Actas de las Jenui, vol. 5. 2020.
Páginas: 205-212

# Influencia de la evaluación entre pares consensuada en la precisión de las autoevaluaciones

Juan Ramón Rico-Juan Cristina Cachero Hermenegilda Macià
Dpto. Lenguajes y Sistemas Informáticos
Universidad de Alicante
03690 Alicante
juanramonrico@ua.es ccachero@dlsi.ua.es

Dpto. Matemáticas
Universidad Castilla-La Mancha
02071 Albacete
hermenegilda.macia@uclm.es

## Resumen

En este trabajo se analiza el impacto del proceso de evaluación entre pares en la precisión con la que el alumno es capaz de autoevaluar su trabajo. El propósito de este análisis es discernir si la modalidad de evaluación entre pares a la que se ve expuesto el alumno (corrección individual, en parejas o en tríos) afecta su capacidad de autoevaluación. Para ello, se presenta un quasi-experimento realizado en la Universidad de Castilla-La Mancha con una muestra de 82 estudiantes de primer curso del grado en Ingeniería Informática divididos en tres grupos (A1, A2, B1). Los estudiantes realizaron tres entregas, y en cada una evaluaron a sus compañeros con una modalidad distinta. Antes de comenzar dicha evaluación, los estudiantes autoevaluaron su propio trabajo. Asimismo, los estudiantes volvieron a autoevaluarse tras evaluar a sus compañeros. La calificación del profesor de esas mismas entregas se usó para calcular la precisión de la autoevaluación. Por último, se analizaron las diferencias en precisión de los estudiantes antes y después de participar en el proceso de evaluación entre pares. Los resultados muestran que la modalidad de evaluación entre pares aplicada no afecta significativamente a la precisión de los alumnos a la hora de evaluar su propio trabajo.

## Abstract

This paper analyzes the impact of the peer evaluation process on the accuracy with which the student is able to self-evaluate his work. The purpose of this analysis is to discern whether the modality of peer evaluation to which the student is exposed (individual correction, in pairs or in trios) affects his self-assessment ability. For this, a quasi-experiment is presented at the Universidad de Castilla-La Mancha. The study used a sample of 82 first year students of the Computer Engineering degree. The students were divided into three groups (A1, A2, B1). Then, they were asked to complete three assignments. In each one they evaluated their classmates with a different peer evaluation modality. Additionally, before beginning each evaluation, the students self-assessed their own work. Likewise, the students reassessed themselves after evaluating their classmates. The teacher's grade of those same assignments was used to calculate the accuracy of the self-assessment. Finally, differences in student accuracy were analyzed before and after participating in the peer review process. The results show that the peer evaluation modality applied does not significantly affect the accuracy of the students when evaluating their own work.

## Palabras clave

Evaluación por pares, EP, estudio empírico, respuestas consensuadas, quasi-experimento, autoevaluación

## 1. Introducción

La labor de evaluación y retroalimentación al alumnado es una tarea intrínsecamente compleja, y que, dependiendo de la actividad, puede además requerir una cantidad de tiempo elevada a los instructores. Si a este hecho le añadimos aulas con un número elevado de alumnos, como es el caso en la universidad española, la revisión y retroalimentación individual por parte del instructor en un intervalo de tiempo aceptable se hace a menudo inviable [6]. Es por ello que, en función del tipo de actividad, es común encontrarnos con formas de evaluación alternativas a dicha revisión individual.

En actividades de respuesta cerrada es habitual que exista una sola respuesta correcta. Esta característica permite realizar una corrección automatizada de manera relativamente fácil [15]. Sin embargo, esta opción no es viable para actividades de respuesta abierta. Es precisamente para este tipo de actividades para las que el proceso de la evaluación por pares (EP) resulta más adecuado. El motivo es que la EP ofrece diferentes ventajas tanto al profesor - instructor como al alumno -

evaluador [10]. Por la parte del instructor, permite realizar una evaluación formativa en la que los alumnos reciban una retroalimentación rápida sobre los trabajos realizados con una inversión razonable de tiempo, dado que su labor se limita a la de supervisión de todo el proceso. Por su parte, el alumno adquiere un rol de evaluador sobre un trabajo cuyo enunciado ya conoce, dado que ha tenido que realizarlo, y tiene que tomar una serie de decisiones sobre otros trabajos que han realizado compañeros, lo que le obliga a ser crítico en diferentes aspectos para determinar el nivel de acierto en cada uno de ellos [12]. Aún más, si la evaluación se hace por grupos es de esperar que las discusiones entre iguales ayuden a entender mejor los conceptos estudiados que si los explicara el profesor [8]. Una manera de comprobar si se ha producido este aprendizaje es midiendo la precisión de la autoevaluación del alumno antes y después del proceso general de EP. De esta forma se pueden detectar problemas en el aprendizaje de ciertos estudiantes y facilitar una intervención adecuada.

