Universidad de Chile
Facultad de Ciencias Sociales
Departamento de Sociología
Carrera de Sociología

Rendimiento, estabilidad y validez de trayectorias de evaluación docente a lo largo de un semestre

Análisis de caso en 23 cursos universitarios de pedagogía en matemática

Tesis para optar al título de Sociólogo

Autor:
Raúl Zamora Zunino

Profesor Guía:
Rodrigo Asun Inostroza

Profesora Co-guía:
Salomé Martínez Salazar

Santiago de Chile, marzo de 2018

Agradecimientos

En primer lugar, agradezco a los integrantes del proyecto FONDEF IT13I10005 "Herramientas para fortalecer la formación de profesores de educación básica basadas en experiencias de enseñanza de la matemática en aula", del Centro de Modelamiento Matemático de la Universidad de Chile, por ser quienes produjeron la valiosa información que fue utilizada para llevar a cabo esta investigación. En especial, agradezco a Salomé Martínez por invitarme a realizar mi Seminario de Licenciatura y Tesis de Grado al interior del proyecto.

Agradezco, en segundo lugar, a dos docentes que marcaron mi paso por la Universidad.

A Gabriela Azócar, quien vio capacidades en mí que sólo descubrí tras su invitación a trabajar como su ayudante. Promovió en mí gran confianza y optimismo sobre mis capacidades para desarrollarme como científico social.

A Rodrigo Asun, por guiarme diligentemente a lo largo de este proceso de Seminario y Tesis. Agradezco más aún sus continuas enseñanzas en metodología cuantitativa, que cada vez más logran incentivar mi interés por la medición de la realidad social. Es en gran parte debido a sus enseñanzas e influencia que poseo actualmente la convicción de que es necesario avanzar hacia una ciencia social que se construya sobre el conocimiento empírico, que deje de lado los relatos ideológicos y los diagnósticos autocomplacientes.

Finalmente, pero no por ello en menor medida, agradezco a mi familia, en especial a mis padres, por proveerme del soporte material y apoyo emocional tan necesarios para estudiar una carrera universitaria y lograr convertirme en un profesional.

Índice

I. Presentación del estudio

El presente estudio se sitúa en el campo de investigación en educación superior, específicamente, en el sub campo de la investigación en evaluación docente en base a cuestionarios respondidos por alumnos (CEDs).

Corresponde a un análisis de caso de las trayectorias longitudinales de evaluación docente de 23 cursos universitarios, dictados por 18 profesores, en 8 universidades chilenas que imparten la carrera de pedagogía básica en matemáticas. Se constituye como la primera investigación longitudinal semestral en base a este tipo de instrumentos (CEDs).

Desde su diseño, implementación, hasta la codificación de datos, la investigación estuvo a cargo de un grupo interdisciplinario de profesionales del Laboratorio de Educación del Centro de Modelamiento Matemático de la Universidad de Chile. Desde el análisis de datos en adelante, el trabajo fue desarrollado por el autor de la presente tesis.

Los principales focos de análisis sobre las trayectorias de evaluación docente fueron su rendimiento, niveles de estabilidad en el tiempo, su relación con aspectos extra-docencia, como también los niveles de validez y fiabilidad de las mediciones. Junto con ese diagnóstico, este trabajo también pone especial atención en los aprendizajes metodológicos y prácticos que pueden extraerse de los resultados del estudio, como de lo que se puede aprender a partir de las limitaciones originadas en su fase de diseño.

II. Introducción al tema de estudio y su relevancia

Una de las funciones sociales del sistema educativo es la formación de la ciudadanía para el ejercicio práctico de las profesiones académicas, es decir, la formación de profesionales especialistas (Parsons & Platt, 1973). Al interior del sistema educativo y sus procesos, la calidad de la docencia es un factor esencial. Su rendimiento tiene incidencia directa sobre la calidad del aprendizaje y las capacidades que adquiere todo capital humano en formación. Por ello, se sitúa como factor determinante del desarrollo productivo, tecnológico, científico y humano en las sociedades modernas. Ello explica que la evaluación de la docencia sea un tema central y de mucha controversia en la política educativa de los países buscando su desarrollo, y al que cada vez se le otorga más atención (OCDE, 2013a; OCDE, 2013b).

La evidencia sobre el importante efecto que posee la calidad de la docencia sobre los procesos de aprendizaje es abundante, e indica que tanto el dominio de la disciplina, el manejo de habilidades pedagógicas y la experiencia docente se asocian positivamente con diversas mediciones del aprendizaje que logran los estudiantes en los contextos educativos en general (Manzi, et al., 2011a). Esta relación se ha probado significativa incluso cuando se aísla la contribución de otros factores que tienen efecto sobre el aprendizaje, como lo son las características propias del establecimiento educativo y los antecedentes socioeconómicos de los estudiantes (Nye, et al., 2004).

Algunas fuentes incluso sitúan a la calidad docente como la variable asociada a la escuela más determinante sobre el rendimiento estudiantil (Rivkin, et al., 2005; Hattie, 2009). Estudios de alto impacto, como el informe McKinsey (Barber & Mourshed, 2007), que han estudiado cuáles son las estrategias de los sistemas educacionales nacionales más exitosos en producir excelencia educativa, indican que el principal impulsor de las variaciones en el aprendizaje escolar es la calidad de los docentes, y que no hay sistemas educativos que provean oportunidades de aprendizaje de calidad que sean diferentes o independientes a las capacidades de sus docentes. El mismo informe a la vez, señala que los sistemas educativos con más alto desempeño reconocen que la única manera de mejorar los resultados es mejorando la enseñanza.

Por estas razones, una de las líneas en investigación educativa más desarrollada a nivel internacional es aquella que indaga en las diferencias de capacidades que poseen los docentes para generar aprendizajes en sus alumnos, analizando si es posible identificar docentes más

efectivos que otros, y aislando cuáles serían los factores que determinarían esta diferencia de logro.

Al alero de este campo de investigación, se han ido desarrollando distintos tipos de procedimientos para evaluar la calidad de la docencia, variando en la forma en que se representan las puntuaciones de rendimiento docente. Ellas pueden considerar puntajes construidos mediante la observación experta del aula, utilizando rúbricas de registro especialmente diseñadas para ello (Praetorius, et al., 2012; Hill, et al., 2012; Ho & Kane, 2013); evaluación según lo que la literatura angloparlante ha denominado como "value-added models" o VAMs, que utiliza los resultados de los estudiantes a pruebas estandarizadas para medir la parte del rendimiento que se considera atribuible a la contribución de los docentes (Boyd, et al., 2009; Darling-Hammond, et al., 2012); o bien se consideran los cuestionarios de evaluación docente respondidos por alumnos (CEDs) para representar el rendimiento del profesorado.

De las tres antes mencionadas, la metodología más extensamente utilizada es esta última, es decir, la aplicación y análisis de cuestionarios de evaluación docente (Marsh, 1984). A nivel internacional, el uso de este tipo de cuestionarios para la evaluación e investigación sobre la calidad de la docencia es extendido, tanto en el mundo anglosajón (Murray, 2007), como en Norteamérica, Europa continental y Asia (Theall & Franklin, 2000). Se ha sugerido que los estudios que utilizan o bien analizan cuestionarios de evaluación docente superan en volumen a todo el resto de las investigaciones que utilizan otras medidas para evaluar la enseñanza (Cashin, 1995).

La docencia y su evaluación se sitúan, también, como una de las principales preocupaciones en la reflexión sobre educación debido a la constante demanda por lograr la excelencia docente, en vista de la constante necesidad de toda institución educativa de tomar decisiones fundamentadas en la promoción y mejoramiento de las plantas docentes (Irby, et al., 1977; Onwuegbuzie, et al., 2007). Con ello, los cuestionarios en evaluación docente respondidos por alumnos han sido los más comúnmente usados al interior de las instituciones de educación superior para examinar el rendimiento de sus docentes (Polikoff, 2015), hasta el punto que en algunas instituciones no sólo se han vuelto la medición más dominante del desempeño docente, sino que también la única (Hornstein, 2017). Esta relativa exclusividad también ocurre para las instituciones de educación superior de nuestro país, Chile (Salazar, 2008; Montoya, et al., 2014).

Ahora bien, pese a su auge en el mundo, a nivel latinoamericano y nacional, el uso extendido de cuestionarios de evaluación docente es más bien reciente, y la investigación asociada a estos

instrumentos es relativamente escaza (Arámburo Vizcarra & Luna Serrano, 2013; Medel & Asun, 2014).

A continuación, se realiza un breve resumen de antecedentes sobre el desarrollo del campo de investigación en evaluación docente en base a cuestionarios respondidos por estudiantes, tanto a nivel internacional como local. Se presenta una identificación de los distintos tópicos sobre los que se ha trabajado el análisis de los cuestionarios de evaluación docente, destacando las preguntas que aún permanecen vigentes, mostrando la evidencia recabada hasta ahora, para finalmente introducir la propuesta de investigación propia de este estudio.

III. Marco de antecedentes

i. Breve historia y mapa del campo de estudio en cuestionarios de evaluación docente

Diversas fuentes reconocen a Herman Remmers como el creador del primer cuestionario de evaluación docente, en la Universidad de Pardue, Estados Unidos, en la década de 1920 (Marsh, 1987; García Garduño, 2000; Salazar, 2008; Cánovas, et al., 2009; Wachtel, 1998). Hasta la década de los 60, Remmers y su equipo fueron los únicos investigadores que realizaron estudios sistemáticos a partir de los resultados de este instrumento (Cánovas, et al., 2009). Sólo a partir de los años 60, otros investigadores se sumaron a la investigación asociada a los CED y su validez, alcanzando esta línea de estudio su mayor apogeo hacia la década de los 80 (Greenwald, 1997). Hacia fines del siglo XX, la investigación en evaluación docente era uno de los temas más estudiados en el campo de la educación superior a nivel internacional, con más de dos mil artículos sobre el tema (García Garduño, 2000; Medel, 2013).

Según Cánovas et al. (2009) muchos autores coinciden en que los propósitos más corrientes de una evaluación docente con base en respuestas de estudiantes son:

a) el diagnóstico y retroalimentación a los profesores sobre su desempeño docente;
b) control administrativo-docente, en cuanto se usa como medida de la calidad docente para ser empleada en decisiones sobre mantenimiento o promoción de cargos;
c) investigación sobre los resultados y procesos docentes.

Dentro de este último propósito, investigativo, también es posible identificar ámbitos de estudio. Según un metaanálisis realizado por Anthony Greenwald (1997), en las últimas cuatro décadas del siglo XX, el campo de investigación sobre este tipo de instrumentos se centró en cuatro preguntas o discusiones:

a) cuál es la estructura conceptual asociada al constructo (¿es la evaluación docente unidimensional o multidimensional?);
b) cuál es el grado de validez convergente de las evaluaciones docentes (¿qué tan bien se asocia con otros indicadores de enseñanza efectiva?);
c) cuál es el grado de validez discriminante que poseen (¿los puntajes se encuentran influenciados por otros factores ajenos a la enseñanza efectiva?);

d) cuál es el grado de validez consecuente de estos instrumentos (¿están siendo los resultados de las investigaciones en este campo efectivamente utilizados para la evaluación y el desarrollo del personal docente?).

En la actualidad, a dos décadas de este metaanálisis que mapeó este campo de estudio, una revisión del estado del arte evidencia que muchas interrogantes en los ámbitos mencionados aún continúan vigentes, como se verá a continuación. La discusión académica aún se pregunta sobre la validez de los instrumentos, dadas distintas acepciones conceptuales de lo que se entiende por "calidad docente", diversos alegatos por los posibles sesgos que incorporarían este tipo de mediciones, la amplia variabilidad de condiciones y contextos de enseñanza, como también, continúa vigente el debate sobre las medidas que se adoptan y el uso que se le da a los resultados que se obtienen de las evaluaciones en los contextos educativos.

A continuación, se presenta una breve revisión sobre el estado investigativo de estos lineamientos, con mayor énfasis en aquellas problemáticas más asociadas a la propia investigación, para posteriormente dar cuenta de las especificaciones de este estudio.

ii. Validez de constructo y de contenido de la evaluación docente

Conceptualmente, son múltiples los aspectos e indicadores que pueden ser considerados como evaluadores de la calidad de la docencia. Según Marsh y Roche (1997), prolíficos investigadores en el ámbito de los instrumentos de evaluación docente, es un acuerdo casi generalizado entre los profesionales e investigadores en educación que la docencia es una actividad compleja, consistente de múltiples dimensiones. Marsh y Roche (1997) insisten que, por ello, sería necesario que toda evaluación docente reflejara esta multidimensionalidad.

De todas formas, hay investigadores con argumentos divergentes, o bien, definitivamente opuestos a esta posición. Autores como Centra (2003) sugieren que la multidimensionalidad no necesariamente refleja las características propias de la buena docencia, sino que sería resultado de que los esfuerzos por medir el constructo tienden a preguntar sobre distintos temas, proceso que inevitablemente haría surgir una estructura dimensional.

Otras posturas son aún más críticas, y sugieren que, además de ser la dimensionalidad un efecto propio de la medición y no del fenómeno, la utilización de estrategias de este tipo introduce errores en la medición. Se alega que lo que reflejarían las medidas dimensionales propuestas para dar cuenta del rendimiento docente, en realidad, no serían más que la "teoría implícita" (Abrami,

et al., 1981; Larson, 1979; Whitely & Doyle, 1976), o las prenociones y creencias (Kember & Wong, 2000) que poseen los estudiantes respondientes respecto a cuáles son las dimensiones de la docencia de calidad, y por consiguiente, los instrumentos de evaluación docente no serían capaces de registrar fielmente el desempeño real de los docentes. Las mediciones no lograrían obtener respuestas espontáneas, sino que rescatarían respuestas construidas (Beecham, 2009). Por ello, algunos investigadores sostienen que la calidad de la docencia debería ser medida utilizando mediciones unidimensionales, y sólo enfatizando el rendimiento general de los docentes (Abrami, et al., 1981; Abrami & D'Apollonia, 1991; Apodaca & Grad, 2005).

También existen críticas que no sólo abordan el problema de contenido conceptual y estructural de las mediciones, sino que abordan la mala formulación de los instrumentos en sí. Algunos autores denuncian una mala calidad de los ítems de muchos de los instrumentos aplicados para evaluar la calidad de la docencia (Tagomori & Bishop, 1995; Johnson, 2000; Knapper, 2001; Asun & Zúñiga, 2017), reprochando su mala formulación, ambigüedad, y poca claridad, además de acusar una generalizada falta de un mínimo de entendimiento sobre qué constituye la buena docencia.

Más allá de hacer el esfuerzo por tomar posición en la discusión sobre la validez de constructo y de contenido conceptual de los instrumentos de evaluación docente, debate que no parece estar resuelto, es posible realizar el ejercicio de identificar bajo cuál de estas ópticas se ha estudiado el fenómeno del rendimiento docente en la investigación empírica que se ha realizado en el campo. Revisando la literatura sobre el tema y los reportes de investigación, es posible asegurar que existe una vasta cantidad de indicadores propuestos como representativos de aquello que correspondería al "buen rendimiento" o "calidad" docente. Esfuerzos por enlistar la amplia cantidad de aspectos que se han considerado como indicadores o conductas de buena docencia en los estudios sobre el tema, han llegado cuantificar dentro de los cientos las distintas dimensiones evaluadas para medir la calidad de la instrucción docente (Chonko, et al., 2002, p. 279).

En efecto, al ser un fenómeno no tangible, la "calidad docente" se convierte en una variable latente, es decir, que puede ser medida y estudiada sólo mediante procesos de especificación en múltiples indicadores. Dichas variables latentes se entienden en la ciencia de la medición como "constructos", al ser conceptos no observacionales de complejo abordaje empírico (Bunge, 1973).

En la investigación empírica existente, la generalidad de los estudios tiende a incorporar distintas dimensiones del rendimiento docente, es decir, se orientan hacia una mirada multidimensional del fenómeno. La investigación de Asun y Zúñiga (2017), concentrada en identificar y cuantificar las definiciones de calidad al momento de definir las "áreas" o "dimensiones" que se incluyen en las

mediciones de calidad docente, analizó 60 instrumentos, que incluyen en total 1608 ítems. Utilizando técnicas de análisis de contenido, sus resultados indican que los ítems analizados se puede agrupar en diez grandes temas que, ordenados en frecuencia de aparición en los instrumentos considerados, son: planificación y conducción del curso (16,6%); interacción con los estudiantes (14,7%); calidad de las evaluaciones (12,1%); uso de metodologías diversas (11,6%); lograr aprendizajes significativos (10,3%); capacidad de motivar (8%); capacidad de comunicación (7%); capacidad de ser guía experto (5,1%); preguntas generales sobre el docente (4,7%); responsabilidades administrativas (3,9%); dominio de contenidos (3,3%); preguntas no relacionadas con la docencia (2,7%)[1].

La revisión de la literatura realizada para la presente investigación evidencia que los aspectos que son propuestos como medidores de este constructo son algunos más recurrentes que otros, siendo los más comunes la pertinencia y relevancia de la instrucción; la capacidad para planificar y organizar las sesiones de clase; la claridad de exposición; el ambiente de aula promovido por el docente; su nivel de responsabilidad; la adecuación de las evaluaciones, entre otros aspectos.

Aspectos menos comunes, pero también involucrados en algunos trabajos que los proponen como medición de la calidad docente, son, entre otros: la capacidad para explicar conceptos especialmente difíciles, la accesibilidad y disposición de los docentes fuera de las aulas de clases (ej. Hanges, et al., 1990), la capacidad de los docentes para relacionarse con los estudiantes (ej. Hativa, 1996), la capacidad de exponer de forma interesante y no aburrida (ejs. Hanges, et al., 1990; Hativa, 1996), el buen uso de medios y la utilidad del material docente (ej. Irby, et al., 1977), la sensibilidad del docente (Pianta, 2010), la capacidad de soporte emocional que provee el docente (ej. Patrick y Mantzicopoulos, 2014), e incluso los niveles de felicidad que genera el ambiente de la clase (ej. Kane y Staiger, 2012).

Otras veces, se proponen medidas aproximativas (o "proxis") como, por ejemplo, consultas sobre la cantidad de aprendizajes que los estudiantes perciben haber logrado con determinado docente, o bien, si se recomendaría el docente a otros estudiantes (Hanges, et al., 1990).

Es evidente que muchos de los indicadores que se proponen para medir la calidad docente pueden ser agrupados, y a la vez, otros pueden ser subdivididos. En la literatura, propuestas de este tipo son comunes. Por ejemplo, se sugiere que la dimensión de "calidad de la instrucción" puede ser descompuesta en distintas sub-dimensiones, como lo son el uso de estrategias de enseñanza y el

[1] Ver Asun y Zúñiga (2017) para detalle y descripción exhaustiva del contenido de cada aspecto (pp. 58-60).

uso de estrategias de monitoreo del aprendizaje; y que la responsabilidad docente puede ser descompuesta, entre otros elementos, en la capacidad de comunicación de información y en el nivel de profesionalismo del docente (Morgan, et al., 2014).

En esta discusión sobre cuáles son las medidas de calidad docente apropiadas a medir, hay incluso posturas que van más allá del rendimiento académico y proponen que la buena docencia no sólo implica enseñar a los alumnos a resolver problemas intelectuales, sino que también involucra promover valores, como la civilidad y la responsabilidad social (Lavigne & Good, 2015).

Pese a las múltiples especificaciones sugeridas en la literatura, algunos investigadores señalan que la evaluación docente carece de validez de contenido, ya que existiría poca claridad y consenso entre la comunidad científica sobre qué es, en sí, la calidad docente (Onwuegbuzie, et al., 2007). Para algunos, la multiplicidad de aspectos propuestos como indicadores de evaluación docente es indicador de esto mismo (Hornstein, 2017). De forma similar, algunos investigadores denuncian que las mediciones a partir de cuestionarios de evaluación docente en realidad no lograrían medir capacidades docentes, sino la satisfacción, popularidad y la afinidad que adquieren los estudiantes con sus docentes (Beecham, 2009).

Pese a la diversidad, un punto en común entre los estudios sobre evaluación docente es que, al proponer un determinado esquema de aspectos o dimensiones del rendimiento docente a evaluar, la mayoría de los trabajos utilizan estrategias de análisis factorial para examinar la validez de sus constructos (Morgan, et al., 2014).

Para sintetizar, es posible señalar que la delimitación sobre qué es la buena docencia, o bien, qué prácticas o conductas la constituyen, es una discusión que no está resuelta. Ello no ha evitado, sin embargo, que se postulen y apliquen una serie de indicadores para su abordaje, y que la investigación con cuestionarios de evaluación docente sea un campo de investigación de mucha productividad. Por ello, pese a la existencia de múltiples críticas, lo cierto es que este tipo de mediciones son vastamente utilizadas en los espacios educativos, sobre todo en los universitarios. Para algunos investigadores, la evaluación docente llegó para quedarse, debido a que está típicamente ligada a los programas de incentivos económicos y al control administrativo del trabajo docente (García Garduño, 2008). Probablemente por la inevitabilidad de aquello, es que algunos autores sugieren que, para atenuar la invalidez estructural y de contenido, toda propuesta de instrumentos que pretenda medir la calidad docente, debe incorporar a todos los actores involucrados en el proceso de enseñanza (docentes, estudiantes, autoridades educativas y

diseñadores de instrumentos), promoviendo una reflexión y buscando acuerdo sobre qué se entiende por "buena docencia" (Spooren, et al., 2013).

iii. Validez discriminante y la existencia de "sesgos" en la evaluación docente

En la investigación internacional asociada a los cuestionarios de evaluación docente de los años 80 y antes, se trabajó mayoritariamente con el supuesto de que las habilidades docentes pueden evaluarse sin mayor consideración a su contexto (Arámburo Vizcarra & Luna Serrano, 2013). Con ello, la aplicación de un instrumento y los consecuentes resultados de rendimiento docente otorgados por los estudiantes se tendían a valorar como mediciones y resultados generalizables.

Con el paso del tiempo, cada vez más aspectos contextuales fueron propuestos como posibles factores que podrían jugar un rol influenciador, y más aún, invalidador de los procesos de evaluación docente (García Garduño, 2000; Rantanen, 2013; Arámburo Vizcarra & Luna Serrano, 2013).

Así, gran parte de la investigación más actual sobre la validez de los cuestionarios de evaluación docente ha puesto el foco en la validez discriminante de los instrumentos utilizados para medir este constructo. Estos estudios se concentran en la identificación y medición del efecto que sobre la evaluación docente poseen determinados "sesgos", entendidos como aspectos que se encuentran más allá de la calidad de la instrucción docente, y que afectan la valoración que hacen los estudiantes de sus profesores.

El volumen de investigación al respecto parece ser tanto, que algunos investigadores partidarios de la validez discriminante de la evaluación docente consideran que al interior del campo ocurriría una "*cacería de brujas*", innecesariamente insistente en identificar potenciales sesgos en las evaluaciones docentes hechas por estudiantes (Marsh, 1987; Theall & Franklin, 2001).

Formalmente, los sesgos en la evaluación docente se definen como las características de los docentes y/o los cursos que no se relacionan con la buena docencia, pero que afectan su medición, ya sea positiva o negativamente (Centra & Gaubatz, 2000). Diversas fuentes (Rantanen, 2013; Arámburo Vizcarra & Luna Serrano, 2013) indican que pueden ser agrupados en tres categorías: variables relativas al curso dictado, relativas al docente, y relativas a los estudiantes.

Encontrándose dentro de las variables relativas al docente, uno de los aspectos que le ha valido más críticas a la validez discriminante de la evaluación docente, es el sesgo de género. En una revisión sistemática de los estudios enfocados a identificar este tipo de sesgo, Centra y Gaubatz

(2000) anuncian que, en suma, el diagnóstico sobre la problemática es contradictorio. Señalan que en 6 estudios sobre el tema realizados entre los años 1974 y 1981, no se encontraron diferencias significativas (o al menos han sido mínimas) en la evaluación de los docentes según su género, mientras que en otros 5 estudios realizados entre los años 1974 y 1987, se reporta un sesgo de género, en específico, de que los estudiantes masculinos tienden a calificar a las docentes femeninas más bajo que a los docentes masculinos. Por otro lado, en su propio estudio empírico, estos mismos investigadores se encuentran con un sesgo a favor de las docentes, que son evaluadas mejor que sus pares masculinos por las estudiantes mujeres.

Otro intento por agrupar los resultados de varios estudios es el trabajo de Feldman (1992), quien realizó un metaanálisis de 17 estudios experimentales, en los que no evidenció diferencias significativas en la evaluación de los docentes según su género.

Sin embargo, investigaciones más recientes realizadas en Francia y Estados Unidos han evidenciado sesgos de género en perjuicio de las mujeres, en una intensidad estadísticamente significativa, y lo suficientemente fuertes como para que, incluso, docentes efectivos tengan menores puntajes que docentes menos efectivos (Boring, et al., 2016). La tendencia sería de mejor evaluación para los docentes hombres, por parte de ambos sexos de estudiantes (Boring, 2016), incluso en algunos contextos en que el estudiantado parece aprender más con las docentes femeninas que con los docentes masculinos (Boring, et al., 2016).

Otro estudio, también reciente, y más relevante aún para nuestro contexto local, ya que fue realizado sobre una muestra chilena, es el de Medel y Asun (2014), quienes reportan una propensión a que las académicas, en comparación con sus colegas masculinos, obtengan mejor evaluaciones en aspectos como la responsabilidad y habilidades pedagógicas, y sean peor evaluadas en el ámbito de dominio disciplinario. A esto, se suma que los patrones de respuesta diferenciados por género en las evaluaciones docentes se manifestarían tanto en evaluadores como en evaluados. Habría una tendencia de las mujeres a evaluar un poco mejor que los hombres. Además, el estudiantado en general tendería a favorecer a las académicas en los primeros años de enseñanza, pero a la vez penalizarlas en los últimos. En esta penalización, serían los varones quienes juegan un rol más decisivo.

Según Boring (2016), el problema con el sesgo de género no es simplemente la mera identificación de rendimientos distintos entre docentes hombres y mujeres, sino las consecuencias negativas reales que tienen estos diagnósticos sobre las docentes. Boring denuncia que entre las consecuencias nocivas de este sesgo, que tendería a afectar generalizadamente a las docentes, se

encuentran: el mayor esfuerzo que tenderían a invertir en mejorar dimensiones de la enseñanza de mayor consumo de tiempo (tales como preparación de la clase o la atención entregada a los estudiantes), lo que por consiguiente también tendería a disminuir su disponibilidad para realizar otras actividades, tales como la investigación; también, en general, habría una menor cantidad de cursos dictados por mujeres al interior de las universidades; una mayor disminución y abandono de la carrera docente por parte de las mujeres, al ver obstaculizadas sus posibilidades de ascender; entre otras desventajas. Con todo, Boring propone que el sesgo en perjuicio de las mujeres presente en la mayoría de los espacios de docencia produciría un contexto académico de inequidad generalizada entre hombres y mujeres.

A la diversidad de diagnósticos sobre si el sesgo se encuentra presente o no en los diversos contextos educativos, también se le suman distintas formas de interpretar estos sesgos de género.

Centra y Gaubatz (2000) señalan que el sesgo en beneficio de las docentes femeninas por parte de las estudiantes del mismo sexo, simplemente se explica por la diferencia de prioridades y concepciones que poseen las estudiantes sobre qué consideran una buena docencia. Proponen en base a teorizaciones de Belenky et al. (1986), que las mujeres serían más receptivas a metodologías participativas y colaborativas, más que a evaluativas y discursivas, y que por ello tenderían a evaluar mejor el tipo de docencia promovida por su mismo sexo. Con ello, sugieren que las diferencias por sexo realmente no serían rasgos invalidantes de las evaluaciones docentes, sino que legítimas diferencias entre concepciones distintas de la docencia, todas medidas de una única forma.

Posturas más críticas, posicionadas desde la teoría del género, señalan que el sesgo de género sería otra manifestación del dominio de los hombres sobre la producción y administración del conocimiento universitario, lo que explicaría por qué las profesoras tienden a ser mejor evaluadas en ámbitos de la docencia considerados “menores” por académicos y estudiantes (como la responsabilidad), en condiciones que los docentes varones tienden a ser mejor evaluados en áreas relativas al conocimiento disciplinario, es decir, en temas mucho más cercanos a las fuentes del prestigio y autoridad académica (Medel & Asun, 2014). En esta misma línea, las investigaciones encabezadas por Boring sugieren que las dimensiones que los estudiantes evalúan de sus docentes tienden a calzar con estereotipos de género, lo que explicaría por qué las mujeres muchas veces son evaluadas peor que sus pares masculinos en aspectos como el conocimiento o el liderazgo (Boring, 2016; Boring, et al., 2016).

Otros rasgos de los docentes también han sido examinados en el análisis de los posibles sesgos de evaluación docente. El prestigio académico previo percibido por los alumnos (Griffin, 2001), los rasgos de personalidad de los docentes (Clayson & Sheffet, 2006; Murray, et al., 1990), e incluso el apropiado uso del humor (Duque, 2013) han sido evidenciados como posibles factores que afectan positivamente la evaluación que los alumnos hacen de sus docentes.

Dentro de las variables asociadas a los cursos que han sido relacionadas con el rendimiento de la evaluación de la calidad docente, está el contenido de las clases. Se ha evidenciado que cursos de diferentes disciplinas o asignaturas poseen evaluaciones docentes distintas entre sí (Acevedo Álvarez & Mairena Rodríguez, 2006; Arámburo Vizcarra & Luna Serrano, 2013), a tal punto que se sugiere que no se deberían realizar comparaciones de rendimiento docente entre diferentes áreas de conocimiento (Ramsden, 1991). Por ejemplo, se ha evidenciado que cursos de matemáticas son evaluados, en general, con menores puntajes que cursos con contenidos en humanidades o artes (Feldman, 1978; Cashin, 1990). Otros estudios empíricos han señalado que, más que el rendimiento de los docentes, lo que explica la evaluación de los alumnos es la satisfacción e interés con el tema revisado en clases (Perry, et al., 1974; Gillmore, et al., 1978).

La electividad u obligatoriedad de los cursos también ha sido examinada como sesgo. Se ha evidenciado que cursos electivos son más favorablemente evaluados que cursos obligatorios (Feldman, 1978; Scherr & Scherr, 1990; Hativa, 1996; Rantanen, 2013), simplemente por una cuestión de mayor interés de los alumnos en los cursos electivos (Wachtel, 1998). También relacionado con las características de los cursos, existe un estudio que incluso ha identificado una relación entre la evaluación docente y el horario en que los cursos son dictados, siendo los cursos de primera hora de la mañana, los inmediatamente después de almuerzo y los de más entrada la tarde, aquellos peor evaluados (Koushki & Kuhn, 1982).

Un aspecto de los cursos sobre el que se ha puesto considerablemente mucha más atención es el tamaño de la clase, y su relación con los puntajes de evaluación docente. La mayoría de las investigaciones han encontrado una relación positiva entre clases con alta cantidad de alumnos, y profesores peor evaluados. Algunos ejemplos son Scott (1977), Feldman (1978), Marsh (1984), McKeachie (1990), Greenwald (1997), Fernández, et al. (1998), Acevedo Álvarez y Mairena Rodríguez (2006), Arámburo Vizcarra y Luna Serrano (2013), entre otros. En un metaanálisis que consideró 52 investigaciones distintas que buscaron examinar esta relación, Feldman (1984) señala que 38% de los estudios reportan una relación lineal inversa estadísticamente significativa entre las variables (a más grande la clase, menor evaluación docente), un 21% reportan una

relación curvilínea (puntajes de evaluación docente más altos para clases levemente más grandes, y levemente más pequeñas) mientras que otro 38% de los estudios no identificaron una relación significativa entre el tamaño de la clase y los resultados de la evaluación docente.

Buscando describir cómo ocurre esta interacción, la investigación de Scott (1977) evidenció que docentes que sentían que sus cursos eran demasiado grandes como para exhibir los contenidos de forma adecuada, obtenían una menor evaluación docente. Según Feldman (1978), esto sugiere que la percepción de los docentes sobre la cantidad de estudiantes en sus clases tiene efecto directo sobre su rendimiento pedagógico. Por su parte, McKeachie (1990) ha interpretado esta relación señalando que mientras más grande la clase, menor es el sentido de responsabilidad personal y proactividad de los docentes, y menor la posibilidad de que estos puedan conocer personalmente a sus estudiantes y adaptar la instrucción a las necesidades individuales de cada uno de ellos.

En Chile, distintos diagnósticos han apuntado hacia conclusiones similares (Cornejo, et al., 2009; Ávalos, 2013). Han señalado que la carga docente (número de alumnos que deben atender los profesores) es uno de los principales factores que afectan negativamente su grado de motivación, percepción de eficacia y sentido del bienestar/malestar. A esto se podría agregar, como argumento, que por el efecto negativo sobre su motivación y bienestar, bien podría ocurrir un empeoramiento en su rendimiento que llevaría a peores puntajes en evaluación docente.

Ahora bien, Marsh (1987) ha argumentado que la relación entre la cantidad de alumnos en las clases y la evaluación docente no constituye un sesgo, sino una auténtica diferencia en el rendimiento de los docentes. Ello, ya que propone que un bajo rendimiento docente en condiciones de alta cantidad de alumnos no debe entenderse como el efecto de una variable ajena o exógena a los docentes afectando su rendimiento, sino que debería ser considerada como una real medida de sus capacidades en condiciones de dificultad. Rechazando la postura de Marsh, Abrami (1989) responde que este argumento no puede ser utilizado para validar los instrumentos, y señala que, en cambio, lo único que demuestra la interacción entre la evaluación docente y el tamaño de la clase, es que todo diagnóstico de calidad docente influenciado por el tamaño de los cursos no debe ser usado para la toma de decisiones de alta connotación al interior de las plantas docentes.

Dentro de las variables asociadas a los estudiantes que se han propuesto como sesgos de la evaluación docente se encuentran, principalmente, las expectativas de notas de los alumnos. Diferentes estudios han demostrado que existe una correlación positiva entre las notas que los

estudiantes esperan conseguir de determinado curso y los puntajes de evaluación docente que otorgan a sus docentes (Feldman, 1976; Centra, 2003; Spooren & Mortelmans, 2006; Acevedo Álvarez & Mairena Rodríguez, 2006; Remedios & Lieberman, 2008; Ewing, 2012). De forma similar, los niveles de carga académica han sido tomados en consideración, con estudios que han evidenciado que menores niveles de carga académica entregada a los alumnos producen mejores puntajes de evaluación docente (Greenwald & Gillmore, 1997; Centra, 2003; Remedios & Lieberman, 2008). Incluso hay estudios que encuentran asociaciones significativas entre bajas notas otorgadas a los alumnos, y bajas evaluaciones docentes (Zabaleta, 2007). Debido a estos sesgos mencionados, algunos investigadores advierten que, por presiones para mejorar su evaluación docente, algunos docentes podrían optar por mayores niveles de indulgencia para sus estudiantes, produciendo esto, entre otras consecuencias negativas y poco éticas, fenómenos de inflación en las notas de los estudiantes y disminución de la exigencia (Eiszler, 2002; Crumbley, et al., 2010).

El año de carrera de los estudiantes también ha sido incorporado al análisis de sesgos, con la mayoría de los estudios apuntando hacia mejores puntajes de evaluación docente en cursos de años finales de carrera (Feldman, 1978; Marsh, 1987). También, algunos estudios presentan evidencia que apunta a que los estudiantes universitarios cambian sus prioridades sobre qué consideran buena docencia a medida que avanza su carrera, teniendo mayor interés por aspectos como la claridad expositiva, apoyo académico y estructuración de los cursos en los primeros años, y un interés más orientado hacia la discusión, debate de ideas e integración de conocimientos en años finales de sus estudios universitarios (Smith & Cranton, 1992; Hills, et al., 2009).

La presente revisión de los principales sesgos que propone la literatura deja en evidencia que las posibles fuentes de invalidez discriminante de los instrumentos de evaluación docente son múltiples. Los diagnósticos sobre si las evaluaciones se encuentran sesgadas o no varían en cada caso, sin embargo, es preocupante la vasta cantidad de trabajos que identifican variables consideradas por los expertos como ajenas a la calidad de la instrucción docente que, de todas formas, se encuentran altamente relacionadas con las medidas de rendimiento docente que los estudiantes otorgan a sus profesores.

Por estos mismos motivos, algunos expertos y docentes son tajantes en apuntar que los cuestionarios de evaluación docente carecen, en general, de validez discriminante (Greenwald & Gillmore, 1997), y que, si bien pueden llegar a identificar atributos que los “buenos” docentes deben tener, se encuentran irremediablemente correlacionados con un gran número de atributos

ajenos con los que no deberían relacionar. Algunas posturas más críticas aseguran que las evaluaciones docentes poseen sesgos más intensos que cualquier conexión que estos instrumentos podrían llegar a tener con la calidad de la docencia, tanto así, que muchas veces su asociación con puntajes convergentes, como el rendimiento de los alumnos, puede llegar a ser inversa (Boring, et al., 2016). Además, estos sesgos serían irremediables, en tanto que su intensidad y direccionalidad dependería de cada contexto educativo, lo que hace imposible que las interpretaciones de las medidas de evaluación docente puedan ajustarse a un punto de observación común.

Pese a las dudas de muchos y el volumen de sesgos propuestos, la comunidad científica que aborda el tema parece, en general, a favor de que las evaluaciones docentes proporcionan al profesorado evaluaciones válidas, fiables, y útiles (Penny, 2003). Si bien nadie niega la existencia de variables ajenas a la buena docencia que podrían afectar las evaluaciones, la mayoría de los expertos considera que las distorsiones no son lo suficientemente intensas como para clasificar erróneamente a los docentes, o bien, invalidar este tipo de mediciones (McKeachie, 1990; Centra & Gaubatz, 2000; Acevedo Álvarez & Mairena Rodríguez, 2006; Medel & Asun, 2014). Algunas visiones más favorables a la validez de estos instrumentos, como la de Cashin (1995) sugieren que a pesar de que los diferentes estudios puedan respaldar casi cualquier conclusión, la tendencia sería a que en general, los cuestionarios de evaluación docente son una medida de calidad docente estadísticamente fiable, válida, y relativamente libre de sesgos, probablemente más que cualquier otro indicador usado para evaluar la calidad docente. Para Medel y Asun (2014), la relativa validez a la que tienden las opiniones de la mayoría de los investigadores no debe excluir que, de todas formas, cada lectura de las evaluaciones docentes deba ajustarse a su situación y contexto, siempre en vigilancia de los factores extra docencia que eventualmente podrían intervenir.

Para Marsh, existe relativo consenso entre los investigadores del campo sobre que determinados sesgos, si bien pueden influenciar la evaluación docente en determinados contextos, las puntuaciones obtenidas mediante estos instrumentos son, de todas formas, una medida representativa del grueso del desempeño docente de los académicos universitarios (Marsh, 2007b). Ahora bien, resulta relevante mencionar que este autor va más allá en la defensa de la validez discriminante de las evaluaciones docentes, al señalar que muchos de los aspectos considerados a priori como sesgos, en realidad son legítimos aspectos que influencian la evaluación docente (Marsh, 1984). Propone que la influencia, por ejemplo, de los rasgos de

personalidad de los profesores y el tamaño de las clases sobre las evaluaciones docentes, más que oponerse a la validez, la avala, en tanto son componentes originalmente vinculados con la capacidad de los docentes para desempeñarse. Por ello, alega además que el fenómeno del rendimiento docente sería mejor comprendido si los investigadores no se concentraran exclusivamente en tratar de interpretar las relaciones entre las variables contextuales y la evaluación docente como sesgos, y se dedicaran a examinar el significado de cada relación específica.

De forma similar, pero escéptica aún, algunos investigadores señalan que el problema con los sesgos que acarrea la evaluación docente no se encuentra necesariamente en la interacción que poseen los puntajes con otras variables ajenas a la calidad de la docencia, sino en la interpretación que se hace de las evaluaciones docentes, y los exagerados alcances que se le han otorgado a este tipo de instrumentos. Por ejemplo, se ha señalado que existe mucha preocupación entre los docentes sobre las posibles diferencias de percepciones que pueden tener los estudiantes sobre qué constituye la buena docencia, y si esas percepciones tienen en absoluto relación con este constructo (Spooren, et al., 2013). Mientras que algunos expertos sugieren que medir la opinión de los alumnos es el mejor procedimiento para evaluar el rendimiento docente, debido a que son permanentes observadores de su labor (Tejedor, 2003), otros critican que las evaluaciones docentes tienden, erróneamente, a dar por sentado que los estudiantes saben qué es lo que requieren de parte de sus docentes, insistiendo que muchas veces se ignora que los estudiantes valoran elementos que pueden ser innecesarios o ajenos a la buena docencia (Chonko, et al., 2002).

Por ello, algunos autores advierten que la evaluación docente no debe ser entendida como medición de la calidad de la enseñanza, sino de la satisfacción de los estudiantes con ella (García Garduño, 2008). De forma similar, se sugiere que no habría que confundir la opinión de los estudiantes, con conocimiento real sobre las capacidades de sus docentes (Ory, 2001). Algunos autores visibilizan posturas de algunos docentes más críticos aún, que califican peyorativamente a las evaluaciones docentes como medidas de “satisfacción del cliente” (Beecham, 2009; Klajman, 1997), “formularios de felicidad” (“happy forms”) (Harvey, 2001, p. 13) o bien como “cuantificaciones sin sentido” usadas para “concursos de personalidad” (Kulik, 2001, p. 10).

Bajo este argumento, es posible proponer que el gran sesgo, o bien, el sesgo “de orden superior” —en la medida que explicaría muchos de los otros sesgos— que podrían estar incorporando la mayoría de los instrumentos de evaluación docente, sería que éstos no sólo miden la calidad de la

docencia, sino que se ven inevitablemente permeados por la satisfacción de los estudiantes, además de sus creencias, opiniones y estereotipos sobre qué conductas constituyen (y cuáles no) buenas prácticas docentes.

Como cierre para la discusión sobre la validez discriminante de los instrumentos de evaluación docente, se puede concluir que no existe un consenso entre los académicos e investigadores acerca de si los cuestionarios de evaluación docente son instrumentos de medición legítimos y válidos de la calidad docente, principalmente debido a los sesgos que involucran (Hornstein, 2017). Según Medel y Asun (2014), no existe actualmente en la comunidad científica un juicio unánime respecto a la validez discriminante de los cuestionarios de evaluación docente, si bien existen investigadores de vasta trayectoria que han apoyado insistentemente su validez (Aleamoni, 1997; Marsh & Roche, 2000; Marsh, 1984; Marsh, 2007a). Ahora bien, pese a la existencia de dudas sobre la validez discriminante que logran los cuestionarios de evaluación docente, el uso de este tipo de instrumentos es vigente, y continúa en aumento.

iv. Validez consecuente y la preocupación por los usos de la evaluación docente

Como ya se ha venido adelantando, existen diferentes y bien fundamentados motivos para dudar de la validez de la evaluación docente. A las imprecisiones y desacuerdos conceptuales al interior de la academia sobre qué significa "calidad docente" y cómo debe ser medida; al posible uso de instrumentos mal formulados; a la relación que tienden a tener estas mediciones con sesgos de todo orden —que se expresan de acuerdo al contexto y son, por ello, incontrolables (Boring, et al., 2016)—; se suma otro problema, en específico, asociado a la interpretación y uso que se hace de los puntajes de evaluación docente. Existen críticas que sugieren que muchas de las autoridades y administradores en el ámbito de la educación no sabrían interpretar los resultados surgidos a partir de estos instrumentos, debido a la falta de entrenamiento para ello (Hornstein, 2017). Reflejo de esto sería, por ejemplo, el cálculo y abuso de medias para puntuar el rendimiento entre el profesorado, y la generalizada práctica —calificada como simplificante y sinsentido— de clasificar a los docentes "efectivos" como aquellos arriba de la media, e "inefectivos" aquellos abajo de la media (Klajman, 1997; Gray & Bergmann, 2003; Montoya, et al., 2014).

Debido a éstas problemáticas, ya esbozadas en este trabajo, al interior de la academia y los espacios de docencia existe gran escepticismo asociado a los CEDs, con muchos docentes que, pese a acceder a formar parte de procesos de evaluación docente, sienten que no corresponde que la opinión de los estudiantes sea la única o más importante instancia evaluativa de su labor

(Salazar, 2008). Reflejando estas dudas, hay expertos que recomiendan precaución al usar las medidas de evaluación docente como guía para el mejoramiento de las habilidades del profesorado (Boring, et al., 2016).

Más aún, existen análisis e investigaciones que evidencian un grado no menor de rechazo por las evaluaciones docentes en muchos espacios educativos. Investigaciones sobre las percepciones que poseen los profesores sobre sus procesos de evaluación muestran que, en algunos contextos, hay docentes se sienten incluso hostiles frente a este tipo de mediciones sobre su desempeño, debido a que, además de percibirse como mediciones sesgadas, mal logradas e inválidas, las evaluaciones docentes serían vistas como un mecanismo de control hacia el profesorado, y por ello más utilizado para atacar su autonomía que como herramienta de mejoría de la enseñanza (Harvey, 2001; Kulik, 2001; Penny, 2003; Silva Montes, 2009). En esta misma línea, se ha criticado que muchas instituciones de educación superior poseen discursos sobre la evaluación docente donde prevalecen los fines de mejora de la calidad docente, pero en realidad sus usos están orientados al control, sirviendo para el mejor funcionamiento institucional, pero sin implicancias reales que ayuden al desarrollo profesional individual (Luna Serrano & Torquemada, 2008).

