1002/tea.20442

Gutiérrez, A. P., y Price, A. F. (2017). Calibration between undergraduate students' prediction of and actual performance: The role of gender and performance attributions. *The Journal of Experimental Education*, *85*(3), 486-500, doi: 10.1080/00220973.2016.1180278

Gutierrez, A. P., Schraw, G., Kuch, F., y Richmond, A. S. (2016). A two-process model of metacognitive monitoring: Evidence for general accuracy and error factors. *Learning and Instruction*, *44*, 1–10, doi: 10.1016/j.learninstruc.2016.02.006

Hacker, D. J., Bol, L., y Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: the effects of incentives, reflection, and explanatory style. *Metacognition and Learning*, *3*(2), 101–121, doi: 10.1007/s11409-008-9021-5

Hacker, D. J., Bol, L., Horgan, D. D., y Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology*, *92*(1), 160–170, doi: 10.1037/0022-0663.92.1.160

Hacker, D. J., Bol, L., y Keener, M. C. (2008). Metacognition in education: A focus on calibration. In J. Dunlosky y R. A. Bjork (Eds.), *Handbook of metamemory and memory* (p. 429455). New York: Taylor & Francis Group.

Hawker, M. J., Dysleski, L., y Rickey, D. (2016). Investigating General Chemistry Students' Metacognitive Monitoring of Their Exam Performance by Measuring Postdiction Accuracies over Time. *Journal of Chemical Education*, *93*(5), 832–840, doi: 10.1021/acs.jchemed.5b00705

Jacobs, J.E. (2005). Twenty-five years of research on gender and ethnic differences in math and science career choices: What havewe learned? En J.E. Jacobs & S.D. Simpkins (Eds.), New Directions for Child and Adolescent Development, 110, 85–94. doi: 10.1002/cd.151

Karatjas, A. G. (2013). Comparing College Students' Self-Assessment of Knowledge in Organic Chemistry to Their Actual Performance. *Journal of Chemical Education*, *90*(8), 1096–1099, doi: 10.1021/ed400037p

Karatjas, A. G. (2014). Use of Student Self-Assessment of Exams To Investigate Student Learning in Organic Chemistry Classes. En Kendhammer, L. K. y Murphy, K. L. (Eds.) *Innovative Uses of Assessments for Teaching and Research* (pp. 133–143). American Chemical Society, doi: 10.1021/bk-2014-1182.ch008

Karatjas, A. G., y Webb, J. (2015). The Role of Gender in Grade Perception in Chemistry Courses. *Journal of College Science Teaching*, *45*(2), 30–35, doi: 10.20429/ijsotl.2017.110214

Kruger, J., y Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134.

Lindsey, B. A., y Nagel, M. L. (2015). Do students know what they know? Exploring the accuracy of students' self-assessments. *Physical Review Special Topics - Physics Education Research*, *11*(2), 20103, doi: 10.1103/PhysRevSTPER.11.020103

Mujtaba, T., y Reiss, M. J. (2013). What Sort of Girl Wants to Study Physics After the Age of 16? Findings from a Large-scale UK Survey. *International Journal of Science Education, 35*(17), 2979–2998, doi: 10.1080/09500693.2012.681076

Nietfeld, J. L., Shores, L. R., y Hoffmann, K. F. (2014). Self-regulation and gender within a game-based learning environment. *Journal of Educational Psychology*, *106*(4), 961–973, doi: 10.1037/a0037116

Palmer T.-A., Burke P. F., y Aubusson P. (2017). Why school students choose and reject science: a study of the factors that students consider when selecting subjects. *Int. J. Sci. Educ.*, **39**(6), 645–662, doi: 10.1080/09500693.2017.1299949

Pirmohamed, S., Debowska, A., y Boduszek, D. (2017). Gender differences in the correlates of academic achievement among university students. *Journal of Applied Research in Higher Education, 9*(2), 313-324. doi: 10.1108/JARHE-03-2016-0015

Potvin P., y Hasni A., (2014). Interest, motivation and attitude towards science and technology at K-12 levels: a systematic review of 12 years of educational research. *Studies in Science Education, 50*(1), 85–129, doi: 10.1080/03057267.2014.881626