Cuando se realiza una EP es habitual proporcionar una rúbrica [11] como guía para estandarizar los criterios y que sea más fácil la evaluación del trabajo [2]. Además, el uso de rúbricas tiene un efecto positivo en el proceso de aprendizaje de los estudiantes [12, 3]. Existen diferentes trabajos que muestran algunas de las posibilidades que abre el uso de la EP en situaciones concretas; por ejemplo, en [14] se presenta una metodología basada en cálculos estadísticos para evaluar tanto las actividades de los alumnos como la calidad del trabajo realizado por los evaluadores. Por su parte, en [7] el autor utiliza técnicas de factorización matricial para proporcionar a los estudiantes calificaciones consistentes junto a una retroalimentación de forma automatizada.

El objetivo general de este trabajo es contribuir a esta línea de investigación de EP, analizando cómo afecta la participación de los evaluadores en distintas modalidades de EP (individual, por parejas y por tríos) en la precisión de su autoevaluación. Para ello, este artículo está estructurado de la siguiente manera: la sección 2 presenta el diseño experimental utilizado en el estudio; los detalles de la ejecución se encuentran en la sección 3; el análisis sobre los resultados obtenidos se muestran en la sección 4; la discusión sobre los resultados más relevantes se comenta en la sección 5 y finalmente, en la sección 6 se exponen las conclusiones y las ideas para trabajos futuros.

## 2. Diseño experimental

Con el fin de aumentar el acervo empírico relacionado con la EP consensuada, entre septiembre y diciembre del 2019 se realizó un quasi-experimento, que se define como un tipo de experimento controlado donde no es posible seleccionar los sujetos de la muestra de manera aleatoria. De este modo es posible estudiar relaciones de causa-efecto en escenarios como el educativo, donde dicha selección aleatoria de sujetos es muy complicada [5].

Siguiendo la estructura de la plantilla GQM [13], el propósito de este estudio fue analizar el impacto de tres modalidades diferentes de EP (individual, parejas, tríos) sobre la precisión de la autoevaluación del trabajo de cada sujeto. El diseño del experimento se basó en el marco experimental propuesto por Wohlin *et al.* [16]. La pregunta de investigación (PI) se diseñó de forma que pudiese ser contestada de manera cuantitativa como sigue:

- PI: ¿Existen diferencias significativas en la precisión de autoevaluación de los sujetos en función de la modalidad de EP en la que han participado?

En este estudio se planeó la recopilación de datos de los 82 estudiantes matriculados en la asignatura de Cálculo y Métodos Numéricos, impartida en el primer cuatrimestre del primer curso del grado en Ingeniería Informática de la Universidad de Castilla-La Mancha, en la Escuela Superior de Ingeniería Informática de Albacete, divididos en tres grupos (A1, A2, B1). Además, se seleccionaron tres actividades abiertas de la asignatura (Act1, Act2, Act3). Para cada una de las actividades, se asignó al azar una modalidad de evaluación a cada grupo, de manera que el diseño estuviera balanceado. El diseño final puede ser visto en el Cuadro 1.

| ModalidadEP | Act1 | Act2 | Act3 |
|---|---|---|---|
| Individual | B1 | A2 | A1 |
| Parejas | A1 | B1 | A2 |
| Tríos | A2 | A1 | B1 |

Cuadro 1: Diseño experimental.

Previo a la EP, para la ejecución del estudio en cada grupo se planificaron los siguientes pasos a seguir con cada actividad:

1. La resolución de uno o dos problemas. Era la misma para todos los grupos y se realizó a la vez en todos los grupos (por la tarde, fuera del horario reglado de clase de la asignatura de cada grupo). Después del plazo establecido para su realización (1h), entregaban la hoja de papel con la solución al instructor.
2. A continuación, el instructor hacía una fotocopia en papel de cada entrega, volvía al aula donde se había realizado la prueba y entregaba a cada estudiante la fotocopia de su trabajo, la cual tenían que convertir en un documento digital (foto,

pdf) que subían a los correspondientes talleres de Moodle.
3. El instructor, antes de la sesión de EP en cada grupo, establecía los grupos de evaluación según la modalidad (ver Cuadro 1) y planificaba en Moodle la asignación de los evaluadores a los trabajos presentados.

A modo de resumen, a continuación se describe en qué consistió cada actividad que se propuso para la EP. También se muestra la fecha de ejecución de cada una de ellas.