Las preocupaciones de los investigadores y docentes son más intensas considerando que en algunos contextos institucionales las evaluaciones docentes se utilizan para diferenciar entre docentes “ineficaces” y “eficaces”, utilizando estos diagnósticos para la toma de decisiones de alta connotación al interior de las plantas docentes, tales como la remoción, o bien la retención o entrega de incentivos monetarios a los docentes (Patrick & Mantzicopoulos, 2016). Como consecuencia de esto, muchos investigadores concuerdan en que los resultados de las evaluaciones docentes pueden tener serios efectos negativos sobre las carreras profesionales de los docentes (Gray & Bergmann, 2003; Kogan, et al., 2010; Spooren, et al., 2013). Por ello, diversas fuentes sugieren cautela en el uso de estas mediciones para fines sumativos sobre el profesorado (García Garduño, 2000), y posiciones más críticas aún señalan que se debería descontinuar totalmente el uso de medidas de evaluación docente en base a respuestas de estudiantes para definir la contratación, despido, promoción, estimulación y en general toda forma de decisión sumativa sobre el personal docente (García Garduño, 2008; Boring, et al., 2016).

Ahora bien, pese a las variadas preocupaciones sobre sus consecuencias, algunos expertos consideran curioso que exista poca investigación sobre cómo interpretar y mejorar el uso de los datos surgidos desde las evaluaciones docentes (Marsh, 1987; Cashin, 1995; Marsh & Roche, 2000).

En efecto, a comienzos de este siglo, en la investigación angloparlante, parecía haber pocos estudios que tratasen sobre este tipo de validez, y a la vez, escasas decisiones en política educativa que se fundamentaran en el análisis de la evaluación docente. Por ejemplo, Rice (2003) señala que es notable la escasez de investigación sobre cómo guiar decisiones tan críticas en los espacios educativos como lo son cuáles profesores contratar, retener, aumentar remuneración, y promover.

En esta línea, el estudio de Solomon et al. (1997) acusa, por ejemplo, que la investigación científica sobre evaluación docente no ha puesto atención suficiente a la que consideran crítica discusión sobre cuántas respuestas a una misma evaluación son necesarias para lograr un determinado nivel de fiabilidad. Para estos investigadores, esta ausencia vendría en detrimento de la capacidad que se tiene de tomar decisiones correctamente informadas a la hora de utilizar las evaluaciones docentes para definir las contrataciones y la asignación de cursos entre el profesorado. Haciéndose cargo del vacío que visibilizan, este mismo estudio concluye que para que la evaluación de un curso logre niveles altos de consistencia interna, se requiere una muestra desde los 30 a 40 casos hacia arriba.

Pese a lo anterior, hay trabajos (Morgan, et al., 2014) que señalan que en el último tiempo el panorama en el sub-campo de investigación sobre la validez consecuente de los cuestionarios de evaluación docente parece estar dando un vuelco progresivo, al observarse una renovación del interés y un consiguiente aumento de la investigación asociada al desempeño y efectividad docente. Las medidas sobre cómo se comporta el rendimiento docente en el tiempo se han comenzado a utilizar cada vez más para hacer juicios más robustos sobre el desempeño de los docentes en las aulas (Guarino, et al., 2015), y más aún, en los últimos cinco a diez años, se ha incentivado la reflexión sobre este tipo de mediciones, sobre todo, en vista de proveer bases sólidas para la toma de decisiones sobre la contratación, permanencia, promoción e incluso la bonificación monetaria de los docentes (Morgan, et al., 2014).

v. Evaluación docente longitudinal

Como ya se ha mencionado, uno de los principales objetivos y usos de la evaluación docente es otorgar retroalimentación, proveniente de los propios alumnos, sobre el rendimiento de cada docente como educador. Bajo este modelo, se asume que los docentes, al recibir su evaluación a cada asignatura dictada, contarían con información útil para identificar y mejorar aquellos aspectos en los que se encuentran débiles.

Ahora bien, lo anterior sólo tiene sentido al asumir que el rendimiento, las fortalezas y debilidades de cada profesor, se comportan de forma estable en el tiempo. Sólo asumiendo esto, las mediciones de rendimiento docente tomadas en un momento determinado (clásicamente al final de los cursos) serían representativas de procesos que comúnmente son semestrales o anuales. Sin embargo, la premisa sobre la estabilidad es algo que no se puede obviar. Por ello, para obtener mediciones de calidad docente fiables, algunos investigadores sugieren que las puntuaciones de evaluación docente se deben obtener en más de una oportunidad, no sólo de forma transversal y de una única vez (West, 1988). Esto, ya que los docentes no sólo podrían variar su rendimiento año a año (Polikoff, 2015), sino que también clase a clase (Patrick & Mantzicopoulos, 2016), pudiendo cualquier docente "tener un mal día" (Kane & Staiger, 2012). Según los investigadores que apoyan la observación de la evaluación docente en el tiempo, usar las medidas de una única oportunidad constituiría un error, ya que clasificaría mal a los docentes en su rendimiento.

Agregando el componente longitudinal al análisis, se reconfigura la pregunta por la fiabilidad de los cuestionarios de evaluación docente. ¿Son las evaluaciones docentes que se realizan una vez por curso, clásicamente al final de estos, una medida representativa de la calidad del docente a lo largo del periodo lectivo? ¿Es la evaluación docente final, por el contrario, una medida poco representativa de la calidad general de la enseñanza impartida, no logrando representar una trayectoria de desempeño docente que es más rica en variabilidad y complejidad?

Al problema de la mala clasificación ocurrido en caso de ignorar que el rendimiento de los docentes puede no ser estable en el tiempo, se le suma un error consecuente más grave aún, que se ha abordado en el apartado anterior: llevar a determinaciones erróneas en el ámbito de la selección docente, pero ahora en específico por motivo de utilizar sólo mediciones transversales. La problemática sobre la estabilidad de la evaluación docente cobra así mucha relevancia, posicionándose como una de las principales preguntas en la investigación sobre la fiabilidad de este tipo de mediciones.

Según Herbert Marsh (2007a), quien es uno de los investigadores más productivos en el campo de la evaluación docente a nivel internacional, una de las mayores limitaciones de la investigación relativa a la experiencia y efectividad de la enseñanza, es que la mayoría de los estudios han considerado las puntuaciones de evaluación docente entregadas por los alumnos en el contexto de un curso específico y en una única ocasión. En esta misma línea, Morgan et al. (2014) señalan que la investigación longitudinal en evaluación docente es muy escasa, sobre todo aquella enfocada en la estabilidad de las trayectorias, y que la mayor cantidad de los trabajos relevantes

que utilizan una metodología longitudinal son la mayoría más bien nuevos, de los últimos cinco a diez años.

Existen, sin embargo, investigadores que reivindican trabajos realizados con anterioridad que utilizan una metodología longitudinal (Konstantopoulos, 2014; Costin, et al., 1971), y otros que aclaran que la situación es de poca visibilidad, y no necesariamente de poco volumen de investigaciones realizadas (Good & Lavigne, 2015). La propia revisión de los estudios longitudinales en evaluación docente lleva a tomar posición acorde a esta última postura. Lo cierto es que existen variadas investigaciones en evaluación docente longitudinal, de distintos tipos, y más o menos similares resultados, como se verá más adelante. De todas formas, en relación a la vasta cantidad de investigaciones que existen en el campo de evaluación docente, el volumen de investigación longitudinal es considerablemente menor, lo que no se condice que con la gran importancia que algunos investigadores le otorgan a este tipo de análisis.

A continuación, se presenta una revisión sobre el estado del arte en investigación longitudinal en evaluación docente, identificando sus principales abordajes, diagnósticos y resultados, como también, finalmente, posibles debilidades y vacíos aún no resueltos.

Para ordenar los diversos estudios en este sub-campo, resulta útil, en primer lugar, distinguir que la forma en que se aborda la longitudinalidad de la evaluación docente puede variar. En términos de periodicidad, algunas investigaciones observan el rendimiento docente a lo largo de años de instrucción impartida, evaluando cómo se modifica y varía la calidad docente según los años de experiencia acumulados. Este tipo de estudios han sido denominados en la literatura de habla inglesa como “year-to-year” (año a año). Por otro lado, las investigaciones de periodo corto, que analizan cómo se comporta la evaluación que hacen los alumnos de sus profesores a lo largo del desarrollo de un mismo curso, semestre o año, se denominan de periodos “within-year” (intra-año).

En la investigación de evaluación docente de periodo largo, el principal diagnóstico al que se ha llegado es a la generalizada estabilidad de las trayectorias de evaluación docente. Algunos ejemplos de estudios con este tipo de resultados son Overall y Marsh (1980), Hanges et al. (1990), Marsh y Hocevar (1991), Krantz-Girod et al. (2004), Marsh (2007a) y Polikoff (2015).

La estabilidad parece ser la norma en los periodos largos. Esto concluyen incluso estudios tan extensos como el de Marsh y Hocevar (1991), que registró puntuaciones de evaluación docente durante un periodo de 13 años, y reporta una ausencia de cambios significativos en la evaluación

que reciben los docentes a lo largo del tiempo. Ahora bien, investigaciones como las de Hanges, et al. (1990) sugieren que la estabilidad anual de la evaluación docente varía según la dimensión medida, con aspectos como la estimulación de la participación y la capacidad de organizar la clase teniendo menores niveles de estabilidad comparados con otras dimensiones de evaluación docente. De todas formas, este mismo estudio concluye que, en la mayoría de los espacios educativos y en la generalidad de las mediciones, se puede asumir que el rendimiento de los docentes es estable con los años.

Para los estudios longitudinales de periodo corto, casi la totalidad de los estudios utiliza un formato de unas pocas mediciones por semestre (comúnmente 2 o 3), y al igual que para el tipo de estudios anterior, la estabilidad de la evaluación docente también parece ser el diagnóstico predominante. Ello evidencian ejercicios de revisión de investigaciones en evaluación docente de este tipo, como el de Costin et al. (1971), que enumera cinco estudios realizados antes de la década de los 70', todos reportando correlaciones altas (típicamente de r= ,7 a ,9) entre distintos momentos de evaluación docente (pp. 512-513). Costin y su equipo interpretan estos niveles de consistencia longitudinal como una característica de los docentes, en general, de ser muy estables en el corto plazo en su calidad pedagógica. Estudios que analizan evaluaciones docentes longitudinales de periodos cortos, posteriores a los revisados por Costin y su equipo, diagnostican similares grados de estabilidad para este tipo de mediciones. Algunos de ellos son Bauswell et al. (1975), Irby et al. (1977), Canaday et al. (1978), Smith (1979), y Hativa (1996).

Estudios más actuales y de alta connotación, como los del MET Project (Bill & Melinda Gates Foundation, 2010), entre sus hallazgos, señalan que las correlaciones entre las puntuaciones de evaluación docente intra año (otoño a primavera) poseen una estabilidad alta (media de r = ,7). Sin embargo, este mismo estudio advierte, de forma similar a Hanges et al. (1990) para el caso de los periodos anuales, que la estabilidad de algunas de las dimensiones específicas de la evaluación docente es sólo intermedia, como lo es para la dimensión de claridad expositiva del docente (r = ,5).

Otras investigaciones de periodo corto han considerado, específicamente, la estabilidad y permanencia de la primera impresión que tiene el docente sobre sus estudiantes en relación a sus niveles de evaluación docente posteriores. Varias investigaciones han evidenciado que la primera evaluación que hacen los alumnos de sus docentes tiende a permanecer en el tiempo, estabilizándose en una intensidad bastante alta. Por ejemplo, la investigación de Buchert, et al. (2008) evidencia que la evaluación docente no cambia significativamente desde lo ya definido en

las primeras dos semanas de clases. Más aún, existen investigaciones que incluso señalan que la evaluación docente semestral es medianamente predecible después de la primera hora de la primera clase del semestre (Kohlan, 1973). Las interpretaciones frente a estos fenómenos indican que esto puede ocurrir debido a que los estudiantes incorporan muy poca nueva información sobre la conducta de sus docentes después de los primeros contactos que tienen con estos, conformándose una percepción sobre sus docentes casi de forma inmediata y prácticamente inamovible; también, se sugiere que los estudiantes ejercerían una estereotipación de sus docentes, que tendería a ser más fuerte que cualquier estímulo o posibilidad de cambio en la apreciación subjetiva que se tiene de éstos.

Pese al diagnóstico de estabilidad al que llegan la mayoría de los estudios que analizan longitudinalmente la evaluación docente en periodos cortos de tiempo, algunas investigaciones obtienen resultados que tienden a poner en duda la presunción de estabilidad general. Overall y Marsh (1979) reportan una estabilidad sólo relativa entre evaluaciones de medio y final de semestre (con correlaciones que pueden variar entre ,3 y ,7). De forma similar, la investigación de West (1988), que consideró mayor cantidad de mediciones en un menor plazo, reporta que las percepciones de los estudiantes sobre la calidad de la enseñanza de sus docentes son sólo moderadamente consistentes en el tiempo, existiendo cambios a lo largo del semestre, y variando la estabilidad en intensidad y sentido según cada curso y docente específico. La investigación de Bauswell et al. (1975), por su parte, evidencia que la estabilidad intersemestral de la evaluación docente es media-alta cuando se observa para docentes que dictan la misma clase (media de r= ,69), sin embargo, la estabilidad es baja cuando se observan profesores que han cambiado la clase que dictan (media de r= ,33), y prácticamente nula para las mismas clases que han cambiado de docente (media de r= ,17). La investigación de Hanges et al. (1990) llega a conclusiones similares a la de Bauswell et al. (1975), también reportando mayores índices de estabilidad para docentes que continúan realizando los mismos cursos, en comparación con aquellos que cambian de curso. Por este motivo el estudio de Hanges et al. (1990) interpreta la estabilidad como un atributo de los docentes, más que de los cursos.

En su conjunto, las investigaciones aludidas en el párrafo anterior evidencian que, si bien las evaluaciones docentes pueden ser inconsistentes en el corto plazo en determinadas circunstancias aún no totalmente estudiadas, generalmente tienden a la estabilidad, sobre todo si se observan indicadores generales de la evaluación docente, y a un mismo educador impartiendo la misma clase en el tiempo.

Otra investigación disidente que sugiere que, más bien, existirían altos niveles de inestabilidad de la calidad docente al interior de los periodos semestrales de los cursos, y ello se evidenciaría correctamente sólo al observar las variaciones de su rendimiento con mayor detalle, es decir, clase a clase (Patrick & Mantzicopoulos, 2016). Con esta proposición, Patrick y Mantzicopoulos llevan a cabo una investigación que utiliza indicadores observacionales para puntuar el rendimiento docente, y en la que diagnostican una marcada variabilidad entre los distintos días de clases, para todas las dimensiones de la calidad docente que analizan[2].

Esta investigación parece relevante debido a que es la única que ha sugerido la necesidad de mediciones clase a clase para observar la inestabilidad longitudinal del rendimiento docente con mayor precisión. Este estudio, sin embargo, trabaja con mediciones observacionales del rendimiento, que son menos fiables en comparación con las mediciones de evaluación docente, y no son inmediatamente comparables con estas (Polikoff, 2015; Kane & Staiger, 2012). De todas formas, hacer alusión a este estudio se considera relevante ya que, como se detalla más adelante, la presente investigación utiliza un abordaje clase a clase, siendo con ello, hasta donde llega el conocimiento del autor, el primer estudio que trabajaría con evaluaciones docentes aplicadas a los alumnos con esta periodicidad.

Habiendo hecho un panorama sobre las principales investigaciones existentes sobre la estabilidad longitudinal de la evaluación docente, y revisados los diagnósticos que realiza cada una al respecto, es posible hacer notar una deficiencia importante que se manifiesta de forma generalizada en las investigaciones que analizan la estabilidad longitudinal de la evaluación docente. Se trata de la falta de interpretabilidad respecto a qué significarían determinados niveles de estabilidad (o bien de inestabilidad) en las trayectorias de evaluación docente, sobre todo en periodos de corto plazo.

Si bien los niveles de estabilidad han sido identificados en cada uno de los estudios analizados y se ha señalado de forma explícita, en la mayoría de los trabajos, que la estabilidad es una medida de fiabilidad para las mediciones de evaluación docente, la investigación existente no se ha hecho cargo de reflexionar el sentido sustantivo de la estabilidad/inestabilidad. Se ha hecho poco por evidenciar cuándo y bajo qué condiciones ocurren fenómenos de inestabilidad, qué factores se relacionan con sus niveles —tanto docentes como extra docentes—, no se han indicado con certeza cuáles serían niveles óptimos de estabilidad/inestabilidad, y más importantemente aún, no

[2] Las dimensiones evaluadas fueron: capacidad de soporte emocional, capacidad de organización de la clase y capacidad de instrucción.

se ha estudiado si este fenómeno afecta el rendimiento general de los docentes, las consecuentes evaluaciones hechas por sus alumnos, y más aún, si es que tiene efecto sobre el logro académico de estos últimos.

Como intentos para delimitar el fenómeno, en la literatura se han hecho distinciones gruesas, por ejemplo, al mencionar que la estabilidad/inestabilidad temporal de toda medición sobre la calidad docente (no sólo de las evaluaciones docentes hechas por alumnos) se compondría de dos elementos: variación temporal de los factores docentes (es decir, variación de la calidad docente en el tiempo), y "ruido estadístico" resultante de errores de medición (McCaffrey, et al., 2009). Aproximaciones como esta son indicio del limitado entendimiento que aún existe sobre el fenómeno de la estabilidad/inestabilidad en las mediciones sobre la calidad docente, ya que con esto sólo se explicita algo que es evidente para toda medición representacional de constructos complejos.

Otros indicios de lo poco que se ha profundizado en la comprensión del fenómeno de la inestabilidad del rendimiento docente son, por ejemplo, el hecho de que se sugiera observar la estabilidad de rendimiento en otras profesiones para poder establecer parámetros de estabilidad aceptables para el rendimiento docente (Polikoff, 2015); que se anuncie que no se tiene claro si es la calidad docente la que sería inconsistente en el tiempo o bien las respuestas de los estudiantes (West, 1988); o que se justifique la existencia de inestabilidad de forma vaga y general, señalando que los espacios de clase manifiestan una variabilidad natural en el tiempo, de forma similar a como se comportan otros fenómenos sociales (Patrick & Mantzicopoulos, 2016).

De todas formas, algunas de las dudas se han clarificado, y ha habido aportes abordando causal e interpretativamente los fenómenos de estabilidad/inestabilidad en las medidas de calidad docente. Por ejemplo, posicionándose sobre la duda a la que hacía alusión West (1988) respecto de si la estabilidad es un atributo de los docentes o de los cursos, la investigación de Hanges et al. (1990) asegura que la estabilidad es efectivamente un atributo de los docentes. También es un avance interpretativo el hecho de que la investigación de Hanges et al. (1990) busque identificar las dimensiones del rendimiento docente que son más y menos estables, al igual como lo hace una de las investigaciones del MET Project (Bill & Melinda Gates Foundation, 2010).

También son aportes parciales lo señalado por Polikoff (2015), quien evidencia que las características de los alumnos, en específico, su género, su grupo étnico, la condición de discapacidad, el logro académico y el uso del inglés como lenguaje no nativo, no explican las diferencias de resultados año a año de las evaluaciones docentes, y que tampoco existen

diferencias de estabilidad entre docentes novatos y experimentados; y el aporte de Hativa (1996), quien señala que las mejorías en el rendimiento docente al interior de los mismos semestres se podrían identificar sólo en docentes que han realizado actividades de mejoramiento durante el tiempo en que son evaluados.

vi. La calidad en la formación docente y el uso de CEDs en el espacio local: Chile y Latinoamérica

Desde principios de este siglo, en Chile se ha observado un progresivo aumento de la preocupación por parte del Estado por la calidad de la formación de profesionales en general, y de docentes en particular, que ha tenido consecuencias directas sobre la atención que han tenido los procesos de evaluación docente en las instituciones de educación superior.

Con la creación de la Comisión Nacional de Acreditación de Pregrado (CNAP) en el año 1999, que dio paso a la Comisión Nacional de Acreditación (CNA) en el año 2006, el Estado comienza a tomar parte cada vez más activa en el aseguramiento y vigilancia de la calidad de la educación superior, tanto en universidades, institutos profesionales y centros de formación técnica, al supervisar periódicamente sus procesos de acreditación (Montoya, et al., 2014). Dentro de las medidas importantes tomadas en el marco de estas políticas, se encuentra la implementación de estrategias de evaluación docente como requisito obligatorio para todas las casas de estudios en educación superior cursando procesos de acreditación institucional (Salazar, 2008; Medel, 2013).

Para las carreras de educación en específico, otro antecedente que explica el aumento de preocupación por la situación y formación de los docentes y el consecuente auge de la evaluación docente, ocurre en el 2004, año en que los niveles de calidad en la formación docente y la preparación de los profesores fueron algunos de los aspectos más criticados del sistema educacional chileno por la OCDE (Manzi, et al., 2011a), en su informe sobre la situación de la política educacional de Chile (OCDE, 2004). Esta preocupación habría tenido respuesta por parte del Estado en el informe del Consejo Asesor Presidencial para la Calidad de la Educación del año 2006, en el que se prestó atención a la necesidad de avanzar hacia la definición de una nueva carrera profesional para los docentes escolares y a mejorar la calidad de la formación inicial de los mismos (Manzi, et al., 2011a). Ese mismo año, con la Ley de Aseguramiento de la Calidad de la Educación Superior, se decretó la obligatoriedad de la acreditación para todas las carreras de pedagogía (Ávalos, 2014). Otras medidas relevantes de la época, que se constituyen como acciones reguladoras por parte del Estado para mejorar la formación inicial docente, son el cierre

de los programas a distancia para carreras de pedagogía, el año 2005, y la aplicación de la Prueba Inicia, inaugurada en el año 2008, con el fin de monitorear el nivel de conocimiento de los docentes recién egresados de las carreras de pedagogía (Ávalos, 2014).

En años más recientes, la preocupación sobre la calidad de la docencia continúa en aumento, dadas las condiciones actuales del sistema de educación superior, en que existe una tendencia a la masificación de la matrícula universitaria en general, y una expansión en el volumen y heterogeneidad de la oferta de matrículas en programas de formación de profesores, algunos de ellos de reducida selectividad y bajos niveles de acreditación (Cox, et al., 2010; Manzi, et al., 2011a).

En efecto, la matrícula total en carreras de pedagogía de universidades privadas entre los años 2000 y 2008 tuvo un aumento de 812,6% en las carreras de educación básica, y de 940,7% en las carreras de educación media, mientras que en las universidades tradicionales tuvieron un aumento de 174,4% y 74,7%, respectivamente (Cox, et al., 2010).

En la actualidad, según datos del proyecto INDICES del Consejo Nacional de Educación (CNED, 2017a), Chile cuenta con una participación en educación superior –o cobertura bruta– en torno al 87%. Con una matrícula total de pregrado de 1.162.306 estudiantes en 2017, nuestro país supera al promedio de los países de la OCDE, con un 68% de cobertura. Dentro de este sistema de educación superior, el área de educación se sitúa en un lugar importante. El 53% de las instituciones de educación superior ofrecen programas relacionados con educación (Peirano, 2009). De la matricula total, un 11% corresponde a carreras de Educación, con 132.473 estudiantes cursando este tipo de carreras para el año 2017, el 61,7% en Universidades, 26,1% en Institutos Profesionales (IP), y 12,2% en Centros de Formación Técnica (CFT) (CNED, 2017b).

Resulta relevante señalar que el estudiantado de las carreras de educación posee una composición predominantemente femenina, a diferencia de lo que sucede con el resto de las carreras universitarias. Para el año 2009, del total de matriculados en el área de educación, dos tercios corresponden a mujeres (Peirano, 2009). Al desagregar las carreras dentro del área se observa que, para el mismo año, las mujeres superan el 95% de la matrícula en las carreras de educación diferencial y parvularia; corresponden al 78% en carreras de educación básica; y sólo se observa una relativa equivalencia entre sexos para la matrícula de las carreras de pedagogía en educación media, donde las mujeres representan un 51% (Peirano, 2009). Estas cifras son consecuentes con las que se observan históricamente en la composición por sexo del profesorado escolar chileno, también de alta composición femenina (Ministerio de Educación, 2017).

Para algunos expertos chilenos en educación, la intensa expansión de la oferta de programas de pedagogía es una respuesta a las demandas de movilidad socioeducativa, que tienden a impulsar el crecimiento del sistema de educación superior en su conjunto, y a la vez, es un fenómeno que responde a un contexto de institucional de educación de mercado escasamente regulado, con actores institucionales privados capacitados económicamente para ofertar nuevos programas y alternativas de estudios superiores, con altas facilidades de crédito (Cox, et al., 2010; Ávalos, 2014).

Lo anterior ha generado un aumento de la preocupación pública por el problema de la formación inicial docente en Chile, en cautela de la calidad de los nuevos docentes egresados de las carreras de educación.

Al respecto, la evidencia indica que el crecimiento de los programas, titulados y matrícula ha sido mayor en instituciones privadas de menor nivel de selectividad, con un gran número de carreras no acreditadas o con pocos años de acreditación y con más bajos resultados al momento del egreso de sus nuevos docentes, según mediciones de la prueba INICIA (Cox, et al., 2010). Sería por efecto de esta expansión de oferta poco selectiva que el rendimiento en las pruebas INICIA resultaría insuficiente, con primeros resultados de la aplicación de esta prueba que indican que la generalidad de los egresados de pedagogía no logra responder correctamente al menos la mitad de las preguntas (Peirano, 2009). A esto se suma que los resultados de la prueba INICIA están altamente relacionados con el puntaje de Prueba de Selección Universitaria (PSU) de ingreso a las carreras, lo que deja en evidencia que los alumnos que ingresan a las carreras de pedagogía con menores conocimientos y habilidades son también quienes demuestran un menor dominio disciplinario al egresar (Peirano, 2009). Dentro de este escenario, Ávalos (2014) también hace notar una distribución inequitativa de profesores en el sistema escolar, en la medida en que quienes supuestamente son mejores profesores por haber sido formados en instituciones más selectivas, no enseñan en establecimientos escolares más vulnerables, de más bajo rendimiento.

Respecto a la formación de docentes en pedagogía en matemáticas, al igual que para la situación de la formación docente en general, los diagnósticos señalan que la situación chilena es preocupante. Dos estudios empíricos que buscaron evaluar la calidad y pertinencia de la preparación para enseñar la asignatura de matemáticas entre los profesionales docentes de esta área (Larrondo, et al., 2007; Varas, et al., 2008), entregan resultados y conclusiones que critican la preparación de los docentes en este ámbito, señalando que en la mayoría de las carreras de formación docente en educación básica existen deficiencias en la cantidad y calidad de cursos de

Matemática y de Didáctica de la Matemática, y que el progreso en el conocimiento en estos ámbitos logrado por los estudiantes entre su ingreso y egreso posee sólo avances modestos. Por ello, estos trabajos insisten en la necesidad de dedicar mayor tiempo a la preparación en matemática y su enseñanza en las carreras dedicadas a ello.

Por lo descrito anteriormente sobre la calidad de la formación docente en Chile y sus dificultades, es que resultan más necesarios y relevantes aún —desde un punto de vista público— los ejercicios de evaluación de la calidad de los docentes chilenos. Al respecto, algunos autores señalan que es precisamente debido al constante aumento de la preocupación pública por la calidad de la docencia, y las políticas públicas que han venido en respuesta de aquello, que la evaluación docente por parte de los estudiantes se ha constituido como la forma más utilizada de evaluar la calidad del trabajo docente en el país (Salazar, 2008; Medel, 2013; Montoya, et al., 2014).

De todas formas, en la mayoría de las instituciones de educación superior chilenas la práctica de evaluación docente en base a CED supera las cuatro décadas, con algunos casos llevados de forma muy rudimentaria (Salazar, 2008). En la actualidad, las universidades chilenas cuentan con sistemas de evaluación del trabajo de sus académicos vinculados a la investigación, a la docencia y a la extensión, con influencia directa, en la mayoría de los casos, en la remuneración del profesorado (González López, et al., 2016). Para la evaluación de la docencia, las instituciones de educación superior declaran considerar, además de la evaluación docente respondida por alumnos, mediciones tales como la autoevaluación; la evaluación por pares; la evaluación por superiores (Salazar, 2008); la cantidad de cursos dictados en un período académico, si son de pre o posgrado; la elaboración de material didáctico; la atención de alumnado; la dirección de tesis y seminarios, y las supervisiones de prácticas (Montoya, et al., 2014).

Hay autores, sin embargo, que denuncian que pese a que los sistemas de evaluación docente al interior de las universidades efectivamente producen información sobre distintos indicadores a partir de la heterogeneidad de fuentes e instrumentos mencionados, este ejercicio es poco extendido y de escaza cobertura, resultando así un escenario en que los sistemas de evaluación docente de muchas instituciones se reducen en la práctica casi exclusivamente a la aplicación de cuestionarios de opinión estudiantil (Montoya, et al., 2014).

A esa crítica se le suman otras, similares a la de la reflexión internacional, que discuten y ponen en duda la trascendencia formativa que poseen estas evaluaciones, en términos de si son utilizadas o no para mejorar la calidad de la enseñanza; denunciando la mala interpretación cuantitativa que se hace de sus resultados; y criticando los efectos sumativos negativos y de control de la labor del

profesorado que implica un mal uso de la evaluación docente en algunos contextos educativos (CINDA, 2007; Salazar, 2008; Montoya, et al., 2014; González López, et al., 2016). A esta crítica, se suma aquella hecha respecto a la escasez de desarrollo investigativo sobre la evaluación docente a nivel de educación superior (Medel & Asun, 2014; Montoya, et al., 2014).

Efectivamente, la propia revisión bibliográfica evidencia que en Chile la investigación empírica en evaluación docente mediante CED ha sido escaza y de ligera especialización. Los pocos estudios que se han podido identificar son los del Centro Interuniversitario de Desarrollo (CINDA, 2007), Salazar (2008), Medel y Asun (2014), González López, et al. (2016) y Asun y Zúñiga (2017), último trabajo que si bien es empírico, corresponde a un metaanálisis de instrumentos. Considerando el volumen y evaluando su contenido, el autor de esta tesis concuerda con Medel y Asun (2014), quienes señalan que el debate respecto de la validez de los CED actualmente constituye un campo de investigación virtualmente inexistente en Chile. A lo anterior, es posible agregar que parece inconveniente la inexistencia de estudios longitudinales sobre evaluación docente en Chile, considerando que su medición es periódica y ampliamente generalizada en el país.

Se debe hacer notar, sin embargo, que la investigación en Chile sobre calidad docente existe, pero típicamente ha utilizado medidas distintas a las generadas a partir de cuestionarios de evaluación docente, y se ha enfocado mayoritariamente en el ámbito de la educación escolar. Según Santelices et al. (2017), las principales fuentes de información utilizadas en la investigación sobre calidad docente y su efecto sobre el rendimiento estudiantil han sido dos: las mediciones SIMCE, calculando con sus resultados indicadores de valor agregado ("value added scores"), y aquellas construidas mediante el Sistema de Evaluación del Desempeño Profesional Docente[3] (Manzi, et al., 2011).

Santelices et al. (2017) señala que entre los hallazgos del conjunto de investigaciones realizadas en el ámbito, se encuentra la identificación de que la calidad y efectividad docente se relaciona con el género del profesor, sus años de experiencia haciendo clases, su grado académico y especialización, en conjunto con el tipo de institución del que provienen, la disciplina que enseñan, sus creencias sobre las habilidades de sus alumnos, la proporción de contenido que declaran haber alcanzado a cubrir en clases, las características del establecimiento en donde

[3] El Sistema de Evaluación del Desempeño Profesional Docente es una evaluación obligatoria para los docentes de aula escolar que se desempeñan en establecimientos municipales en Chile. Es una medida conjunta de cuatro abordajes para medir la calidad docente: 1) autoevaluación del docente; 2) evaluaciones de superiores jerárquicos (denominados Informes de Referencia de Terceros); 3) evaluaciones de pares (evaluación de colegas); 4) Portafolio (evaluación a partir de evidencia directa del trabajo en aula).

enseñan, así como su localización. Otros hallazgos son que el rendimiento estudiantil también se encuentra asociado con el nivel de educación de los padres, el ingreso familiar, el ingreso familiar de los estudiantes pares y el grado de ruralidad de la escuela.

La propia investigación de Santelices et al. (2017), es, por lo demás, la única investigación latinoamericana publicada hasta el momento que utiliza una estrategia longitudinal para evaluar la calidad docente. Utilizando metodologías observacionales de prácticas en el aula basadas en estándares profesionales y medidas de contribución de los docentes al aprendizaje de los alumnos ("value added scores"), su investigación revela que gran parte de la efectividad docente se explica por la escuela y la municipalidad en que esta se encuentra.

Por otro lado, en latinoamérica, si bien existe una amplia y generalizada aplicación de CED, los puntajes obtenidos son poco usados en investigación educativa (Arámburo Vizcarra & Luna Serrano, 2013; García Garduño, 2000). De todas formas, su volumen es algo mayor que la investigación realizada en Chile. Ahora bien, la propia revisión del estado del arte en latinoamérica evidencia que el volumen de investigación es menor al de la producción angloparlante, con la mayoría de los estudios empíricos tendiendo a reproducir las preguntas y problemáticas de investigación que se trabajan en la investigación internacional.

Junto con el recurrente ejercicio de realizar diagnósticos de rendimiento docente en lugares o muestras específicas[4], la mayoría de los estudios empíricos trabajan objetivos relativamente básicos, como la validación psicométrica de escalas (Acevedo Álvarez & Fernández Díaz, 2004), o bien se dedican a replicar estudios ya realizados en la investigación internacional, comparando los resultados obtenidos en sus contextos locales (García Garduño, 2003; González López, et al., 2016). Sólo una parte limitada de la totalidad de trabajos poseen objetivos más especializados, como el estudio de posibles sesgos en la evaluación docente (Acevedo Álvarez & Mairena Rodríguez, 2006; Medel & Asun, 2014). Estas investigaciones, sin embargo, no ofrecen hallazgos o suponen propuestas muy distintas a lo ya visto en la investigación internacional.

Vale la pena mencionar que, al igual que en la producción internacional, existen muchas reflexiones analíticas y algunos estudios empíricos que problematizan sobre la validez consecuente de la evaluación docente, específicamente, cuestionando la real trascendencia sobre el mejoramiento de la educación que poseen este tipo de evaluaciones, y visibilizando las preocupaciones, tanto de expertos como del profesorado, acerca de los efectos que tienen sus

[4] Ver Arbesú, et al. (2006) para un recuento sobre este tipo de estudios más exhaustivo que el de la presente investigación.

resultados sobre el bienestar, condiciones laborales y la labor educativa de los docentes en el contexto latinoamericano (Arbesú & Rueda, 2003; CINDA, 2007; Salazar, 2008; García Garduño, 2008; Luna Serrano & Torquemada, 2008; Silva Montes, 2009).

vii. Síntesis de la revisión de antecedentes y avance hacia una propuesta de estudio

De la revisión sobre el estado del arte de las problemáticas centrales sobre la validez de los CEDs, se evidencia que no parece haber acuerdo sobre su validez de contenido, de constructo y discriminante, a la vez que existen distintos diagnósticos sobre sus niveles de estabilidad longitudinal, e indeterminaciones sobre el significado y efectos que posee este aspecto. Con todo, se puede asegurar que existen muchas interrogantes y nudos por resolver en este campo. No se esperaría, de todas formas, un escenario distinto.

Respecto a la validez de contenido, parece apropiado que aún se discuta el sentido sustantivo del concepto de calidad docente. Ello evidencia algo que es innato en la medición de constructos complejos en ciencias sociales, que es el hecho de que sólo es posible abordar de forma representacional los fenómenos. El desacuerdo no es necesariamente un indicador de invalidez, sino que lo es de la intensa reflexión y vasta atención que se les ha dado a los cuestionarios de evaluación docente como medida de calidad de la enseñanza.

A su vez, las distintas posiciones sobre la validez de constructo de los cuestionarios de evaluación docente son sólo efecto de lo anterior, es decir, de la diversidad de propuestas que se han hecho para delimitar el concepto de calidad docente. Tal como mencionan otros autores, la presente tesis postula que las discrepancias sobre la validez de constructo se deben simplemente a divergencias entre los distintos estudios sobre las dimensiones que se consideran apropiadas para comprender el fenómeno de la calidad docente. Es una cuestión de medición, y no propia del fenómeno. Sí sería preocupante que no existieran regularidades y acuerdos relativamente generales sobre qué aspectos e indicadores miden la calidad docente, que no es el caso. Como ya se ha mencionado anteriormente, el estudio de Asun y Zúñiga (2017) apunta a que sí hay elementos que son típicamente recurrentes en las distintas fórmulas de contenido para abordar la calidad docente.

En el presente estudio, se busca validar un instrumento de medición nunca antes aplicado, identificando si es posible validar una estructura dimensional de los aspectos que pretenden medir la calidad de la docencia. Una vez hecha esta tarea, se dará énfasis en diagnosticar las trayectorias

de rendimiento docente de una muestra de profesores universitarios de carreras de educación básica. Cabe clarificar que, para la presente investigación, el constructo a medir es el rendimiento docente, no la efectividad docente. En la literatura se entienden de distinta forma, siendo la primera la medida de la calidad de la enseñanza, y la segunda la medida del rendimiento de los alumnos que se considera atribuible a la contribución de los docentes. Como se ha mencionado anteriormente, este segundo tipo de medición se realiza utilizando indicadores denominados "value-added scores", y se obtienen a partir de las calificaciones de los estudiantes, en distintos tipos de test y pruebas estandarizadas (Boyd, et al., 2009).

Sobre la temática de la validez discriminante de los cuestionarios de evaluación docente revisada en los antecedentes, se puede señalar que, al igual que para el tipo de validez anterior, tampoco se pretende un acuerdo entre las distintas investigaciones. Un diagnóstico absoluto sobre la validez discriminante de los distintos tipos de CED es inalcanzable, no sólo debido a las distintas formas de medir el fenómeno, sino también producto de la variabilidad de contextos educativos. El efecto que los factores extra docencia poseen sobre las evaluaciones docentes radica en el hecho de que las interacciones de enseñanza se producen en determinados contextos de aula, entre ellos, de distinta composición estudiantil y carácter institucional, con diverso tipo de prácticas pedagógicas, de variado espectro sociocultural y socioeconómico, etc. Muchos investigadores consideran que todos estos elementos contextuales conforman distintos escenarios de enseñanza, o "idiosincrasias" (González López, et al., 2016), que terminan por mediar la evaluación que los estudiantes hacen del desempeño de sus docentes (Vásquez Rizo & Gabalán Coello, 2006; Arámburo Vizcarra & Luna Serrano, 2013; Medel & Asun, 2014; Good & Lavigne, 2015; Boring, et al., 2016).

Con esto en consideración, la postura que deben adquirir los estudios sobre validez discriminante de las medidas de evaluación docente es de indagación sobre los efectos de sesgos en contextos específicos, no buscando la generalización de resultados, sino la precisión de un diagnóstico de caso. Este tipo de análisis es precisamente el que se lleva a cabo en este estudio. Con los datos que se tienen, se buscará identificar el posible efecto de sesgo de cinco aspectos extra docentes, tanto a nivel transversal como longitudinal. Estos elementos son: sexo del docente, la cantidad de alumnos inscritos en los cursos, la cantidad de horas de clase realizadas en el semestre, el nivel de pérdida de clases y el formato programático de clases. Los dos primeros elementos han sido estudiados con anterioridad en varias investigaciones, como ya se ha visto. Los tres restantes, constituyen una innovación.

La cantidad de horas realizadas, al igual que la pérdida de clases, se postulan como dos aspectos relevantes a explorar como posibles sesgos de evaluación docente, debido a que los procesos pedagógicos universitarios en Chile han estado en los últimos años muy sujetos a la contingencia política. De forma progresiva en los últimos años, ha habido un aumento en el número de eventos de protesta estudiantil y el número de asistentes a estos eventos (Somma, 2017). La protesta estudiantil típicamente va acompañada de paros o tomas de establecimientos educacionales, incluyendo los universitarios, procesos que tienen como una de sus consecuencias la no realización de las clases programadas. La intención es, con esto, identificar si es que la pérdida de clases afecta la forma en que los docentes son evaluados. De la misma forma, y considerando que una cantidad de carga académica bien delimitada es vista como una medida de la calidad de la educación por el CRUCh (Pey, et al., 2012), resulta relevante identificar si es que el rendimiento de la evaluación docente se relaciona con la cantidad de horas de clase realizadas en los cursos analizados.

Respecto al análisis de la inestabilidad de la evaluación docente, pese a que la mayoría de las investigaciones han identificado una relativa estabilidad de la evaluación docente al interior de periodos cortos de tiempo, se ha considerado, al igual como lo hacen Patrick y Mantzicopoulos (2016), que es necesario mirar con mayor detalle las variaciones clase a clase del rendimiento docente. Para el presente caso, se realiza este ejercicio con medidas de rendimiento docente en base a CEDs aplicados clase a clase, durante todo un semestre. Aumentando la periodicidad de la medición se logra, a su vez, aumentar nuestra capacidad de identificar la medida en la que los niveles de estabilidad/inestabilidad se relacionan con el rendimiento docente, como también con factores extra docencia. Medel y Asun (2014) aseguran que nuevos estudios de carácter longitudinal sobre la evaluación docente se constituyen como una necesidad para profundizar en el campo de investigación en evaluación docente chileno. Por ello, se puede señalar que la investigación propuesta constituye un análisis novedoso, relevante y necesario.

Con todo, la presente investigación tiene como objeto de análisis una serie de cuestionarios de evaluación docente aplicados en 23 cursos de 8 universidades chilenas, a lo largo de todas las clases de un mismo semestre lectivo, realizado entre agosto y diciembre de 2014. Se realiza un análisis del comportamiento de las trayectorias de evaluación, a la luz del paso del tiempo y en relación a una serie de aspectos extra docencia que, se presume, podrían influenciar las distintas formas en que la evaluación docente se comporta a lo largo de un semestre.

Todas las evaluaciones docentes analizadas fueron aplicadas en cursos de carreras de formación inicial de pedagogía, en particular, en cursos de enseñanza de las matemáticas. En consideración de lo ya mencionado anteriormente en los antecedentes, ello implica que los resultados de este estudio no son totalmente extrapolables a los resultados de evaluaciones docentes realizados en base a clases que se pudieran realizar en otra disciplina universitaria, ya sea de pedagogía o no, debido a que la disciplina tiende a ser un factor de sesgo sobre las evaluaciones docentes, y más aún, las evaluaciones en cursos de matemáticas tienden a ser menores en comparación con cursos de contenidos distintos, como por ejemplo, humanidades o artes.

Para cerrar la problematización, cabe señalar que la investigación propuesta es doblemente relevante en el ámbito de la evaluación docente, ya que se tiene por objeto de análisis no sólo el rendimiento docente en sí, sino la calidad de cómo están siendo formados nuevos profesores. Considerando que, en la mayor parte de los países del mundo, las políticas educativas están otorgando una importancia creciente a los docentes, no sólo en relación a la influencia de su desempeño sobre los aprendizajes de sus alumnos, sino también respecto a la calidad de su propia formación inicial (Manzi, et al., 2011a), esta investigación cobra más relevancia aún, ya que se relaciona con ambos aspectos.

IV. Pregunta y objetivos de investigación

Pregunta de investigación:

¿Cómo se comportan las trayectorias de evaluación docente semestrales de los 23 cursos de formación inicial en pedagogía en matemáticas de la muestra del proyecto FONDEF IT13I10005?

Objetivo general:

Caracterizar las trayectorias de evaluación docente semestrales de los 23 cursos de formación inicial en pedagogía en matemáticas de la muestra del proyecto FONDEF IT13I10005.

Objetivos específicos:

1. Clasificar las trayectorias de evaluación docente según niveles de mejoría y empeoramiento para los 23 cursos de formación inicial en pedagogía en matemáticas de la muestra del proyecto FONDEF IT13I10005.
2. Identificar los niveles de inestabilidad de las trayectorias de evaluación docente de los 23 cursos de formación inicial en pedagogía en matemáticas de la muestra del proyecto FONDEF IT13I10005.
3. Medir el efecto que poseen el sexo del docente, la cantidad de alumnos inscritos, la cantidad de horas de clase realizadas, la pérdida de clases y el formato programático de clases sobre el comportamiento de las trayectorias de evaluación docente de los 23 cursos de formación inicial en pedagogía en matemáticas de la muestra del proyecto FONDEF IT13I10005.
4. Dar cuenta del rendimiento de los ítems y los niveles de la validez y fiabilidad de la escala de evaluación docente utilizada en el proyecto FONDEF IT13I10005.
5. Formular recomendaciones para mejorar el rendimiento de los profesores en su evaluación docente a partir de los resultados de las mediciones semestrales de los 23 cursos de formación inicial en pedagogía en matemáticas de la muestra del proyecto FONDEF IT13I10005.

V. Marco Metodológico

i. Tipo y características del estudio

Esta investigación es de tipo cuantitativo, longitudinal y no experimental. Es de carácter descriptivo, ya que busca identificar y caracterizar trayectorias longitudinales, y a su vez de carácter relacional, ya que busca observar el comportamiento de una serie de variables dependientes en relación a la influencia que sobre ellas tienen otro conjunto de variables independientes.