Schraw, G., Potenza, M. T., y Nebelsick-Gullet, L. (1993). Constraints on the calibration of performance. *Contemporary Educational Psychology*, *18*(4), 455–463, doi: 10.1006/ceps.1993.1034

Sharma, M. D., y Bewes, J. (2011). Self-monitoring: Confidence, academic achievement and gender differences in Physics. *Journal of Learning Design*, *4*(3), 1–13, doi: 10.5204/jld.v4i3.76

Schumm, M. F., y Bogner, F. X. (2016). Measuring adolescent science motivation. *International Journal of Science Education*, *38*(3), 434–449, doi: 10.1080/09500693.2016.1147659

Vázquez Alonso, Á., y Manassero Más, M. (2015). La elección de estudios superiores científico-técnicos: análisis de algunos factores determinantes en seis países. *Revista Eureka sobre enseñanza y divulgación de las ciencias, 12*(2), 264-277.

Volz-Sidiropoulou, E., y Gauggel, S. (2012). Do subjective measures of attention and memory predict actual performance? Metacognition in older couples. *Psychology and Aging*, *27*(2), 440–450, doi: 10.1037/a0025384

Zamora Á., y Ardura D., (2014). ¿En qué medida utilizan los estudiantes de Física de Bachillerato sus propios errores para aprender? Una experiencia de autorregulación en el aula de secundaria. *Enseñanza de las ciencias*, *32*(2), 253–268, doi: 10.5565/rev/ensciencias.1067

Zamora Á., Suárez J. M., y Ardura D., (2018). Error detection and self-assessment as mechanisms to promote self-regulation of learning among secondary education students. *Journal of Educational Research*, *111*(2), 175–185, doi: 10.1080/00220671.2016.1225657

Anexo I

Prueba de evaluación y guion de evaluación empleados en 3º de ESO

PREGUNTA 01 (2 puntos). Indica si las siguientes afirmaciones son verdaderas o falsas. Razona la respuesta.

(a) Todas las hipótesis se convierten, tarde o temprano, en leyes.

Indica que la afirmación es FALSA	0.3p
No todas las hipótesis llegan a ser leyes, pues un experimento podría demostrar que algunas son falsas	0.7p

(b) La expresión correcta en notación científica del número 0,000782 es $78{,}2 \cdot 10^{-5}$

Indica que la afirmación es FALSA	0.3p
En notación científica sólo debe haber un dígito distinto de cero a la izquierda de la coma. Por tanto, la expresión correcta sería: $7{,}82 \cdot 10^{-4}$	0.7p

PREGUNTA 02 (1 punto). Enumera las fases del método científico.

El proceso de medida implica definir una unidad y …	0.3p
… compararla con la magnitud que se desea medir	0.3p
Se puede poner un ejemplo con la longitud en el que se muestra la longitud que deseamos medir, como definimos una unidad y la comparamos con la magnitud a medir	0.4p

PREGUNTA 03 (3 puntos). Rodea con un círculo la respuesta correcta de entre las que se dan en cada pregunta.

No es una magnitud fundamental del sistema internacional de Unidades:
(a) La masa (b) El tiempo (c) La superficie (d) La temperatura

El valor verdadero de una medida es 4,00 cm. ¿Cuál será el error relativo cometido si una balanza mide 4,50 cm?
(a) 1,25% (b) 0,125% (c) 25,0% (d) 12,5%

Las unidades de densidad en el Sistema Internacional son:
(a) g/cm^3 (b) m^3 (c) kg/L (d) kg/m^3

PREGUNTA 04 (4 puntos). El cobre tiene una densidad de 9 g/cm^3. Sabiendo que el volumen de un trozo de cobre es de 2L, ¿cuál será su masa?

Toma los datos y asocia las variables correspondientes a los mismos	0.5p
Lleva a cabo los cambios de unidades oportunos. Lo más sencillo sería para los 2L a cm^3. Alternativamente, se podrían pasar todas al Sistema Internacional	1p
Escribe correctamente la fórmula de la densidad	0.5p
Despeja adecuadamente la masa	1p
Sustituye correctamente los datos y obtiene la solución (m=18000 g ó m=18kg), indicando correctamente sus unidades.	1p

Fecha de recepción: 17 de junio de 2019.
Fecha de revisión: 4 de julio de 2019.
Fecha de aceptación: 17 de julio de 2019.