- Actividad 1 (Act1): un solo ejercicio donde se tenía que aplicar el método de inducción matemática para comprobar que se cumplía una fórmula. (30 septiembre 2019)
- Actividad 2 (Act2): un ejercicio de optimización y otro ejercicio donde había que utilizar el polinomio de Taylor para aproximar el valor de una función en un punto. (5 noviembre 2019)
- Actividad 3 (Act3): dos ejercicios de calcular áreas de regiones planas, uno siendo el área comprendida entre dos curvas, y el otro entre la gráfica de una función y el eje OX. (10 diciembre 2019)

Para cada actividad, se preparó un documento que incluía la solución detallada de la misma, así como la rúbrica para su evaluación. Las rúbricas de corrección se diseñaron de modo que su comprensión y aplicación por parte de los alumnos tuviese una complejidad similar. Todos los documentos facilitados a los estudiantes para la realización de la EP pueden consultarse en `http://bit.ly/2Rueucl`. Las líneas generales de la rúbrica eran conocidas de antemano, pero no el detalle para cada caso concreto de la solución.

El proceso de EP para cada actividad se realizó independientemente en cada grupo (entre 1 y 4 días después de la entrega) en una sesión (1h 30m) de su horario de clase. Se empezó con la entrega en papel a cada estudiante de la solución de la actividad y la rúbrica (que también estaba disponible previamente en Moodle) y se siguieron estos 5 pasos:

1. Explicación y detalles de la rúbrica por parte de la instructora.
2. Autoevaluación individual de la actividad aplicando la rúbrica.
3. División de los alumnos en equipos según la modalidad de evaluación.
4. EP de los trabajos asignados con la modalidad asignada para esa actividad en ese grupo.
5. Nueva autoevaluación individual aplicando los conocimientos adquiridos durante la EP.

**Variables y escalas de medición**

- VD: Precisión (PREC): Medida de tipo ratio. Refleja la diferencia (en valor absoluto) entre la autoevaluación del alumno y la calificación otorgada por el profesor. Las calificaciones se apoyan en rúbricas específicas de actividad. Rango de la medida [0..10].
- VI:
  1. ModalidadEP: factor intra-sujeto, de tipo nominal, que refleja la Modalidad de EP en la que han participado para corregir las actividades de los compañeros: 1 (individual), 2 (parejas) y 3 (tríos).
  2. Tiempo: factor intra-sujeto. Se han definido dos puntos de tiempo: 1 (Pre, es decir, antes de participar en la EP) y 2 (Post, después de participar en la EP). Estos dos puntos de tiempo están representados en los datos por dos medidas: PREC_Pre y PREC_Post.
  3. Grupo: factor inter-sujeto que refleja el grupo de prácticas en el que está integrado el sujeto. Variable de tipo nominal con tres valores posibles: A1, A2, B1.

**Hipótesis**

A partir de la pregunta de investigación formulada y las variables y medidas definidas, se estableció la siguiente hipótesis nula, susceptible de ser refutada estadísticamente mediante el método de Refutación de Hipótesis:

- $H1_0$: el impacto de las distintas modalidades de EP en la evolución de la precisión de autoevaluación de los estudiantes es el mismo independientemente del grupo al que pertenezcan ($\mu 1 = \mu 2 = \mu 3$).

# 3. Ejecución del estudio

La recogida de datos tuvo lugar durante las sesiones de clase donde se realizaba la EP. El Cuadro 2 muestra, para cada grupo (#Grupo) su número de alumnos matriculados (#Matriculados) (82 en total), el número de estudiantes en cada grupo que realizó cada actividad y las subió al Moodle (#Alumnos), cuántos de ellos participaron en cada modalidad del proceso de EP (columna #PRE-POST) y la Actividad con la que realizaron tanto su autoevaluación como la evaluación de los compañeros (#Actividad). Solamente aquellos estudiantes que habían realizado la actividad podían participar en la EP. Un análisis de los alumnos que no se presentaron a la actividad o que, habiéndose presentado, no realizaron posteriormente la autoevaluación y la EP, descartó que el motivo fuera la propia EP, por lo

que se asume que estos abandonos no afectan a los resultados obtenidos. Los alumnos no recibieron ningún tipo de ayuda durante sus autoevaluaciones, más allá de la proporcionada por la rúbrica. La instructora que supervisó cada sesión también se encargó de solucionar cualquier incidencia, como el reajuste de los equipos evaluadores si no se había presentado algún miembro, o el proporcionar copias de las tareas si la copia digital era defectuosa, así como también se ocupó de evitar cualquier tipo de interacción entre los distintos equipos de evaluadores.

| Grupo | #Matriculados | #Alumnos | #ModalidadEP | #PRE-POST | #Actividad |
|---|---|---|---|---|---|
| A1 | 28 | 25 | Individual | 25 | Act3 |
| | | 28 | Parejas | 28 | Act1 |
| | | 28 | Tríos | 27 | Act2 |
| A2 | 28 | 21 | Individual | 21 | Act2 |
| | | 18 | Parejas | 17 | Act3 |
| | | 24 | Tríos | 24 | Act1 |
| B1 | 26 | 25 | Individual | 21 | Act1 |
| | | 25 | Parejas | 20 | Act2 |
| | | 21 | Tríos | 18 | Act3 |

Cuadro 2: Número de sujetos participantes en cada celda del estudio.