La fuente de información del estudio es primaria. Se elaboró a partir de un cuestionario de evaluación docente[5], previamente diseñado y distribuido de forma auto-aplicada por la institución a cargo del proyecto FONDEF en el que se enmarca nuestra investigación (Laboratorio de Educación – Centro de Modelamiento Matemático de la Universidad de Chile). Esta misma institución estuvo a cargo del trabajo en terreno y la digitación de los datos.

La aplicación del instrumento fue realizada entre los meses de agosto y diciembre del año 2014. Se aplicó a una muestra de 481 estudiantes de 23 cursos universitarios, dictados por 18 profesores, en 8 universidades chilenas que imparten la carrera de pedagogía básica[6]. Dichas universidades fueron incluidas para participar en el proyecto de acuerdo con su interés en el estudio, y según criterios fijados por la dirección del proyecto relacionados con el componente de intervención de la investigación[7]. Se trata, por consiguiente, de una muestra preseleccionada y no probabilística. La serie de diagnósticos surgidos a partir de este estudio son, por consiguiente, relativos a este grupo docente específico (estudio de caso), no generalizables directamente al universo de docentes universitarios chilenos.

La unidad de análisis del estudio son las trayectorias longitudinales semestrales de la evaluación docente de cada curso. La unidad de información son los alumnos asistentes a estos cursos, a quienes se les aplicó el instrumento de evaluación docente al final de cada clase del semestre.

[5] Ver anexo i. Encuesta clase a clase.

[6] Pontificia Universidad Católica de Valparaíso (PUCV), Universidad Alberto Hurtado (UAH), Universidad Católica de la Santísima Concepción (UCSC), Universidad Católica Silva Henríquez (UCSH), Universidad de las Américas (UDLA), Universidad Diego Portales (UDP), Universidad de Playa Ancha (UPLA), Universidad Santo Tomás (UST).

[7] Cantidad de años de acreditación; satisfacción de los criterios de elegibilidad para la obtención de la beca vocación de profesor; número de cursos del área de matemática (incluyendo didáctica específica) mayor o igual a 4; participación en la prueba INICIA de los estudiantes de la carrera.

ii. Plan de validación del instrumento de medición

El instrumento utilizado como cuestionario de evaluación docente posee diez preguntas. Por ello, resultaría un producto poco parsimonioso si se analizaran diez series temporales por cada uno de los veintitrés cursos constitutivos de la muestra. Asimismo, sería un ejercicio redundante, ya que es posible analizar una menor cantidad de series construyendo trayectorias según cada una de las dimensiones latentes que representan los ítems. Una lectura inductiva del test[8] permite identificar que busca medir cuatro dimensiones de la evaluación de la docencia de pedagogía en matemáticas:

1) (P1) Capacidad de la clase para generar aprendizaje en conocimientos en matemáticas (1 ítem, n°1);
2) (F1) Capacidad de la clase para desarrollar aprendizaje en habilidades pedagógicas (4 ítems, n°2 al n°5);
3) (F2) Capacidad de la clase para motivar la participación (4 ítems, n°7 al n°10);
4) (P6) Calidad general de la clase (1 ítem, n°6).

Cuatro indicadores longitudinales para caracterizar la evaluación docente semestral de cada curso constituyen un volumen de información viable para analizar, y aprehensible a la hora de dar cuenta de los resultados y diagnosticar el estado de la evaluación docente de la muestra.

Para probar en qué medida los ítems que conforman las dimensiones F1 y F2 efectivamente representan los constructos latentes que se han propuesto, se utilizará la técnica de análisis factorial confirmatorio (AFC) para datos ordinales. Como ya se ha evidenciado mediante la revisión de antecedentes, la herramienta considerada apropiada para estimar la validez de constructo de los CEDs es comúnmente, el análisis factorial. Este procedimiento permitirá una validación de constructo de las dos dimensiones propuestas ya mencionadas. Con ello, habrá claridad sobre en qué medida el instrumento mide realmente aquello que pretende medir (Latiesa, 2000).

Para estimar los parámetros del modelo factorial, se utilizará el estimador ULS, que es el apropiado para procesar ítems de nivel de medida ordinal, ya que trabaja con correlaciones policóricas (Freiberg Hoffmann, et al., 2013). Como el carácter de los datos es longitudinal, se pretende realizar tres análisis factoriales confirmatorios. Un primer análisis para los datos de la primera clase del semestre, un segundo para los datos de la última clase, y un tercero para los

[8] Ver anexo ii. Operacionalización inductiva de la dimensionalidad del test.

datos de todas las clases en conjunto. Adicionalmente, se consideró la existencia de correlación entre ambas dimensiones, al tratarse de aspectos teórica y estadísticamente interrelacionados.

También se buscará identificar los niveles de fiabilidad de la escala, entendiendo por esto el grado de constancia de las observaciones que produce el instrumento de medida (Latiesa, 2000). Los niveles de alfa de Cronbach ordinal (Domínguez, 2012) permitirán evaluar la fiabilidad de las mediciones identificando el nivel de acuerdo relativo entre los estudiantes sobre la evaluación que hacen de sus docentes para cada dimensión. Los niveles de estabilidad longitudinal de las trayectorias de evaluación docente también podrán ser interpretados, alternativamente, como una medida de fiabilidad, en tanto reflejan el nivel de consistencia en el tiempo que logran las mediciones. Serán utilizados para representar la fiabilidad sólo de forma alternativa debido a que la inestabilidad no es solo una medida de los errores de medición, sino también de los cambios reales ocurridos en el tiempo de los fenómenos medidos (Latiesa, 2000).

Cabe señalar que la decisión sobre cómo agrupar los ítems en distintos aspectos de la evaluación docente implica que no se podrá calcular los niveles de validez de constructo y fiabilidad que sí se pueden obtener para las dimensiones constituidas de más de una variable. Esta es una dificultad ineludible, por lo que no se niega la posibilidad de que las mediciones sobre la capacidad de la clase para generar aprendizaje en conocimientos en matemáticas y la calidad general de la clase contengan errores de medición cuya intensidad no estará cuantificada.

iii. Definiciones operacionales sobre las variables

Para la caracterización de cada uno de los cuatro indicadores de evaluación docente, se pretende construir un conjunto de variables con el fin de caracterizar las trayectorias que conforman las series semestrales. Dichos indicadores son:

a) Los **puntajes promedio de la primera y la última clase**, como indicadores de la evaluación que reciben el conjunto de docentes al comienzo y final de semestre.
b) La **media** de puntuaciones de cada serie semestral, como indicador de logro general en la evaluación semestral recibida por cada docente.
c) El **porcentaje de cambio semestral**. Este indicador ha sido generado a partir de la pendiente de la recta lineal de regresión de cada serie, como indicador de mejoramiento o

empeoramiento de la trayectoria de evaluación que recibe cada docente a lo largo del semestre[9].

d) La **media de las diferencias absolutas entre los puntos observados de la serie y los puntos correspondientes en la recta de regresión lineal**, como indicador de inestabilidad de la evaluación que recibe cada docente en el transcurso de su semestre. Este valor indica qué tanto el conjunto de las puntuaciones obtenidas durante el semestre por los docentes se aleja de su propia recta de regresión lineal. Se considera, por esto, como el indicador oportuno para identificar la inestabilidad de cada serie, ya que refleja qué tanto las trayectorias de evaluación docente se desvían de una trayectoria que sea perfectamente estable (lineal).

Por su parte, las variables independientes cuyo efecto sobre las trayectorias de evaluación docentes se buscará medir, son:

a) El sexo del docente.
b) La cantidad de horas de clases realizadas en el semestre.
c) El porcentaje de clases perdidas proporción a las clases realizadas durante el semestre.
d) La cantidad de alumnos inscritos por ramo, como indicador aproximado de la cantidad de estudiantes asistentes periódicamente a los cursos analizados.
e) El formato programático de las sesiones de clases, es decir, si se trata de ramos con una o dos clases por semana.

iv. Técnicas de análisis

Para el análisis de los datos, junto con el uso de estadísticos descriptivos cuando corresponda (media, moda, mediana, desviación estándar, gráfico de distribución), se considera el uso de las siguientes técnicas de análisis estadístico:

- Estadístico de correlación de Pearson para identificar y estimar intensidad de asociación entre dos variables cuando ambas son cuantitativas.
- Prueba t de Student para identificar diferencias de medias de las variables dependientes cuantitativas en los grupos conformados a partir de las variables independientes de carácter nominal. Esta es la prueba estadística clásica utilizada para contrastar la hipótesis nula de igualdad de medias entre dos muestras o grupos. Ahora bien, para su aplicación, se

[9] Considerando que la cantidad promedio de clases realizadas por el conjunto de docentes estudiados es de 21, se consideró que un punto de pendiente correspondería a un cambio de 20% de logro. Esto, ya que una mejoría de 100% en un periodo de 21 clases corresponde matemáticamente a una pendiente de 5 puntos.

recomienda un n muestral no inferior a 30 casos (Rubio Hurtado & Berlanga Silvente, 2012). Debido a que en nuestro estudio esta condición tiene dificultades de cumplimiento (n=23), se ha decidido utilizar de forma complementaria la prueba de U de Mann-Whitney, prueba no paramétrica de diferencia de medias que opera comparando los rangos de distribución de la variable. Al ser no paramétrica, no posee supuestos de cantidad de casos, sin embargo, es más exigente al rechazar la hipótesis nula de igualdad y por tanto tiene menos posibilidades de acertar cuando no es rechazada (más posibilidades de cometer un error tipo II).

- Análisis Factorial Confirmatorio (AFC) para ítems ordinales aplicando el estimador de parámetros ULS. Como ya se ha mencionado antes, se plantea utilizar esta técnica multivariada para estimar los niveles de validez del instrumento utilizado para medir el rendimiento docente.
- Regresión lineal para representar cada trayectoria de evaluación docente semestral, estimando la pendiente de cada recta para construir los indicadores de mejoría y empeoramiento, y calculando la media de las diferencias absolutas entre los puntos observados de la serie y los puntos correspondientes en la recta de regresión lineal, como indicador de inestabilidad de la evaluación que recibe cada docente en el transcurso de su semestre.

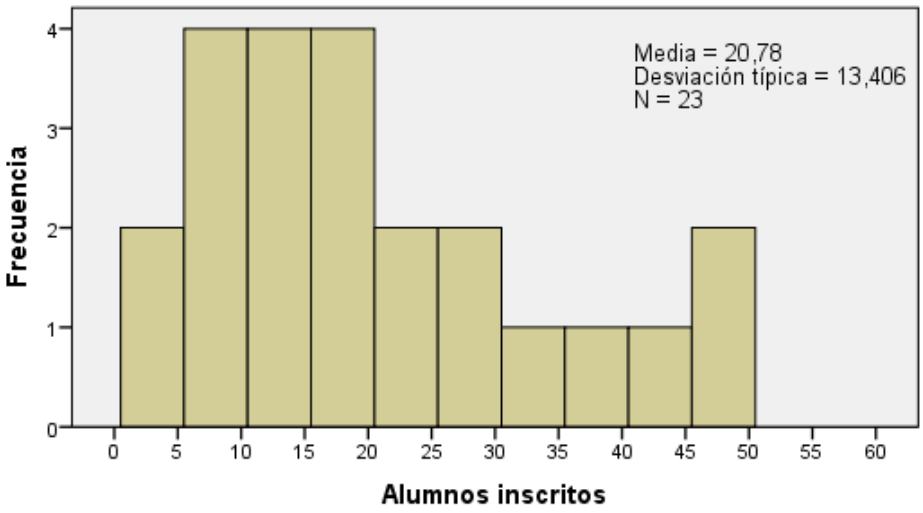
VI. Análisis estadístico y resultados del estudio

i. Caracterización de la muestra

Como ya se ha señalado anteriormente, se analizaron las respuestas a un cuestionario de evaluación docente (CED) de 481 estudiantes distribuidos en 23 cursos universitarios de pedagogía en matemática en educación básica. El instrumento fue aplicado al final de todas las clases de un semestre lectivo, realizado entre los meses de agosto y diciembre de 2014, resultando un total de 6580 encuestas aplicadas. En suma, 18 docentes fueron evaluados, 9 mujeres y 9 hombres, en 8 Universidades distintas.

Se observa a partir del histograma a continuación, que la cantidad promedio de alumnos inscritos por curso corresponde a aproximadamente 21 personas.

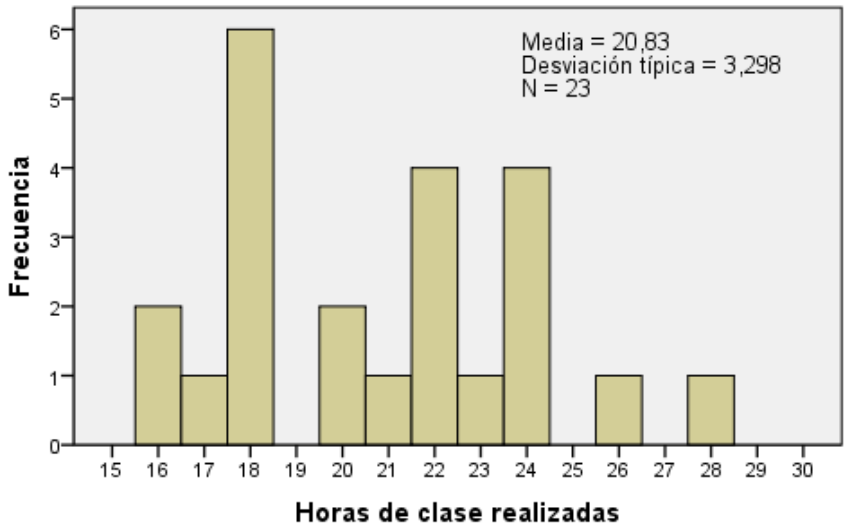
Histograma 1: Cantidad de alumnos inscritos por asignatura



Con una desviación estándar de 13,4 puntos, la dispersión de este indicador es intensa. Ello también se observa en el rango de alumnos inscritos en el grupo de cursos estudiados, que es bastante amplio, con una variación de entre 3 a 49 inscritos. La mayoría de los cursos posee baja cantidad de alumnos: 6 cursos tienen entre 3 a 10 inscritos, 8 cursos entre 11 a 20 inscritos, 4 cursos entre 21 a 30 inscritos, y sólo 5 cursos superan los 30 alumnos.

En relación a la cantidad de horas de clases realizadas, se observa a partir del histograma siguiente que la media corresponde a aproximadamente 21 horas de clase.

Histograma 2: Horas de clase realizadas



La moda, sin embargo, corresponde a 18 horas, con 6 cursos correspondientes a esa cantidad. Adicionalmente, es posible señalar que la menor cantidad de horas de clases realizadas en los cursos evaluados fueron 16, y la mayor cantidad 28. La alta variabilidad de este indicador llama la atención, considerando que se trata de cursos orientados a un currículo similar. Considerando que una cantidad de carga académica bien delimitada es vista como una medida de la calidad de la educación por el CRUCh (Pey, et al., 2012), resulta relevante identificar más adelante si es que el rendimiento de la evaluación docente y otros indicadores a observar se relacionan con la cantidad de horas de clase realizadas en los cursos analizados.

En relación a las horas de clases perdidas, es decir, aquellas clases que, pese a estar incorporadas dentro de la planificación, no se realizaron, es posible señalar que, de los 23 cursos analizados, 11 cursos realizaron la cantidad de clases de acuerdo a lo programado, mientras que 12 perdieron al menos una hora de clases en el semestre. En este segundo grupo, la media de pérdida de clases es de un 18,5% de total de las horas de clase del semestre, con una desviación estándar de 11,5. El curso que más pérdida de clases posee, no realizó el 37,5% del total de sus clases. Al igual que para el caso anterior, se buscará más adelante identificar si la pérdida de clases se encuentra asociada con las evaluaciones que los docentes logran.

Adicionalmente, resulta relevante señalar que, entre los cursos, existen distintos formatos programáticos: uno de dos clases por semana, y otro formato de solo una clase por semana de dos bloques horarios seguidos. La cantidad de cursos con una clase por semana corresponden a 15, y los 8 restantes poseen un formato de dos clases por semana.

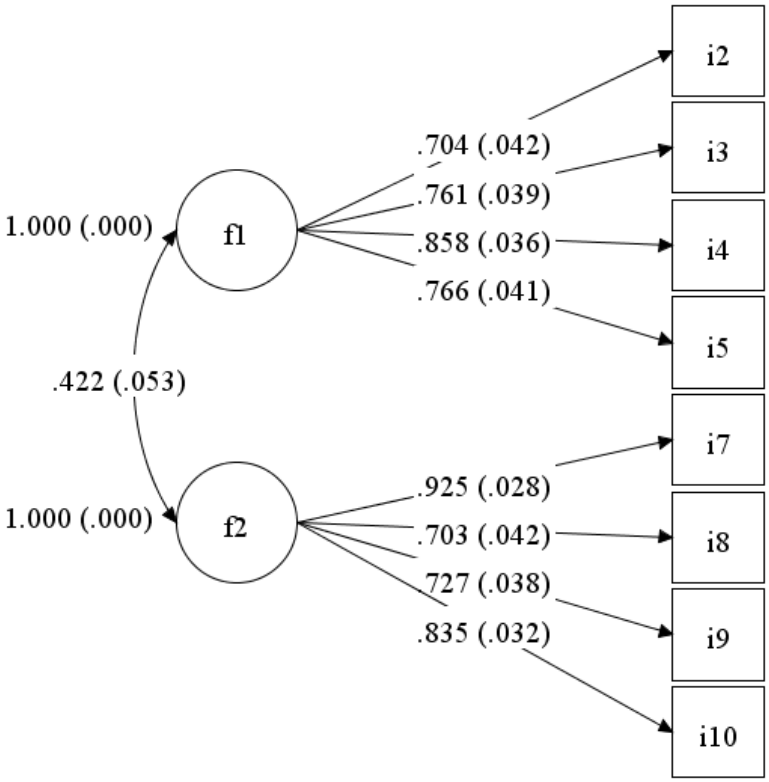
En síntesis, se tiene una muestra bastante heterogénea en relación a todos los aspectos revisados: alta variabilidad de alumnos inscritos en los cursos, alta variabilidad en horas de clases efectivamente realizadas a lo largo del semestre, distintos niveles de intensidad de pérdida de clases programadas, y dos tipos de formato de clases según cantidad de sesiones por semana.

Esta heterogeneidad resulta interesante en dos aspectos: en primer lugar, porque se trata de cursos orientados a un currículo similar de enseñanza, por lo que se esperaría relativa similitud entre ellos y, en segundo lugar, porque la existencia de heterogeneidad otorga la posibilidad de observar en qué medida estos aspectos se relacionan con las trayectorias de evaluación docente, y en específico, con qué dimensiones de ella.

ii. Validación del instrumento de medición (CED) y construcción de índices

A continuación, se presentan tres diagramas de análisis factoriales confirmatorios de ítems ordinales, correspondientes a la primera clase, la última clase, y a todas las clases en conjunto. Cabe recordar, como se planteó en el marco metodológico, que la necesidad de llevar estos análisis a cabo se encuentra en evaluar la capacidad del test para evaluar las dimensiones de evaluación docente que se ha propuesto medir.

Diagrama 1: Modelo de Análisis Factorial para Primera Clase

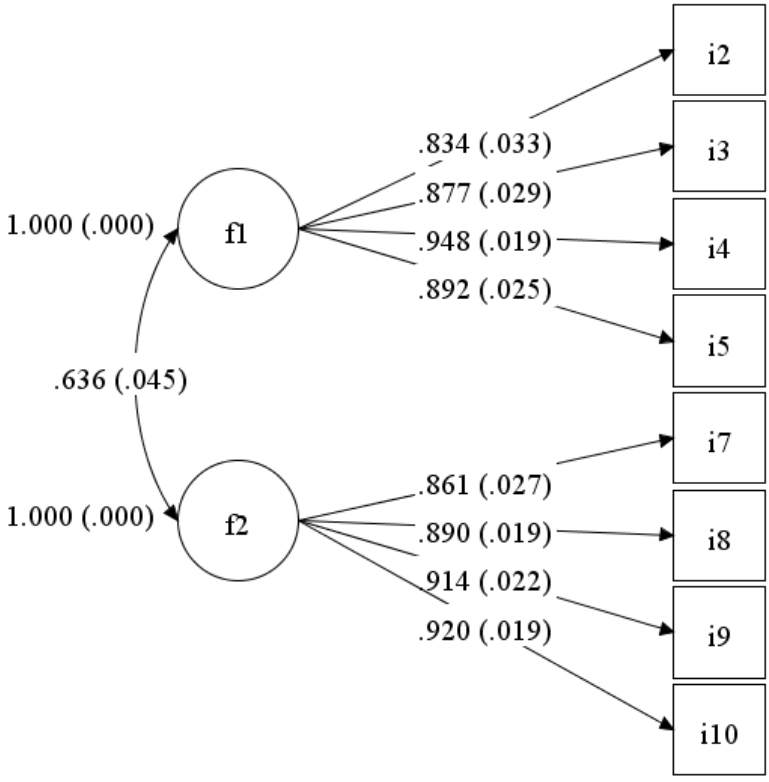


RMSEA	0,037
CFI	0,992
TLI	0,988

α ordinal F1	0,856
α ordinal F2	0,877

N válido = 349 casos

Diagrama 2: Modelo de Análisis Factorial para Última Clase

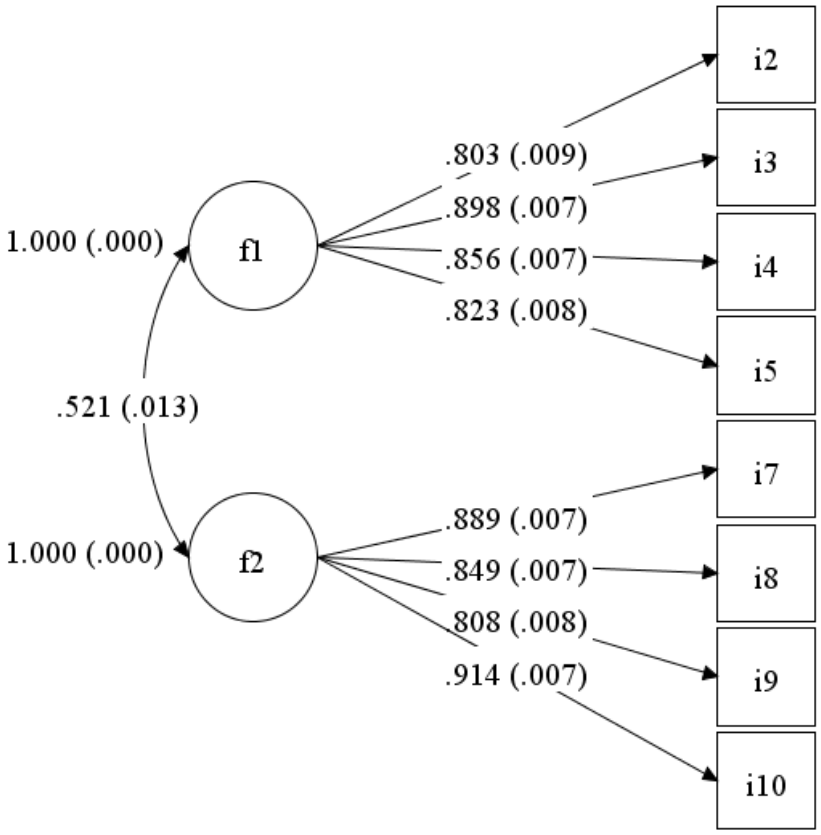


RMSEA	0,053
CFI	0,994
TLI	0,991

α ordinal F1	0,937
α ordinal F2	0,942

N válido = 269 casos

Diagrama 3: Modelo de Análisis Factorial para todas las clases



RMSEA	0,072
CFI	0,980
TLI	0,971

α ordinal F1	0,909
α ordinal F2	0,923

N válido = 4530 casos

A partir de los diagramas y estadísticos de ajuste anteriores se confirma que, en los tres momentos medidos, las dimensiones F1 y F2 corresponden a constructos latentes que miden un fenómeno común. Todos los estadísticos de ajuste arrojan buenos resultados para la validación de la escala: indicadores RMSEA[10] son menores a 0,07, y CFI[11] y TLI[12] superan el valor 0,97. A su vez, se obtienen lambdas mayores a 0,7, lo que corresponde a cargas factoriales de muy buena calidad.

Junto con ello, los indicadores de alfa de Cronbach ordinal (Domínguez, 2012) superan el valor 0,88 en ambas dimensiones de los tres análisis factoriales confirmatorios, lo que indica buena fiabilidad de la escala para ambas dimensiones.

Así, teniendo evidencia de que ambos conjuntos de ítems miden dos dimensiones latentes de la evaluación docente, tanto de forma transversal como longitudinal, los análisis posteriores de estos

[10] "Root Mean Square Error of Approximation". Valores de RMSEA ≤ 0,05 indican buen ajuste; valores ≤ 0,08 indican ajuste aceptable (Kline, 2011, pp. 205-206) (Albright & Park, 2009, p. 7).
[11] "Bentler Comparative Fit Index". Valores de CFI ≥ 0,95 indican ajuste "aceptable" (Kline, 2011, p. 208) (Albright & Park, 2009, p. 7).
[12] "Tucker-Lewis Index". Valores cercanos a 0,95 indican buen ajuste (Albright & Park, 2009, p. 7).

ítems se hicieron utilizando índices calculados en base a la sumatoria simple de sus puntuaciones según dimensión. Junto con ello, todas las puntuaciones, de los cuatro aspectos antes mencionados, fueron estandarizadas de 0 a 100, para una mayor capacidad de interpretación en una escala de logro[13].

En síntesis, con la validación parcial de la escala y la modificación de la métrica de medición hacia una escala de logro, se poseen cuatro indicadores longitudinales en métrica de porcentaje (de 0 a 100), para evaluar el desempeño docente de la muestra en los cursos analizados. Dichos análisis se presentan en el apartado a continuación.

iii. Análisis de las trayectorias longitudinales en evaluación docente

Como ya se ha mencionado, la medición del desempeño docente percibido por los estudiantes se realizó en base a una serie de mediciones a lo largo de un semestre. Ello resulta en un conjunto de datos con un formato longitudinal. Por ello, es que en el presente estudio se ha trabajado con distintos indicadores para aproximarse a una descripción de las trayectorias semestrales de evaluación docente. A continuación, dichos indicadores se encuentran desarrollados uno a uno.

- ***Puntajes de inicio, final, media semestral y desviación estándar del conjunto de trayectorias de evaluación docente.***

En primer lugar, se analiza la información del conjunto de los cursos, en sus puntuaciones de la primera y última clase, junto con la media de su trayectoria semestral (y la desviación estándar entre ellos), en los cuatro indicadores de evaluación docente.

Tabla 1: Puntajes medios de inicio, final, y media semestral con desviación estándar por dimensión de evaluación docente (n=23)

	Primera clase	Última clase	Media semestral	Desv. Est. de Media Semestral
Capacidad de la clase para generar aprendizaje en conocimientos en matemáticas (P1)	80,4	78,9	80,5	10,5
Capacidad de la clase para desarrollar aprendizaje en habilidades pedagógicas (F1)	68,3	72	70,6	12,2
Capacidad de la clase para motivar la participación (F2)	80,2	80,1	80,8	7,9
Calidad general de la clase (P6)	80,7	82,5	82,3	8,7

[13] La fórmula utilizada para ello fue: [(puntuación bruta – puntuación mínima posible del rango) / (puntuación máxima posible del rango - puntuación mínima posible del rango)] x 100.

Con los datos anteriores es posible señalar que la media grupal de puntuaciones en los cuatro indicadores de evaluación docente analizados es bastante alta, ya que todos superan los 70 puntos de logro en una escala de 100. Los puntajes de la primera y última clase son bastante similares entre sí, y a la vez similares a la media, lo que anuncia de forma preliminar la existencia de una relativa estabilidad del logro a lo largo del semestre en las evaluaciones que hacen los estudiantes de sus clases y de sus docentes. Lo anterior, al menos, considerando los cursos y docentes revisados como un conjunto.

Si bien todos los indicadores son altos, el indicador con menor rendimiento entre los docentes es aquel que describe qué tanto perciben los alumnos que las clases facilitan el desarrollo de habilidades pedagógicas (F1). A su vez, este mismo indicador posee los mayores niveles de heterogeneidad entre los profesores analizados.

El indicador con mejor rendimiento es la evaluación general de la clase (P6), que, en conjunto con la evaluación de la capacidad de la clase para motivar la participación en los alumnos (F2), se sitúan como los indicadores que presentan los mayores niveles de homogeneidad entre el conjunto de cursos analizados.

- ***Análisis de mejoría y empeoramiento en las trayectorias de evaluación docente.***

Como ya se ha mencionado anteriormente, el análisis de la mejoría y el empeoramiento de las trayectorias de evaluación docente se ha representado a partir de la pendiente de la recta lineal de regresión de la serie de datos. Mediante un cálculo aritmético[14], la pendiente ha sido transformada a un porcentaje de cambio semestral, que representa la tendencia hacia la mejoría o el empeoramiento, dependiendo del caso.

La siguiente tabla contiene información sobre los porcentajes de mejoría y empeoramiento de las trayectorias semestrales del conjunto de los cursos estudiados. Adicionalmente a la información específica para cada indicador de evaluación docente, se entrega un quinto dato, correspondiente al promedio de los cuatro indicadores (representada como "Σ/4").

[14] Considerando que la cantidad promedio de clases realizadas por el conjunto de docentes estudiados es de 21, se estableció que un punto de pendiente correspondería a un cambio de 20% de logro. Esto, ya que una mejoría de 100% en un periodo de 21 clases corresponde matemáticamente a una pendiente de 5 puntos.

Tabla 2: Promedios y número de casos de mejoría y empeoramiento de las trayectorias de evaluación docente (n=23)

	P1	F1	F2	P6	Σ/4
Promedio de cambio neto	0,4%	2,5%	2,1%	2,0%	1,8%
Promedio para casos de mejoría	9,4%	11,5%	8,7%	6,1%	7,2%
Promedio para casos de empeoramiento	-9,4%	-7,4%	-6,5%	-12,7%	-6,7%
Cantidad de casos de mejoría	12	12	13	18	14
Cantidad de casos de empeoramiento	11	11	10	5	9

De modo general, se observa a partir de la tabla anterior que en el conjunto de las asignaturas no existen cambios demasiado intensos durante el transcurso del semestre. Para el indicador conjunto de los cuatro aspectos de evaluación docente (Σ/4), el promedio de mejoría para los cursos que aumentan su rendimiento es de 7,2%, y el promedio de empeoramiento para los cursos que lo disminuyen es de 6,7%.

La misma tendencia se presenta al observar los cambios de cada indicador de evaluación docente por separado. En ello, se evidencia que los promedios de mejoría son relativamente leves, cercanos al 10%. A su vez, los promedios de empeoramiento bordean la misma cifra en sentido opuesto.

Se evidencia también que existe una relación relativamente equilibrada entre la cantidad de cursos que mejoran y los que empeoran, salvo para el caso del indicador sobre la calidad general de la clase (P6), donde se tiene que 18 de 23 cursos mejoran, y sólo 5 empeoran. El promedio de aquellos que empeoran, sin embargo, es el más intenso de los cuatro indicadores, con un promedio de empeoramiento de 12,7%. Ocurre lo inverso con aquellos que mejoran: poseen la menor intensidad de mejoría, con un 6,1% grupal en el indicador sobre la calidad general de la clase.

Esta baja magnitud en los cambios semestrales se confirma al observarse las mejorías y empeoramientos de forma individual por curso[15]. Al hacerlo, se evidencia que pocos casos superan un dígito de porcentaje de cambio en el promedio de los cuatro indicadores de evaluación docente (Σ/4).

[15] Ver Anexo, iii. Tablas: Tabla Anexo 1.

Pese a la baja intensidad de cambio, es posible clasificar a los cursos evaluados según su intensidad de cambio general en los cuatro indicadores (Σ/4)[16], como se observa en la siguiente tabla de agrupación:

Tabla 3: Cantidad de casos según tipo de trayectoria de indicador conjunto de evaluación docente (Σ/4) (n=23)

Tipo de trayectoria[17]	Cantidad de casos
Mejoría media	2
Mejoría leve	5
Mejoría mínima	7
Empeoramiento mínimo	6
Empeoramiento leve	1
Empeoramiento medio	2

La tabla anterior refuerza lo antes mencionado. En ella se observa que la distribución del indicador de cambio tiende hacia una distribución simétrica y leptocúrtica, de valores muy concentrados hacia la media, ubicándose la mayoría en la categoría de cambios "mínimos" y "leves", y no observándose casos consistentes de mejorías o empeoramientos de alta intensidad.

Con todo lo anterior, se tiene evidencia inicial para establecer que, si bien existen mejorías y empeoramientos en el rendimiento docente percibido por los alumnos durante el transcurso del semestre, estos cambios son leves, existiendo más intensamente a nivel de indicadores y cursos específicos que como tendencia semestral generalizada. Por ejemplo, si se analizan datos individuales por curso, se observan, por ejemplo, tres cursos en que ocurren mejorías de alta intensidad (de 25% a 49,9%) en al menos uno de los indicadores de evaluación docente. De la misma forma, existen dos cursos en que ocurren empeoramientos de alta intensidad (de -49,9% a -25%) en al menos uno de los indicadores[18]. Ello se observa en los gráficos que siguen:

[16] Ídem.

[17] Para una agrupación más optimizada de los promedios en categorías tipo, se trabajó con categorías intervalares no equidistantes. Ello, debido a que la variable Σ/4 no posee una distribución normal, sino un comportamiento leptocúrtico. Así, la categorización de los porcentajes de mejoría/empeoramiento el tipo de trayectoria fue realizada de la siguiente forma: Empeoramiento alto: de -49,9% a -25%. Empeoramiento medio: de -24,9% a -15%. Empeoramiento leve: de -14,9% a -5%. Empeoramiento mínimo: de -4,9% a -0,1%. Estabilidad absoluta: pendiente igual a 0. Mejoría mínima: de 0,1% a 4,9%. Mejoría leve: de 5% a 14,9%. Mejoría media: de 15% a 24,9%. Mejoría alta: de 25% a 49,9%.

[18] Para ver detalle de mejoría/empeoramiento por cada indicador de la evaluación docente para todos los casos, ver Tabla Anexo 1**:**

Gráfico 1: Casos con mejoría alta en alguno de los indicadores de evaluación docente (n=4)

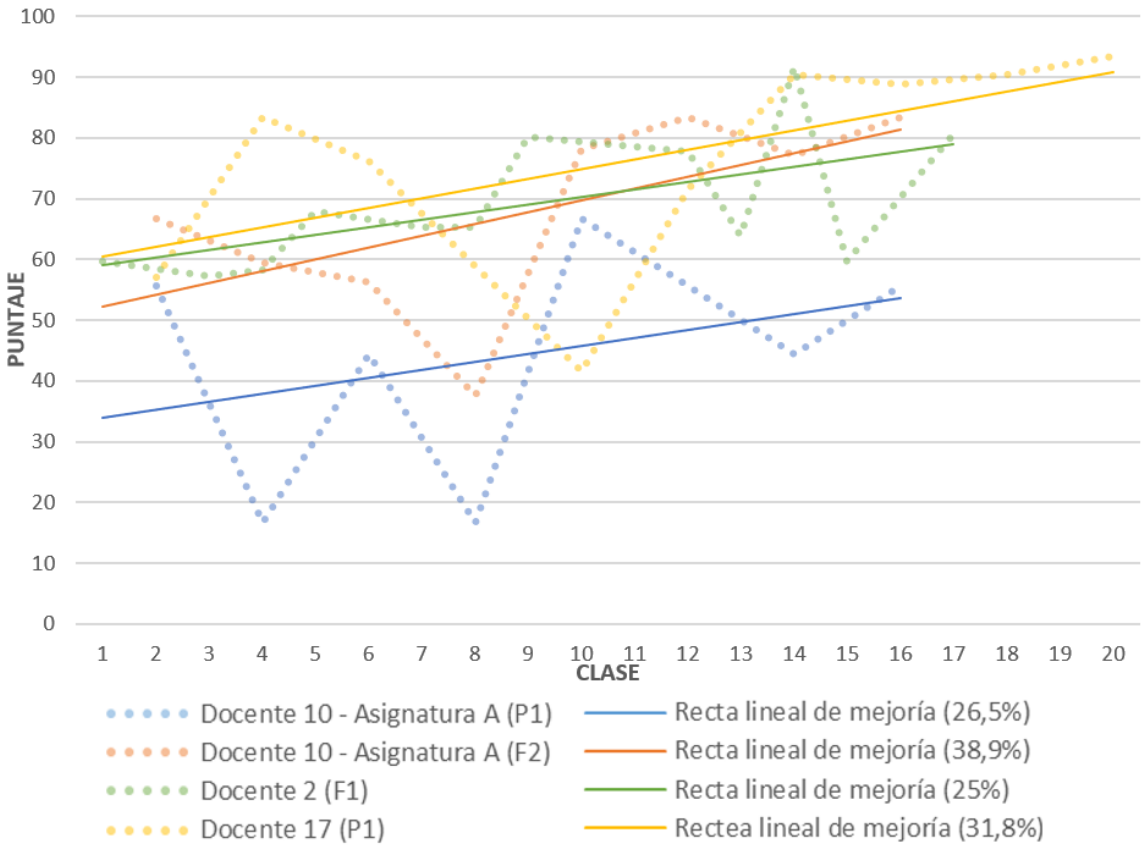
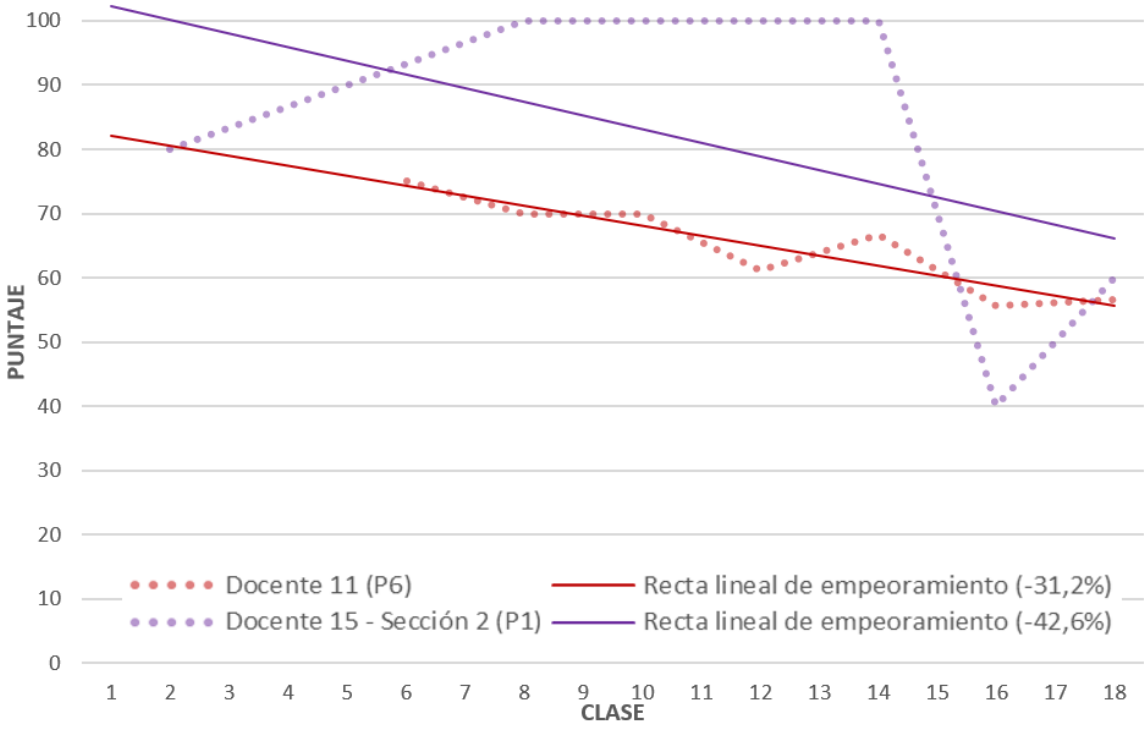


Gráfico 2: Casos con empeoramiento alto en alguno de los indicadores de evaluación docente (n=2)



Como aspecto aparte del análisis individual, es decir, aislando por curso y por cada indicador de evaluación docente por separado, es posible detectar también que los cuatro indicadores de evaluación docente tienden a encontrarse interrelacionados[19]. Esta tendencia se confirma mediante un análisis de correlación entre los cuatro indicadores de mejoría/empeoramiento, presentados en la siguiente tabla:

Tabla 4: Correlación de Pearson entre indicadores de mejoría/empeoramiento (n=23)

	P1	F1	P6	F2
P1	1			
F1	,391	1		
P6	**,501***	**,531****	1	
F2	**,523***	**,472***	**,841****	1

*. La correlación es significante al nivel 0,05 (bilateral).

**. La correlación es significativa al nivel 0,01 (bilateral).

A partir de la matriz de correlaciones anterior, se observa que las tendencias hacia la mejoría y el empeoramiento se encuentran, por lo general, significativamente correlacionadas, al menos con una intensidad media. La interrelación de mejoramiento más intensa se da para el caso del indicador de la calidad general de la clase (P6) y el indicador de la capacidad de la clase para motivar la participación (F2), con una correlación notablemente alta (r=,84).

La única excepción es la correlación entre los indicadores sobre la percepción de los alumnos respecto a la capacidad de la clase para generar aprendizaje en conocimientos en matemáticas (P1) y la capacidad de la clase para generar aprendizaje en habilidades pedagógicas (F1), en cuyo caso la correlación, si bien es de intensidad media, no logra ser significativa. Ahora bien, teniendo presente la escaza cantidad de casos con la que se generó la matriz (n=23), y considerando que en una muestra pequeña la potencia de la relación debe ser bastante alta para lograr niveles significativos de correlación, no se descarta la posibilidad de que exista efectivamente una asociación entre ambos indicadores, si bien no sería tan intensa como aquellas que si alcanzaron un nivel de confianza mayor a 95% en esta prueba.

Con la evidencia de la matriz anterior y su consecuente análisis, se puede sintetizar que tanto la mejoría como los empeoramientos que experimentan los cursos en los distintos indicadores de

[19] Al realizar un análisis visual de los colores de la Tabla Anexo 1, que representan la mejoría y empeoramiento, se evidencia que un empeoramiento o mejoría en uno de los indicadores se tiende a asociar con un cambio en el mismo sentido para el resto de los indicadores: los colores verdes (mejoramiento) tienden agruparse en la parte de superior de la tabla, y los colores amarillos (empeoramiento) tienden agruparse hacia la parte inferior.

evaluación docente, están relacionados entre sí, lo que posiblemente se debe a que todos responden a un fenómeno común, presumiblemente, a la capacidad y desempeño docente que ha sido medido mediante la retroalimentación de los alumnos.

Finalmente, cabe mencionar que el indicador porcentual de mejoría/empeoramiento calculado a partir de la pendiente de la recta de regresión lineal de la serie de datos posee una validez convergente con otro indicador que se aproxima hacia una identificación de la mejoría semestral, como lo es la diferencia de puntuaciones entre la última y la primera clase del semestre. Ambos indicadores correlacionan de forma alta y significativa en cada una de las dimensiones que evalúan el rendimiento docente (rP1=,65; rF1=,79; rP6=,82; rF2=,81). Con ello, se puede señalar que el indicador mediante el cual se representa la tendencia hacia la mejoría o empeoramiento semestral posee un suficiente grado de validez, pese a ser una linealización de una serie no lineal de datos.

- ***Análisis de inestabilidad de las trayectorias de evaluación docente.***

A continuación, se presentan los resultados sobre la inestabilidad de las trayectorias de evaluación docente. Como se ha mencionado anteriormente, en el marco metodológico, el indicador de inestabilidad identifica en qué intensidad las trayectorias de evaluación docente se desvían de una trayectoria perfectamente estable, vale decir, una recta lineal.

Tabla 5: Medias, desviación estándar, mínimos y máximos de inestabilidad de las trayectorias de evaluación docente (n=23)

	P1	F1	F2	P6	Σ/4
Media grupal de inestabilidad	7,36	6,56	5,35	4,10	5,84
D.E. grupal de inestabilidad	3,40	2,04	2,28	1,53	1,89
Inestabilidad mínima	3,31	3,38	2,09	2,09	3,03
Inestabilidad máxima	15,99	11,16	10,12	6,94	10,46

En esta tabla se observa la comparación de inestabilidad entre los cuatro indicadores de evaluación docente. La media grupal nos indica el promedio grupal de inestabilidad de los 23 cursos analizados, dato que, a su vez, es un promedio de la inestabilidad a lo largo del semestre para cada curso.

Al respecto, se evidencia en primer lugar que la mayoría de los valores son más bien leves, tanto para el promedio de los cuatro indicadores (Σ/4) como para cada indicador por separado. Los promedios grupales de inestabilidad por indicador de evaluación docente van del 4,1% al 7,4%, siendo el más inestable la evaluación de los estudiantes respecto a la capacidad de la clase para

generar aprendizaje en conocimientos en matemáticas (P1), y el más estable la evaluación de la calidad general de la clase (P6).

Tomando la misma escala utilizada para clasificar los cursos según su mejoría o empeoramiento, y aplicándola para clasificar la inestabilidad en el indicador conjunto (Σ/4), se tienen 9 casos que poseen una inestabilidad mínima (de 0,1 a 4,9) y 14 casos que poseen una estabilidad leve (de 5 a 14,9). No hay casos cuya inestabilidad se clasifique en una intensidad media o mayor.