# 4. Análisis de datos

Para el análisis estadístico de los datos se utilizó el software SPSS Statistics v.26.

El tipo de análisis elegido en un principio, dado que tenemos dos factores intra-sujeto (ModalidadEP y Tiempo) y un factor inter-sujeto (Grupo), fue un Anova Mixto de Tres Factores (Three Way Mixed Anova). Este test tiene siete asunciones que deben ser comprobadas:

- Existe (1) una variable dependiente continua (PREC), (2) dos factores intra-sujetos categóricos con dos o más niveles en cada factor (Tiempo y ModalidadEP) y (3) un factor inter-sujeto (Grupo) que es categórico.
- (4) Valores atípicos. La inspección de los boxplots detectó la existencia de 11 valores atípicos. Una inspección de diez de los valores (con diferencias con respecto al experto superiores a cuatro puntos que los sujetos repetían en varias modalidades de EP) mostró que se había producido un problema con la comprensión del instrumento de medición por parte de seis sujetos: 2 del grupo A1, 3 del grupo A2 y 1 del grupo B1. Estos sujetos acumulaban 10 de los 11 valores atípicos. Es por ello que sus datos fueron descartados de este análisis. El undécimo valor sin embargo se mantuvo, al considerarse un valor genuinamente inusual.
- (5) Normalidad: sólo una de las celdas del diseño presentaba una distribución normal. Una transformación de tipo SQRT (la más adecuada para la forma de la distribución) no consiguió normalizar la distribución. No obstante, dado que los ANOVAs son robustos con respecto a la violación de normalidad, se decidió seguir con el análisis.
- (6) La asunción de homogeneidad de varianzas también se violaba en varias de las celdas de diseño, tal y como mostró la aplicación de un test de Levene para la igualdad de varianzas ($p<0.05$).
- (7) Por último, el test de esferidad de Mauchly indicó que la asunción de esferidad se cumplía, $\chi^2(2) = 1.288$, $p = .525$.

.

La ejecución del test no detectó una interacción significativa entre Tiempo, ModalidadEP y Grupo, F(4,86)=0,445 p=0,776. Por tanto, se procedió a analizar las tres posibles interacciones de los factores dos a dos. Los resultados fueron como sigue:

- Los datos muestran una interacción significativa entre Tiempo y Grupo, $F(2, 43) = 4{,}926$, $p = .012$. Esto significa que las diferencias entre las autoevaluaciones antes (pre) y después (post) de haber aplicado cada modalidadEP varían en función del grupo.
- Los datos también muestran una interacción significativa entre ModalidadEP y Grupo, $F(4, 86) = 5{,}762$, $p < 0.005$. Esto significa que la ModalidadEP aplicada ha afectado de manera distinta a cada grupo.
- Por último, los datos no muestran que exista una interacción significativa entre Tiempo y ModalidadEP, F(2,43)=1,877, p=0.159. Esto significa que el tipo de modalidad aplicada no parece afectar (considerada en los tres grupos en conjunto) a la evolución de la precisión.

Lo más destacable de estos resultados es el impacto del grupo sobre el resto de variables. Una posible explicación a este hecho es que, dado que los grupos estaban predefinidos (en lugar de haber sido formados de manera aleatoria) el nivel y el perfil de los alumnos en cada grupo probablemente era distinto. Es por ello que se ha optado por analizar la evolución de la precisión en función de la modalidadEP por separado para cada grupo.

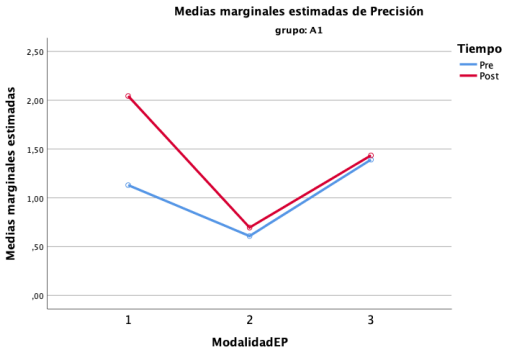


Figura 1: Grupo A1: Evolución de la precisión del alumno gracias al proceso EP

## 4.1. Diferencias por Grupo en la precisión de la autoevaluación en función de la ModalidadEP

Para evaluar este aspecto se aplicó un ANOVA de dos factores con medidas repetidas. Para estos análisis se consideraron exclusivamente los sujetos que habían participado en todas las modalidadesEP y que, en caso de mostrar valores atípicos, éstos se habían considerado genuinos (23 en el grupo A1, 16 en el grupo A2 y 13 en el grupo B1).

### 4.1.1. Grupo A1

El Cuadro 3 muestra los principales estadísticos descriptivos -Media (M) y desviación típica (DT)- de las dos medidas de precisión incluidas en este estudio para el grupo A1 (con autoevaluaciones en todas las ModalidadesEP).

| A1 (23 sujetos) | PREC_Pre | | PREC_Post | |
|---|---|---|---|---|
| | M | DT | M | DT |
| Individual (Act3) | 1,13 | 1,05 | 2,04 | 1,92 |
| Parejas (Act1) | 0,60 | 0,99 | 0,69 | 1,02 |
| Tríos (Act2) | 1,39 | 1,31 | 1,43 | 1,44 |

Cuadro 3: Grupo A1: Valores descriptivos de la Precisión antes y después de EP.

En la Figura 1 se puede observar cómo, en el grupo A1, los alumnos se autoevalúan peor (es decir, la diferencia con respecto al experto en valor absoluto es mayor) después de haber participado en todas las modalidades, aunque a medida que aumenta el número de correctores la diferencia va disminuyendo.

Para ver si estas diferencias son significativas hemos aplicado un ANOVA de dos factores con medidas repetidas. Los resultados muestran que la asunción de esferidad, comprobada mediante la prueba de esferidad de Mauchly, se viola para la interacción de los dos factores ( $\chi^2(2) = 6{,}218$, p = .045). Por tanto, se ha aplicado la corrección de Greenhouse-Geisser. Con esta corrección, la interacción Tiempo*ModalidadEP es significativa (F(2,35)=3,558, p=0.049), por lo que, en lugar de estudiar los efectos principales, hay que estudiar los efectos simples principales. El efecto principal simple del Tiempo (i.e. de haber participado en el proceso de EP), medido mediante un RM T-Test, muestra que, para la ModalidadEP 1 (individual), el empeoramiento que se produce entre la precisión del estudiante antes y después de haber participado en el proceso es significativo, t(24)=-2,441, p=0.023. Sin embargo, ni el empeoramiento en la condición 2 (parejas) (t(25)=0, p=1) ni en 3 (tríos) (t(26)=-0.214, p=0.832) es significativo. Por otro lado, el efecto principal de ModalidadEP sobre la medida PREC_post de los alumnos, testado mediante un one-way RM ANOVA, es significativo (F(2,44)=5,256, p=0.009). Dicho de otro modo, a pesar del empeoramiento generalizado en la PREC_Post frente a la PREC_Pre, los alumnos son significativamente más precisos cuando han participado en la ModalidadEP 2 (parejas), seguido por la modalidad 3 (tríos) y, por último, la modalidad 1 (individual).

### 4.1.2. Grupo A2

Por lo que respecta al grupo A2, el Cuadro 4 muestra los principales estadísticos descriptivos -M y DT- de las dos medidas de precisión.

| A2 (16 sujetos) | PREC_Pre | | PREC_Post | |
|---|---|---|---|---|
| | M | DT | M | DT |
| Individual (Act2) | 0,69 | 0,79 | 0,75 | 0,77 |
| Parejas (Act3) | 1,44 | 1,31 | 1,25 | 0,93 |
| Tríos (Act1) | 1,12 | 1,67 | 0,93 | 1,48 |

Cuadro 4: Grupo A2: Valores descriptivos de la Precisión antes y después de EP.

En la Figura 2 se puede observar cómo, en este caso, los alumnos se autoevalúan ligeramente peor después de haber participado en la modalidad individual, y mejoran ligeramente (disminuye su diferencia respecto al evaluador experto) tras haber participado en la modalidad de parejas y tríos de la EP.

Los resultados del ANOVA de dos factores con medidas repetidas muestran que la asunción de esferidad, comprobada mediante la prueba de esferidad de Mauchly, se cumple para la interacción de los dos factores ( $\chi^2(2) = 0{,}409$, p = .815).