Ahora bien, si se observan los casos más en detalle, de forma individual (tabla 6), se evidencia que las medias individuales de inestabilidad por curso, por lo general, no son siempre representativas del conjunto de la inestabilidad de la trayectoria. Esto, ya que las desviaciones estándar para cada media individual de inestabilidad son más bien altas, encontrándose la mayoría cercanas al valor del promedio y en algunos casos incluso sobrepasando este valor. Este nivel de diferencias con la media grupal de inestabilidad anuncia que, si bien en conjunto la inestabilidad no parece ser muy intensa, a nivel de trayectorias individuales podrían observarse casos en que la inestabilidad sí posee intensidades considerables.

Tabla 6: Inestabilidad de trayectorias de evaluación docente según promedio y desviación estándar de diferencias entre puntos reales y recta lineal de regresión

	P1		F1		P6		F2		Σ/4	
Docente - Curso	$\overline{X}$	**D.E.**	$\overline{X}$	**D.E.**	$\overline{X}$	**D.E.**	$\overline{X}$	**D.E.**	$\overline{X}$	**D.E.**
Docente 1	3,31	2,97	6,66	4,59	2,82	1,86	2,86	2,08	3,91	2,87
Docente 2	5,49	3,87	6,51	5,85	4,32	2,71	7,78	4,81	6,03	4,31
Docente 3	5,75	3,79	9,56	5,36	4,53	3,16	8,15	5,59	7,00	4,47
Docente 4	11,63	15,09	11,16	12,22	6,45	4,50	7,65	6,84	9,22	9,66
Docente 5	4,98	5,19	6,09	4,79	3,20	2,73	3,11	2,22	4,34	3,73
Docente 6	7,55	4,83	6,56	3,25	6,07	4,54	5,83	3,66	6,50	4,07
Docente 7	7,43	6,34	7,51	5,85	3,71	2,45	5,46	3,81	6,03	4,61
Docente 8	4,67	4,31	7,78	5,38	2,09	1,95	3,31	2,68	4,46	3,58
Docente 9 - Sección 1	5,48	3,21	4,98	3,36	3,32	2,69	4,01	2,78	4,45	3,01
Docente 9 - Sección 2	7,74	3,98	4,98	3,24	3,20	2,00	3,24	1,44	4,79	2,66
Docente 10 - Asignatura A	13,89	9,14	9,24	4,90	6,94	4,72	9,44	8,27	9,88	6,76
Docente 10 - Asignatura B	6,35	5,55	8,96	5,99	5,82	3,81	4,20	2,84	6,33	4,55
Docente 11	13,05	9,49	5,39	3,84	2,57	2,06	2,82	2,71	5,96	4,52
Docente 12 - Asignatura A	6,44	5,98	3,38	2,95	5,42	3,72	7,27	6,33	5,63	4,74
Docente 12 - Asignatura B	6,45	7,14	5,54	4,70	3,31	3,64	4,86	4,74	5,04	5,05
Docente 13	4,76	2,29	5,41	5,15	2,79	2,10	2,87	2,78	3,96	3,08
Docente 14	5,05	2,10	4,50	2,33	5,02	2,76	6,90	4,40	5,37	2,90
Docente 15 - Sección 1	5,39	4,72	5,69	3,08	2,64	2,62	3,80	2,51	4,38	3,24
Docente 15 - Sección 2	15,99	10,79	8,95	5,77	6,77	5,03	10,12	5,16	10,46	6,69
Docente 16	8,38	5,67	7,90	5,48	4,83	1,96	5,63	2,70	6,69	3,95
Docente 17	10,49	9,40	6,34	6,33	2,26	2,25	5,97	3,01	6,27	5,25
Docente 18 - Sección 1	5,67	4,06	3,63	2,63	3,82	2,83	5,61	2,59	4,69	3,03
Docente 18 - Sección 2	3,39	2,43	4,15	2,18	2,50	1,56	2,09	1,69	3,03	1,96
Media	7,36	5,75	6,56	4,75	4,10	2,94	5,35	3,72	5,84	4,29
D.E.	3,40	--	2,04	--	1,53	--	2,28	--	1,89	--

Lo anterior nos lleva a considerar que, si bien en el conjunto de cursos la inestabilidad no es alta, existen momentos específicos de inestabilidad a lo largo de las trayectorias de muchos de los cursos. La inestabilidad, si bien posee magnitudes leves en promedio a nivel grupal, e incluso a nivel individual, tiene máximos y mínimos considerablemente altos al interior de los casos en sí. Ello queda en evidencia al observar cada caso con más detalle aún, como lo es mediante los datos de inestabilidad mínima y máxima (tabla 7), que corresponden a los datos más extremos de inestabilidad de todas las clases de un mismo curso. A partir de los datos de la tabla 7 se evidencia que los mínimos de inestabilidad por lo general son leves, la mayoría no alcanzando el 1%, sin embargo, de forma inversa, los máximos son considerablemente altos, teniendo la mayoría dos dígitos.

Tabla 7: Puntos mínimo y máximo de inestabilidad para las trayectorias de evaluación docente

	P1		F1		P6		F2	
Docente - Curso	**Min.**	**Max.**	**Min.**	**Max.**	**Min.**	**Max.**	**Min.**	**Max.**
Docente 1	0,37	12,88	0,11	16,08	0,12	7,02	0,02	8,79
Docente 2	0,03	11,68	0,06	17,71	0,74	10,83	1,43	15,04
Docente 3	0,01	13,83	0,21	21,03	0,10	11,72	0,05	21,58
Docente 4	0,08	72,11	0,65	55,24	0,14	19,29	0,04	29,69
Docente 5	0,18	19,75	0,09	19,52	0,04	9,76	0,01	9,11
Docente 6	0,61	12,79	1,36	12,77	1,40	15,37	0,56	13,19
Docente 7	0,77	29,40	0,21	22,65	0,01	8,75	0,80	15,15
Docente 8	0,02	16,36	0,65	20,39	0,05	6,93	0,03	9,75
Docente 9 - Sección 1	1,36	11,18	0,76	10,43	0,32	8,00	1,19	8,67
Docente 9 - Sección 2	0,54	16,63	0,65	11,49	1,25	6,91	0,77	5,44
Docente 10 - Asignatura A	3,17	25,13	0,69	16,15	0,93	12,04	1,59	26,31
Docente 10 - Asignatura B	0,29	17,21	1,09	17,49	0,96	12,76	2,06	9,95
Docente 11	2,46	30,83	2,43	12,97	0,33	5,45	0,12	7,14
Docente 12 - Asignatura A	0,16	16,30	0,82	10,55	0,96	11,39	0,68	21,18
Docente 12 - Asignatura B	0,67	21,87	1,21	16,13	0,08	9,36	0,41	13,80
Docente 13	2,52	11,11	0,19	17,66	0,01	7,70	0,03	11,63
Docente 14	3,08	7,70	1,22	7,11	0,70	8,02	1,50	11,71
Docente 15 - Sección 1	0,20	13,47	1,26	10,99	0,20	7,24	0,49	9,73
Docente 15 - Sección 2	0,43	32,47	4,22	17,50	1,13	14,41	2,72	18,64
Docente 16	0,05	20,16	0,78	17,31	1,79	7,74	2,10	11,60
Docente 17	3,40	31,61	0,13	20,49	0,05	6,54	2,76	11,80
Docente 18 - Sección 1	0,67	12,06	0,96	8,46	0,83	7,99	1,37	10,84
Docente 18 - Sección 2	0,02	7,40	1,54	8,92	0,54	5,37	0,47	6,71
Promedio	0,92	20,17	0,93	16,91	0,55	9,59	0,92	13,37
Valores extremos	0,01	72,11	0,06	55,24	0,01	19,29	0,01	29,69

Finalmente, y de forma adicional, es posible señalar que no existe una relación entre la inestabilidad y la mejoría que experimentan los cursos a lo largo del semestre. En ninguno de los componentes de evaluación docente estudiados existe una correlación estadísticamente significativa entre la mejoría y la inestabilidad (rP1=-,06 p=,78; rF1=-,31 p=,89; rP6=,15 p=,5; rF2=,325 p=,13). Esto constituye un hallazgo interesante, ya que deja en evidencia que las

tendencias de mejoría o empeoramiento ocurridos en los procesos de enseñanza no se relacionan con el nivel de estabilidad con el que se desarrolla esta trayectoria de cambio.

En su conjunto, la evidencia anterior sobre a la inestabilidad sirve para constatar que ésta no se constituye como un rasgo que defina cierto tipo de cursos, y está lejos de ser una anomalía o una característica propia del mal rendimiento docente. Tiende a ser una característica presente en algunos momentos de la generalidad de las trayectorias semestrales, sin ser constante y a todo momento en las trayectorias de evaluación docente.

Sería esperable, por lo tanto, que los cursos pasaran por fluctuaciones en la retroalimentación que hacen los estudiantes, pero que no responden a la calidad de las clases, sino a la complejidad de un proceso de interacción social en los contextos de aprendizaje, que son espacios de sesiones periódicas que se comportan de forma dinámica y no se siguen bajo la forma de una constante rutina de acumulación lineal de logro en la obtención de conocimientos y aprendizajes.

- ***Capacidad predictiva de la primera y última clase sobre la media de evaluación docente.***

A continuación, se realiza un análisis de las correlaciones entre las puntuaciones de la primera, la última y el promedio de puntuaciones de las trayectorias de evaluación docente, con el fin de diagnosticar si es que la primera y última evaluación son capaces de explicarse entre sí, y también, si es que las mediciones inicial y final son representativas del promedio obtenido del conjunto de mediciones hechas a lo largo del semestre.

Tabla 8: Correlación de Pearson entre momentos de evaluación docente (n=23)

	Primera y Última Clase	Primera Clase y Promedio	Última Clase y Promedio
P1	,228	**,746****	**,615****
F1	,383	**,808****	**,638****
F2	,352	**,784****	,405
P6	**,727****	**,872****	**,839****

**. La correlación es significativa al nivel 0,01 (bilateral).

La primera columna de la tabla evidencia que no existe una correlación significativa entre la primera y la última evaluación del semestre en los elementos más específicos del rendimiento docente (P1, F1 y F2). Sin embargo, a nivel del indicador de calidad general de la clase (P6), sí existe una notable capacidad predictiva de la primera clase sobre la última (r=,727).

A partir de este resultado, se evidencia que existe un cambio, mediado por el transcurso del semestre, en la evaluación que poseen los alumnos de los cursos a los que asisten. Esto se complementa con lo ya visto respecto a las trayectorias de mejoría y empeoramiento de los cursos, donde se observa que, desde un mismo punto inicial, algunos docentes tienden a bajar sus evaluaciones y otros a subirlas durante el semestre de clases.

Ahora bien, la alta correlación en el caso del indicador sobre la percepción de calidad general de las clases (P6) demuestra que sería posible, para el caso de este indicador, predecir con relativa exactitud el resultado de la clase final del semestre mediante la medición de inicio de semestre. Esto se debe a las características que ya se han revisado de este indicador: la mayoría de los docentes tienden a mejorarlo a lo largo del semestre y es el más estable de todos. Ambas características serían propias de este indicador, que no comparte con los demás indicadores de la evaluación docente.

La segunda columna de la tabla otorga información sobre la representatividad que logra la evaluación inicial sobre el promedio de la evaluación semestral. La prueba de asociación entre la primera evaluación y el promedio de ellas siempre indica la existencia de correlación, para todos los indicadores, de forma alta y significativa (r>,7). Lo mismo tiende a ocurrir con la correlación entre la última clase y el promedio de ellas, situación en que, para tres de los cuatro indicadores, la evaluación docente final tiende a ser relativamente representativa del promedio de todas las evaluaciones realizadas en el semestre. Sólo en el indicador asociado a la capacidad de la clase para motivar la participación (F2), si bien la fuerza de la relación es intermedia, se evidencia que no existe una representatividad significativa de la evaluación final por sobre la media semestral.

En síntesis, los resultados anteriores apuntan a que la evaluación docente, como producto terminal de los cursos semestrales, corresponde a una medición relativamente representativa del promedio del proceso, a nivel de la evaluación general de la clase, y de su capacidad para entregar conocimientos. No logra ser representativa de forma significativa, sin embargo, de la medición de las capacidades motivacionales de la clase considerando la media general del transcurso del semestre. Adicionalmente, llama la atención que la capacidad predictiva sobre la media semestral es mayor en el caso de la medición al inicio del semestre, que la medición realizada al final de este. Se estaría en presencia, entonces, de trayectorias bastante estables donde la primera impresión logra ser, en mayor o menor medida, definitiva.

Ahora bien, cabe recordar que el instrumento de evaluación docente refleja las percepciones de rendimiento en los cuatro indicadores según los alumnos para cada clase individual. Por ello, la

evaluación de la última clase no correspondería a un diagnóstico del conjunto del semestre. Con ello presente, no resulta extraño que la última clase no logre representar el conjunto del proceso semestral en su totalidad, en específico, que la motivación que logra la última clase discrepe del conjunto semestral, y que esta clase en sí no sea tan representativa del semestre como lo es la primera. Esto puede deberse a que la última clase es también la clase de cierre, por lo que los procesos de enseñanza y aprendizaje al interior de ella no son tan similares al resto del proceso semestral; se incluyen otras dinámicas, como podrían ser reflexiones de síntesis y otras similares, y a su vez, no logran ser lo suficientemente satisfactorias para los cursos en que no se han logrado revisar todos los contenidos, siendo en este momento de cierre semestral en que quedaría en evidencia para los alumnos este tipo de carencias, que influirían negativamente sobre todo en la evaluación motivacional que hacen de sus docentes.

iv. Análisis de asociación entre variables independientes y las trayectorias de evaluación docente

A continuación, se encuentra el apartado de análisis de asociación entre el comportamiento de las trayectorias de evaluación docente, y las variables que se han medido y considerado como aspectos relevantes a considerar en relación con el rendimiento de la evaluación docente. En este sentido, este apartado se constituye no sólo como la descripción del comportamiento longitudinal de las trayectorias de evaluación docente, sino como la exploración de qué factores se encuentran asociados a su comportamiento.

- ***Diferencias según Sexo del docente.***

Para estudiar las diferencias según sexo (masculino/femenino) en las trayectorias de evaluación docente, se utilizaron las pruebas t de Student y U de Mann-Whitney. La muestra de docentes contó con 9 hombres y 9 mujeres. Ahora bien, en los cursos analizados, no se detectaron diferencias significativas según sexo del docente en las trayectorias de evaluación docente, en ninguno de los cuatro componentes evaluados. En base a la evidencia recolectada y analizada para este caso, es posible señalar que el sexo del docente no se encuentra asociado a los puntajes de evaluación de comienzo o final de semestre, ni con la media semestral, ni con la inestabilidad, mejoría o empeoramiento que experimenta la evaluación docente de los cursos analizados.

- ***Horas de clase realizadas***

Tabla 9: Correlaciones de Pearson entre horas de clase realizadas y variables dependientes (n=23)

	P1	F1	F2	P6
Puntaje primera clase	,382	**,614****	,396	,295
Puntaje última clase	,387	,124	,305	,291
Media semestral	**,509***	**,414***	**,595****	,403
Porcentaje de mejora semestral	-,040	-,358	-,141	,016
Indicador de Inestabilidad	-,355	-,010	**-,472***	-,283

*. La correlación es significante al nivel 0,05 (bilateral).
**. La correlación es significativa al nivel 0,01 (bilateral).

A partir de la tabla anterior, se observa, en primer lugar, que existe una correlación positiva de intensidad medianamente alta (r=,614) entre la cantidad de horas de clase realizadas en el semestre, y el grado en que los estudiantes percibieron que la primera clase logró generar en ellos un aprendizaje en habilidades pedagógicas. Esto podría encontrar su causa en que, desde el punto de vista del docente, encontrarse en un contexto institucional que asegura más horas de clase, supone una mayor certeza sobre lo que puede hacer en el semestre y mejor capacidad para planificar las asignaturas, logrando con ello realizar una clase de introducción y presentación de los cursos de mejor calidad.

En caso de ser esto cierto, se estaría en presencia de un fenómeno novedoso: una mejor introducción del semestre afecta directamente a la percepción que poseen los alumnos de la medida en que logran desarrollar sus habilidades pedagógicas, en esa clase en específico. Esto significaría que la actividad de presentación semestral sería, a su vez, una actividad de desarrollo de las habilidades pedagógicas de los estudiantes, no cumpliendo una función únicamente formal al interior de los cursos; sería una actividad de aprendizaje en sí misma.

También se observan correlaciones positivas de intensidad media entre la cantidad de horas de clase realizadas en el semestre, y la media semestral de percepción de la capacidad de los cursos para generar aprendizaje en conocimientos de matemáticas (r=,509), desarrollar aprendizaje en habilidades pedagógicas (r=,414) y motivar la participación (r=,595). A mayor cantidad de clases en el semestre, la evaluación docente tiende a lograr mejores resultados de promedio semestral en sus aspectos específicos, pero no así en su evaluación general, con la cual no existe una asociación significativa, pese a que la intensidad de la relación entre ambas variables es intermedia y solo un poco más baja que con las otras variables.

En tercer lugar, se observa que existe una correlación significativa negativa de intensidad media (r=-,472) entre la cantidad de horas de clase realizadas en el semestre, y la inestabilidad semestral de la evaluación sobre la capacidad de los cursos para motivar la participación en los estudiantes. Esto podría interpretarse de la siguiente forma: a mayor duración de los semestres de clases, los estudiantes experimentan una estabilización de la sensación de motivación que poseen con las asignaturas. Llama la atención que se trata de una estabilización, sin embargo, no una disminución o aumento en la motivación a medida que transcurre el semestre. Teniendo esto último en consideración, este fenómeno puede interpretarse como de sensación de rutinización que podrían experimentar los asistentes a los cursos que duran más. A mayor cantidad de clases, los alumnos conocen de mejor forma los aspectos más procedimentales de la rutina de las sesiones, por lo que experimentan menor sensación de inestabilidad motivacional a lo largo del semestre.

- ***Pérdida de clases***

Para analizar el efecto sobre las trayectorias de evaluación docente que posee la pérdida de clases, se ha trabajado con dos variables para representar este último elemento. En primer lugar, se ha utilizado una variable cuantitativa discreta que enumera la cantidad de clases perdidas en el semestre que poseen los cursos y, en segundo lugar, una variable dicotómica que distingue entre el grupo de cursos que perdieron al menos una clase en el semestre, y el grupo de cursos que no perdieron clases.

Utilizando la primera variable de pérdida de clases, cuantitativa discreta, no se ha identificado una correlación lineal significativa entre la proporción de clases perdidas y las variables que describen las trayectorias de evaluación docente. Se han identificado, sin embargo, diferencias significativas utilizando la variable dicotómica, que distingue la presencia o ausencia de pérdida de clases: entre los cursos que pierden clases y los que no (tabla 10). Existe una diferencia leve, pero significativa, en la inestabilidad de la percepción que poseen los alumnos sobre la capacidad de las clases para generar aprendizaje en matemáticas (P1) y motivar la participación (F2). La ausencia de asociación en el primer caso y la presencia de ella en el segundo refleja, entonces, una relación no lineal entre la pérdida de clases y la inestabilidad de la trayectoria de evaluación docente en dos de sus aspectos.

Tabla 10: Diferencias de medias significativas en inestabilidad según pérdida de clases

	Pierde clases	N° casos	Media	Diferencia de medias	Sig. t de Student	Sig. U de Mann-Whitney
Inestabilidad de "Capacidad de la clase para generar aprendizaje en conocimientos en matemáticas" (P1)	No	11	6,13	2,35	0,098	0,031
	Si	12	8,49			
Inestabilidad de "Capacidad de la clase para motivar la participación" (F2)	No	11	4,31	1,98	0,034	0,023
	Si	12	6,29			

Los valores p de la tabla anterior apuntan hacia diferencias de medias estadísticamente significativas a un buen nivel de confianza ($p < ,05$). Para la primera diferencia de medias, sin embargo, sólo se alcanza un nivel de confianza de un 90%. Ahora bien, considerando que se tiene una cantidad de casos por debajo de lo recomendado y la distribución de la variable dependiente (P1) no es normal (p de Shapiro-Wilk=,04) parece ser más apropiado, en este caso, guiarse por la significación de la prueba no paramétrica (U de Mann-Whitney), que sí indica una diferencia de rangos estadísticamente significativa a un nivel de confianza mayor a 95%. Para la diferencia de inestabilidad sobre la motivación que logran las clases, la significación otorga un nivel de confianza mayor a 95% para ambas pruebas, lo que resulta consistente.

Con lo anterior, es posible sintetizar que tanto la estabilidad de la percepción de aprendizaje en matemáticas que logran las clases, como la estabilidad de la motivación que estas generan en los alumnos, se ve ligeramente afectada por el hecho de que en los cursos se pierdan clases o no. Es una relación dada por una dicotomía, y no se constituye como una relación lineal: la diferencia de inestabilidad en estos dos aspectos la hace el hecho de perder clases o no, y no el volumen con el cual se pierden. Cabe mencionar, al igual que en otros resultados, que sólo se trata de un efecto sobre la estabilidad longitudinal del rendimiento de las clases en los aspectos antes mencionados, y no así de un efecto de la no realización de clases por sobre la evaluación del logro de las clases en sí, en cuyo caso no se observan efectos significativos.

Se puede concluir entonces, respecto a este elemento —la pérdida de clases—, que tiene efectos sobre la inestabilidad de la evaluación de los docentes, pero no sobre la media de su evaluación o su trayectoria en el tiempo.

- ***Cantidad de alumnos inscritos***

En la sección de datos descriptivos sobre la muestra, se observó alta heterogeneidad existente en la cantidad de alumnos inscritos en los cursos evaluados. Con una media de 21 personas inscritas y una desviación estándar de 13.4 en un grupo de 23 cursos, la dispersión resultó ser bastante alta. Resulta relevante entonces, bajo estas condiciones, observar si este aspecto se relaciona con alguno de los indicadores de evaluación docente evaluados en el presente estudio.

Tabla 11: Correlaciones de Pearson entre Cantidad de alumnos inscritos y variables dependientes (n=23)

	P1	F1	F2	P6
Puntaje primera clase	,401	,122	-,176	,114
Puntaje última clase	,087	,042	-,012	,013
Media semestral	,392	,102	-,002	,130
Porcentaje de mejora semestral	-,262	-,100	-,094	-,108
Indicador de Inestabilidad	**-,486***	-,242	**-,440***	-,393

*. La correlación es significante al nivel 0,05 (bilateral).

Es posible observar a partir de la tabla anterior que el único indicador sobre el comportamiento de la trayectoria semestral de evaluación docente que se relaciona con la cantidad de alumnos inscritos en los cursos analizados es la inestabilidad de la evaluación de los cursos.

En específico, se observa una correlación negativa de intensidad media entre la cantidad de alumnos inscritos en los cursos, y la inestabilidad semestral de la percepción de los estudiantes respecto de haber desarrollado aprendizajes en matemáticas (r=-,486) y haber estado motivados para participar durante las clases (r=-,44).

Esto indica que la cantidad de alumnos inscritos en las asignaturas afecta medianamente la estabilidad de la trayectoria de evaluación: a mayor cantidad de alumnos inscritos, más estable es la forma en que se percibe que se aprende matemáticas y se participa a lo largo del semestre. Este fenómeno puede tener que ver con que, a menor cantidad de alumnos en sala, existe una tendencia hacia la personalización de la enseñanza. Así, en contextos de menor cantidad de alumnos, el enfoque docente se torna más personalizado, a la vez que los alumnos preguntan y resuelven sus dudas sobre matemáticas de forma más participativa y con mayor facilidad. De forma opuesta, las clases con más alumnos presentes tienden a adquirir un formato más de cátedra dictada frente a una audiencia, donde los espacios de preguntas y participación en general son menores y se encuentran más normalizados.

Ahora bien, la cantidad de alumnos si bien interviene sobre la estabilidad de estos indicadores, no interfiere sobre la percepción de la calidad resultante de las clases en los dos ámbitos antes mencionados. La cantidad de alumnos no influye sobre las trayectorias de evaluación ni, en la evaluación media. Esto es observable en la ausencia de relación entre los demás indicadores presentes en la tabla de correlación, distintos a la inestabilidad.

Se puede establecer, por lo tanto, que el cambio en el formato de la clase —dado por la cantidad de alumnos—, que intervendría en la forma en que participan y aprenden matemáticas, sólo estaría afectando la percepción de la inestabilidad de estos procesos, no así la percepción sobre la media semestral de logro en estos indicadores de evaluación docente. Es un cambio de formato de clases que, si bien facilita la intervención ocasional de los alumnos, no afecta la precepción de logro y aprendizaje resultante del total del semestre que poseen los alumnos en los indicadores discutidos.

Llama la atención, adicionalmente, que la relación entre la cantidad de alumnos inscritos en los cursos y la inestabilidad de las trayectorias sólo se dé para el caso de la percepción de aprendizaje en matemáticas y la motivación de la participación. La cantidad de alumnos en clases no estaría relacionada con la estabilidad de la percepción sobre la capacidad de la clase para desarrollar aprendizaje en habilidades pedagógicas, ni con la estabilidad de la percepción de la calidad general de la clase. El fenómeno de personalización de la enseñanza antes mencionado, dado por la baja cantidad de alumnos en sala, no tendría efecto en la estabilidad con la que los alumnos evalúan en general las clases, ni en la forma en que perciben que la clase aporta desarrollar sus conocimientos en pedagogía a lo largo del semestre.

- ***Formato programático de las sesiones de clases***

Finalmente, se tienen resultados respecto a la influencia que posee el formato programático de las sesiones de clases, es decir, si se trata de ramos con una o dos clases por semana, por sobre las trayectorias de evaluación docente.

Al respecto, de todas las variables utilizadas para caracterizar las trayectorias, sólo se identifica un efecto por sobre la inestabilidad de la evaluación de las capacidades que poseen las clases para desarrollar aprendizaje en habilidades pedagógicas (tabla 12).

Tabla 12: Diferencias de medias significativas en inestabilidad según formato programático de las sesiones de clase

	Formato de clase	N° casos	Media	Diferencia de medias	Sig. T de Student	Sig. U de Mann-Whitney
Inestabilidad de "Capacidad de la clase para desarrollar aprendizaje en habilidades pedagógicas" (F1)	Una clase por semana (bloques seguidos)	15	5,93	-1,79	0,042	0,024
	Dos clases por semana	8	7,72			

La diferencia detectada es leve, sin embargo, significativa. Los cursos con dos clases por semana tenderían hacia una ligera mayor inestabilidad en su capacidad percibida para desarrollar habilidades pedagógicas, que aquellos con una clase por semana.

Así, es posible señalar que para los estudiantes que poseen más clases por semana, existe una percepción de menor estabilidad con la que perciben su propio aprendizaje en habilidades pedagógicas. Esta diferencia, al igual que en caso anteriores, se trata de un efecto por sobre la estabilidad de las trayectorias de evaluación docente, no así del logro conjunto percibido por los estudiantes.

VII. Discusión de resultados y conclusiones del estudio

i. Síntesis de resultados

Lo primero que resalta tomando el conjunto de resultados, son los altos niveles en el rendimiento de los cuatro indicadores de evaluación docente, y a su vez, la homogeneidad que existe en este aspecto entre los distintos cursos integrados en el análisis. Los puntajes de los cursos parecen ser bastante buenos, condición que es relativamente generalizada entre los cursos revisados. Se tiene así, una calidad de la docencia que —al menos desde la retroalimentación y juicio que entregan sus alumnos— es bastante satisfactoria.

Ello se acompaña de trayectorias que, en el tiempo, se mantienen relativamente estables, sin alteraciones de muy alta intensidad entre clase y otra, ni empeoramientos o mejoramientos a lo largo del semestre que se proyecten más allá de cambios mínimos y/o leves. Existen, por cierto, situaciones puntuales de empeoramiento, mejorías y/o inestabilidades que sobrepasan la tendencia generalizada hacia la estabilidad, sin embargo, ocurren más bien a nivel de casos aislados o indicadores de evaluación docente específicos, no constituyendo una tendencia generalizada.

A esta tendencia a la mantención, se suma el hecho de que no se evidenció asociación entre la mejoría —o bien el empeoramiento— con los factores extra-docencia revisados. Al respecto, cabe señalar que la tendencia generalizada a la mantención dificulta la capacidad de observar cuáles son las determinantes sobre las posibilidades de variación de las trayectorias, ya sea hacia la mejoría, o bien hacia el empeoramiento. Esto lleva a considerar la hipótesis de que la mejoría o el empeoramiento intra-semestral son fenómenos que tienen que ver, eminentemente, con el cambio en las capacidades docentes ocurridos exclusivamente al interior de los periodos semestrales. Para resolver dicha interrogante, correspondería estudiar en qué medida las trayectorias de evaluación docente experimentan cambios de pendiente ocasionados, por ejemplo, por capacitaciones o mejoramientos curriculares que experimentan los docentes *durante* los periodos semestrales. De todas formas, como ya fue revisado anteriormente, otras investigaciones ya han apuntado a que las mejorías en el rendimiento docente al interior de los mismos semestres se identificarían sólo en docentes que han realizado actividades de mejoramiento durante el tiempo en que son evaluados (Hativa, 1996).

Por otro lado, y como ya se menciónó, han existido pocas investigaciones que avancen en la interpretabilidad del fenómeno de la estabilidad/inestabilidad, especificando qué significa que una trayectoria de evaluación docente a lo largo del tiempo sea inestable o estable, qué causa este comportamiento o con qué se relaciona, y si produce un efecto positivo o negativo sobre la calidad de la enseñanza.

Al respecto, a partir de esta investigación empírica se tiene evidencia para señalar que, al menos, la inestabilidad no es un fenómeno generalizado ni constante en las trayectorias de evaluación docente analizadas, y no pareciera afectar el logro de estas. No se relaciona ni con las medias semestrales, ni con las trayectorias de empeoramiento o mejoramiento de los docentes en los periodos semestrales. Tiende a ser un indicador de intensidades generalmente bajas que parece reflejar las interrupciones naturales de un proceso de evaluación docente que no es continuo, sino que es periódico, en este caso, semanal.

Ahora bien, se observa que ciertos aspectos extra-docencia producen inestabilidad en las trayectorias semestrales en sus distintos aspectos medidos, sin embargo, estos efectos se juegan más bien en el orden de lo procedimental de las clases, no teniendo relación con el logro en la generación de aprendizajes, ni con el mejor o peor rendimiento de las clases en términos del desenvolvimiento docente. A lo más, estos resultados confirman la proposición anterior, relativa a que la inestabilidad es solo reflejo de la periodicidad de las sesiones de clases.

ii. Factores que afectan el rendimiento docente

Respecto a los factores extra-docencia que se analizaron en relación con los puntajes transversales de evaluación docente, y también con las trayectorias longitudinales, lo primero que resalta es la ausencia de asociación entre el rendimiento y el sexo del docente.

Estos resultados se alejan de lo obtenido por otras investigaciones. Hubiera sido esperable llegar a conclusiones similares a las de Medel y Asun (2014), quienes también trabajaron con una muestra chilena y reportan una propensión a que las académicas, en comparación con sus colegas masculinos, sean mejor evaluadas en aspectos como la responsabilidad y habilidades pedagógicas, y peor evaluadas en el ámbito de dominio disciplinario. En nuestro caso, no se observan dichas diferencias en estos aspectos dadas por el sexo del docente, que podrían haberse evidenciado mediante el cruce entre los indicadores medios de F1 (para evaluar las “habilidades pedagógicas”) y P1 (para “dominio disciplinario”) con la variable Sexo.

La hipótesis frente a la ausencia de relación encuentra su justificación en la especificidad de la muestra. Dicho sesgo de género evidenciado en otros estudios puede darse en algunos contextos de enseñanza, incluyendo los chilenos, pero no se logra percibir en este estudio específico ya que se analizan cursos de docencia universitaria orientados a la pedagogía escolar, espacio que históricamente en Chile se encuentra altamente feminizado, como ya se ha mencionado previamente (Ministerio de Educación, 2017). Por este motivo, los contextos educativos analizados en el presente estudio son espacios donde, en mayor medida, se han abandonado bastantes de las creencias que consideran que la calidad y rendimiento docente se comportan de forma diferenciada según el sexo del docente. Una hipótesis alternativa frente a la ausencia de relación es la falta de casos suficientes para obtener resultados positivos, lo que se torna más problemático aún si se cumple la hipótesis anterior, en que la relación entre el sexo y el rendimiento docente reportado por los alumnos tendería a ser leve.

Cabe señalar, adicionalmente, que este estudio no incorporó el sexo del estudiante como variable de análisis. Esto, debido a que el estudio trabajó con las clases y sus docentes como unidad de análisis. No hubo espacio, por tanto, para evidenciar o descartar el efecto de un sesgo de género entrecruzado estudiante-docente.

Otro hallazgo relevante respecto a los factores extra-docencia analizados, es el hecho de que una alta cantidad de clases realizadas se encuentra relacionada con una mejor planificación inicial del semestre, y con un mejor rendimiento de los ámbitos disciplinares, pedagógicos y motivacionales que logran los docentes a lo largo de toda su asignatura semestral. A este respecto, resulta relevante destacar que, toda vez que la mayor cantidad de clases se relaciona con medias de evaluación semestral más altas, no existe asociación significativa entre la cantidad de clases y la mejoría o empeoramiento semestral. Ello implica que los cursos más largos no generan necesariamente trayectorias ascendentes de evaluación, sino medias de evaluación más altas. En otras palabras, el efecto positivo sobre la evaluación docente que otorga mayor cantidad de clases no tiene que ver con que estas vayan en progresiva mejoría, sino que se relaciona con su aporte al conjunto de todas las clases del semestre.

Si se considera lo anterior, en conjunto con el hecho de que la pérdida de sesiones de clases a lo largo del semestre no se encuentra significativamente asociada con el rendimiento docente o su mejoría/empeoramiento semestral, se pude concluir que la obtención de evaluaciones docentes más positivas no se encuentra en el efecto que produce únicamente el aumento de la cantidad de clases, o en el hecho de evitar perder clases programadas, sino que se explica por el efecto de una

programación planificada de las clases, que dispone espacios para una revisión reposada de los contenidos, evitando la reducción, "compresión", revisión superficial o "a la rápida" de estos, prácticas que tenderían a reducir el puntaje docente semestral en todos sus ámbitos: disciplinar, pedagógico y motivacional.

A propósito de este hallazgo en particular, cabe hacer una reflexión sobre la validez consecuente de la evaluación docente, asunto que, como se vio al esbozar este campo de estudio, muchos investigadores apuntan que requiere especial atención. El efecto positivo de la alta cantidad de clases sobre el logro docente lleva a advertir sobre uso de la información que se hace a partir de las evaluaciones docentes en los espacios institucionales reales. Es posible dudar de la representatividad de las evaluaciones docentes hechas en contextos donde las medidas de las capacidades de los profesores son 'castigadas' por el efecto que tienen semestres cortos, donde no se logran pasar los contenidos, y donde muchas veces estos son impuestos institucionalmente mediante currículos de enseñanza ajenos a las decisiones pedagógicas de los docentes.

Para finalizar la reflexión sobre los efectos que tienen los aspectos extra-docencia sobre las trayectorias de evaluación docente, cabe hacer notar la ausencia de efecto de la pérdida de clases, la cantidad de alumnos inscritos, y el formato programático. Estos factores extra-docentes solo se relacionan con la inestabilidad de algunos aspectos de las trayectorias de evaluación docente (disciplinar, pedagógica, motivacional) pero no con la calidad de la enseñanza y su correspondiente evaluación. Estos son hallazgos interesantes, siendo algunos opuestos a lo que podría considerarse de forma anticipada.

La pérdida de clases, a veces imputada como el factor culpable del no logro de los objetivos docentes dentro del ambiente académico, no evidenció en esta investigación efecto sobre el rendimiento docente. Ahora bien, cabe advertir que la causa de la falta de clases no fue tipologizada en el recogimiento de datos por parte del equipo encargado de esto, por lo que fenómenos como cierres anticipados o comienzos retrasados de semestre; la falta programada de clases; paros por parte de los alumnos; ausencia individual del docente; u otros casos específicos de falta de clases, resultaron agrupados todos en una misma categoría de "clase perdida", no pudiendo ser diferenciados entre sí. Una aproximación que estudie el rendimiento docente, transversal y longitudinal, y su relación con cada uno de estos tipos de ausencia de clases por separado, podría eventualmente hallar relaciones que en el presente estudio no son posibles de someter a prueba.

Por su parte, la cantidad de alumnos inscritos en los cursos tampoco evidenció efecto sobre el logro de la evaluación docente. Esto resulta interesante, debido a que no sólo discrepa de los resultados de muchas otras investigaciones, sino que va a contrasentido: se suele creer que clases más reducidas en cantidad de alumnos logran mejores resultados, al posibilitar una mejor atención docente a las dificultades individuales de los estudiantes y al mejorarse las capacidades de los profesores para poder monitorear y manejar el desenvolvimiento de las clases. El hallazgo de esta investigación, sin embargo, pese a su contrasentido, se complementa con lo indicado por organismos internacionales como la OCDE, que mencionan que, hasta la fecha, no hay pruebas de peso que demuestren que las diferencias en el tamaño de las clases afecten al rendimiento de los estudiantes (OCDE, 2015, p. 463). Este argumento —del rendimiento de los estudiantes asociados a la cantidad de alumnos en sala— es extensible ahora también hacia el rendimiento docente.

Cabe mencionar de todas formas, que un estudio más preciso que el nuestro sería aquel que considere la asistencia de los alumnos a cada sesión, y no sólo la cantidad de inscritos en las asignaturas, como indicador de la cantidad de alumnos en sala. Ello tendría un mayor nivel de fiabilidad, tanto a nivel longitudinal como transversal, y por consiguiente, se aumentarían las posibilidades de hallar relaciones significativas de la cantidad de alumnos en el aula con las trayectorias de evaluación docente, en la eventualidad de que sean fenómenos relacionados.

Finalmente, respecto al formato programático de las clases, es posible decir que lo único que cambia entre clases divididas en dos sesiones por semana, y una sesión por semana, es la inestabilidad del rendimiento pedagógico de los docentes. En este caso, el formato de mayor cantidad de clases por semana parece producir trayectorias ligeramente más inestables en el ámbito de rendimiento pedagógico que el formato de una clase por semana. Como ya se ha mencionado, la inestabilidad de las trayectorias no refleja su rendimiento o calidad, sino el propio carácter periódico de sesiones separadas. La relación que se ha identificado estaría evidenciando el efecto que tienen mayor cantidad de interrupciones entre clase y clase sobre la estabilidad longitudinal de la evaluación docente, pero no tendría relación con su rendimiento.

iii. Estrategias para mejorar las trayectorias de evaluación docente

Tomando en consideración el conjunto de resultados anteriores, es posible generar recomendaciones en torno a cuáles serían condiciones positivas para la generación de buenas trayectorias de evaluación docente. Mediante esto, se pretende agotar la explicación sobre cuáles son las condiciones para una trayectoria ideal, o bien, hacer juicio sobre efectos que vayan más

allá de las asociaciones estadísticas entre las variables empíricas que involucró este estudio. A continuación se ofrecen, acotadamente, recomendaciones sobre cómo generar mejores trayectorias semestrales de evaluación docente a partir de la evidencia recabada.

Como ya se ha comprobado, contar con más clases en el semestre afecta positivamente el rendimiento medio de todas las clases del conjunto del semestre, en los tres ámbitos de rendimiento estudiados: disciplinar, pedagógico y motivacional. Esto se debe al efecto que posee realizar clases en que los contenidos no son comprimidos o "pasados a la rápida". Se puede decir así, que se obtienen mejores resultados para la evaluación docente cuando se realizan asignaturas con una carga de materia prudente y revisada de forma paulatina.

Mayor cantidad de clases en los semestres no solo posibilita mejores condiciones de planificación de contenidos para los docentes, y un aumento en la calidad en que esta materia es revisada en clases, sino que también permite a los docentes hacer una mejor introducción de los contenidos que se entregarán a los alumnos. Esta actividad formal de presentación, además de clarificar la programación de contenidos, es también una actividad de aprendizaje pedagógico en sí para los estudiantes. Se puede declarar con ello, que la planificación del aprendizaje es también un proceso de aprendizaje en sí mismo.

Finalmente, cabe señalar que para que las condiciones anteriores tengan posibilidad real de llevarse a cabo, deben darse ciertas circunstancias en los contextos institucionales educativos. Específicamente, que los docentes deben poder gozar de autonomía a la hora de determinar el volumen de contenido que resultaría viable revisar en cada asignatura, en vista de la cantidad de clases que se les otorga para el semestre que deben llevar a cabo. Un excesivo aumento en la cantidad de contenidos solicitados a revisar por parte de las instituciones, de forma obligatoria sobre la planificación de los docentes, tendería inevitablemente a generar lo que ya se ha descrito como "compresión" de contenidos.

iv. Rendimiento y limitaciones de los ítems utilizados para medir la calidad docente

Comenzando con los aspectos positivos del rendimiento de los indicadores utilizados para medir la calidad docente, se puede señalar que el instrumento aplicado posee buenos estándares de validez y fiabilidad. Siendo aplicado sobre una muestra robusta, de forma consistente y repetida en el tiempo, resultó capaz de medir cuatro aspectos de la evaluación docente que demostraron

ser dimensiones autónomas, con comportamientos claramente diferenciables entre sí, si bien no totalmente independientes.

Como ya se ha reportado, todos los indicadores de evaluación docente poseen puntajes altos y con baja heterogeneidad entre los distintos cursos evaluados. Revisando su rendimiento específico, llama la atención, en primer lugar, determinadas características del indicador que logra el mejor rendimiento medio: la evaluación general de la clase (P6).

Este indicador es, a su vez, el más estable, pudiendo incluso predecirse con relativa exactitud el resultado medio del semestre mediante la medición del inicio de semestre. También, es el indicador en que menos se evidencian pendientes negativas, y, sin embargo, aquellos casos que empeoran lo hacen en una intensidad mucho mayor a la de otros indicadores.

Se postula que el comportamiento de este indicador, que es distinto al de los otros en las características antes mencionadas, tiene su origen en que, al ser una evaluación más general y no específica del desempeño docente, no rescata el detalle de los comportamientos docentes específicos, y refleja más bien una primera impresión de los alumnos que no tiende a cambiar demasiado, toda vez que los indicadores más específicos logran un juicio que puede variar con mayor facilidad en el tiempo.

Junto con lo anterior, resulta muy clarificadora la alta correlación entre la evaluación general y la evaluación motivacional de la clase —la correlación entre indicadores más alta entre todas—, que deja en evidencia que las preguntas enfatizadas sobre el rendimiento general de las clases se apegan más a elementos motivacionales que a la evaluación del aprendizaje logrado o el desempeño directo del docente.

Es catalogable, por consiguiente, como un indicador poco exigente, que la mayoría de los docentes aprueba muy bien, pero que resulta castigador en caso de que no se cumpla con las expectativas de los estudiantes. Se constituye, sin embargo, como una medición que refleja la impresión con la que se quedan los estudiantes de sus asignaturas, cuestión que define la opinión general de largo plazo.

Cabe señalar que otras investigaciones también han observado la mayor estabilidad de los ítems que reflejan impresiones más generales de la docencia (Hativa & Raviv, 1993; Marsh, 1987), y a su vez, han destacado la capacidad predictiva que poseen estos ítems sobre las medias de rendimiento docente en sus distintas dimensiones (Hativa, 1996). El aporte adicional en este

punto que haría esta investigación sería evidenciar la capacidad predictiva de estos ítems ahora también desde una perspectiva longitudinal.

Toda vez que el indicador general de la calidad de la clase es el de mejor rendimiento, el indicador que rescata la capacidad de la clase para desarrollar aprendizajes en habilidades pedagógicas (F1) es aquel con menor puntaje medio. Es de todas formas un indicador de rendimiento alto, sin embargo, menor en alrededor de un 10% a los otros tres indicadores.

Este resultado avala una proposición hecha con anterioridad, que hace relación a que los ítems más generales y menos focalizados en acciones concretas tienden a tener mejor evaluación y a ser más estables en el tiempo. Al tratarse de una evaluación que solicita a los alumnos hacer retrospectiva de los beneficios concretos de la sesión de clase, en este caso, sobre los aprendizajes pedagógicos adquiridos, la evaluación del rendimiento en habilidades pedagógicas resulta un indicador más exigente y por consiguiente posee un rendimiento menor.

Cabe señalar que otras investigaciones ya habían detectado que, entre ellos, los distintos componentes de la docencia no son igualmente estables, siendo algunos mucho más estables que otros (Polikoff, 2015). La investigación propia también llega a esa conclusión, y adicionalmente, entrega una interpretación a este comportamiento, punto que se acaba de exponer.

Tomando ahora en cuenta las dificultades de los indicadores utilizados, cabe señalar que uno de los principales defectos de este estudio se encuentra precisamente en la formulación de los indicadores generados para medir el rendimiento docente. Éstas son, en concreto, deficiencias en el instrumento de medición de origen. Este tiene problemas tanto en términos de la cantidad de preguntas, como también en la calidad de su contenido.