No existe una interacción significativa entre Tiempo y ModalidadEP, F(2, 30) = 0,556, p = .580. Por tanto, se puede proceder a analizar los efectos principales

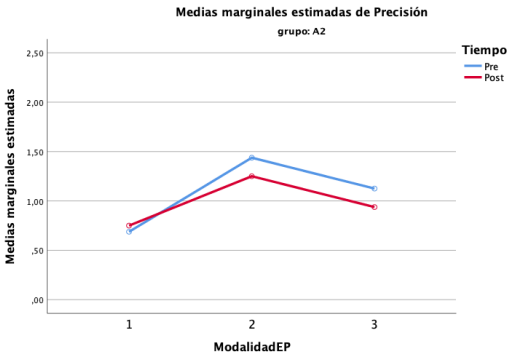


Figura 2: Grupo A2: Evolución de la precisión del alumno gracias al proceso EP

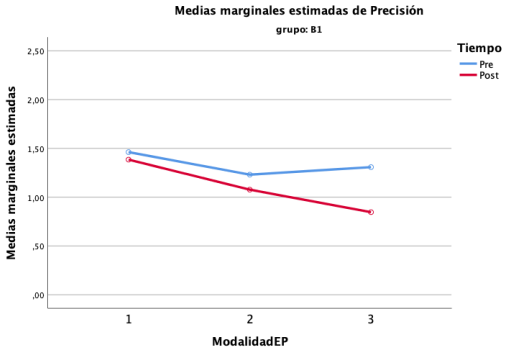


Figura 3: Grupo B1: Evolución de la precisión del alumno gracias al proceso EP

de Tiempo y de ModalidadEP. En el caso del grupo A2, ni el Tiempo (F(1,15)=1,334, p=0.264) ni la modalidad (F(2,30)=1,227, p=0.308) muestran un efecto significativo. Esto quiere decir que, en el grupo A2, los alumnos no mejoraron de manera significativa sus autoevaluaciones tras haber participado en una EP, y que tampoco el corregir de manera individual, en pares o en tríos tuvo un efecto significativo en el grado de mejora de esa autoevaluación.

### 4.1.3. Grupo B1

Por último, el Cuadro 5 muestra los principales estadísticos descriptivos -M y DT- de las dos medidas de precisión en el grupo B1.

| B1 (13 sujetos) | PREC_Pre | | PREC_Post | |
|---|---|---|---|---|
| | M | DT | M | DT |
| Individual (Act1) | 1,46 | 1,27 | 1,38 | 1,39 |
| Parejas (Act2) | 1,23 | 1,01 | 1,08 | 1,12 |
| Tríos (Act3) | 1,31 | 1,03 | 0,85 | 0,80 |

Cuadro 5: Grupo B1: Valores descriptivos de la Precisión antes y después de EP.

En la Figura 3 se puede observar cómo, en este caso, los alumnos se autoevalúan mejor independientemente de la modalidad de EP en la que participan.

Los resultados del ANOVA de dos factores con medidas repetidas muestran que la asunción de esferidad, comprobada mediante la prueba de esferidad de Mauchly, se cumple para la interacción de los dos factores ( $\chi^2(2) = 2{,}994$, p = .224).

Al igual que ocurría en el grupo A2, no existe una interacción significativa entre Tiempo y ModalidadEP, F(2, 24) = 1.348, p = .279. Por tanto, se puede proceder a analizar los efectos principales de Tiempo y de ModalidadEP. En el caso del grupo B1, ni el Tiempo (F(1,12)=3.984, p=0.069) ni la modalidad (F(2,24)=0.342, p=0.714) muestran un efecto significativo. Esto quiere decir que, en el grupo B1, los alumnos no mejoraron de manera significativa sus autoevaluaciones tras haber participado en una EP, y que tampoco el corregir de manera individual, en pares o en tríos tuvo un efecto significativamente distinto en una posible mejora de esa autoevaluación.

## 5. Discusión

Tal y como se ha comentado en la sección 4, los análisis estadísticos no han permitido rechazar la hipótesis nula, y por tanto tenemos que asumir que, en el contexto de este estudio, la participación de los sujetos en un proceso de EP no ha mejorado estadísticamente su precisión a la hora de evaluar su rendimiento, y tampoco el hecho de haber participado de manera individual, en parejas o en tríos ha tenido un efecto significativo sobre dicha precisión. Estos resultados no son los esperados acorde con otros trabajos como [1, 4, 9].

Sin embargo, sí se pueden observar ciertas tendencias. Si unimos las observaciones de todos los grupos, se puede constatar que todos empeoraron más, o mejoraron menos, con la modalidad de EP Individual:

- A1 En la modalidadEP de parejas (Act1) y en la de tríos (Act2), aunque se empeora siempre con la post, la pre y la post tienen resultados parecidos (las diferencias son 0.09 y 0.04). Sin embargo, en la modalidad individual (Act3) sí que se obtienen unos resultados mucho peores en la post.
- A2 lo hizo mejor en la post que en la pre, salvo en la modalidad individual (Act2) donde la PREC_Pre y la Prec_Post fueron muy parecidas, con una diferencia de +0.06 en favor de la PREC_Pre.
- B1 siempre obtuvo resultados más precisos en la post que en la pre, pero la menor mejora (y la peor precisión) se dio con la modalidadEP individual (Act1).