En primer lugar, respecto a la cantidad de preguntas, es problemático que el instrumento tenga distinta cantidad de ítems para cada dimensión del rendimiento docente. Para las dimensiones de “evaluación de la capacidad de la clase para generar aprendizaje en conocimientos en matemáticas” (P1) y “evaluación de la calidad general de la clase” (P6), se tiene solo una pregunta en el cuestionario para cada dimensión, mientras que para la “evaluación de la capacidad de la clase para desarrollar aprendizaje en habilidades pedagógicas” (F1) y la “evaluación de la capacidad de la clase para motivar la participación” (F2) se tienen cuatro preguntas para cada dimensión.

Como ya se ha discutido[20], se debilita la fiabilidad de la medición en aquellos aspectos de evaluación docente para los cuales existe una sola pregunta. Este error se debe a la ausencia de un proceso de operacionalización riguroso en la planificación y producción del instrumento, idealmente mediante el uso de una tabla de especificaciones. El autor de la presente tesis puede suponer que, en principio, se redactó el instrumento con la intención de cuantificar el rendimiento de las clases, como indicador del rendimiento docente, pero se hizo con una limitada reflexión conceptual, y sin considerar que el proceso de formulación de los ítems debe emplear una cantidad similar entre dimensiones y suficientes para someter a los ítems a pruebas de validez y fiabilidad, elementos necesarios para evaluar la calidad de los instrumentos posterior a su aplicación sobre la muestra.

No solo son de cantidad de ítems los problemas del instrumento utilizado, sino también de la calidad del contenido de estos. Existen problemas en la escala de los ítems, problemas de imprecisión conceptual de aquello que se pretende medir, como también problemas en el fraseo de los enunciados de las preguntas.

Un primer problema de contenido de los ítems es la disparidad en la escala utilizada. Para la "evaluación de la capacidad de la clase para generar aprendizaje en conocimientos en matemáticas" (P1) y la "evaluación de la capacidad de la clase para desarrollar aprendizaje en habilidades pedagógicas" (F1), se utiliza una escala ordinal de puntaje mínimo 0 y máximo 3 (rango de 3), con fraseos de intensidad para cada una de las opciones; para la "evaluación de la calidad general de la clase" (P6) se tiene una escala continua discreta de "notas", con puntaje mínimo 1 y máximo 10 (rango de 9); y para la "evaluación de la capacidad de la clase para motivar la participación" (F2) se utiliza una escala Likert —con enunciados para los niveles de acuerdo sólo en los extremos— de puntaje mínimo 1 para el "muy en desacuerdo" y puntaje máximo de 5 para el "muy de acuerdo" (rango de 4).

La utilización de escalas de distinto rango genera distorsión de las respuestas entregadas por quien responde, como también, afecta los análisis que se hacen de dichos datos. Las respuestas entregadas se distorsionan debido a que los encuestados no responden igual forma a escalas que están medidas de forma distinta. Es por ello que en psicometría se privilegia la utilización de cuantificadores equivalentes para una serie de preguntas con distinto contenido (Cañadas Osinski & Sánchez Bruno, 1998). En términos de análisis, la utilización de escalas de distinto rango afecta la comparación entre los indicadores de evaluación docente, al poseer varianzas que son

[20] Ver marco metodológico.

artificialmente distintas, con origen en la diferencia de rango, y no necesariamente en la dispersión que tengan los fenómenos en la realidad. La lectura y análisis de la varianza y sus comparaciones, por lo tanto, no se puede hacer con única atribución al fenómeno medido —la evaluación docente—, sino que se encuentra contaminada por la forma en que se midió. Este defecto de la medición, junto con la utilización de escalas que son distintas también en fraseo, son dos factores que potencialmente distorsionan el proceso de construcción de dimensiones que se arman a partir de los análisis factoriales confirmatorios que se realizan para este propósito. Como se puede evidenciar mediante la lectura de resultados de los análisis factoriales, las dimensiones resultantes siempre agrupan ítems que fueron preguntados de la misma forma, esto es, con la misma escala. Pese a que lo anterior podría considerarse como evidencia para invalidar y seguidamente desechar los constructos armados mediante los análisis factoriales, se ha decidido conservarlos y continuar los análisis posteriores siguiendo estos resultados, ya que no sólo los estadísticos de ajuste resultantes de los análisis cumplen los requerimientos, sino que también las dimensiones resultantes tienen un sentido sustantivo a la hora de interpretarse como aspectos reales de la evaluación docente.

En relación con la imprecisión conceptual del fenómeno a medir, se puede decir que los conceptos utilizados denotan que aquello que se está midiendo no es directamente el rendimiento "*docente"*, sino el rendimiento "*de la clase"*. Al postularse como evaluación docente, se esperaría que el test y los conceptos utilizados en este aludieran específicamente al rendimiento de los profesores, cuestión que no se cumple. El rendimiento "*de la clase"* es un aspecto mucho más general del proceso de enseñanza, aunque se puede considerar que involucra de forma parcial y se relaciona directamente con el rendimiento docente en sí. Esto, ya que además de evaluar al docente, la evaluación de la clase arrastra consigo una serie de apreciaciones sobre otros elementos exógenos a la calidad docente, como lo son, eventualmente, aspectos contextuales, como el entorno de enseñanza, el propio interés y participación en la clase de los estudiantes, la preferencia de los estudiantes por los contenidos revisados en las clases, etc. Más aún, para la evaluación de la capacidad de la clase para motivar la participación (F2) se pude establecer, sin lugar a dudas, que no se trata de una apreciación que se hace del docente, sino una evaluación de la propia actitud del alumno frente a la clase. Estos ítems se refieren de forma explícita a la participación y motivación personal de los estudiantes, más que a la capacidad que tenga el docente para generar este efecto. Si bien ambos aspectos pueden estar relacionados —y de ahí su capacidad de dar cuenta indirectamente del rendimiento docente—, no son constitutivamente lo mismo.

Al problema conceptual mencionado, se le suman, como ya se ha adelantado, dificultades a nivel de fraseo. El instrumento cuenta con múltiples enunciados innecesariamente extensos (ej. ítem 5) y enunciados con palabras dobles y repetición con sinónimos (ej. ítems 2, 3 y 4). Esta sobre-especificación no contribuye a clarificar ni especificar las preguntas para los respondientes, sino todo lo contrario, da cabida a que se generen distintas interpretaciones sobre qué es aquello que se les pregunta, lo que es un problema, ya que merma la capacidad de que cada respuesta a un mismo ítem mida el mismo fenómeno. Para lograr mediciones con buena fiabilidad y validez, se requieren enunciados sucintos, parsimoniosos, y con la menor cabida posible a la reinterpretación.

v. Otras limitaciones del estudio

El presente estudio trabajó con una muestra pequeña de docentes, altamente heterogénea. Por este motivo, algunos de los análisis bivariados que se pensaron realizar en instancias de diseño resultaron ser inviables en la práctica, debido a la insuficiencia de casos, o bien, concluyeron en la identificación de relaciones entre variables que se encuentran en el borde de la no significancia y/o son de baja intensidad bajo los estándares tradicionales, pero que podrían llegar a evidenciar relaciones más intensas y estadísticamente significativas en condiciones de mayor cantidad de casos.

Un análisis relevante que no se pudo efectuar por el primer motivo antes mencionado, fueron pruebas de asociación entre el tipo de curso (asignaturas de enseñanza de pedagogía, denominadas “didácticas”, de enseñanza de matemáticas, denominadas “disciplinares”, o de tipo “mixto”) y las trayectorias docentes. Este análisis fue requerido expresamente por la parte solicitante del presente estudio, sin embargo, la insuficiencia de casos en cada uno de los tres tipos de cursos hizo inviable los análisis a este respecto.

Además de haber sido una solicitud explícita, un análisis como este resulta atingente debido a que, como se mencionó en los antecedentes, hay evidencia que apunta a que existen sesgos en la evaluación docente según el contenido de sus asignaturas. Sería relevante, por lo tanto, estudiar cómo se expresa este fenómeno entre los distintos cursos de pedagogía en matemáticas, clasificados bajo la forma antes mencionada. Un análisis tal sería de gran utilidad para los docentes del área.

Otro análisis también solicitado, pero que fue irrealizable debido a las condiciones de la muestra, fue el de diferencias en las trayectorias docentes agrupando los cursos por la universidad en la que son ofrecidos. Debido a la alta cantidad de universidades involucradas en el estudio (8) en

proporción a la cantidad de casos (23), este análisis no se pudo realizar. Tampoco fue posible agregar las universidades en "tipos de universidades", ya que no existían criterios de agrupación evidentes para ello.

Otros análisis también relevantes, sin embargo, no considerados por la parte diseñadora del presente estudio, corresponde a la identificación de otros sesgos de evaluación docente. Como se ha visto mediante la revisión de antecedentes, son múltiples los aspectos que ha propuesto la literatura como factores extra docentes que tienden a relacionarse con los puntajes de evaluación docente en base a CED. La influencia de estos aspectos se encuentra íntimamente relacionada con las condiciones de cada contexto de enseñanza específico, por lo que resulta necesario que cada evaluación de rendimiento docente específica considere identificar los niveles de sesgo que posee, al menos, de parte de los principales elementos que han sido propuestos por la literatura. Aspectos como la cantidad de carga académica, expectativas de notas de los alumnos, nivel de especialización de los cursos, su obligatoriedad o electividad, la percepción del prestigio previo de los docentes, sus rasgos de personalidad, y otros elementos, son evidenciados en algunas investigaciones como fuentes de sesgo que para el presente estudio no fueron considerados. Estas omisiones son limitaciones del estudio que hubiesen sido fácilmente superables en caso de haberse reflexionado en este ámbito en fases de diseño o trabajo de campo.

Como ya se ha mencionado anteriormente, este estudio carece de capacidad de entregar medidas de validez de constructo e indicadores de fiabilidad mediante alfa de Cronbach para dos de los cuatro aspectos que se utilizaron para medir el rendimiento docente. Ello, debido a que en fases de diseño no se consideró que era necesario contar con más de un ítem para validar la medición de cada dimensión de constructo. Este estudio también posee la dificultad de asegurar un buen nivel de fiabilidad intra-evaluadores para las evaluaciones de los cursos con poca cantidad de alumnos inscritos. Como ya se revisó, evidencia de otros estudios (Solomon, et al., 1997) indica que se requieren al menos unos 30 casos para poder contar con una evaluación docente con consistencia interna intra-evaluadores alta, y en este caso, tres cuartos del total de los cursos poseen menos de 30 alumnos inscritos. Ambas deficiencias mencionadas implican que se ignoran los niveles de error que poseen las mediciones en estos aspectos. No se descarta la posibilidad de que algunos docentes involucrados en este estudio fuesen mal clasificados en relación a su rendimiento docente real, ya sea siendo clasificados como mejores docentes de lo que son, o bien peores. Pese a ello, esta limitación del estudio no se constituye como un problema totalmente invalidante. Como bien señalan Kane y Staiger (2012), no existe un estándar uniforme de fiabilidad

que se aplique para los diferentes usos que se les dan a estas pruebas. Diferentes usos de la información requieren diferentes estándares de fiabilidad. En este caso, el diagnóstico no posee consecuencias reales sobre los docentes involucrados, ni tampoco pretende ser el producto mediante el cual se les entregue formalmente su retroalimentación sobre su rendimiento docente, por lo que deficiencias en los niveles de validez y fiabilidad alcanzados no son de gran gravedad.

Cabe también aclarar que este estudio no logra, ni tampoco pretende ofrecer, un diagnóstico del desempeño docente generalizado en la educación superior chilena. Ello, porque corresponde a una pequeña muestra de docentes universitarios, que no es posible asegurar que represente el rendimiento general de los docentes universitarios chilenos, y también, debido a que esta evaluación se ajusta a un currículo de enseñanza específico, de educadores en pedagogía básica en matemáticas.

vi. Recomendaciones para el mejoramiento de los cuestionarios de evaluación docente

Con los resultados que posee este estudio, se podría fácilmente establecer que existe un escenario de muy buen rendimiento docente entre los cursos estudiados. Sin embargo, se adoptará ahora una postura escéptica sobre este punto, poniendo en tela de juicio los buenos resultados, y sugiriendo mayor exigencia a los instrumentos de medida.

Dándole cabida a esta postura crítica, es posible señalar que existe una necesidad de formular mediciones más exigentes, que logren diferenciar rendimientos al interior de un grupo que, tal como está siendo medido actualmente, aparece como demasiado bueno y muy homogéneo. En efecto, como bien señalan esfuerzos por mejorar la calidad de la instrucción docente tan importantes como el proyecto MET, financiado por la Bill & Melinda Gates Foundation, indicar que la totalidad de los docentes son satisfactorios no genera beneficios para nadie, ni siquiera para los docentes, no produce un efecto de mejoría sobre la enseñanza, y resulta una real inconsistencia a la luz de las disparidades existentes en el rendimiento de los alumnos (Kane & Staiger, 2012).

A favor de la postura escéptica que se propone, se ha evidenciado que mediciones más genéricas, que no consultan sobre elementos de la docencia o el aprendizaje logrado en específico, tienden a tener mejor evaluación y mayor estabilidad. Por ello, se postula que una aproximación que exija a los respondientes evaluar enunciados menos abstractos sobre el rendimiento docente, y más bien evalúen a los docentes en situaciones concretas, a nivel de comportamientos y conductas docentes esperables, podría tener una capacidad mayor de discriminar entre los docentes.

Formular instrumentos donde se evalúen aspectos concretos, como hechos o prácticas docentes (ej. “suele ejemplificar al explicar”), en vez de aspectos más genéricos (ej. “explica con claridad”) produce, además, una mejoría en los niveles de validez y fiabilidad de las mediciones.

Se mejora la validez de contenido de los instrumentos ya que en la medida que una determinada comunidad educativa (docentes, estudiantes, autoridades educativas y diseñadores de instrumentos) busca encontrar acuerdos sobre qué se entiende por “buena docencia” y con ello logra delimitar una serie de conductas a medir, se pueden formular instrumentos de evaluación docente que al evaluar hechos o prácticas concretas dan menor cabida múltiples interpretaciones conceptuales sobre qué es la calidad docente, eventualmente divergentes. Junto con ello, se mejoran los niveles de fiabilidad de las medidas, ya que hay menor cabida a que los estudiantes respondan a las distintas preguntas ofreciendo su propia interpretación del rendimiento de sus docentes. Deben evaluar hechos, y no aspectos conceptuales, lo que logra rebajar el efecto de su satisfacción global/superficial con el docente o de sus creencias idiosincráticas sobre qué es la buena enseñanza.

Ahora bien, la estrategia de formulación de indicadores específicos iría en contraposición con posturas como las de Abrami et al. (1981) o Apodaca y Grad (2005), quienes sugieren que para rescatar fielmente las capacidades docentes, se debe cuestionar a los estudiantes respecto al rendimiento general del docente, no focalizando según aspectos. Ahora bien, la postura del presente estudio es que el problema al que apuntan los autores que alegan sobre la invalidez conceptual de las propuestas dimensionales, no es distinto al que podría atribuírsele a toda medición cuantitativa, referente a la real cercanía que tiene toda medición con el fenómeno que busca rescatar. Epistemológicamente, este debate no posee una solución incondicional, ya que toda observación científica posee, inevitablemente, un punto ciego (Maturana, et al., 1994) y es intrínsecamente una observación situada (García, 2012).

Desde la postura de esta investigación, se ignora si realmente se rescatan con las mediciones dimensionales lo que intrínsecamente es la calidad docente. Sin embargo, se halla en este tipo de mediciones un valor práctico que no poseen las mediciones generales y unidimensionales. La postura que se ha adquirido en este estudio, basada en la evidencia empírica recabada, es que los instrumentos que incluyen dimensionalidad permiten representar las diferencias entre los docentes con mayor claridad y, por consiguiente, indican hacia qué aspectos de la docencia podrían concentrarse los esfuerzos por mejorar. No se descarta que las mediciones sobre calidad docente hechas por alumnos posean sesgos extra docentes como los que se han discutido a lo

largo de este trabajo, o bien, se encuentren contaminadas por apreciaciones cercanas —pero no realmente reflejo— de la calidad docente, como podría ser la satisfacción con la clase. De hecho, el autor de la presente tesis ya ha reconocido con anterioridad, cuando se ha debatido sobre la imprecisión conceptual de los ítems, que las mediciones con las que se trabaja incorporan este problema, sin embargo, no por ello se ha descartado que las mediciones posean utilidad.

Tal como ha sugerido la literatura, se propone que frente a esta duda por la validez de los instrumentos de evaluación docente, en primer lugar, se aumenten los esfuerzos por formular buenos instrumentos de medición, consensuados entre los distintos actores del proceso de enseñanza, y las determinaciones que se tomen con los resultados se acompañen de más evidencia e información cuando se trate de determinaciones sumativas de alta connotación, como por ejemplo, la contratación, recompensación, o despido de docentes.

Respecto a nuestra insistencia por formular mediciones más exigentes, se propone la estrategia de utilizar preguntas redactadas de forma opuesta a como se acostumbra a hacer: ya no consultando por el buen rendimiento, sino por las carencias en el rendimiento de cada docente, en caso de haberlas. Si bien puede haber posturas en contra de esta fórmula, la postura de esta investigación es que la esencia de todo esfuerzo por mejorar la enseñanza es la identificación y posterior trabajo sobre los aspectos problemáticos de la instrucción, como bien señala Hativa (1996). Con ello en mente, fácilmente se puede considerar que todo lo que se suele hacer con los instrumentos de evaluación tradicionales es buscar los aspectos problemáticos de forma inversa, es decir, observando el buen rendimiento y atribuyéndole a las diferencias de rendimiento el espacio de las falencias. La fórmula que aquí se propone podría parecer controversial, pero en contextos donde la docencia ha alcanzado buenos niveles aparentes, la forma de identificar las carencias que aún permanecen sería buscándolas de forma explícita.

Ejemplo de lo que se propone pueden ser mediciones de la frecuencia o acuerdo con que “el docente fue poco claro al explicar contenidos”, en reemplazo de ítems como “el docente es claro al explicar contenidos”, o medidas de la frecuencia en que “el docente tuvo poca disposición para resolver dudas, en reemplazo de medidas como “el docente tiene buena disposición para resolver dudas”.

Se propone como hipótesis, que los niveles de fiabilidad de las medidas que se lograrían con nuestra formula serían mayores que los logrados mediante los fraseos típicos, ya que para que un respondiente puntúe un ítem formulado bajo este formato, requiere de su parte mayor nivel de reflexión sobre su experiencia con el docente y mayor uso de la memoria que lo requerido con la

fórmula tradicional. Adicionalmente, esta propuesta podría tener no sólo un efecto sobre la mejor identificación de las carencias docentes, sino también sobre la consecuente mejoría que tendría el rendimiento de sus estudiantes. Hay evidencia de otros estudios que señala que los docentes que han identificado de forma explícita sus debilidades y las han trabajado, mejoran sus posibilidades de afectar positivamente el rendimiento de sus estudiantes (Marsh & Roche, 1993).

Finalmente, otro beneficio más de esta fórmula es que evita la duda sobre la presunta inviabilidad de medir la calidad docente debido a la ausencia de claridad y consenso sobre qué es la calidad docente en sí. Como ya se trató anteriormente, para algunos investigadores, la falta de claridad sobre el concepto de calidad docente invalida la mayoría de las formas de medir el rendimiento docente. Ahora bien, siempre que resulta complejo definir un concepto, resulta mucho más fácil delimitar qué no es. Con ello en mente, parece una fórmula más fiable identificar las conductas no buscadas o reprobables en la práctica de la docencia, más que buscar definir qué es la buena docencia y observar si es que los docentes cumplen con ello.

Respecto a en qué momento realizar las evaluaciones docentes, también se obtienen aprendizajes a partir del presente estudio. Para los cursos en que las evaluaciones docentes se hacen una vez en el semestre, como es en la mayoría de los casos, se advierte que una eventual aplicación de estos instrumentos inmediatamente sobre los cierres de semestre podría incorporar una alteración en las respuestas. Esto, ya que como se ha visto, las respuestas a las evaluaciones docentes hechas en las últimas clases rescatarían con mayor intensidad la apreciación que se tiene de estas clases en sí, y el rendimiento de la evaluación docente de las últimas clases tiende a ser anómalo si se revisa en contraste con su comportamiento medio semestral. Se recomienda, entonces, no aplicar instrumentos de evaluación docente justo luego de los cierres de semestre, toda vez que lo que se busca evaluar con este tipo de instrumentos es la apreciación que se tiene del conjunto del proceso semestral. Serían mejores momentos para aplicar evaluaciones docentes, a mitad de semestre, o bien a comienzos de un siguiente semestre, donde la apreciación de los estudiantes de los cursos se haya cristalizado, eventualmente, en una apreciación del conjunto y no exclusivamente del final de los cursos.

También, se ofrecen a continuación propuestas para los estudios que contemplen mediciones con un formato similar al del presente estudio, es decir, intra-semestre. Para estudios tales, cabe reflexionar sobre la necesidad de realizar mediciones de evaluación docente cada clase. Desde una observación de los propios datos, no hay motivo para pensar que las variaciones en el rendimiento docente se dan fuerte o prioritariamente entre clase y otra. En caso de haber variaciones, los

cambios de trayectorias son más bien procesos lentos. Ello se evidencia en la medida que los indicadores de inestabilidad y mejoría/empeoramiento analizados poseen magnitudes leves. Por lo demás, realizar mediciones sólo cada cierto tiempo, y no todas las clases, parece ser un abordaje suficiente y aceptable dentro del campo de estudio, no habiendo críticas o discusiones acerca de la necesidad de realizar mediciones más frecuentes. La tendencia identificada mediante la propia revisión de los estudios longitudinales en evaluación docente es que los estudios de pocos años (ej. 2 años) han trabajado comúnmente con dos mediciones por semestre, una a mitad y otra final, y los estudios de mayor cantidad de años tienden a efectuar una sola medición por semestre.

Cabe advertir que hay posturas que señalan que aplicaciones demasiado recurrentes de los mismos cuestionarios sobre una misma muestra, pueden introducir problemas de exceso o "evaluation overkill" (Irby, et al., 1977), que refiere a la saturación y acostumbramiento que adoptan los respondientes frente a los cuestionarios repetitivos, problemas que reducen las posibilidades de medir de forma fiable y lograr registrar un eventual cambio en los puntajes. Haciéndose parte de la advertencia, el autor presente estudio considera que la medición día a día debe realizarse cuando sea justificada de forma necesaria frente a otras mediciones más aplazadas como sería, por ejemplo, cuando las investigaciones sean de periodo corto y se requieran aun así múltiples mediciones en el tiempo. No valdría la pena, por ejemplo, efectuar mediciones todos los días en estudios que se llevan durante años.

Ahora bien, para estudios que, de todas formas, contemplen mediciones cada día de clase, también se postulan recomendaciones. Con el fin de obtener trayectorias longitudinales más fiables, se propone un fraseo que, en vez de preguntar por "la clase del día de hoy", se refiera a cómo se evalúa el logro de los cursos o el rendimiento de los docentes "en lo que lleva el semestre". Con este formato de fraseo, las diferencias entre las puntuaciones indicarían cómo ha variado el rendimiento en relación a todo el proceso semestral anterior, lográndose así percibir con mayor precisión los cambios de trayectoria en comparación con lo que logran las mediciones como están formuladas actualmente en este estudio, en que el cambio de trayectoria se analiza mediante el diferencial de logro entre clase y otra. Dicho cambio de fraseo otorgará la posibilidad de aprehender de mejor forma las variaciones reales de las trayectorias evaluación a lo largo del semestre, colaborando a evitar el desequilibrio de las puntuaciones entre clase y otra que, como se ha visto mediante los análisis de inestabilidad de trayectorias, no parece ser atribuible a cambios en el desempeño docente.

vii. Ideas y propuestas para estudios a futuro

Para finalizar este estudio, se exponen breves sugerencias e ideas sobre la forma en que esta investigación podría continuarse, o bien complementarse.

La aproximación a los datos que tuvo este estudio fue considerar a los cursos como unidad de análisis. Ello, si bien permite abordar la generalidad de los docentes en su conjunto, constituye solamente un primer acercamiento de lo que se podría lograr con los datos generados en la investigación en que se enmarcó esta tesis. Observar a los alumnos respondientes, cada uno como unidad de análisis, es un procedimiento que lograría mejores niveles de fiabilidad (Canaday, et al., 1978; West, 1988), además de proporcionar más conocimientos, como, por ejemplo, en la búsqueda de efectos de sesgo de género cruzado (estudiante-docente) y otros sesgos asociados específicamente a características de los estudiantes (expectativas de nota, percepción de carga académica, etc.). Sería posible, también, enriquecer algunas de las variables extra docencia, como la asistencia longitudinal de cada estudiante a las clases, en vistas de observar su relación con la evaluación docente.

También resulta muy relevante para la muestra específica de este estudio, comparar su rendimiento docente transversal y longitudinal con otras carreras o disciplinas. Ello, a propósito de que existe evidencia que específicamente los cursos de matemáticas son evaluados con menores puntajes que cursos con otros contenidos (Feldman, 1978; Cashin, 1990).

Cabe agregar también que, a propósito de que Chile es un país donde el rendimiento educacional de los estudiantes se encuentra muy correlacionado con su estatus socioeconómico, y el sistema educacional, sobre todo el básico, se encuentra muy segregado en este sentido (Valenzuela, et al., 2014), resulta relevante incorporar a los estudios de sesgos sobre la evaluación docente, elementos socioeconómicos. Elementos como el ingreso del hogar, el ingreso de los pares, el grado de ruralidad y el tipo/localización de establecimiento educativo han probado ser elementos que diferencian los rendimientos entre el estudiantado chileno, por lo que bien podrían ser sesgos que se relacionen con la evaluación docente en la educación superior.

Finalmente cabe abordar una solución bastante satisfactoria a la constante problemática de asegurar la validez y la fiabilidad de las mediciones de rendimiento docente, encontrada en la literatura más reciente y especializada sobre el tema. Estudios recientes han trabajado complementando los CED con medidas observacionales y VAMs (“value added scores”). Estos estudios han evidenciado que los niveles de fiabilidad que se logran son mucho más altos que

cualquiera de los de estas medidas por separado, tanto que logran una gran capacidad predictiva sobre el rendimiento estudiantil (Kane & Staiger, 2012; Polikoff, 2015). Además, estas mediciones se complementan, reforzándose entre ellas en las debilidades de cada una, y colaborando con sus especificidades. Por ejemplo, las medidas construidas mediante CED ofrecen una fiabilidad más alta, en tanto cada puntaje es construido en base a juicios de múltiples estudiantes. Las medidas observacionales, si bien se construyen en base a una sola observación y por ello poseen los menores índices de fiabilidad de las tres medidas, además de registrar el rendimiento docente, logran aprehender cualitativamente el proceso de enseñanza, no sólo identificando el buen o mal rendimiento docente, sino también describiéndolo e identificando sus componentes. Ello es muy útil cuando se quiere pasar del diagnóstico a la capacitación. Por otro lado, las medidas en base a VAMs, toda vez que estudian directamente el aprendizaje de los estudiantes y poseen los mejores índices de fiabilidad de las tres medidas (Polikoff, 2015), no ofrecen mayor conocimiento sobre el proceso de enseñanza, sus aspectos, o bien lineamientos sobre cómo mejorarlo.

Estudios futuros podrían adoptar esta metodología integradora, tanto para validar mejor sus resultados, como también para poseer más entendimiento sobre cómo funciona el proceso de la docencia e identificar qué produce mejores resultados en los estudiantes, que es la finalidad última de todo estudio que se realiza en función de evaluar la calidad de la docencia.

VIII. Bibliografía

Abrami, P. C., 1989. Review: SEEQing the Truth about Student Ratings of Instruction. *Educational Researcher,* 18(1), pp. 43-45.

Abrami, P. C., Leventhal, L. & Dickens, W. J., 1981. Multidimensionality of student ratings of instruction. *Instructional Evaluation,* 6(1), pp. 12-17.

Abrami, P. & D'Apollonia, S., 1991. Multidimensional students' evaluations of teaching effectiveness – generalizability of "N=1" research: comment on Marsh (1991). *Journal of Educational Psychology,* 83(3), pp. 411-415.

Acevedo Álvarez, R. & Fernández Díaz, M. J., 2004. La percepción de los estudiantes universitarios en la medida de la competencia docente: validación de una escala. *Revista de Educación,* 28(2), pp. 145-166.

Acevedo Álvarez, R. & Mairena Rodríguez, N., 2006. Factores de sesgo asociados a la validez de la evaluación docente universitaria: un modelo jerárquico lineal. *Archivos Analíticos de Políticas Educativas,* 14(34), pp. 1-22.

Albright, J. J. & Park, H. M., 2009. *Confirmatory Factor AnalysisUsing Amos, LISREL, Mplus, and SAS/STAT CALIS (Working Paper),* Indiana University: The University Information Technology Services (UITS) Center for Statistical and Mathematical Computing.

Aleamoni, L. M., 1997. Student Rating Myths Versus Research Facts from 1924 to 1998. *Journal of Personnel Evaluation in Education,* 13(2), pp. 153-166.

Apodaca, P. & Grad, H., 2005. The dimensionality of student ratings of teaching: integration of uni- and multidimensional models. *Studies in Higher Education,* 30(6), p. 723–748.

Arámburo Vizcarra, V. & Luna Serrano, E., 2013. La influencia de las características, profesor y del curso en los puntajes de evaluación docente. *Revista mexicana de investigación educativa,* 18(58 julio-septiembre), pp. 949-968.

Arbesú, I. y otros, 2006. La evaluación de la docencia universitaria en México: Un estado de conocimiento del período 1990-2004. *Perspectiva Educacional ,* Volumen 48, pp. 27-58.

Arbesú, M. I. & Rueda, M., 2003. La evaluación de la docencia desde la perspectiva del propio docente. *Reencuentro,* Volumen 36, pp. 56-64.

Asun, R. & Zúñiga, C., 2017. Evaluación docente universitaria: Hacia una perspectiva unificada. *Revista de Sociología,* 32(1), pp. 50-70.

Ávalos, B., 2013. *¿Héroes o villanos? La profesión docente en Chile.* Santiago de Chile: Editorial Universitaria.

Ávalos, B., 2014. La formación inicial docente en Chile: Tensiones entre políticas de apoyo y control. *Estudios pedagógicos (Valdivia),* 40(Especial), pp. 11-28.

Barber, M. & Mourshed, M., 2007. *How the World's Best-Performing School Systems Come Out On Top,* McKinsey & Company: Social Sector Office.

Bausell, R. B., Schwartz, S. & Purohit, A., 1975. An Examination of the Conditions under Which Various Student Rating Parameters Replicate across Time. *Journal of Educational Measurement,* 12(4), pp. 273-280.

Beecham, R., 2009. Teaching quality and student satisfaction: nexus or simulacrum?. *London Review of Education,* 7(2), pp. 135-146.

Belenky, M. E., Clinchy, B. M., Goldberger, N. R. & Tarule, J. M., 1986. *Women's ways of knowing.* New York: Basic Books.

Bill & Melinda Gates Foundation, 2010. *Learning about Teaching. Initial Findings from the Measures of Effective Teaching Project,* Bill & Melinda Gates Foundation: Research Paper - MET Project.

Boring, A., 2016. Gender biases in student evaluations of teaching. *Journal of Public Economics,* Volumen 145, pp. 27-41.

Boring, A., Ottoboni, K. & Stark, P. B., 2016. Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research – Section: SOR-EDU.*

Boyd, D. y otros, 2009. Teacher Preparation and Student Achievement. *Educational Evaluation and Policy Analysis,* 31(4), pp. 416-440.

Buchert, S., Laws, E. L., Apperson, J. M. & Bregman, N. J., 2008. First impressions and professor reputation: influence on student evaluations of instruction. *Social Psychology of Education,* 11(4), p. 397–408.

Bunge, M., 1973. *La ciencia, su método y filosofía.* Buenos Aires: Siglo XX.

Canaday, S. D., Mendelson, M. A. & Hardin, J. H., 1978. The effect of timing on the validity of student ratings. *Academic Medicine,* Volumen 53, pp. 958-964.

Cánovas, L. y otros, 2009. *Desempeño docente: elementos conceptuales, indicadores y estrategias de la encuesta a estudiantes de la Universidad Nacional de Cuyo,* Universidad Nacional de Cuyo - Secretaría de Ciencia, Técnica y Posgrado: Proyectos de Investigación 2007-2009.

Cañadas Osinski, I. & Sánchez Bruno, A., 1998. Categorías de respuesta en escalas tipo Likert. *Psicothema,* 10(3), pp. 623-631.

Cashin, W., 1990. Students do rate different academic fields differently. *New Directions for Teaching and Learning,* Volumen 43, p. 113–121.

Cashin, W. E., 1995. *Student Ratings of Teaching: The Research Revisited. IDEA Paper No. 32.,* Manhattan: Center for Faculty Evaluation and Development in Higher Education, Kansas State University.

Centra, J. A., 2003. Will Teachers Receive Higher Student Evaluations by Giving Higher Grades and Less Course Work?. *Research in Higher Education,* 44(5), p. 495–518.

Centra, J. A. & Gaubatz, N. B., 2000. Is There Gender Bias in Student Evaluations of Teaching?. *The Journal of Higher Education,* 71(1), pp. 17-33.

Chonko, L. B., Tanner, J. F. & Davis, R., 2002. What Are They Thinking? Students' Expectations and Self-Assessments. *Journal of Education for Business,* 77(5), pp. 271-281.

CINDA, 2007. *Evaluación del desempeño docente y calidad de la docencia universitaria,* Centro Interuniversitario de Desarrollo (CINDA). Grupo operativo de Universidades Chilenas - Fondo de Desarrollo Institucional: Mineduc - Santiago de Chile.

Clayson, D. E. & Sheffet, M. J., 2006. Personality and the Student Evaluation of Teaching. *Journal of Marketing Education,* 28(2), pp. 149-160.

CNED, 2017a. *Tendencias de Educación Superior 2017,* Consejo Nacional de Educación (CNED): Junio de 2017.

CNED, 2017b. *Matrícula total por Área del Conocimiento, años 2005-2017.* [En línea] Available at: https://www.cned.cl/indices/matricula-sistema-de-educacion-superior [Último acceso: 01 02 2018].

Cornejo, R. y otros, 2009. *Bienestar/malestar docente y condiciones de trabajo en profesores de enseñanza media de Santiago,* Ministerio de Educación, Gobierno de Chile: Fondo de Investigación y Desarrollo de la Educación, FONIDE, Proyecto nº59.

Costin, F., Greenough, W. T. & Menges, R. J., 1971. Student Ratings of College Teaching: Reliability, Validity, and Usefulness. *Review of Educational Research,* 41(5), pp. 511-535.

Cox, C., Meckes, L. & Bascopé, M., 2010. La institucionalidad formadora de profesores en Chile en la década del 2000: velocidad del mercado y parsimonia de las políticas. *Pensamiento Educativo,* 46(1), pp. 205-245.

Crumbley, L. C., Flinn, R. E. & Reichelt, K. J., 2010. What is ethical about grade inflation and coursework deflation?. *Journal of Academic Ethics,* Volumen 8, p. 187–197.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E. & Rothstein, J., 2012. Evaluating Teacher Evaluation. *Phi Delta Kappan ,* 93(6), pp. 8-15.

Domínguez, S., 2012. Propuesta para el cálculo del Alfa Ordinal y Theta de Armor. *Revista IIPSI,* 15(1), pp. 213-217.

Duque, L. C., 2013. A framework for analyzing higher education performance: Students' satisfaction, perceived learning outcomes, and dropout intention. *Total Quality Management and Business Excellence,* 25(1-2), pp. 1-21.

Eiszler, C. F., 2002. College Students' Evaluations of Teaching and Grade Inflation. *Research in Higher Education,* 43(4), p. 483–501.

Ewing, A. M., 2012. Estimating the impact of relative expected grade on student evaluations of teachers. *Economics of Education Review,* 31(1), pp. 141-154.

Feldman, K., 1978. Course characteristics and college students' ratings of their own teachers: What we know and what we don't. *Research in Higher Education,* Volumen 9, p. 199–242.

Feldman, K. A., 1976. Grades and college students' evaluations of their courses and teachers. *Research in Higher Education,* 4(1), p. 69–111.

Feldman, K. A., 1984. Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher Education,* 21(1), p. 45–116.

Feldman, K. A., 1992. College Students' Views of Male and Female College Teachers: Part I: Evidence from the Social Laboratory and Experiments. *Research in Higher Education,* 33(3), pp. 317-375.

Fernández, J., Mateo, M. A. & Muniz, J., 1998. Is there a Relationship between Class Size and Student Ratings of Teaching Quality?. *Educational and Psychological Measurement,* 58(4), pp. 596-604.

Freiberg Hoffmann, A., Stover, J. B., De la Iglesia, G. & Fernández Liporace, M., 2013. Correlaciones policóricas y tetracóricas en estudios factoriales exploratorios y confirmatorios. *Cienc. Psicol.,* 7(2), pp. 151-164.

García Garduño, J. M., 2000. ¿Qué factores extraclase o sesgos afectan la evaluación docente en la educación superior?. *Revista Mexicana de Investigación Educativa,* 5(10 julio-diciembre), pp. 303-325.

García Garduño, J. M., 2003. Profesores universitarios y su efectividad docente: Un estudio comparativo entre México y Estados Unidos.. *Perfiles educativos,* 25(100), pp. 42-55.

García Garduño, J. M., 2008. El proceso perverso de la evaluación de la docencia en las universidades: un balance inicial y apuntes para mejorarlo. *Reencuentro,* Volumen 53, pp. 9-19.

García, J. F., 2012. *Hacia una razón stuada.* Santiago de Chile: LOM Ediciones.

Gillmore, G. M., Kane, M. T. & Naccarato, R. W., 1978. The generalization of student ratings of instruction: Estimation of the teacher and course components. *Journal of Educational Measurement,* 15(1), p. 1–13.

González López, I., López Cámara, A. & Nail Kroyer, I., 2016. Claves de Compround para la redefinición del modelo de evaluación de la calidad docente en la Universidad de Concepción. *Estudios Pedagógicos,* 50(4), pp. 69-85.

Good, T. L. & Lavigne, A. L., 2015. Issues of teacher performance stability are not new: Limitations and possibilities. *Education Policy Analysis Archives,* 23(2).

Gray, M. & Bergmann, B. R., 2003. Student Teaching Evaluations: Inaccurate, Demeaning, Misused. *Academe,* 89(5), pp. 44-46.

Greenwald, A., 1997. Validity Concerns and Usefulness of Student Rating of Instruction. *American Psychologist,* 52(11), pp. 1182-1186.

Greenwald, A. & Gillmore, G., 1997. No pain, no gain? The importance of measuring course workload in student ratings of instruction. *Journal of Educational Psychology,* 89(4), pp. 743-751.

Griffin, B. W., 2001. Instructor Reputation and Student Ratings of Instruction. *Contemporary Educational Psychology,* 26(4), pp. 534-552.

Guarino, C. M., Reckase, M. D. & Wooldridge, J. M., 2015. Can Value-Added Measures of Teacher Performance Be Trusted?. *Education Finance and Policy,* 10(1), pp. 117-156.

Hanges, P. J., Schneider, B. & Niles, K., 1990. Stability of Performance: An Interactionist Perspective. *Journal of Applied Psychology,* 75(6), pp. 658-667.

Harvey, L., 2001. *Student Feedback: a report to the Higher Education Funding Council for England,* Birmingham: Centre for Research into Quality, University of Central England.

Hativa, N., 1996. University instructors' ratings profiles: Stability over time, and disciplinary differences. *Research in Higher Education,* 37(3).

Hativa, N. & Raviv, A., 1993. Using a single score for summative teacher evaluation by students. *Research in Higher Education,* 34(5), pp. 625-646.

Hattie, J. A. C., 2009. *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* London: Routledge.

Hill, H. C., Charalambous, C. Y. & Kraft, M. A., 2012. When Rater Reliability Is Not Enough: Teacher Observation Systems and a Case for the Generalizability Study. *Educational Researcher,* 41(2), pp. 56-64.

Hills, S., Naegle, N. & Bartkus, K., 2009. How Important Are Items on a Student Evaluation? A Study of Item Salience. *Journal of Education for Business,* 84(5), pp. 297-303.

Ho, A. D. & Kane, T. J., 2013. *The Reliability of Classroom Observations by School Personnel,* Bill & Melinda Gates Foundation: MET Project.

Hornstein, H. A., 2017. Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education,* 4(1).

Irby, D. y otros, 1977. The use of student ratings in multiinstructor courses. *Journal of Medical Education,* 52(8), pp. 668-673.

Johnson, R., 2000. The Authority of the Student Evaluation Questionnaire. *Teaching in Higher Education,* 5(4), pp. 419-434.

Kane, T. J. & Staiger, D. O., 2012. *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper,* Bill & Melinda Gates Foundation: MET Project.

Kember, D. & Wong, A., 2000. Implications for evaluation from a study of students' perceptions of good and poor teaching. *Higher Education,* 40(1), p. 69–97.

Klajman, G., 1997. *Nightmares of academic assessment. ASSESS - Assessment in Higher Education. ASSESS Archives.* [En línea].

Kline, R. B., 2011. *Principles and Practice of Structural Equation Modeling.* 3era ed. New York - London: The Guilford Press.

Knapper, C., 2001. Broadening our approach to teaching evaluation. *New Directions for Teaching and Learning,* Volumen 88, pp. 3-9.

Kogan, L. R., Schoenfeld-Tacher, R. & Hellyer, P. W., 2010. Student evaluations of teaching: perceptions of faculty based on gender, position, and rank. *Teaching in Higher Education,* 15(6), pp. 623-636.

Kohlan, R. G., 1973. A Comparison of Faculty Evaluations Early and Late in the Course. *The Journal of Higher Education,* 44(8), pp. 587-595.

Konstantopoulos, S., 2014. Teacher effects, value-added models, and accountability. *Teachers College Record,* 116(1).

Koushki, P. A. & Kuhn, H. A. J., 1982. How reliable are student evaluations of teachers?. *Engineering Education,* Volumen 72, pp. 362-367.

Krantz-Girod, C. y otros, 2004. Stability of Repeated Student Evaluations of Teaching in the Second Preclinical Year of a Medical Curriculum. *Assessment & Evaluation in Higher Education,* 29(1), pp. 123-133.

Kulik, J. A., 2001. Student Ratings: Validity, Utility, and Controversy. *New Directions for Institutional Research,* 2001(109), p. 9–25.

Larrondo, T. y otros, 2007. Desarrollo de habilidades básicas en lenguaje y matemáticas en egresados de pedagogía. Un estudio comparativo. *Calidad en la Educación,* Volumen 27, pp. 150-176.

Larson, J. R., 1979. The Limited Utility of Factor Analytic Techniques for the Study of Implicit Theories in Student Ratings of Teacher Behavior. *American Educational Research Journal,* 16(2), pp. 201-211.

Latiesa, M., 2000. Validez y fiabilidad de las observaciones sociológicas. En: M. García Ferrando, J. Ibáñez & F. Alvira, edits. *El análisis de la realidad social. Métodos y técnicas de investigación.* 3ª ed. Madrid, España: Alianza Editorial, p. 409.

Lavigne, A. & Good, T., 2015. *Improving Teaching Through Observation and Feedback: Beyond State and Federal Mandates.* New York: Routledge.

Luna Serrano, E. & Torquemada, A. D., 2008. Los cuestionarios de evaluación de la docencia por los alumnos: balance y perspectivas de su agenda. *Revista electrónica de investigación educativa,* Volumen 10, pp. 1-15.

Manzi, J., González, R. & Sun, Y., 2011. *La Evaluación Docente en Chile,* Facultad de Ciencias Sociales, Escuela de Psicología - Universidad Católica: Mide UC.

Manzi, J. y otros, 2011a. *¿Qué características de la formación inicial de los docentes se asocian a mayores avances en su aprendizaje de conocimientos disciplinarios?,* Ministerio de Educación - Chile: Fondo de Investigación y Desarrollo en Educación - FONIDE.

Marsh, H., 1984. Students' evaluations of university teaching: Dimensionality, reliability, validity, potential baises, and utility. *Journal of Educational Psychology,* 76(5), pp. 707-754.

Marsh, H., 1987. Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational,* 11(3), pp. 253-387.

Marsh, H., 2007a. Do University Teachers Become More Effective With Experience? A Multilevel Growth Model of Students' Evaluations of Teaching Over 13 Years. *Journal of Educational Psychology,* 99(4), pp. 775-790.

Marsh, H., 2007b. Students' evaluations of university teaching: A multidimensional perspective. En: R. P. Perry & J. C. Smart, edits. *The scholarship of teaching and learning in higher education: An evidence based perspective.* New York: Springer, p. 319–384.

Marsh, H. W. & Hocevar, D., 1991. Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching and Teacher Education,* 7(4), pp. 303-314.

Marsh, H. W. & Roche, L. A., 1993. The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal,* 30(1), pp. 217-251.

Marsh, H. W. & Roche, L. A., 1997. Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist,* 52(11), pp. 1187-1197.

Marsh, H. W. & Roche, L. A., 2000. Effects of grading leniency and low workload on students' evaluations of teaching: popular myth, bias, validity, or innocent bystanders?. *Journal of Educational Psychology,* 92(1), pp. 202-228.

Maturana, H., Varela, F. & Behncke, R., 1994. *El árbol del conocimiento: las bases biológicas del entendimiento humano.* Santiago de Chile: Universitaria.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R. & Mihaly, K., 2009. The Intertemporal Variability of Teacher Effect Estimates. *Education Finance and Policy,* 4(4), pp. 572-606.

McKeachie, W. J., 1990. Research on college teaching: The historical background. *Journal of Educational Psychology,* 82(2), pp. 189-200.

Medel, R., 2013. *Sesgos en la evaluación docente: factores de género y ciclo de estudio. El caso de la facultad de ciencias sociales en la Universidad de Chile.* Santiago de Chile: Tesis para optar al título de Sociólogo.

Medel, R. & Asun, R., 2014. Encuestas de evaluación docente y sesgos de género: un estudio exploratorio. *Calidad en la Educación,* 2014(julio 40), pp. 171-199.

Ministerio de Educación, 2017. *Educación para la Igualdad de Género - Plan 2015-2018,* Santiago de Chile: Unidad de Equidad de Género - Ministerio de Educación - Gobierno de Chile.

Montoya, J., Arbesú, I., Contreras, G. & Conzuelo, S., 2014. Evaluación de la docencia universitaria en México, Chile y Colombia: Análisis de experiencias. *Revista Iberoamericana de Evaluación Educativa,* 7(2e), pp. 15-42.

Morgan, G. B., Hodge, K. J., Trepinksi, T. M. & Anderson, L. W., 2014. The stability of teacher performance and effectiveness: Implications for policies concerning teacher evaluation. *Education Policy Analysis Archives,* 22(95).

Murray, H., 2007. Low-inference teaching behavior and college teaching effectiveness: recent developments and controversies. En: *The scholarship of teaching and learning in Higher education: an evidence-based perspective.* Dordrecht, Países Bajos: Springer, pp. 145-183.

Murray, H. G., Rushton, J. P. & Paunonen, S. V., 1990. Teacher personality traits and student instructional ratings in six types of university courses. *Journal of Educational Psychology,* 82(2), pp. 250-261.

Nye, B., Konstantopoulos, S. & Hedges, L., 2004. How Large Are Teacher Effects?. *Educational Evaluation and Policy Analysis,* XXVI(3), pp. 237-257.

OCDE, 2004. *Reviews of National Policies for Education: Chile,* Organisation for Economic Co-operation and Development: OCDE Publishing.

OCDE, 2013a. *Synergies for Better Learning. An International Perspective on Evaluation and Assessment,* Organisation for Economic Co-operation and Development: Reviews of Evaluation and Assessment in Education.

OCDE, 2013b. *Teachers for the 21st Century: Using Evaluation to Improve Teaching,* Organisation for Economic Co-operation and Development: Background Report for the 2013 International Summit on the Teaching Profession.

OCDE, 2015. *Panorama de la educación 2015 - Indicadores de la OCDE,* España: Fundación Santillana.

Onwuegbuzie, A. J. y otros, 2007. Students' Perceptions of Characteristics of Effective College Teachers: A Validity Study of a Teaching Evaluation Form Using a Mixed-Methods Analysis. *American Educational Research Journal ,* 44(1), pp. 113-160.

Ory, J. C., 2001. Faculty Thoughts and Concerns About Student Ratings. *New Directions for Teaching and Learning,* Volumen 87, p. 3–15.

Overall, J. U. & Marsh, H. W., 1979. Midterm feedback from students: Its relationship to instructional improvement and students' cognitive and affective outcomes. *Journal of Educational Psychology,* 71(6), pp. 856-865.

Overall, J. U. & Marsh, H. W., 1980. Students' evaluations of instruction: A longitudinal study of their stability. *Journal of Educational Psychology,* 72(3), p. 321–325.

Parsons, T. & Platt, G., 1973. *The American University.* Harvard: Harvard University Press.

Patrick, H. & Mantzicopoulos, P., 2016. Is Effective Teaching Stable?. *The Journal of Experimental Education ,* 84(1), pp. 23-47.

Peirano, C., 2009. *Las carreras de pedagogía y sus desafíos,* Seminario Internacional 2009: Calidad de los egresados responsabilidad institucional ineludible. Consejo Nacional de Educación.

Penny, A. R., 2003. Changing the Agenda for Research into Students' Views about University Teaching: Four shortcomings of SRT research. *Teaching in Higher Education,* 8(3), pp. 399-411.

Perry, R. P., Niemi, R. R. & Jones, K., 1974. Effect of prior teaching evaluations and lecture presentation on ratings of teaching performance. *Journal of Educational Psychology,* 66(6), pp. 851-856.

Pey, R., Durán, F. & Jorquera, P., 2012. *Informe para la toma de decisiones sobre duración de las carreras de pregrado en el CRUCh,* Chile: CRUCh.

Pianta, R. C., 2010. *Standardized Classroom Observations from PreK to Third Grade: A Mechanism for Improving Quality Classroom Experiences During the P-3 Years,* University of Virginia: Foundation for Child Development.

Polikoff, M. S., 2015. The Stability of Observational and Student Survey Measures of Teaching Effectiveness. *American Journal of Education,* 121(2), pp. 183-212.

Praetorius, A.-K., Lenske, G. & Helmke, A., 2012. Observer ratings of instructional quality: Do they fulfill what they promise?. *Learning and Instruction,* 22(6), pp. 387-400.

Ramsden, P., 1991. A performance indicator of teaching quality in higher education: The Course Experience Questionnaire. *Studies in Higher Education,* 16(2), pp. 129-150.

Rantanen, P., 2013. The number of feedbacks needed for reliable evaluation. A multilevel analysis of the reliability, stability and generalisability of students' evaluation of teaching. *Assessment & Evaluation in Higher Education,* 38(2), pp. 224-239.

Remedios, R. & Lieberman, D. A., 2008. I liked your course because you taught me well: The influence of grades, workload, expectations and goals on students' evaluations of teaching.. *British Educational Research Journal,* 34(1), p. 91–115.

Rice, J. K., 2003. *Teacher quality: Understanding the effectiveness of teacher attributes.* Washington, DC: Economic Policy Institute.

Rivkin, S. G., Hanushek, E. A. & Kain, J. F., 2005. Teachers, schools, and academic achievement. *Econometrica,* Volumen 73, p. 417–458.

Rubio Hurtado, M. J. & Berlanga Silvente, V., 2012. Cómo aplicar las pruebas paramétricas bivariadas t de Student y ANOVA en SPSS. Caso práctico. *REIRE, Revista d'Innovació i Recerca en Educació,* 5(2), pp. 83-100.

Salazar, J., 2008. Diagnóstico Preliminar sobre Evaluación de la Docencia Universitaria. Una Aproximación a la Realidad en las Universidades Públicas y/o Estatales de Chile. *Revista Iberoamericana de Evaluación Educativa,* 1(3e), pp. 67-84.

Santelices, M. V., Valencia, E., Gonzalez, J. & Taut, S., 2017. Two teacher quality measures and the role of context: evidence from Chile. *Educational Assessment, Evaluation and Accountability,* 29(2), p. 111–146.

Scherr, F. C. & Scherr, S. S., 1990. Bias in student evaluations of teacher effectiveness. *Journal of Education for Business,* 65(8), pp. 356-358.

Scott, C. S., 1977. Student ratings and instructor-defined extenuating circumstances. *Journal of Educational Psychology,* 69(6), pp. 744-747.

Silva Montes, C., 2009. Las encuestas de opinión en la Universidad Autónoma de Ciudad Juárez: ¿un caso de exclusión del profesorado?. *Archivos Analíticos de Políticas Educativas,* Volumen 17, pp. 1-31.

Smith, P. L., 1979. The stability of teacher performance in the same course over time. *Research in Higher Education,* 11(2), p. 153–165.

Smith, R. A. & Cranton, P. A., 1992. Students' perceptions of teaching skills and overall effectiveness across instructional settings. *Research in Higher Education,* 33(6), p. 747–764.

Solomon, D., Speer, A., Rosebraugh, C. & DiPette, D., 1997. The reliability of medical student ratings of clinical teaching. *Evaluation & Health Professions,* 20(3), p. 343–352.

Somma, N., 2017. Protestas y conflictos en el Chile contemporáneo: quince tesis para la discusión. En: R. Araya & F. Ceballos, edits. *Conflictos, controversias y disyuntivas.* Santiago: Ediciones Abierta, Serie IDRC vol. 1, pp. 37-86.

Spooren, P., Brockx, B. & Mortelmans, D., 2013. On the Validity of Student Evaluation of Teaching: The State of the Art. *Review of Educational Research,* 83(4), pp. 598-642.

Spooren, P. & Mortelmans, D., 2006. Teacher professionalism and student evaluation of teaching: will better teachers receive higher ratings and will better students give higher ratings?. *Educational Studies ,* 32(2), pp. 201-214.

Tagomori, H. T. & Bishop, L. A., 1995. Student Evaluation of Teaching: Flaws in the Instruments.. *Thought & Action,* 11(1), pp. 63-78.

Tejedor, F. J., 2003. Un modelo de evaluación del profesorado universitario. *Revista de Investigación Educativa,* 21(1), pp. 157-182.

Theall, M. & Franklin, J., 2000. Creating responsive student ratings systems to improve evaluation practice. *New directions for teaching and learning,* 2000(83), pp. 95-107.

Theall, M. & Franklin, J., 2001. Looking for Bias in All the Wrong Places: A Search for Truth or a Witch Hunt in Student Ratings of Instruction?. *New Directions for Institutional Research,* 2001(109), p. 45–56.

Valenzuela, J. P., Bellei, C. & Ríos, D., 2014. Socioeconomic school segregation in a market-oriented educational system. The case of Chile. *Journal of Education Policy,* 29(2), pp. 217-241.

Varas, L. y otros, 2008. Oportunidades de preparación para enseñar matemática de futuros profesores de educación general básica en Chile. *Calidad en la Educación,* Volumen 29, pp. 63-88.

Vásquez Rizo, F. E. & Gabalán Coello, J., 2006. Percepciones estudiantiles y su influencia en la evaluación del profesorado. Un caso en la Universidad Autónoma de Occidente, Cali-Colombia. *Relieve,* 12(2), pp. 219-245.

Wachtel, H. K., 1998. Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education,* 23(2), p. 191–211.

West, R. F., 1988. The short-term stability of student ratings of instruction in medical school. *Medical Education,* 22(2), p. 104–112.

Whitely, S. E. & Doyle, K. O., 1976. Implicit Theories in Student Ratings. *American Educational Research Journal,* 13(4), pp. 241-253.

Zabaleta, F., 2007. The use and misuse of student evaluations of teaching. *Teaching in Higher Education,* 12(1), pp. 55-76.

IX. Anexos

i. Encuesta clase a clase

ENCUESTA ESTUDIANTE

Fondef IT13I10005
"Herramientas para fortalecer la formación de profesores de educación básica basadas en experiencias de enseñanza de la matemática en aula"

IDENTIFICACIÓN DE ESTUDIANTE	RUT	

ENCUESTA DE FINAL DE CLASE

Nos gustaría saber **cómo la clase de hoy contribuyó a tu formación como profesor**. Por favor contesta las siguientes preguntas considerando tu experiencia en la clase de hoy y marcando la alternativa que más se asemeja a tu opinión.

La clase de hoy me dio la oportunidad de...	No se dio (0)	Sí, un poco (1)	Sí, media-namente (2)	Sí, bastante (3)
... adquirir conocimientos de matemática que no sabía o no tenía totalmente claros.				
... analizar y/o diseñar tareas/actividades matemáticas escolares.				
... aprender a analizar y anticipar producciones matemáticas de niños y niñas.				
... aprender a promover discusiones y conversaciones matemáticas en niños y niñas.				
... aprender a gestionar la conducta de niños y niñas, para promover un ambiente propicio para la enseñanza de las matemáticas.				

En términos generales, ¿cómo evaluarías la clase que tuviste en una escala de 1 a 10, donde 10 es la nota máxima? (marca tu preferencia con un círculo).

1 2 3 4 5 6 7 8 9 10

Respecto a tu participación en la clase de hoy, ¿cuán de acuerdo estás con las siguientes frases? Considera que 1 significa "Muy en desacuerdo" y 5 significa "Muy de acuerdo".

	Muy en Desacuerdo (1)	(2)	(3)	(4)	Muy de Acuerdo (5)
Aproveché las oportunidades para aprender que me entregó la clase.					
Mantuve atención hacia el profesor y las actividades de la clase.					
Participe activamente de las discusiones y actividades propuestas por el profesor.					
Estuve motivado durante la clase.					

ii. Operacionalización inductiva de la dimensionalidad del test

CONSTRUCTO	DIMENSIÓN	INDICADOR	ÍTEM
EVALUACIÓN DOCENTE DE LA CLASE	*Evaluación de la capacidad de la clase para generar aprendizaje en conocimientos en matemáticas (P1)*	Capacidad de la clase de entregar conocimientos en matemáticas	1. La clase de hoy me dio la oportunidad de **adquirir conocimientos de matemática que no sabía o no tenía totalmente claros**.
	Evaluación de la capacidad de la clase para desarrollar aprendizaje en habilidades pedagógicas (F1)	Capacidad de la clase de entregar habilidades para el desarrollo de actividades escolares	2. La clase de hoy me dio la oportunidad de **analizar y/o diseñar tareas/actividades matemáticas escolares**.
		Capacidad de la clase de entregar habilidades para el análisis de producciones escolares	3. La clase de hoy me dio la oportunidad de **aprender a analizar y anticipar producciones matemáticas de niños y niñas**.
		Capacidad de la clase de entregar habilidades para generar discusiones escolares	4. La clase de hoy me dio la oportunidad de **aprender a promover discusiones y conversaciones matemáticas en niños y niñas**.
		Capacidad de la clase de entregar habilidades para generar buen ambiente de enseñanza escolar	5. La clase de hoy me dio la oportunidad de **aprender a gestionar la conducta de niños y niñas, para promover un ambiente propicio para la enseñanza de las matemáticas**.
	Evaluación de la calidad general de la clase (P6)	Evaluación general de la clase	6. En términos generales, ¿cómo evaluarías la clase que tuviste en una escala de 1 a 10, donde 10 es la nota máxima?
	Evaluación de la capacidad de la clase para motivar la participación (F2)	Aprovechamiento de clase	7. Respecto a tu participación en la clase de hoy, ¿cuán de acuerdo estás con las siguientes frases? **Aproveché las oportunidades para aprender que me entregó la clase**.
		Atención en clases	8. Respecto a tu participación en la clase de hoy, ¿cuán de acuerdo estás con las siguientes frases? **Mantuve atención hacia el profesor y las actividades de la clase**.
		Participación en clases	9. Respecto a tu participación en la clase de hoy, ¿cuán de acuerdo estás con las siguientes frases? **Participe activamente de las discusiones y actividades propuestas por el profesor**.
		Motivación en la clase	10. Respecto a tu participación en la clase de hoy, ¿cuán de acuerdo estás con las siguientes frases? **Estuve motivado durante la clase**.

iii. Tablas

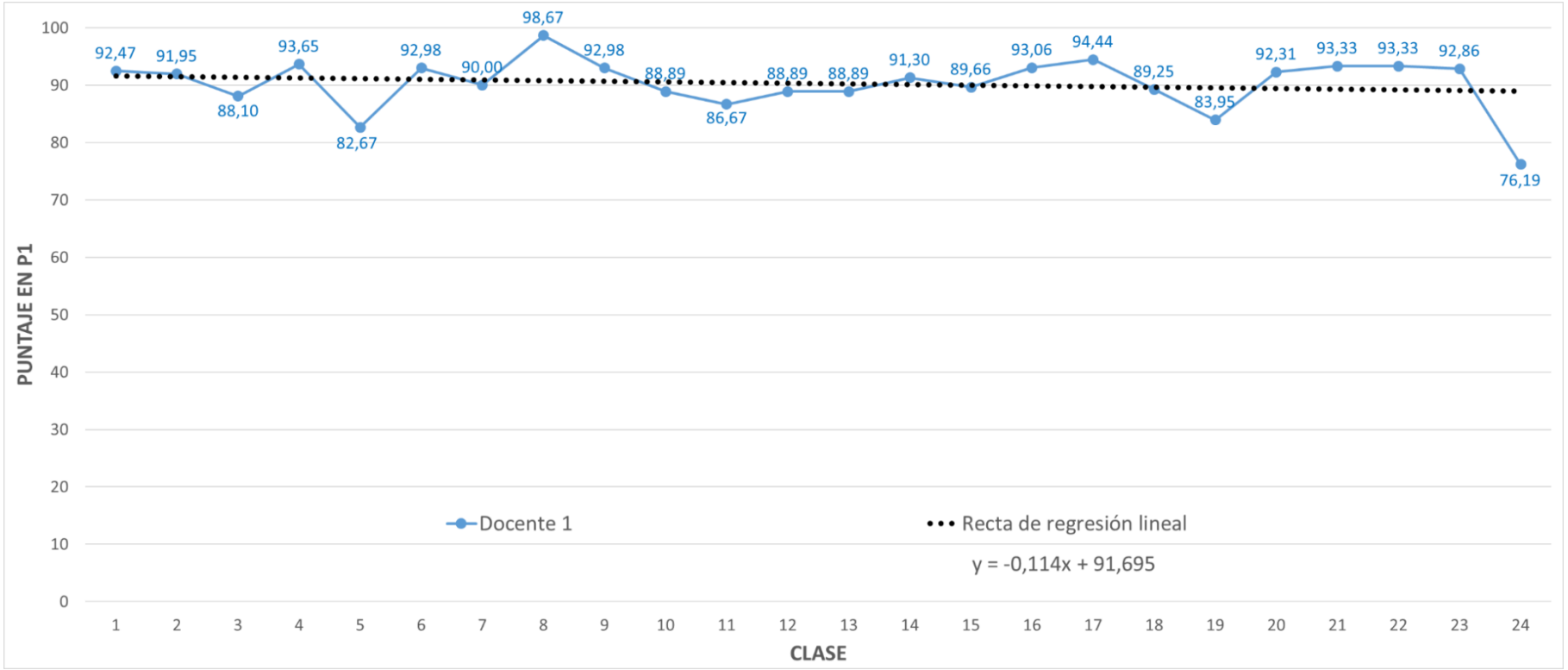
Tabla Anexo 1: Porcentajes individuales y categorización de mejorías/empeoramientos semestrales en las trayectorias de evaluación docente

Curso evaluado	Porcentaje de mejoría					Tipo de trayectoria
	P1	F1	P6	F2	Σ/4	
Docente 10 - Asignatura A	26,5%	7,8%	18,5%	38,9%	22,9%	Mejoría media
Docente 2	13,2%	25,0%	20,6%	24,8%	20,9%	Mejoría media
Docente 17	31,8%	24,0%	2,8%	-2,8%	14,0%	Mejoría leve
Docente 12 - Asignatura A	6,5%	23,5%	9,8%	4,0%	10,9%	Mejoría leve
Docente 3	8,1%	8,2%	3,8%	5,8%	6,5%	Mejoría leve
Docente 5	2,7%	12,5%	4,8%	0,9%	5,2%	Mejoría leve
Docente 9 - Sección 1	7,8%	1,6%	6,5%	4,0%	5,0%	Mejoría leve
Docente 18 - Sección 1	-6,5%	8,3%	2,1%	11,2%	3,8%	Mejoría mínima
Docente 8	-1,2%	9,0%	2,6%	2,9%	3,3%	Mejoría mínima
Docente 12 - Asignatura B	-1,1%	-3,4%	8,4%	6,5%	2,6%	Mejoría mínima
Docente 10 - Asignatura B	-10,6%	10,5%	6,4%	1,9%	2,0%	Mejoría mínima
Docente 6	2,5%	-7,3%	7,6%	4,1%	1,7%	Mejoría mínima
Docente 18 - Sección 2	-3,4%	-1,3%	1,7%	7,2%	1,0%	Mejoría mínima
Docente 14	5,0%	-2,1%	3,7%	-3,2%	0,8%	Mejoría mínima
Docente 9 - Sección 2	-5,4%	5,0%	-1,4%	-0,3%	-0,5%	Empeoramiento mínimo
Docente 16	-1,1%	-7,3%	4,6%	-0,2%	-1,0%	Empeoramiento mínimo
Docente 13	0,4%	-10,0%	2,7%	-2,3%	-2,3%	Empeoramiento mínimo
Docente 7	0,6%	-12,1%	1,3%	0,7%	-2,3%	Empeoramiento mínimo
Docente 4	-7,2%	3,2%	-4,4%	-3,0%	-2,8%	Empeoramiento mínimo
Docente 1	-2,3%	-12,5%	2,0%	-1,4%	-3,5%	Empeoramiento mínimo
Docente 15 - Sección 1	-21,7%	-3,0%	-4,5%	-13,6%	-10,7%	Empeoramiento leve
Docente 11	8,1%	-16,7%	-31,2%	-22,1%	-15,5%	Empeoramiento medio
Docente 15 - Sección 2	-42,6%	-5,4%	-22,1%	-16,6%	-21,7%	Empeoramiento medio
Promedio	0,4%	2,5%	2,0%	2,1%	1,8%	
Promedio de positivos	9,4%	11,5%	6,1%	8,7%	7,2%	
Promedio de negativos	-9,4%	-7,4%	-12,7%	-6,5%	-6,7%	
n positivos	12	12	18	13	14	
n negativos	11	11	5	10	9	

iv. Gráficos de trayectorias de evaluación docente[21]

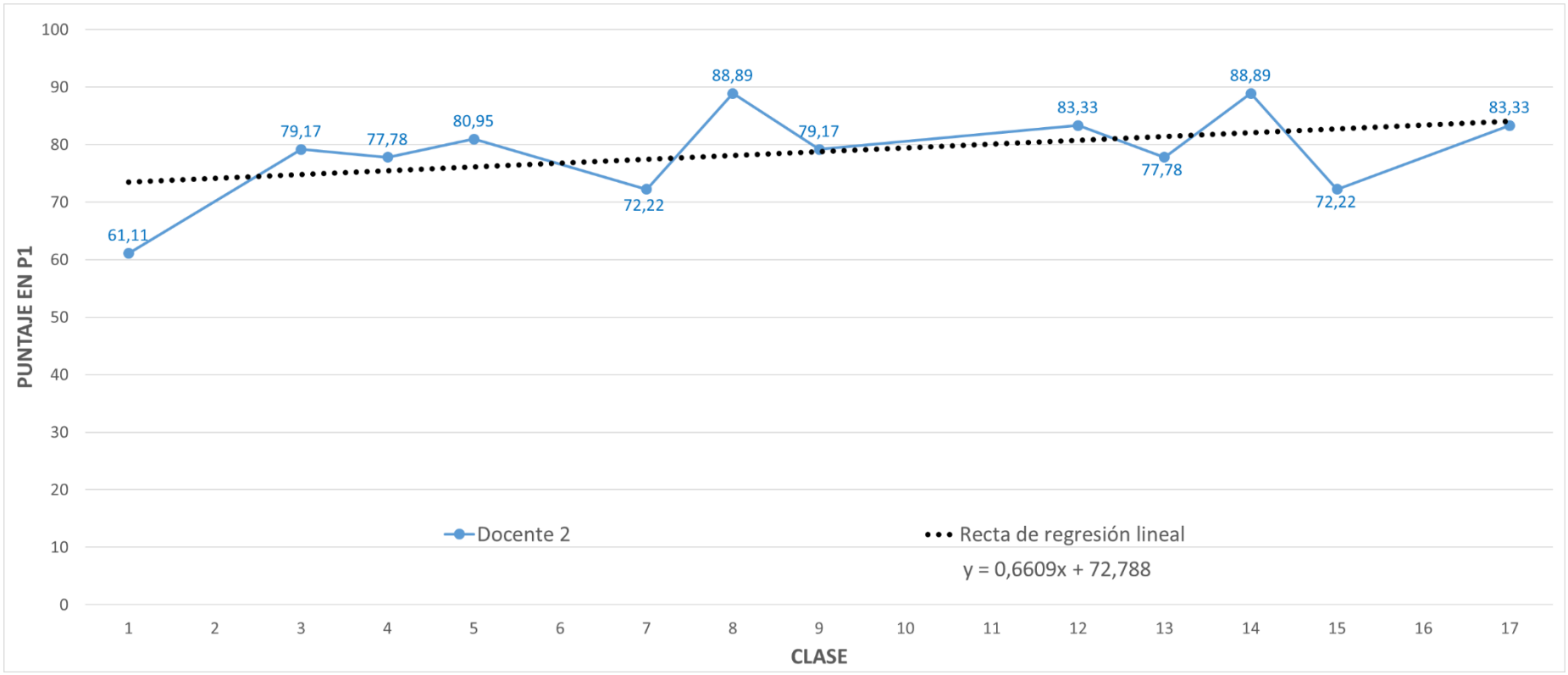
- ***Series de ítem P1: "Capacidad de la clase para generar aprendizaje en conocimientos en matemáticas"***

a) P1 - Docente 1 (n=47, $\overline{X}$=90,3)

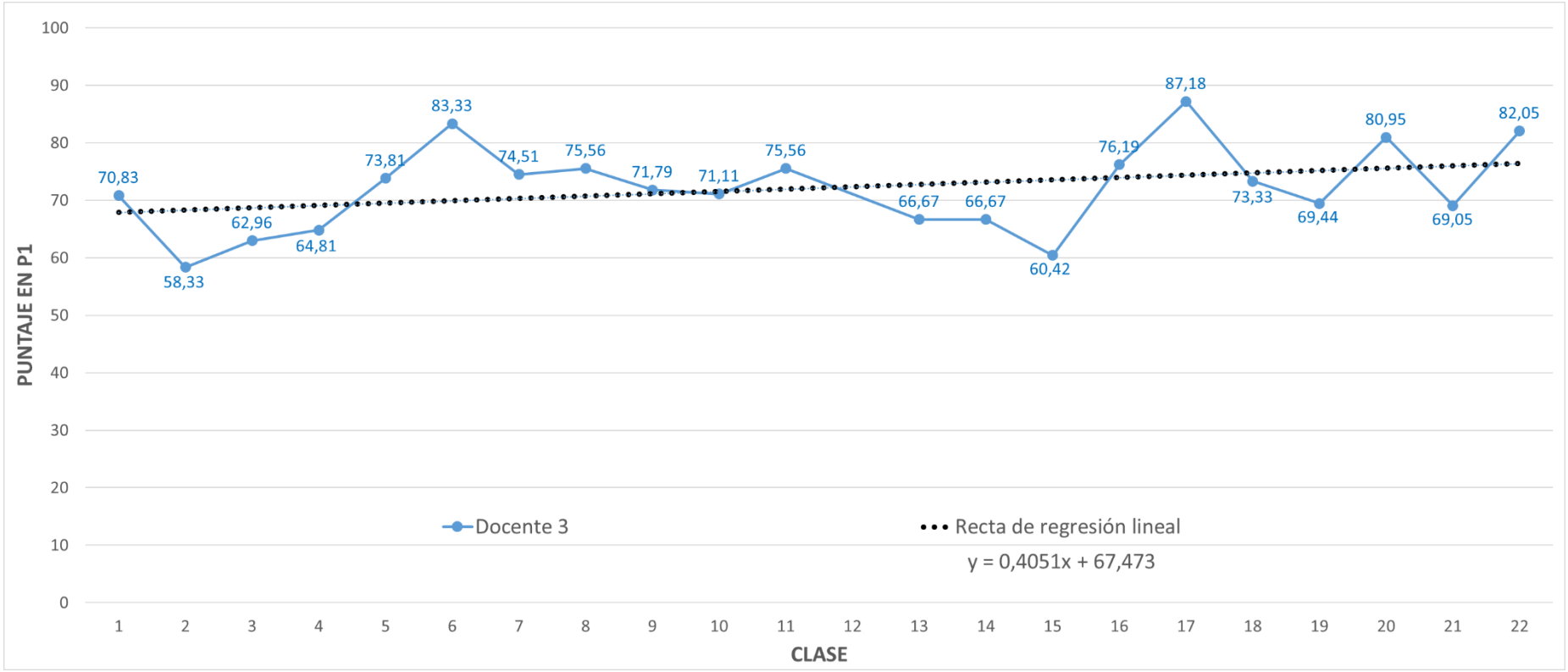


[21] Para todos los gráficos, se indica un "n". Este corresponde a la cantidad de alumnos inscritos en las asignaturas, no a la cantidad por punto de la serie, que varía según la cantidad de clases realizadas en cada curso, y que se puede observar en el máximo del eje y.

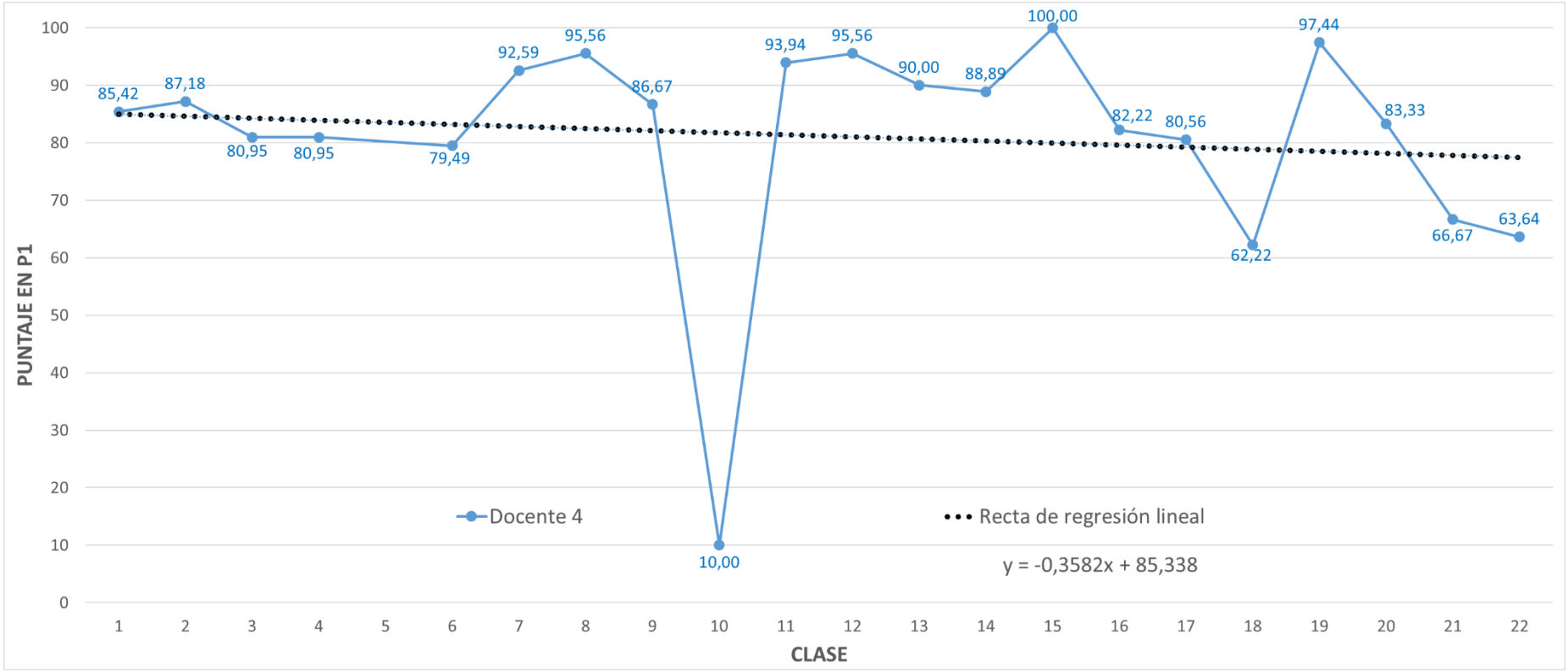
b) P1 - Docente 2 (n=8, ?=78,7)



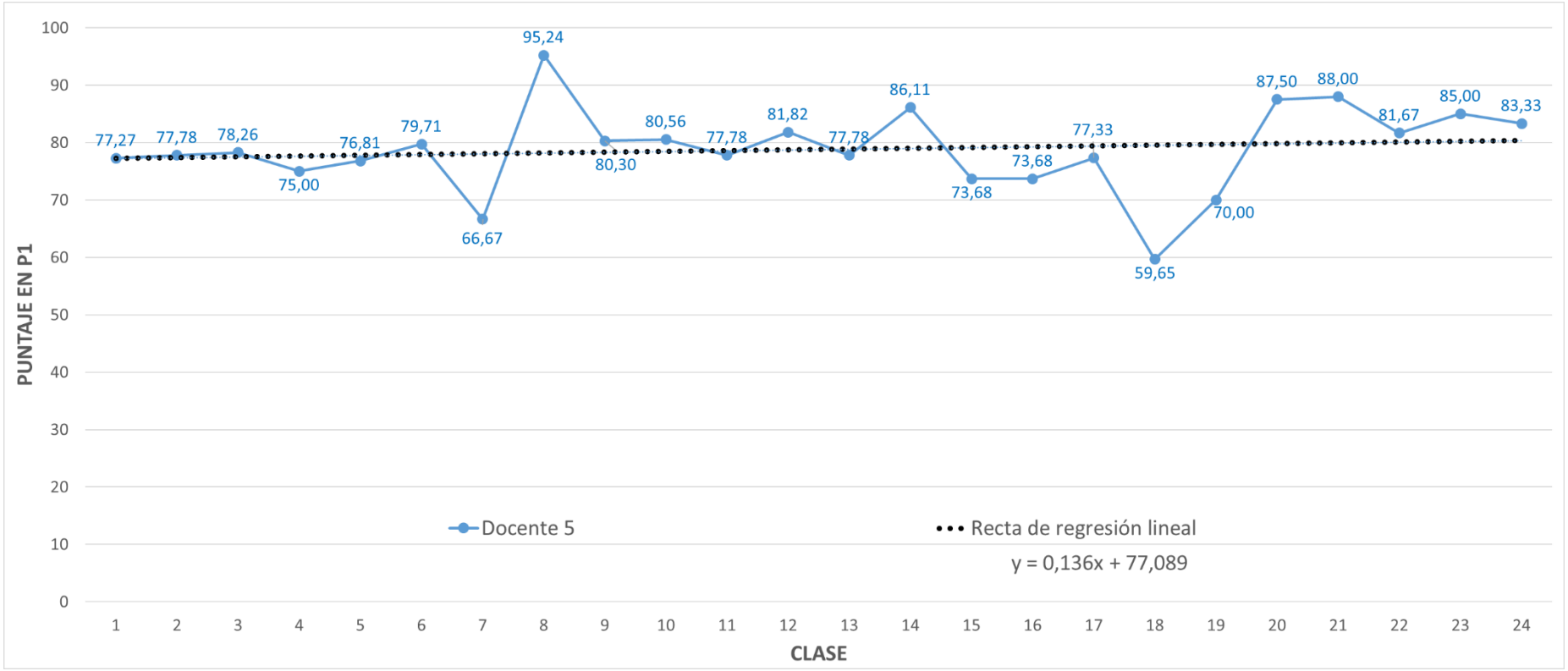
c) P1 - Docente 3 (n=22, $\overline{X}$=72,1)



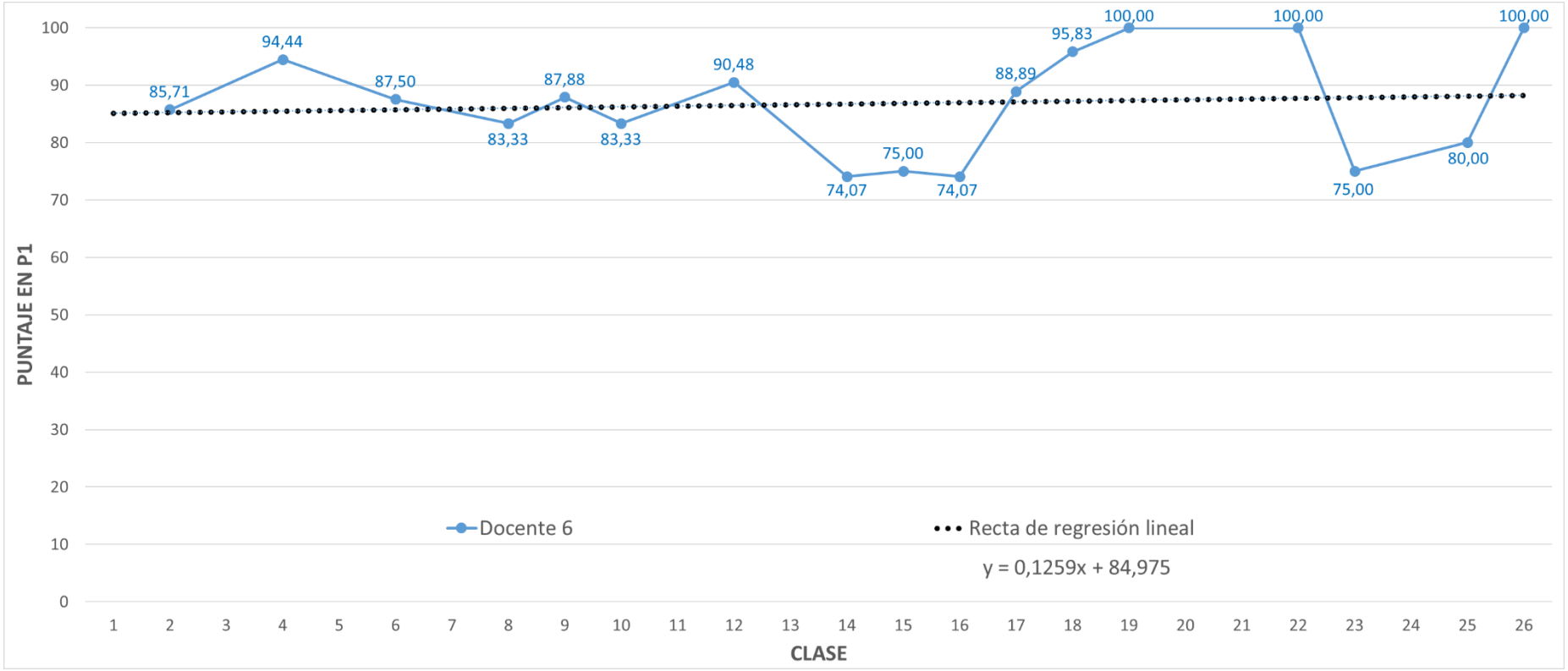
d) P1 - Docente 4 (n=17, $\overline{X}$=81,1)



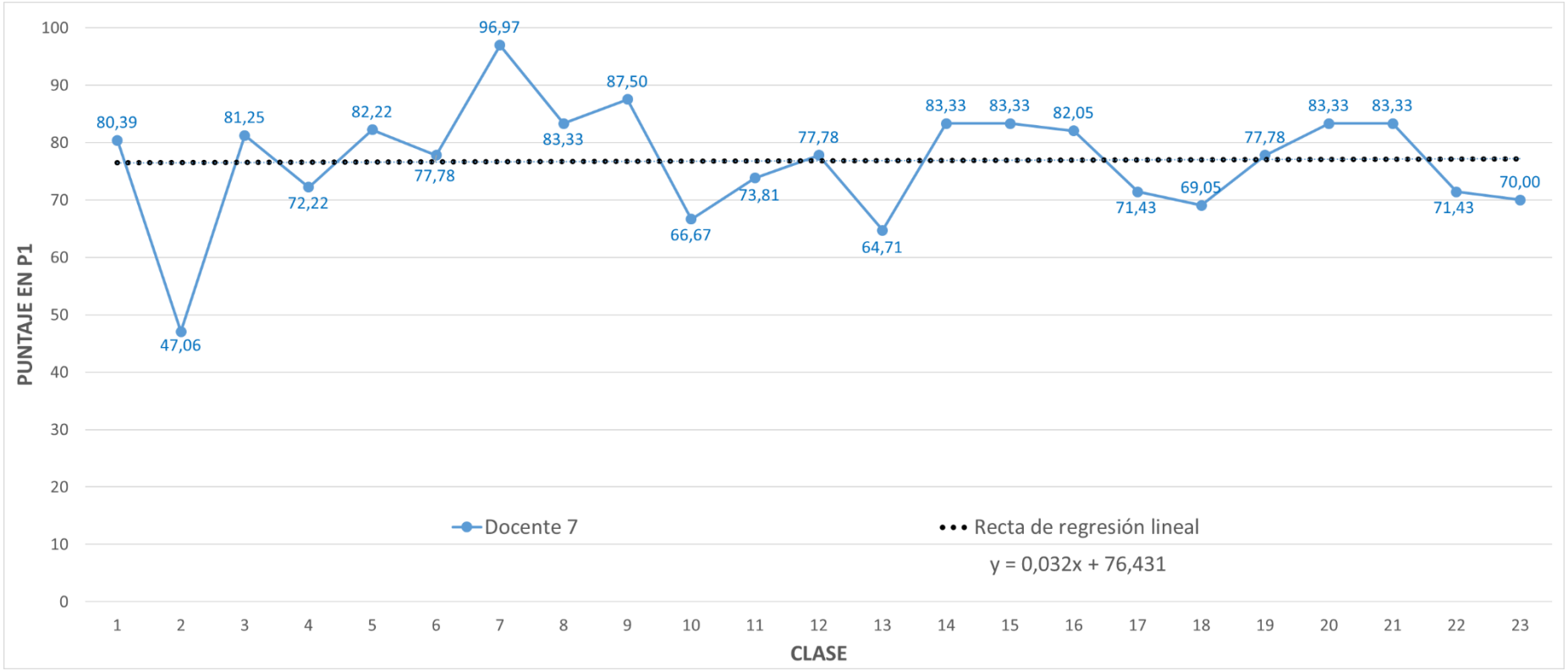
e) P1 - Docente 5 (n=28, $\overline{X}$=78,8)



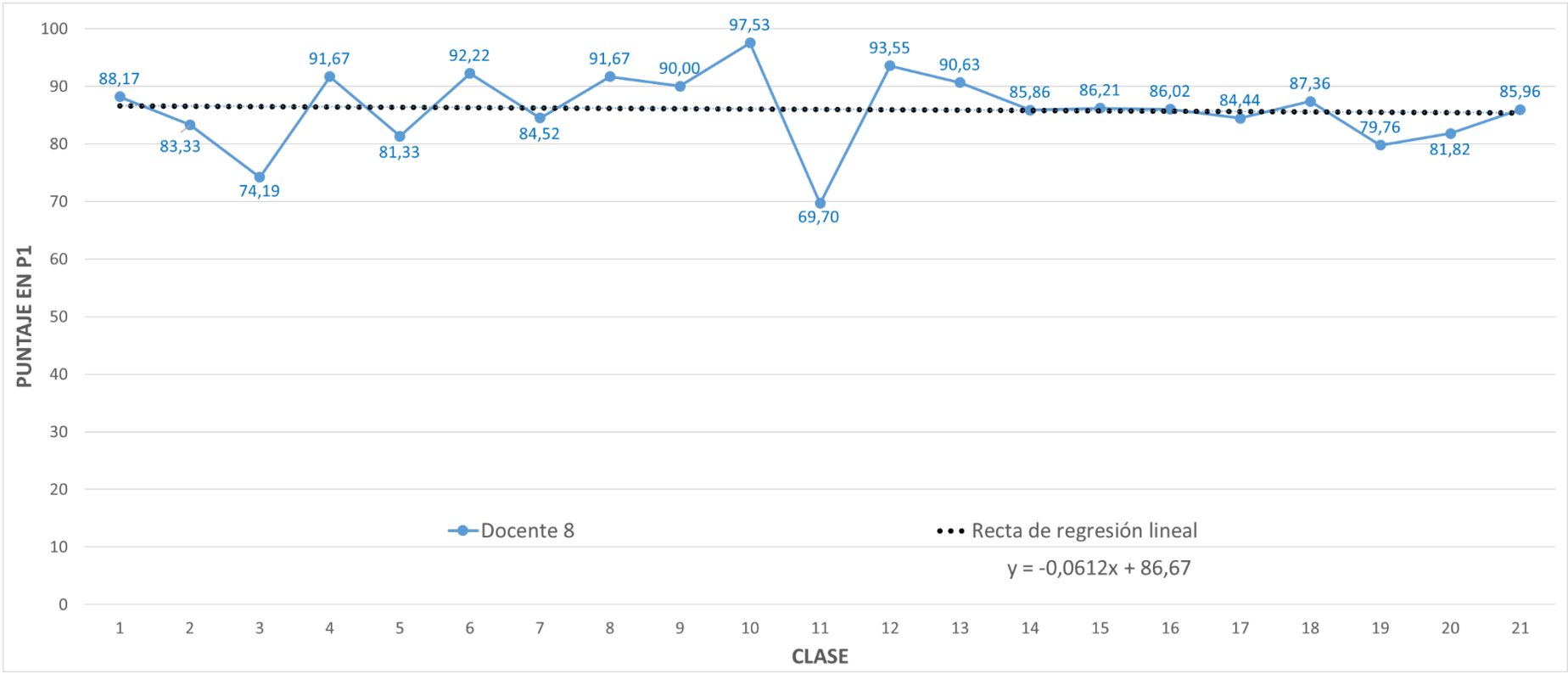
f) P1 - Docente 6 (n=14, $\overline{X}$=86,8)