Con el fin de buscar una posible explicación a los resultados, en cada grupo, parte de la última clase del curso se dedicó a una puesta en común sobre las EP que se habían realizado. Todos los grupos opinaron que la Actividad 1 había sido, subjetivamente, la más fácil de realizar y corregir. Aunque las notas obtenidas corroboran que sí fue la más fácil de realizar, un análisis de las precisiones (cuadros 3, 4 y 5) muestra que no en todos los grupos fue la más fácil de autoevaluar desde el punto de vista de la precisión. Solo en el grupo A1 fue la actividad que tuvo la mejor precisión, mientras que en el grupo A2 fue la actividad media y en el grupo B1 fue la que obtuvo una peor precisión media (más alejada de la opinión del experto).

En este sentido, se les preguntó explícitamente a los estudiantes si pensaban que la dificultad de la actividad influía a la hora de realizar una EP más o menos precisa. A este respecto los estudiantes comentaron:

- *la EP variaba, más o menos somos una piña, cuánto más difícil era la Actividad, más intentábamos ayudar al otro, decíamos: vamos a hacer para que no saque menos de un 4, menos de un 5, vamos a ver si le podemos sacar algo. Y cuánto más fácil era, pues ibas más al aire. Dices: esto lo tiene mal, pues mal. También eramos más duros al principio, pero conforme conoces a la gente...* (J.O., grupo B1).
- *Si la Actividad es más fácil, puedes apañarte con EP Individual, pero si la Actividad es más compleja, hacer la EP en parejas o en tríos, ayuda* (M.A., grupo A1).
- *Si La Actividad es más fácil o más complicada, no repercute en la EP, ya que había que seguir la rúbrica* (P.A., grupo A2).

Si atendemos al comentario de J.O. (grupo B1), la actividad más precisa debería haber sido la Actividad 1, cosa que, como hemos comentado antes, no ha sido el caso.

Por lo que respecta a qué modalidad de EP les había gustado más, todos los grupos estaban de acuerdo en que fue la modalidad que, en cada caso, habían aplicado a la Act1. Una explicación posible puede ser que ellos opinaban que era la Actividad más fácil, pero creemos que también puede estar influenciado por la novedad que suponía para ellos la EP, que les hizo a todos afrontar el proceso de evaluación con más ilusión, independientemente de la actividad, tal y como se desprende de sus comentarios:

- *Comienzas con ilusión con la primera EP en la Actividad 1, pero luego ya en las otras Actividades...te vas cansando...* (M.A., grupo B1).
- *La EP individual es lenta y aburrida...* (T.O., grupo A1)

Estos comentarios pueden explicar, en parte, el caso más llamativo de empeoramiento de la precisión del grupo A1 en la Actividad 3 (modalidad Individual). Pensamos que puede atribuirse a que era al final del curso, y este grupo estaba especialmente cansado de realizar EP.

Por último, es importante destacar que hay otros elementos, además de la dificultad de la tarea, que se han quedado fuera del estudio y que para su análisis requerirían por tanto realizar nuevas réplicas para controlar cada uno de estos aspectos. Quizás el más relevante sea el formato de la rúbrica. Parece razonable pensar que tanto el número y tipo de ítems incluidos como su redacción pueden estar influyendo en la precisión de la evaluación por parte de los alumnos.

# 6. Conclusiones

La EP es una alternativa a la evaluación tradicional que, en entornos masificados como los de la universidad española, permite proporcionar retroalimentación a los alumnos con un esfuerzo asumible por el profesor. Sin embargo, no está claro de qué manera influye la participación del alumno en una EP en su capacidad de autoevaluarse. Con el fin de estudiar este aspecto, en este trabajo se ha presentado un quasi-experimento donde se ha pedido a los alumnos que se autoevalúen antes y después de participar en tres modalidades distintas de EP: individual, en parejas y en tríos.

Los resultados estadísticos no permiten afirmar que exista una influencia de dicha participación en la precisión de los alumnos en su autoevaluación, al menos en el contexto de este estudio. No obstante, esto no significa necesariamente que la EP no influya en su aprendizaje, ya que puede haber otros factores, no controlados en el estudio, que proporcionen explicaciones alternativas a los resultados.

Así estos resultados, más que contradecir lo comúnmente aceptado en la comunidad educativa, abren un interesante debate sobre cómo y cuándo debe realizarse una EP para que de verdad se produzca el efecto de una mejor precisión en la autoevaluación. En este cómo se incluyen aspectos como tipos de actividades con las que resulta más efectiva, tipos de rúbricas que más favorecen el autoaprendizaje, etc.