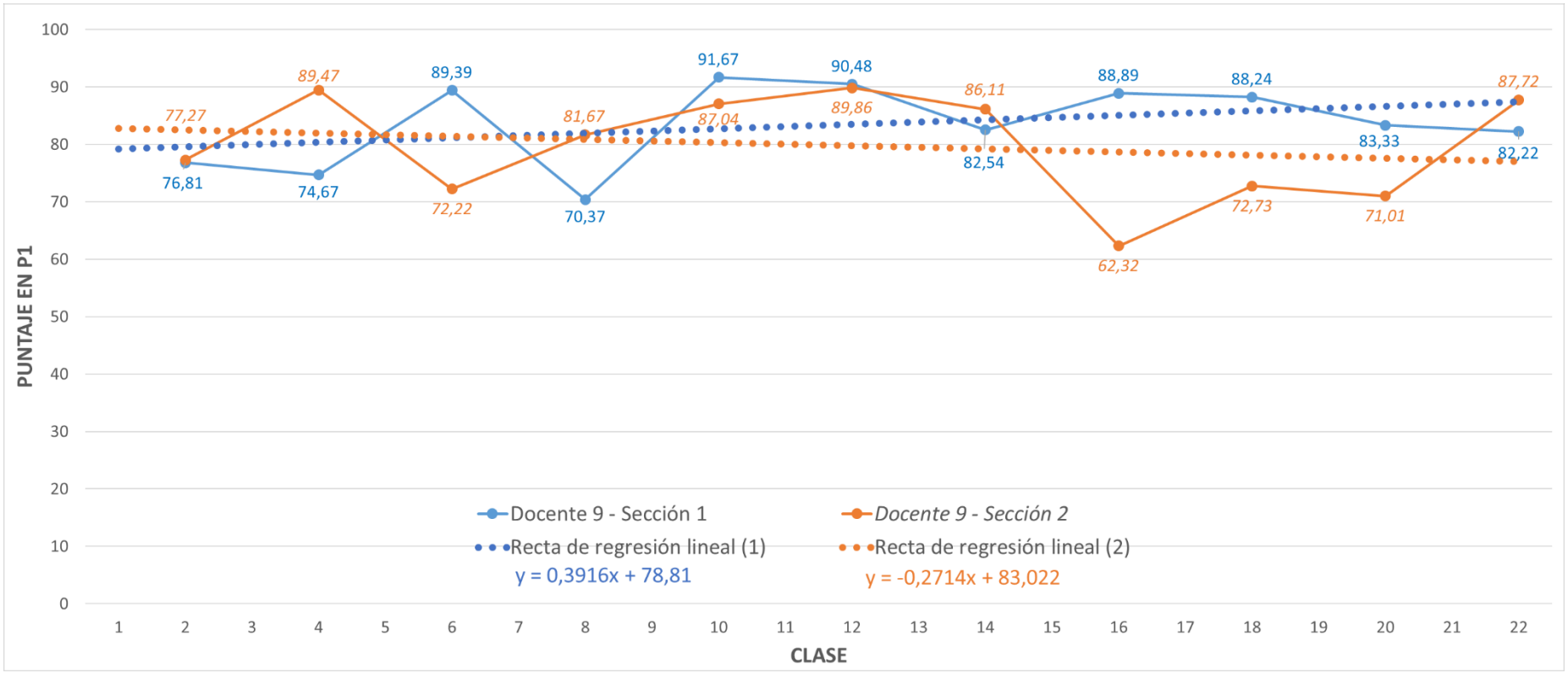
g) P1 - Docente 7 (n=18, $\overline{X}$=76,8)



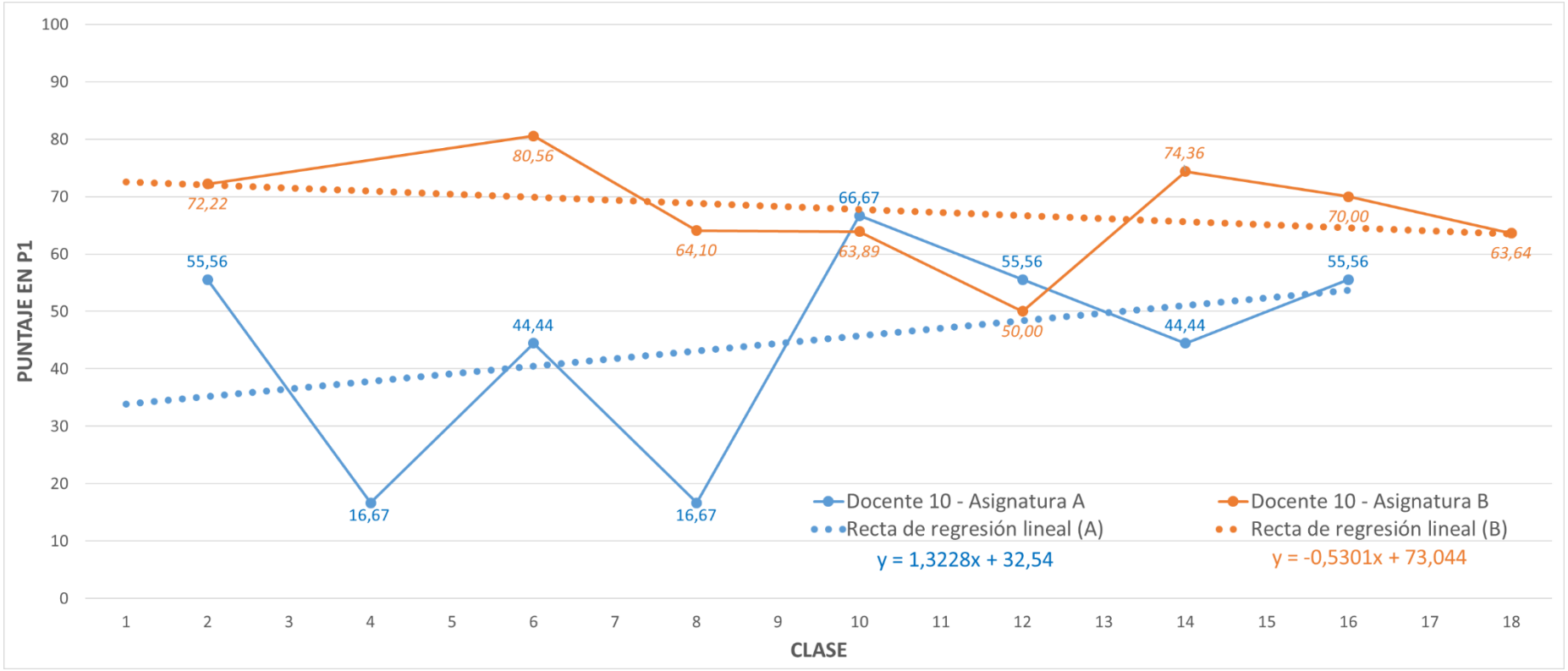
h) P1 - Docente 8 (n=43, $\overline{X}$=86)



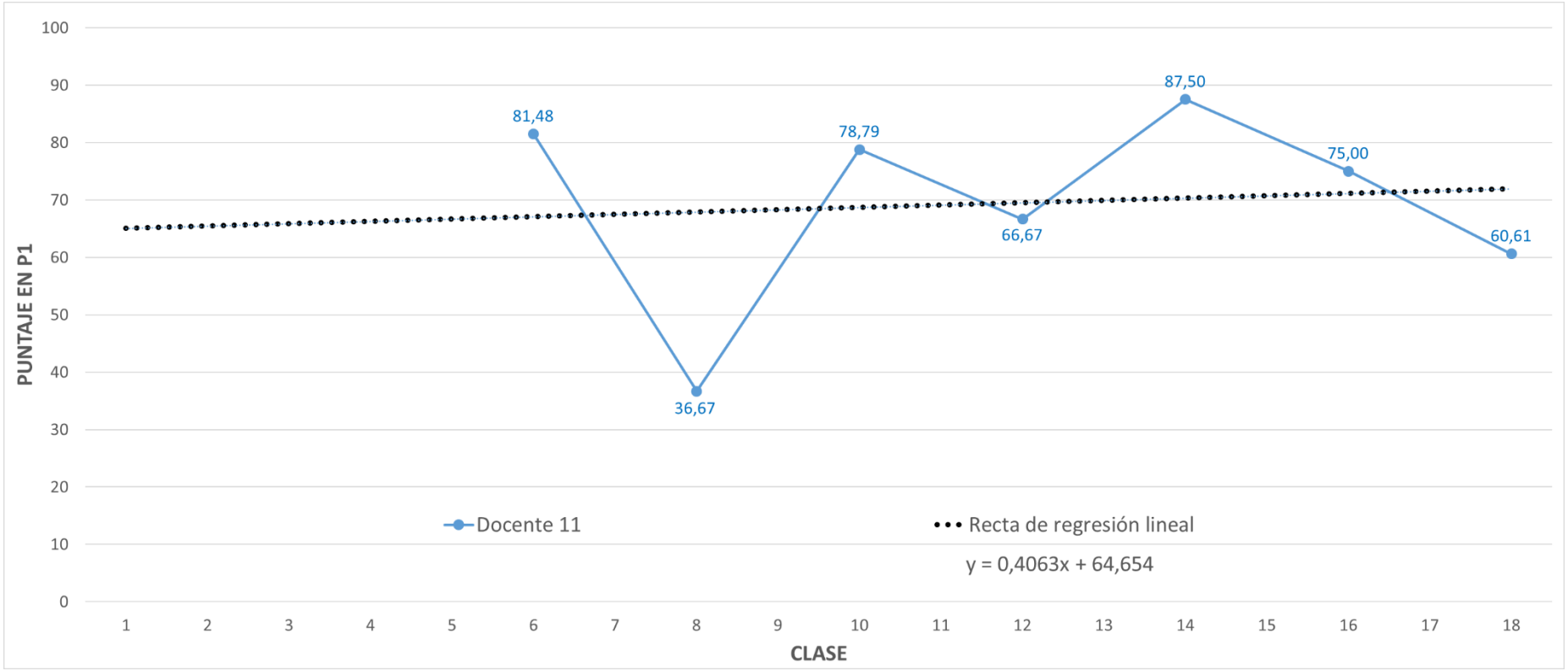
i) P1 - Docente 9 - Sección 1 (n=28, $\overline{X}$=83,5) y Sección 2 (n=33, $\overline{X}$=79,8)



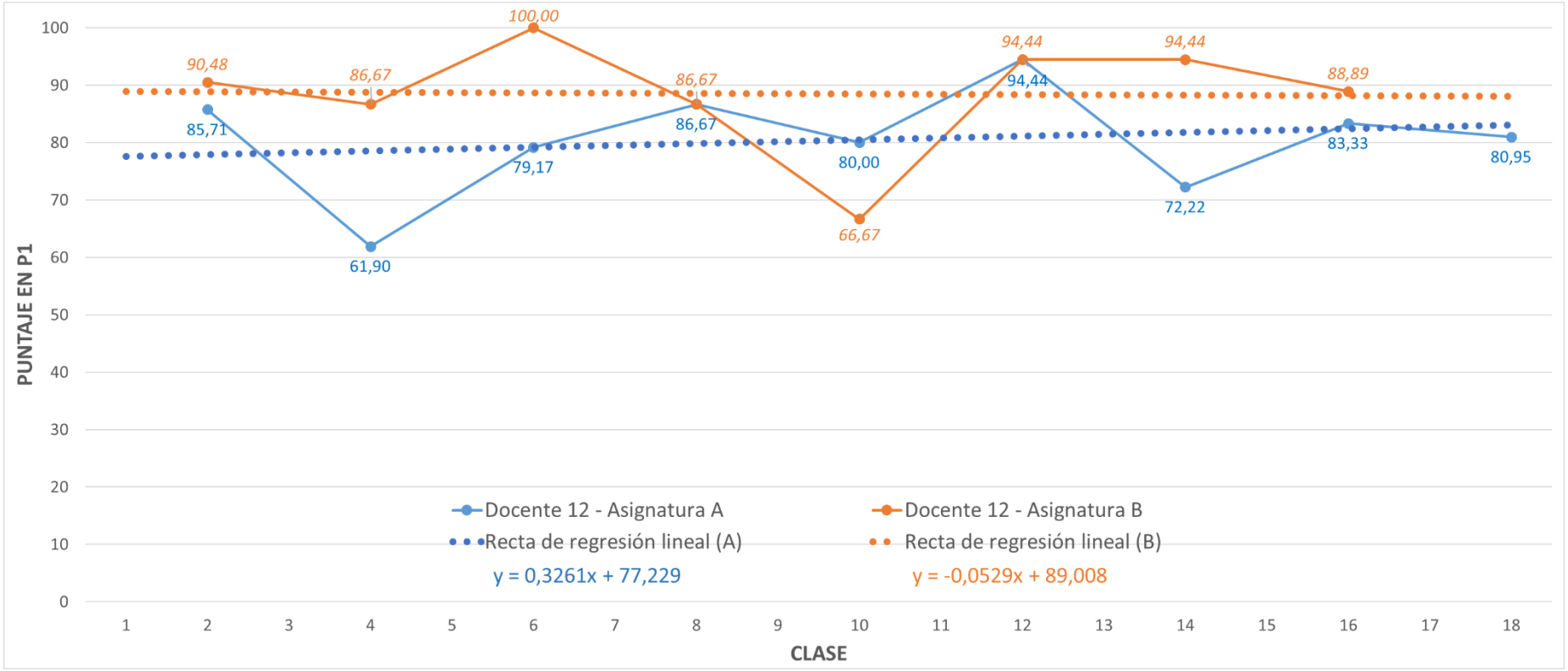
j) P1 - Docente 10 - Asignatura A (n=3, $\overline{X}$=44,4) y Asignatura B (n=16, $\overline{X}$=67,3)



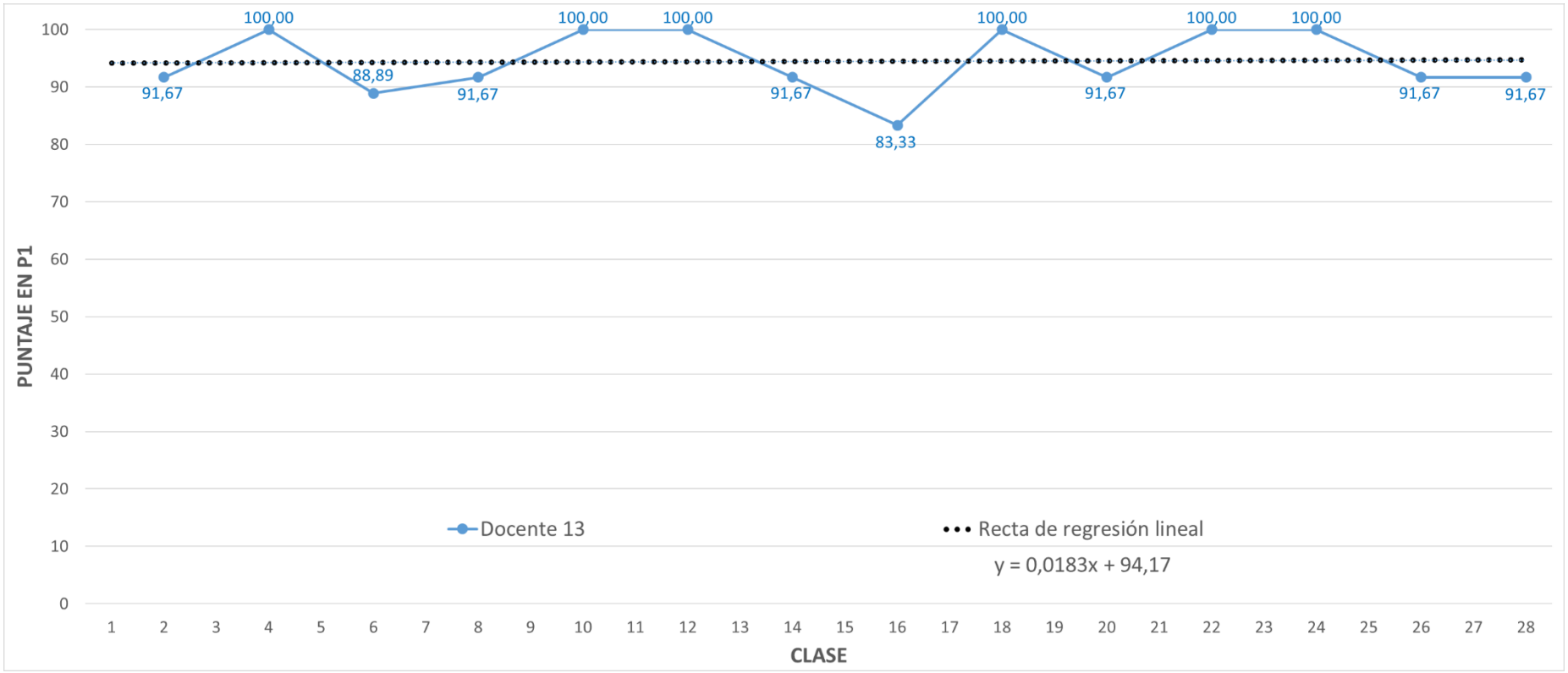
k) P1 - Docente 11 (n=13, $\overline{X}$=69,5)



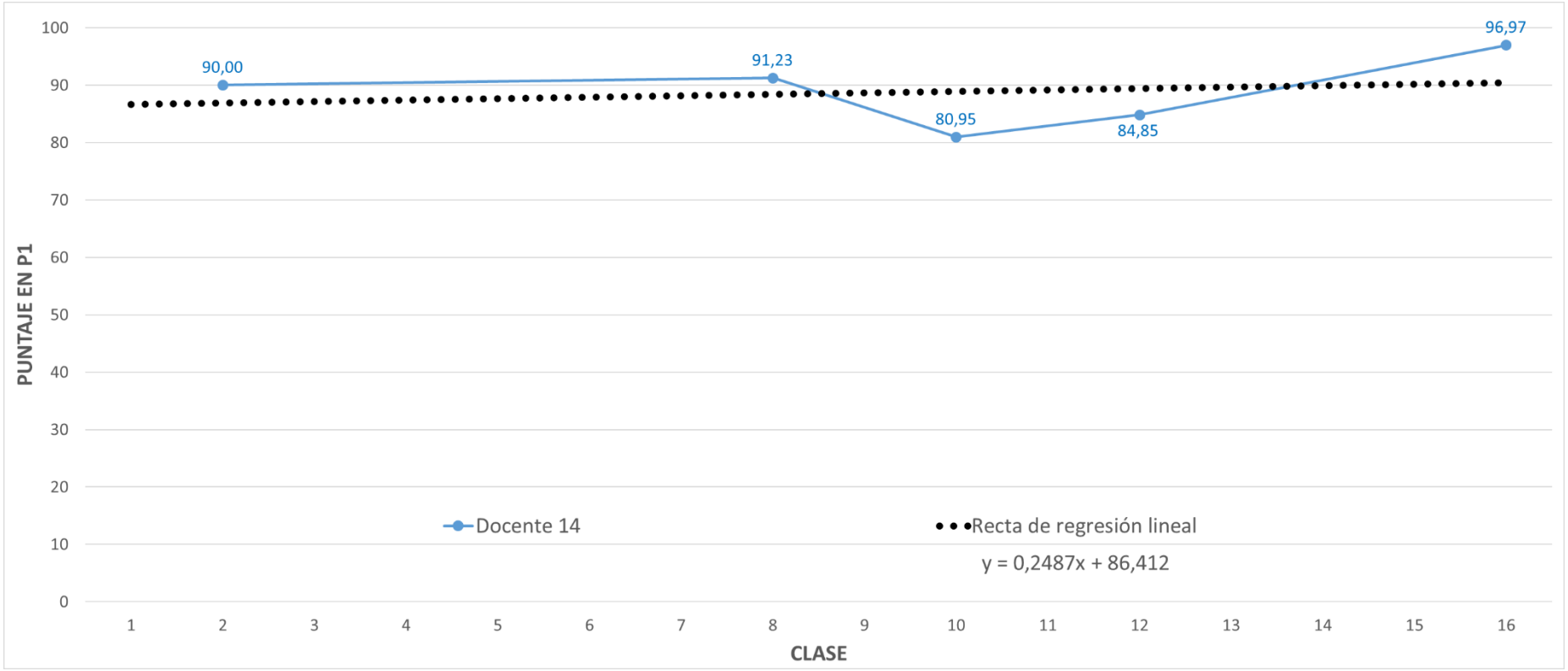
l) P1 - Docente 12 - Asignatura A (n=9, $\overline{X}$=80,5) y Asignatura B (n=7, $\overline{X}$=88,5)



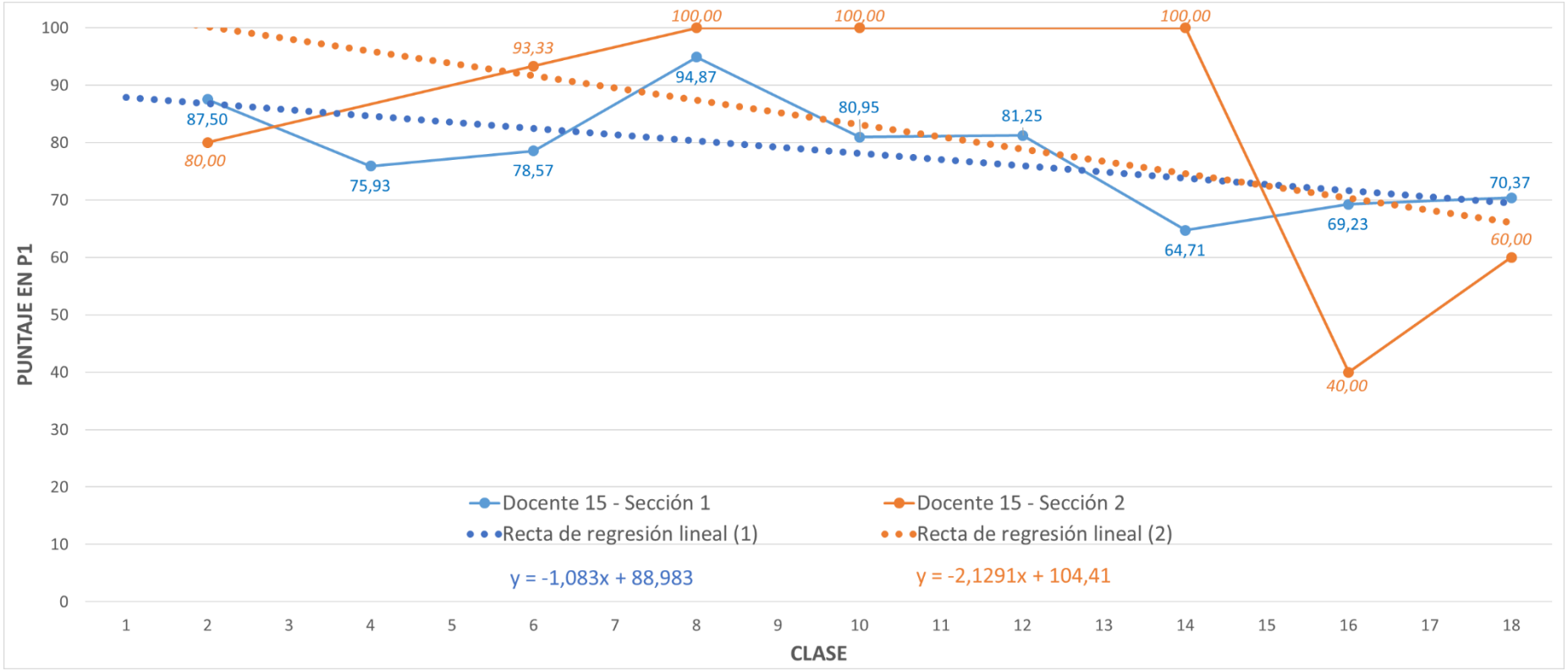
m) P1 - Docente 13 (n=4, $\overline{X}$=94,4)



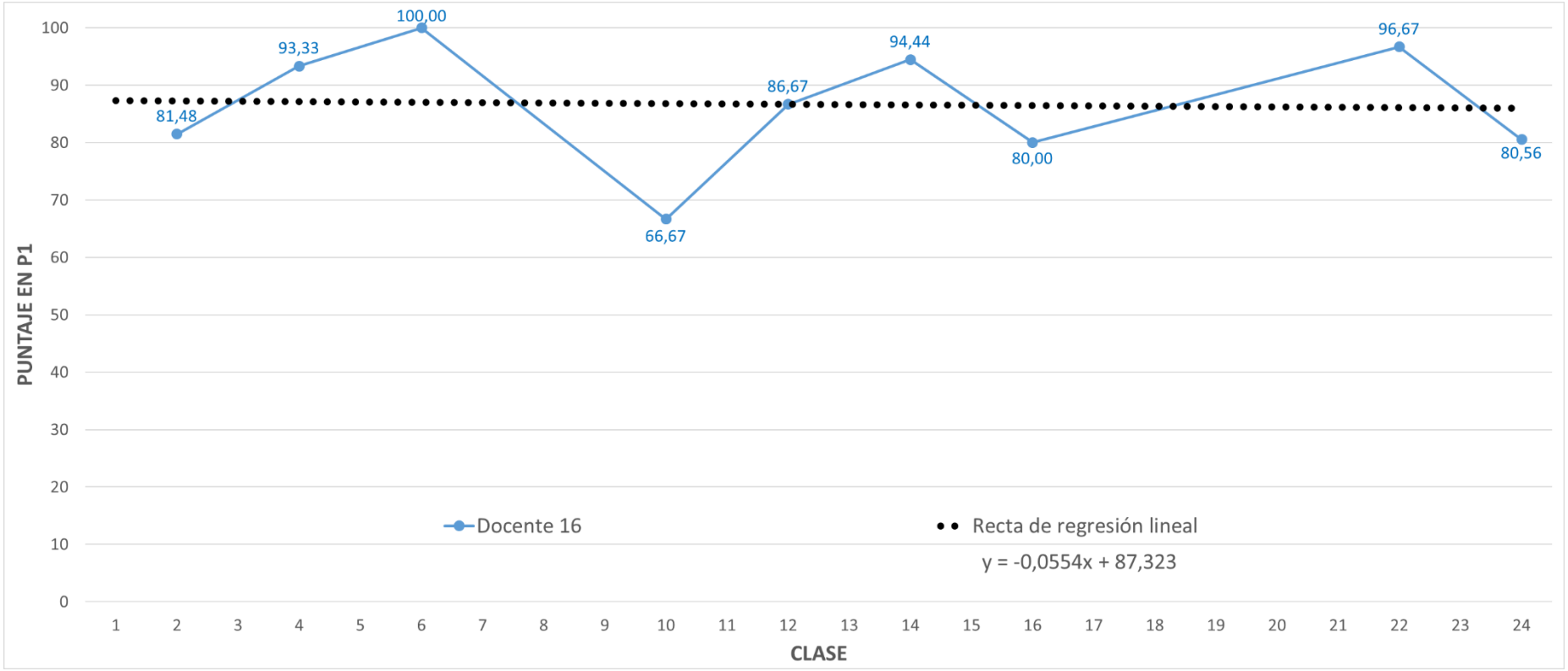
n) P1 - Docente 14 (n=25, $\overline{X}$=88,8)



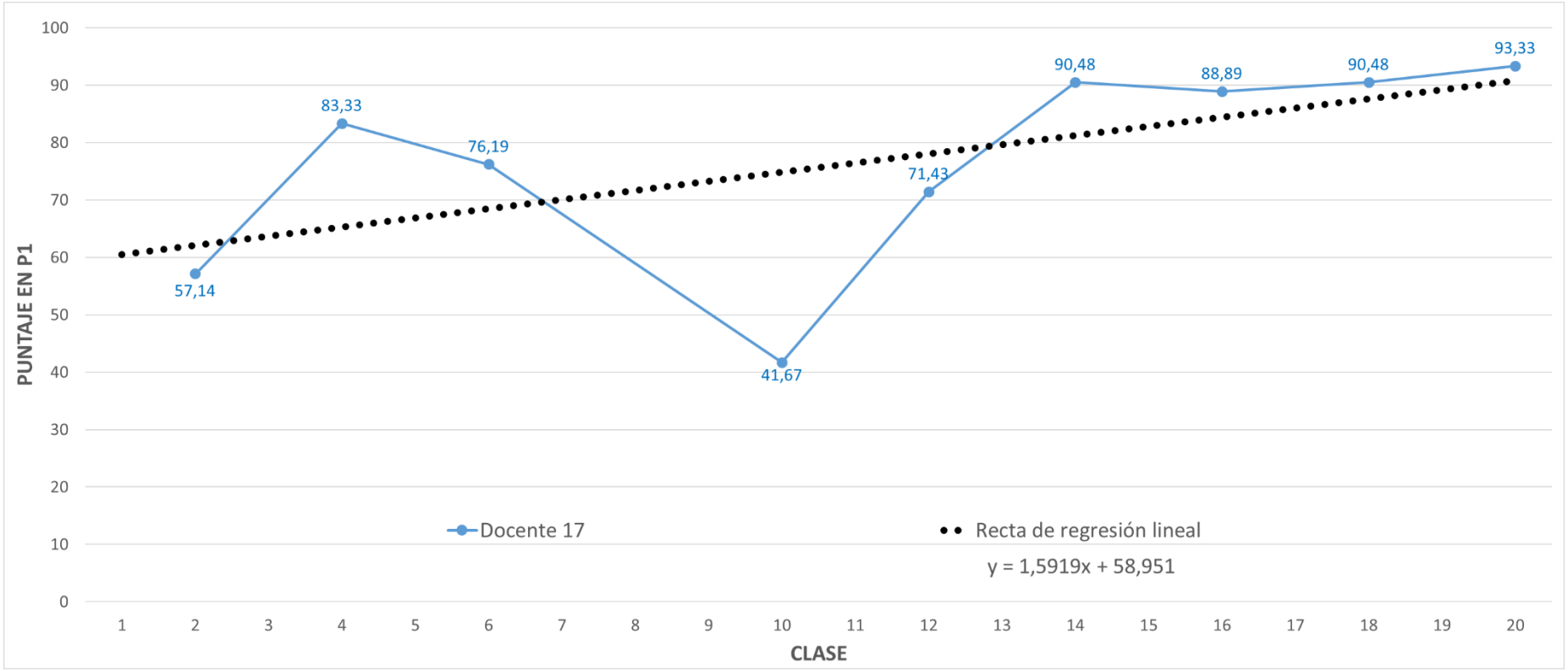
o) P1 - Docente 15 - Sección 1 (n=20, $\overline{X}$=78,2) y Sección 2 (n=15, $\overline{X}$=81,9)



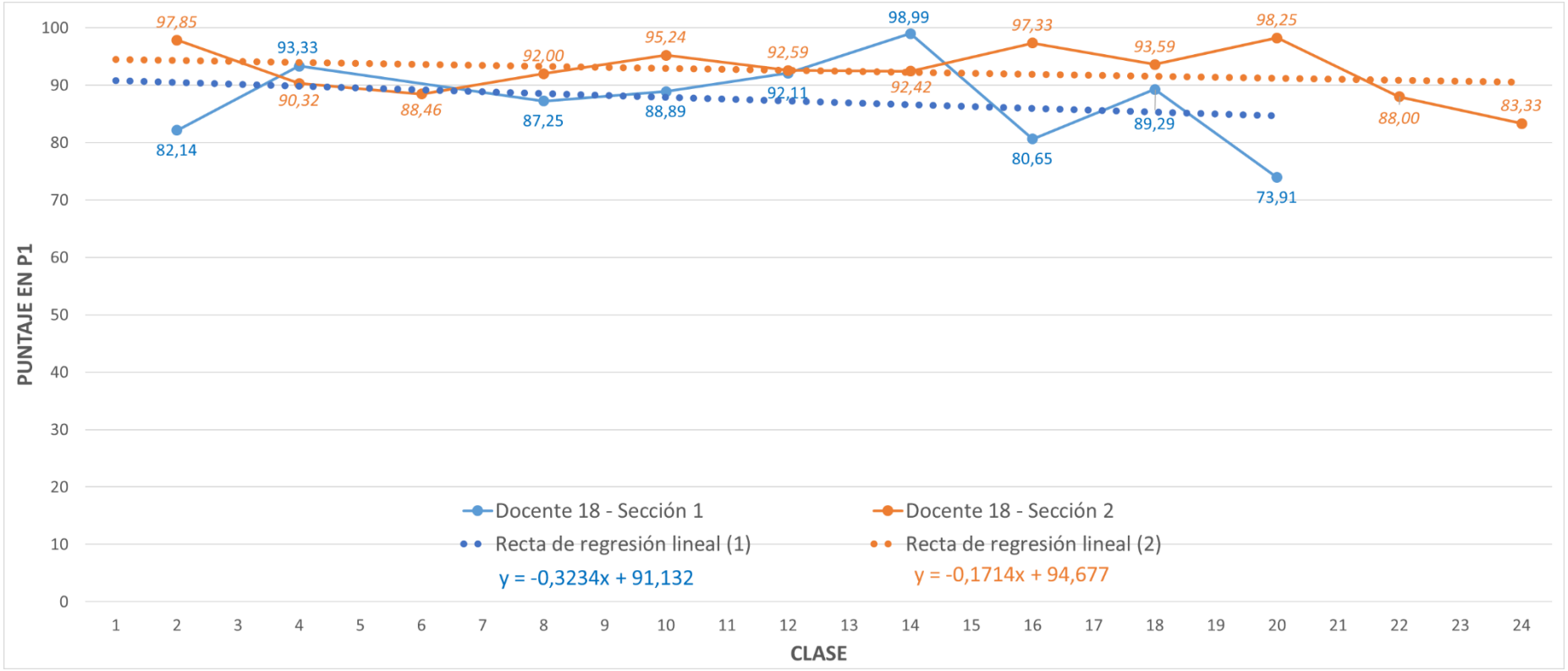
p) P1 - Docente 16 (n=15, $\overline{X}$=86,6)



q) P1 - Docente 17 (n=8, $\overline{X}$=77)

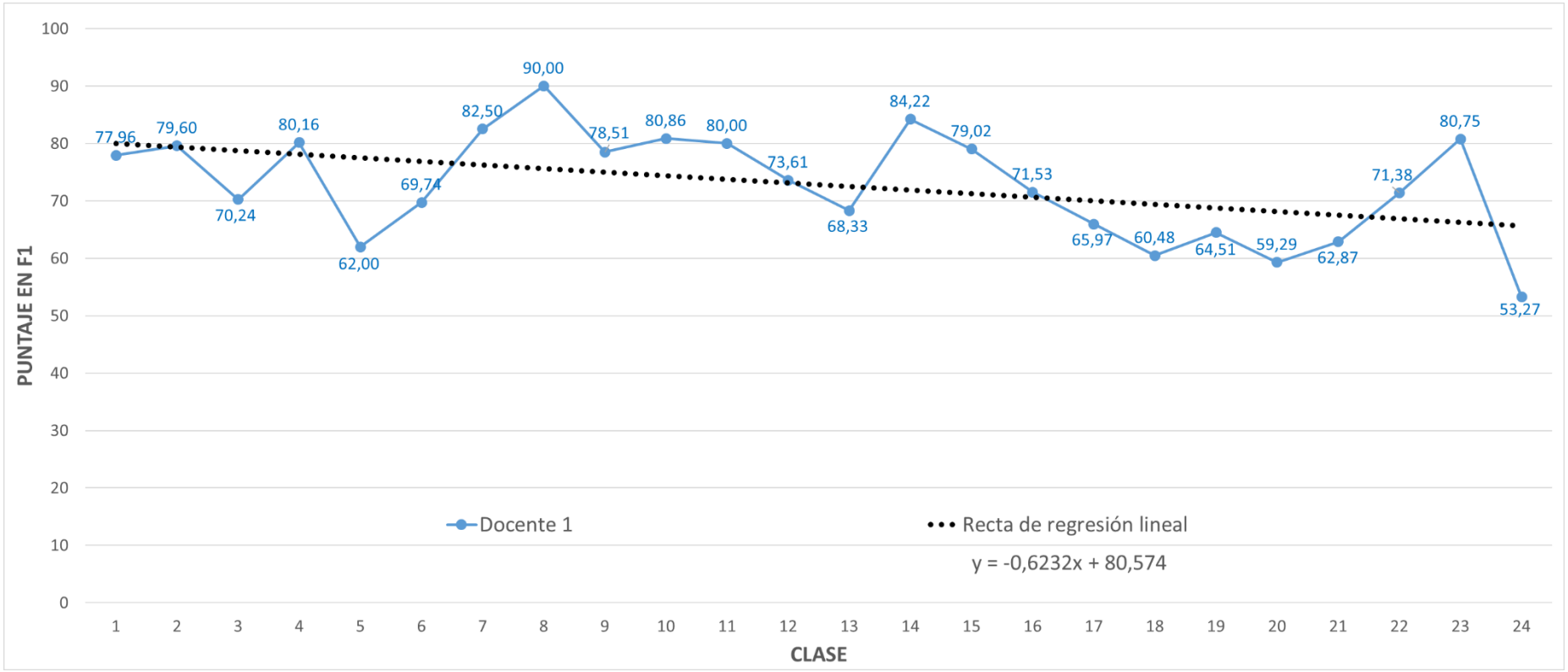


r) P1 - Docente 18 - Sección 1 (n=49, $\overline{X}$=87,4) y Sección 2 (n=36, $\overline{X}$=92,4)

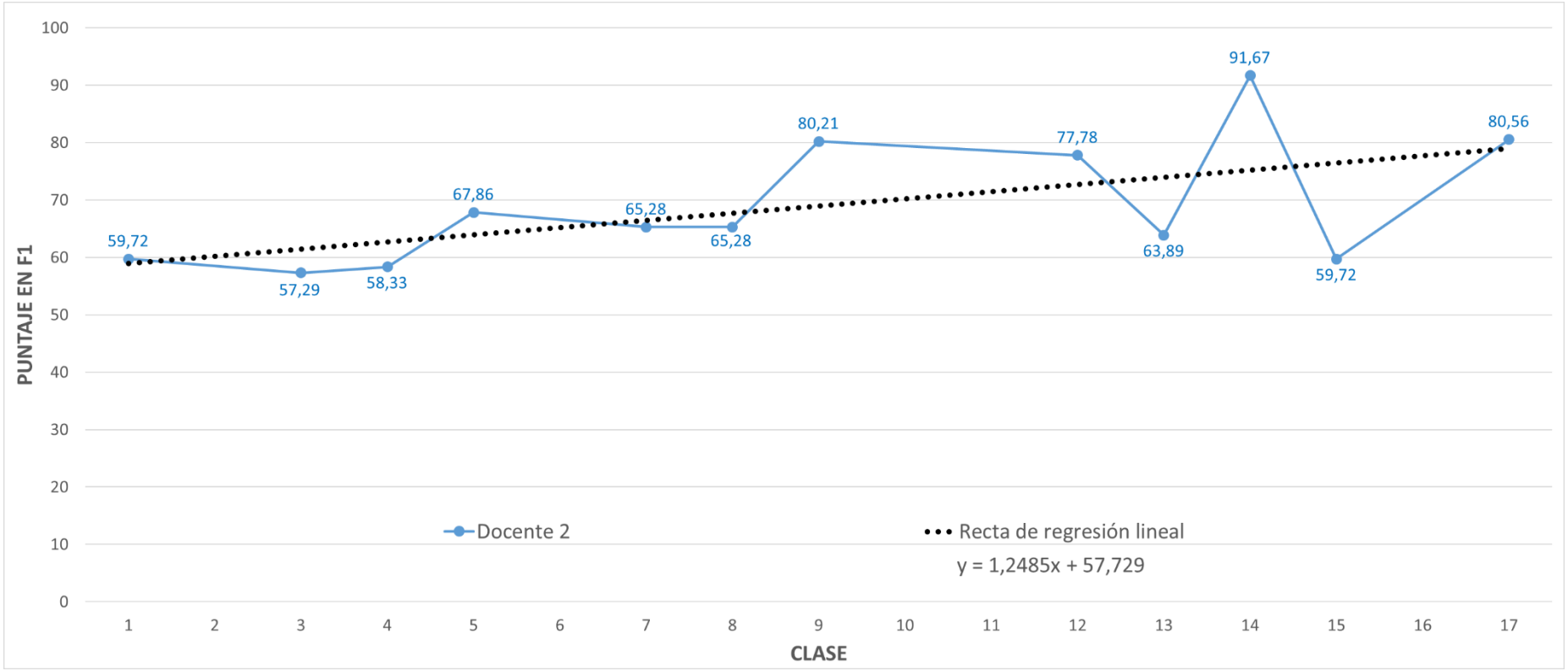


- ***Series de dimensión F1: "Capacidad de la clase para desarrollar aprendizaje en habilidades pedagógicas"***

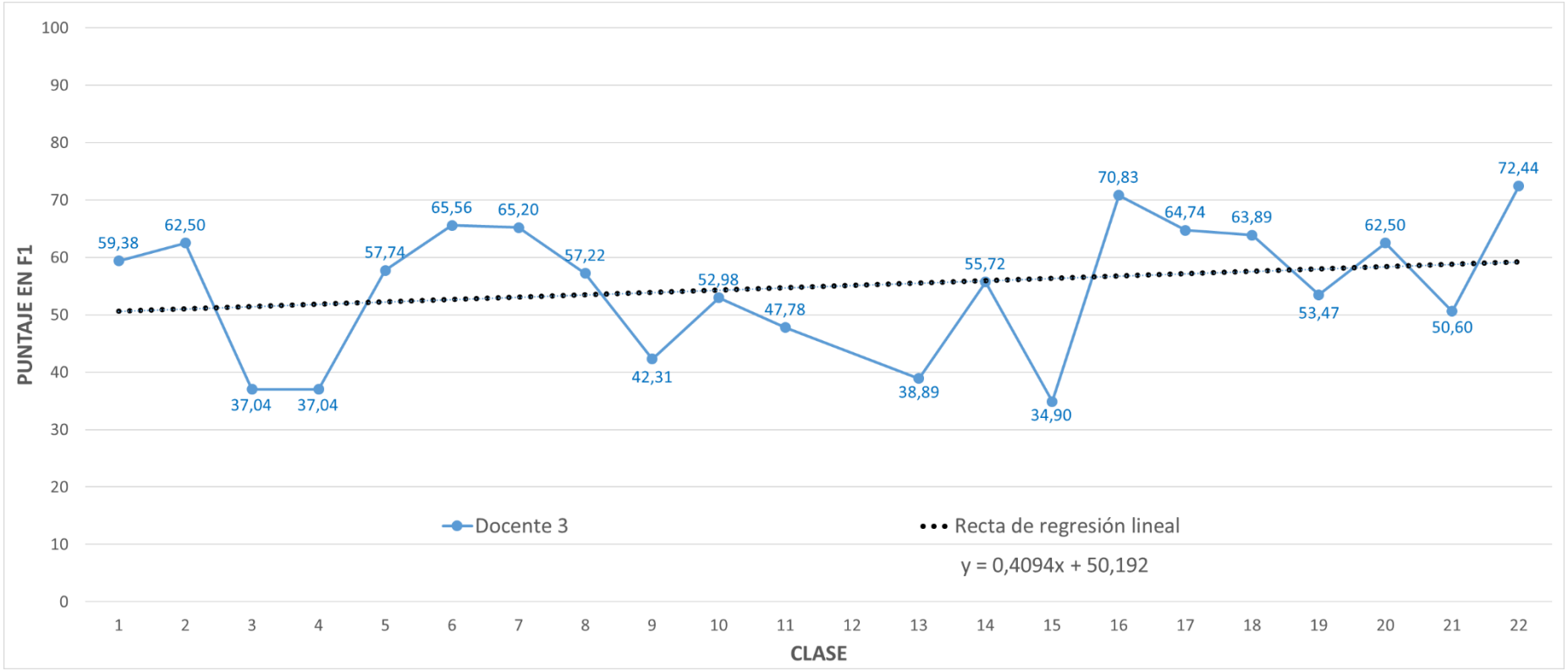
a) F1 - Docente 1 (n=47, $\overline{X}$=72,8)



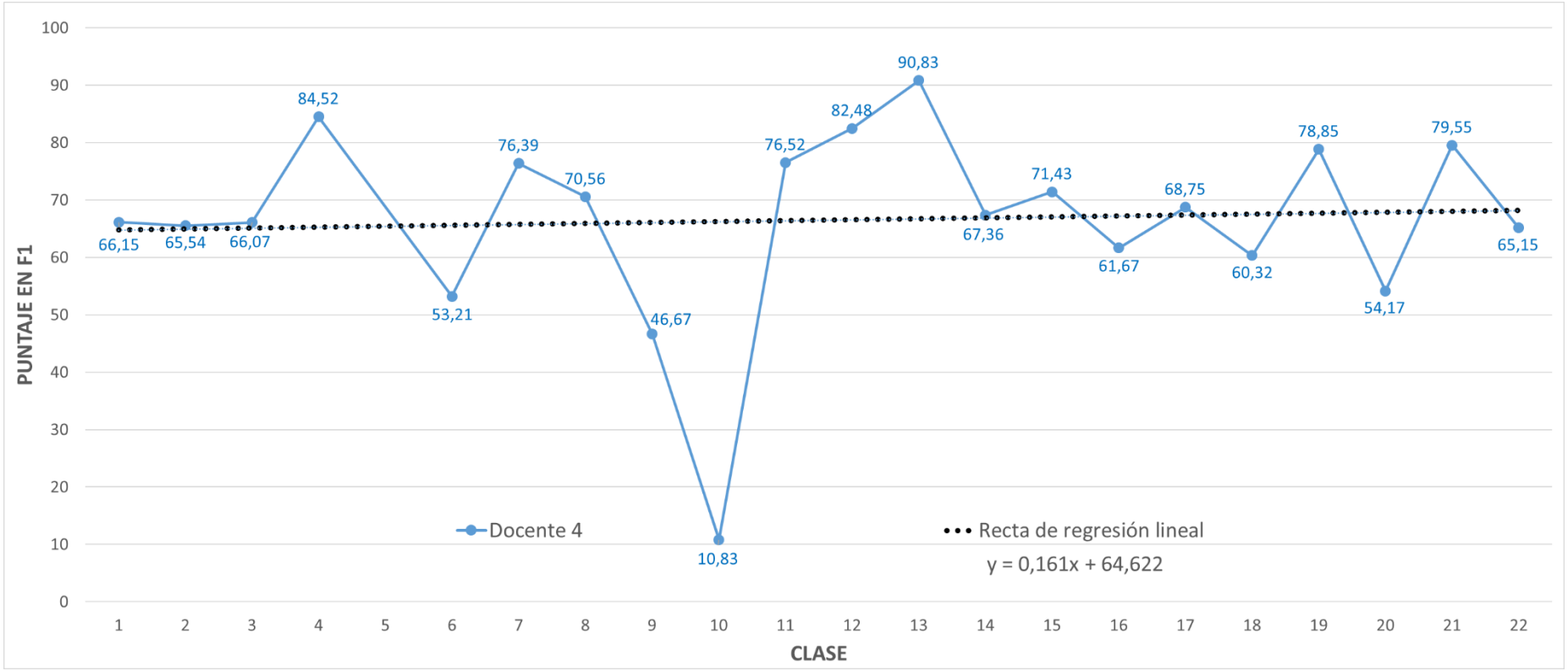
b) F1 - Docente 2 (n=8, $\overline{X}$=69)