Nuestro estudio también sugiere (aunque, de nuevo, los datos no son estadísticamente significativos) que, cuando se aplica este tipo de evaluación, una corrección por parejas o en tríos resulta ligeramente más efectiva que una corrección individual, independientemente de la complejidad de la tarea.

Como trabajo futuro se pretende replicar este estudio controlando la dificultad de la tarea. Asimismo, se pretende trabajar en el diseño de rúbricas con distintas características, para intentar discernir cómo el tipo y la redacción de la rúbrica afecta el autoaprendizaje y la

precisión del alumno durante su aplicación.

## Consideraciones éticas

La participación en este estudio de los sujetos ha sido voluntaria, libre e informada. Así mismo, todos los sujetos dieron su consentimiento explícito para el tratamiento de sus datos de manera anónima y agregada, preservando así su derecho a la intimidad.

## Agradecimientos

El presente trabajo ha sido parcialmente financiado por el programa Redes-I$^3$CE de investigación en docencia universitaria del Instituto de Ciencias de la Educación de la Universidad de Alicante (REDES-I3CE-2019-4607) y por el proyecto EduTech (609785-EPP-1-2019-1-ES-EPPKA2-CBHE-JP), cofinanciado por el Programa Erasmus+ de la UE.

Los investigadores agradecen además la colaboración de todos los sujetos que accedieron a participar en este estudio de manera desinteresada.

# Referencias

[1] Heidi Andrade y Ying Du: *Student perspectives on rubric-referenced assessment*. Practical Assessment, Research, and Evaluation, 10(1):3, 2005.

[2] Linda Anglin, Kenneth Anglin, Paul L. Schumann y John A. Kaliski: *Improving the Efficiency and Effectiveness of Grading Through the Use of Computer-Assisted Grading Rubrics*. Decision Sciences Journal of Innovative Education, 6(1):51–73, 2008.

[3] Susan M. Brookhart y Fei Chen: *The quality and effectiveness of descriptive rubrics*. Educational Review, 67(3):343–368, 2015.

[4] Anders Jonsson y Gunilla Svingby: *The use of scoring rubrics: Reliability, validity and educational consequences*. Educational research review, 2(2):130–144, 2007.

[5] Vigdis By Kampenes, Tore Dybå, Jo E Hannay y Dag IK Sjøberg: *A systematic review of effect size in software engineering experiments*. Information and Software Technology, 49(11):1073–1086, 2007.

[6] Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller y Scott R. Klemmer: *Peer and Self Assessment in Massive Online Classes*. ACM Transactions on Computer-Human Interaction, 20(6):1–31, 2013.

[7] Oscar Luaces, Jorge Díez y Antonio Bahamonde: *A peer assessment method to provide feedback, consistent grading and reduce students' burden in massive teaching settings*. Computers & Education, páginas 283–295, 2018.

[8] Eric Mazur y Robert C Hilborn: *Peer instruction: A user's manual*. Physics Today, 50(4):65, 1997.

[9] Michael J McCormick, Kim E Dooley, James R Lindner y Richard L Cummins: *Perceived Growth versus Actual Growth in Executive Leadership Competencies: An Application of the Stair-Step Behaviorally Anchored Evaluation Approach.* Journal of Agricultural Education, 48(2):23–35, 2007.

[10] Ernesto Panadero y Gavin TL Brown: *Teachers' reasons for using peer assessment: positive experience predicts use*. European Journal of Psychology of Education, 32(1):133–156, 2017.

[11] Ernesto Panadero y Anders Jonsson: *The use of scoring rubrics for formative assessment purposes revisited: A review*. Educational Research Review, 9:129–144, 2013.

[12] Ernesto Panadero, Margarida Romero y Jan Willem Strijbos: *The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort*. Studies in Educational Evaluation, 39(4):195–203, 2013.

[13] Dewayne E. Perry, Adam A. Porter y Lawrence G. Votta: *Empirical studies of software engineering: a roadmap*. En *Proceedings of the conference on The future of Software engineering*, páginas 345–355. ACM, 2000.

[14] Juan Ramón Rico-Juan, Antonio Javier Gallego, Jose J Valero-Mas y Jorge Calvo-Zaragoza: *Statistical semi-supervised system for grading multiple peer-reviewed open-ended works*. Computers & Education, 126:264–282, 2018.

[15] Hao Chuan Wang, Chun Yen Chang y Tsai Yen Li: *Assessing creative problem-solving with automated text grading*. Computers & Education, 51(4):1450–1466, 2008.

[16] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell y Anders Wesslén: *Experimentation in software engineering*. Springer Science & Business Media, 2012.