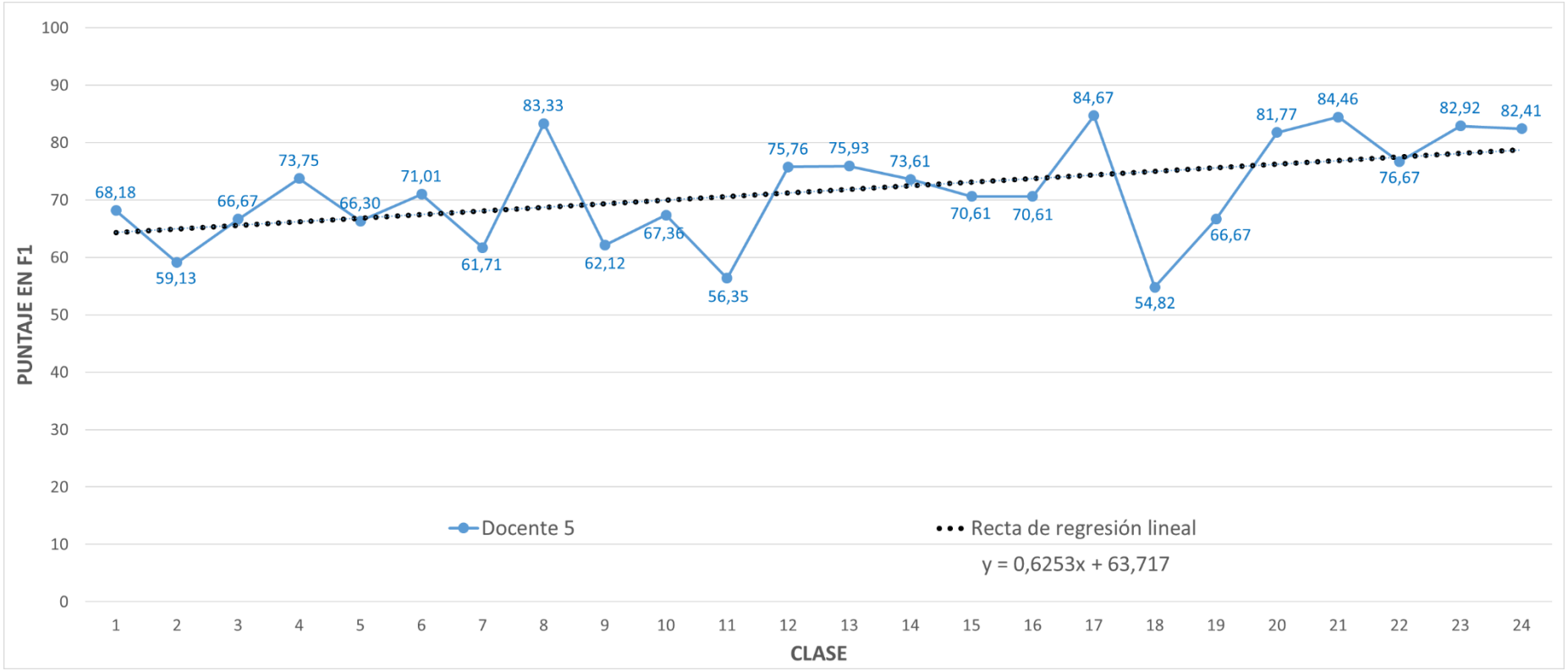
c) F1 - Docente 3 (n=22, $\overline{X}$=54,9)



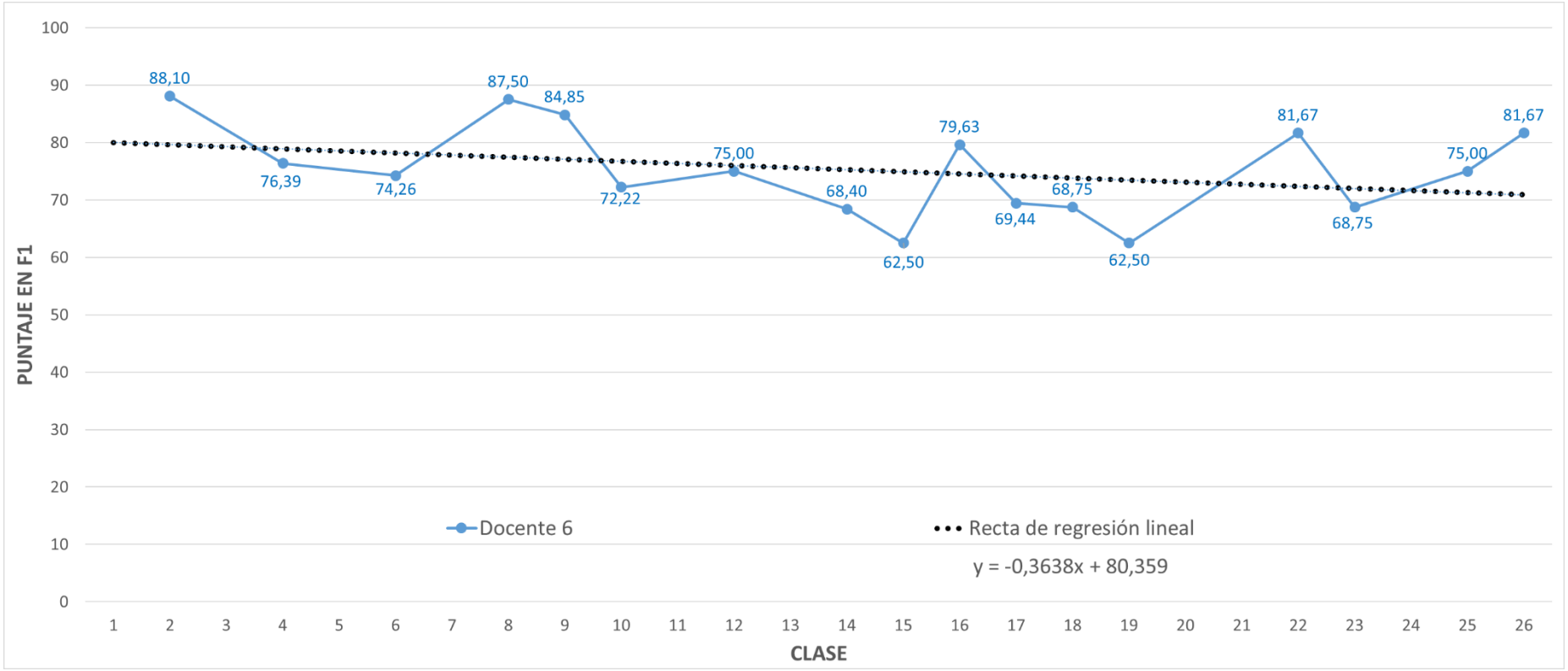
d) F1 - Docente 4 (n=17, $\overline{X}$=66,5)



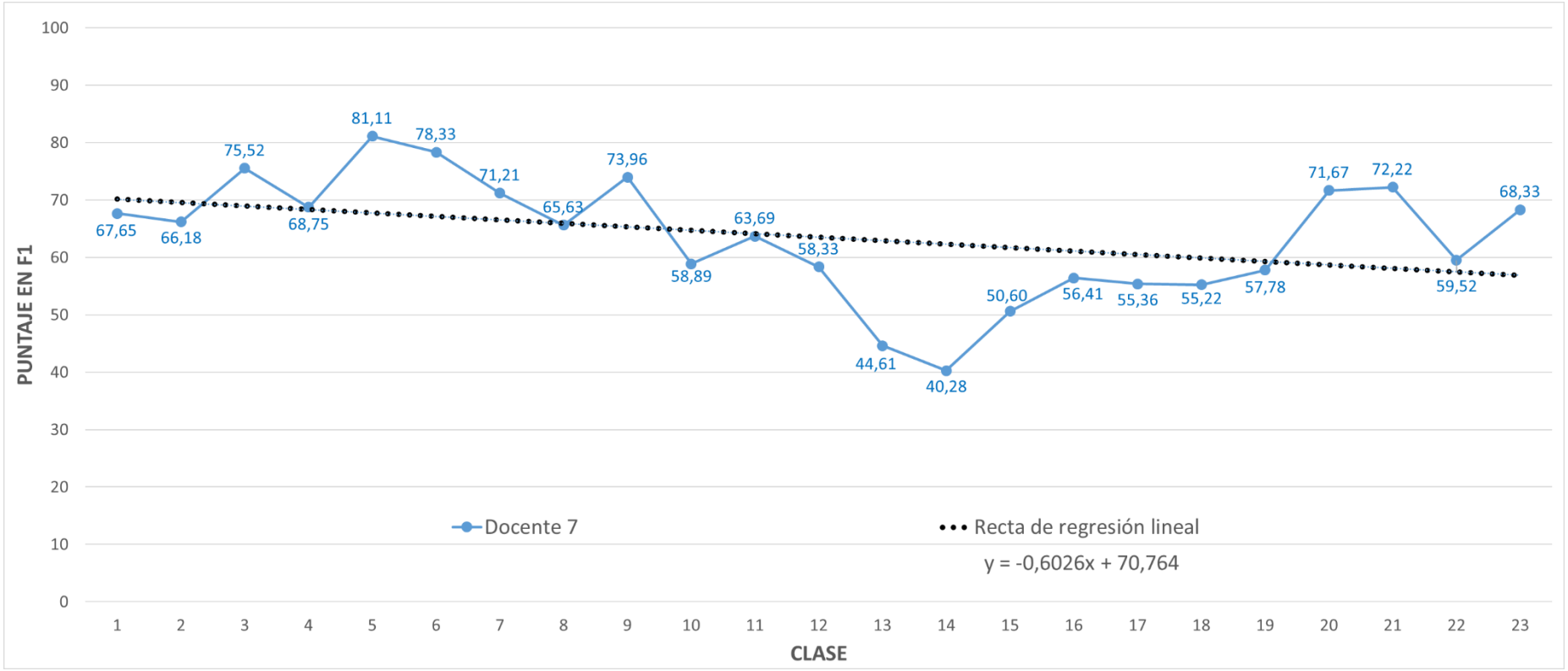
e) F1 - Docente 5 (n=28, $\overline{X}$=71,5)



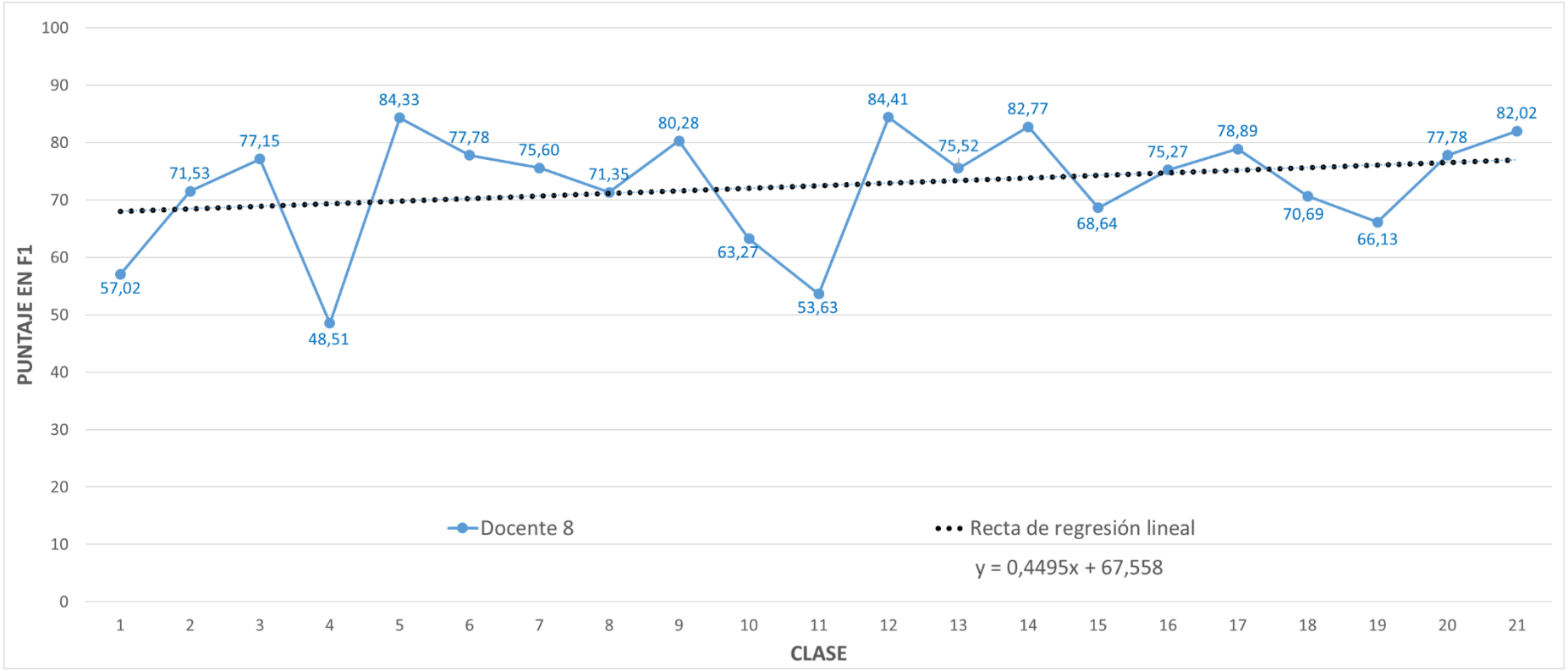
f) F1 - Docente 6 (n=14, $\overline{X}$=75,1)



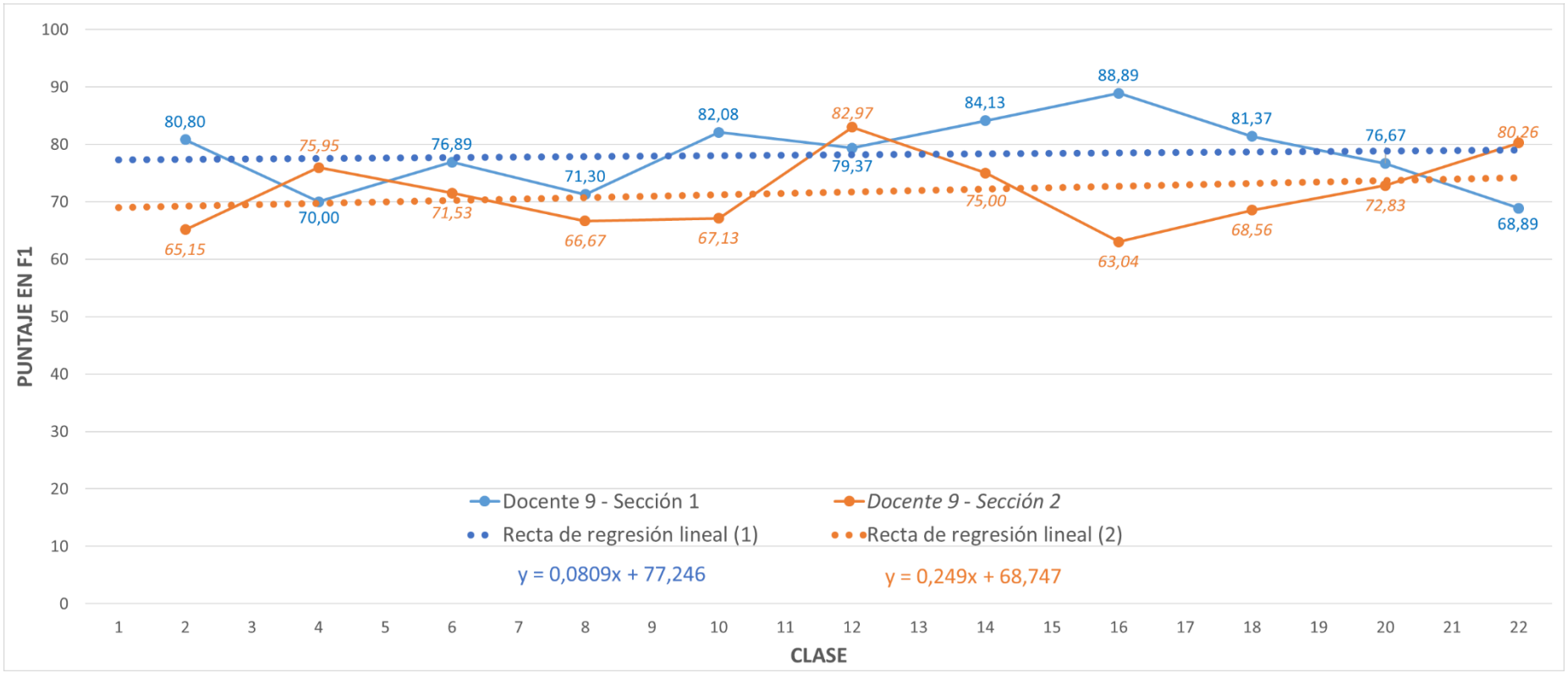
g) F1 - Docente 7 (n=18, $\overline{X}$=63,5)



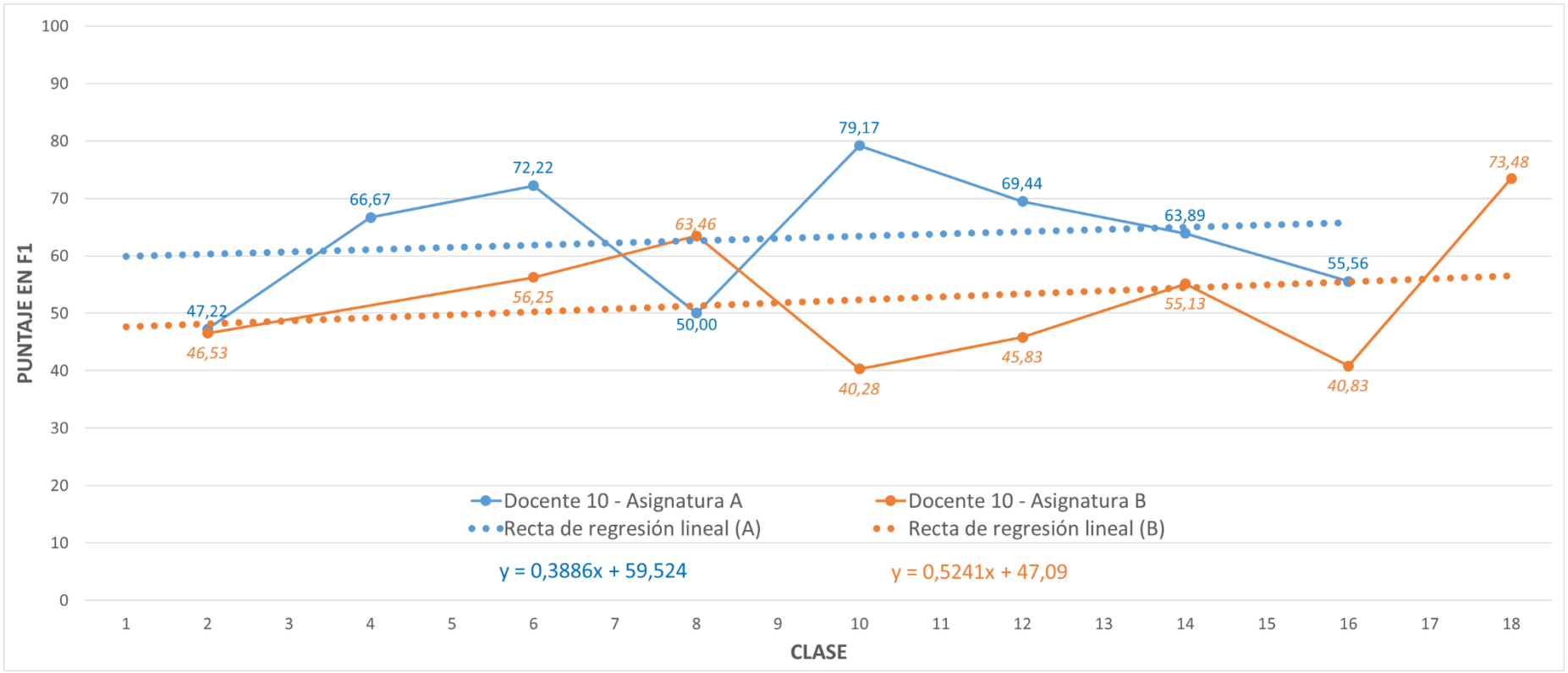
h) F1 - Docente 8 (n=43, $\overline{X}$=72,5)



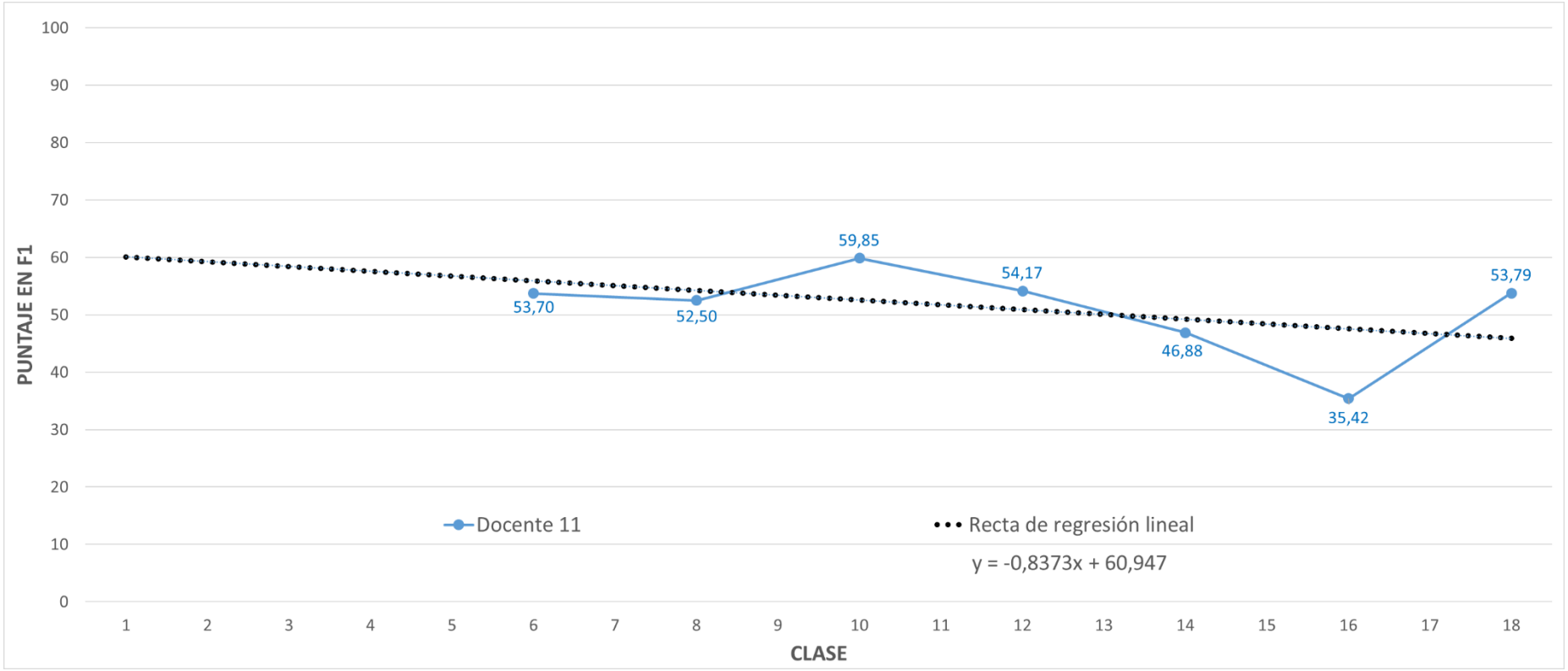
i) F1 - Docente 9 - Sección 1 (n=28, $\overline{X}$=78,2) y Sección 2 (n=33, $\overline{X}$=71,7)



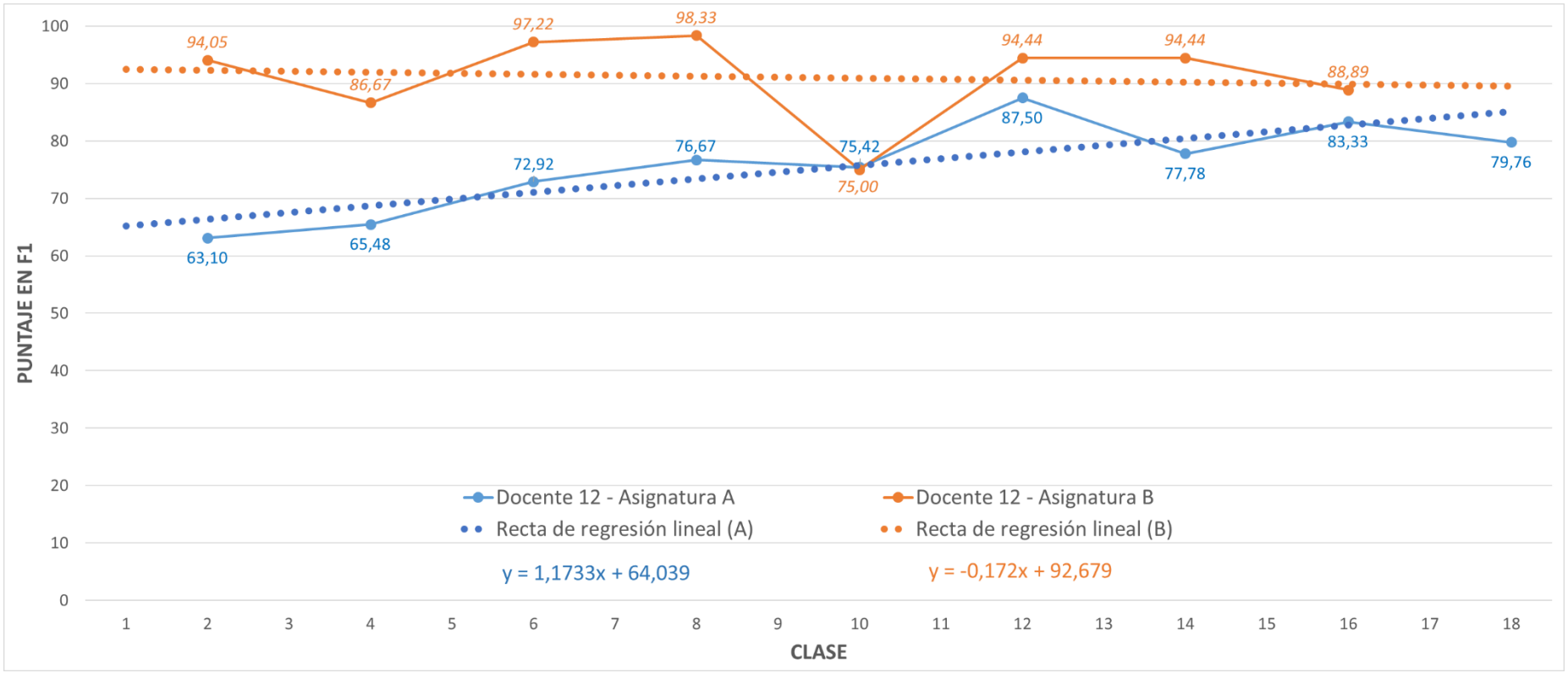
j) F1 - Docente 10 - Asignatura A (n=3, $\overline{X}$=63) y Asignatura B (n=16, $\overline{X}$=52,7)



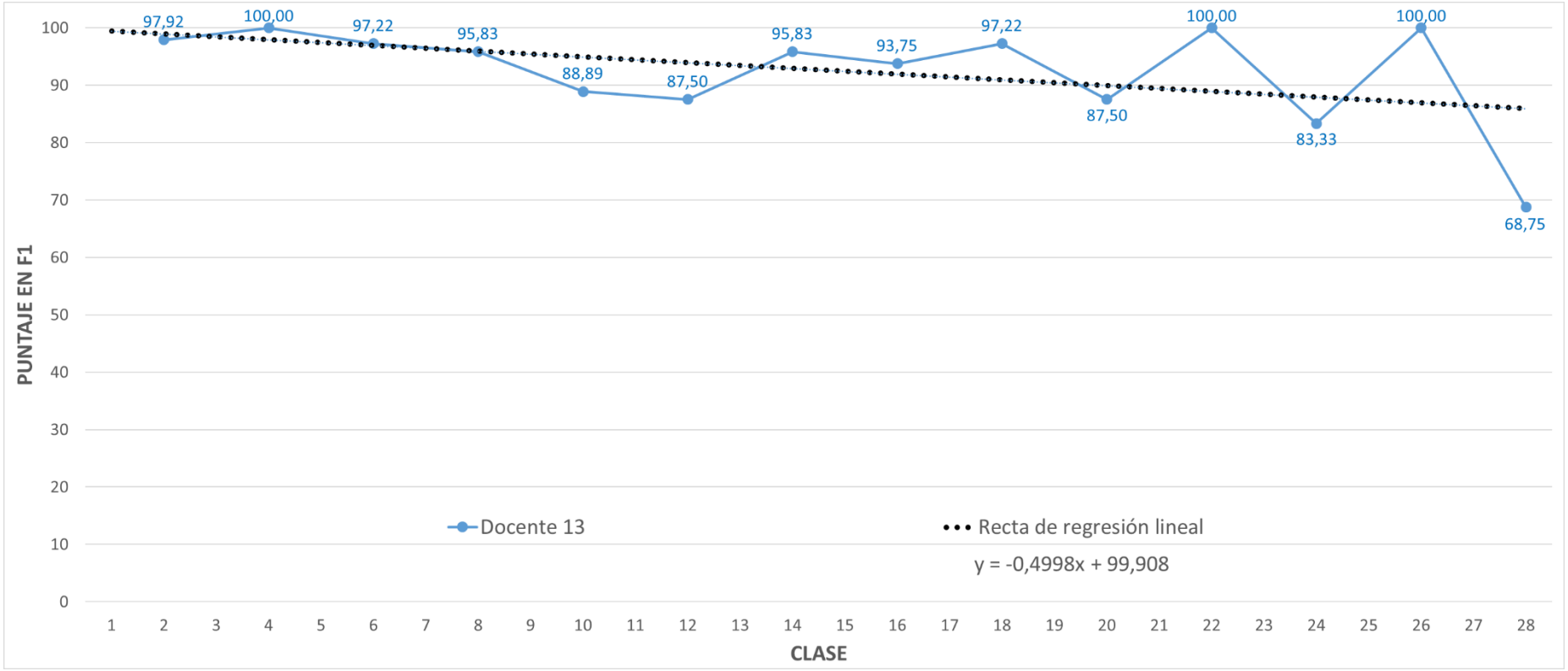
k) F1 - Docente 11 (n=13, $\overline{X}$=50,9)



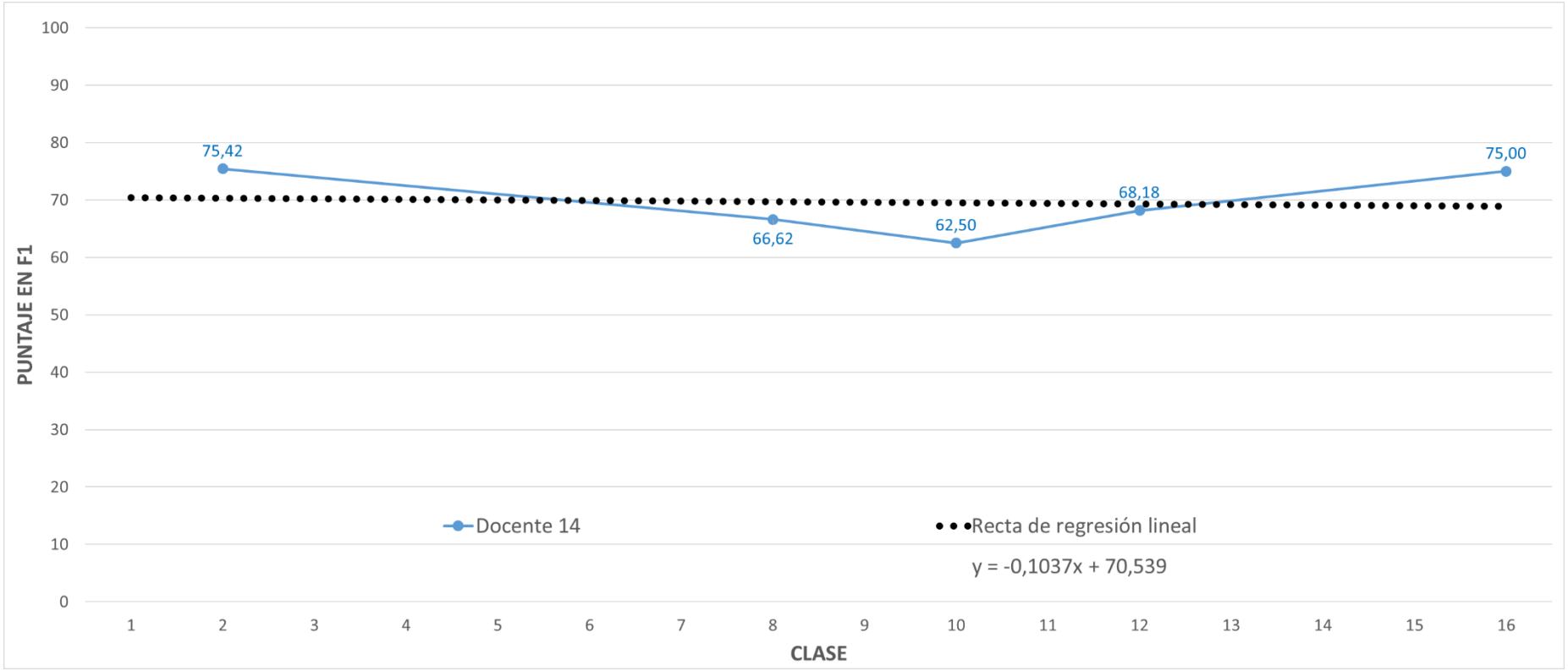
l) F1 - Docente 12 - Asignatura A (n=9, $\overline{X}$=75,8) y Asignatura B (n=7, $\overline{X}$=91,1)



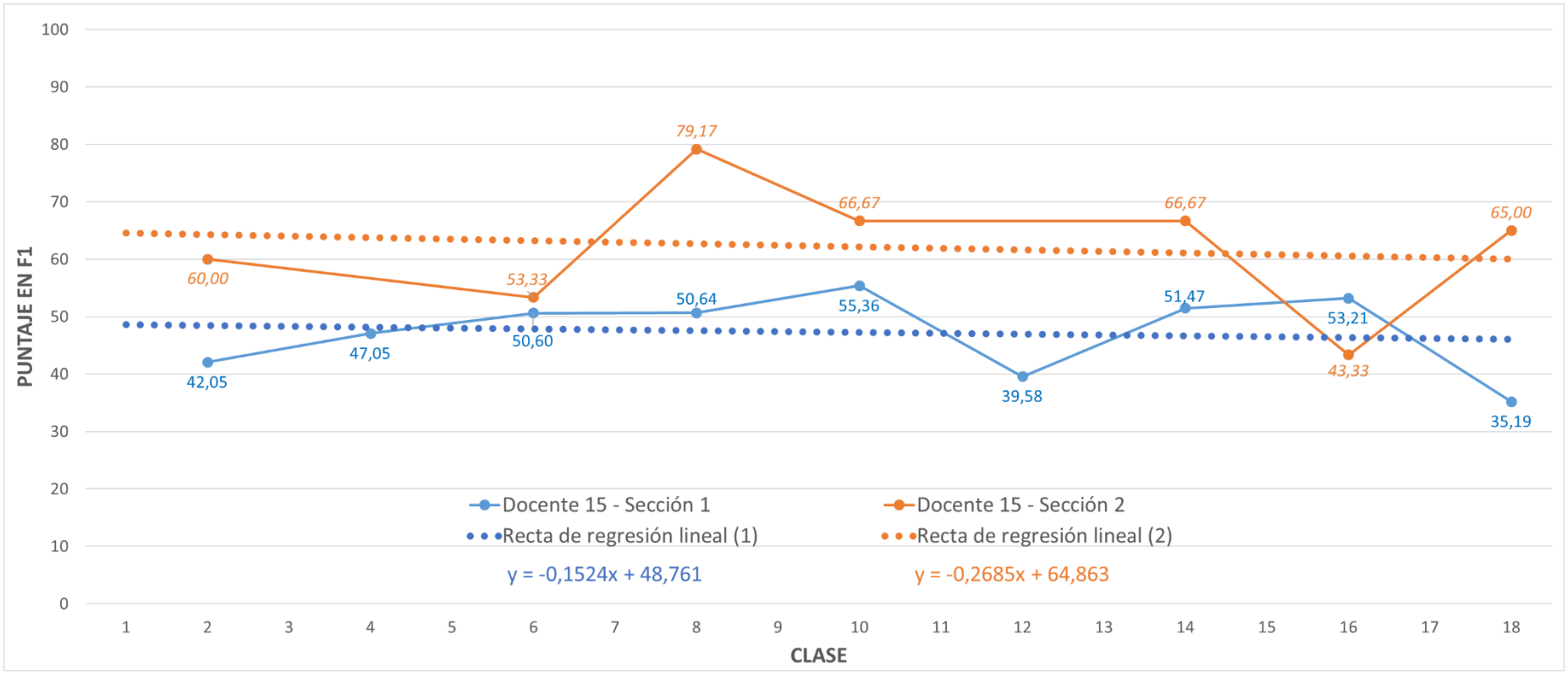
m) F1 - Docente 13 (n=4, $\overline{X}$=92,4)



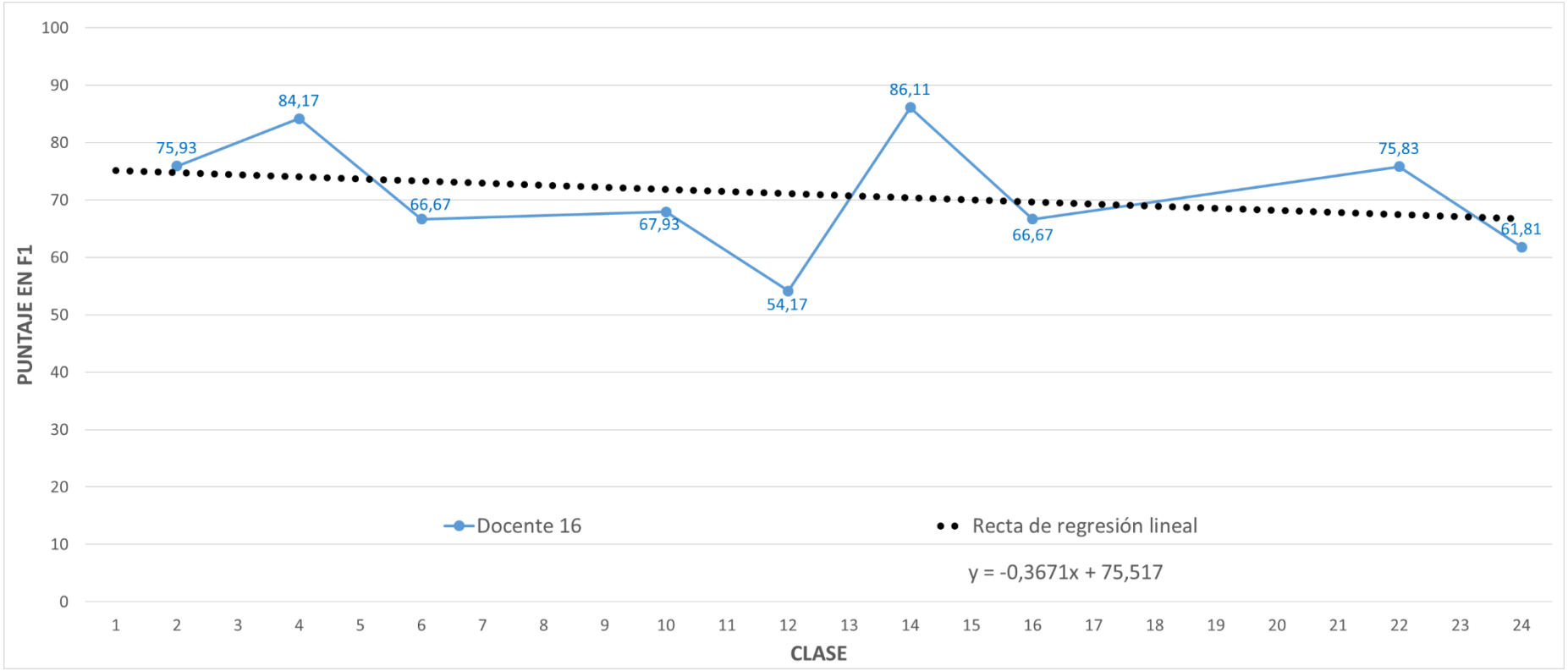
n) F1 - Docente 14 (n=25, $\overline{X}$=69,5)



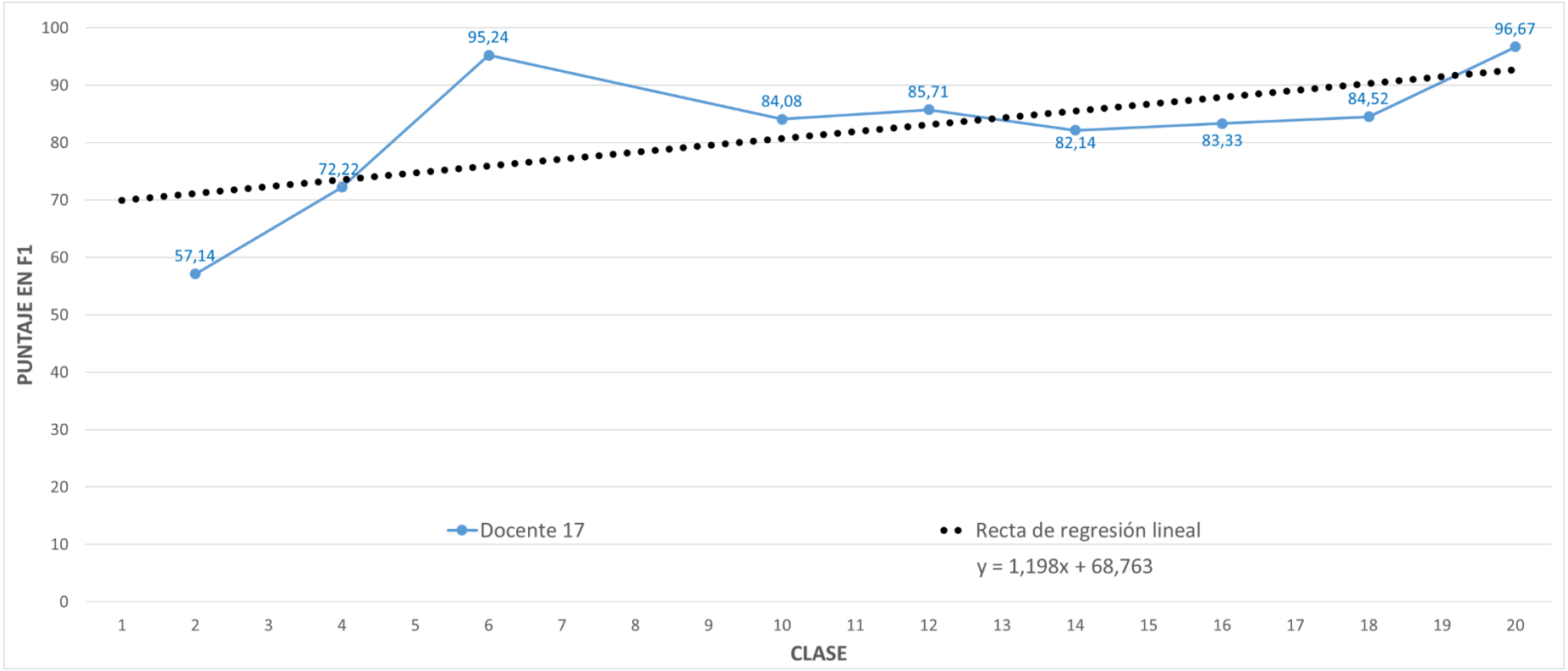
o) F1 - Docente 15 - Sección 1 (n=20, $\overline{X}$=47,2) y Sección 2 (n=15, $\overline{X}$=62)



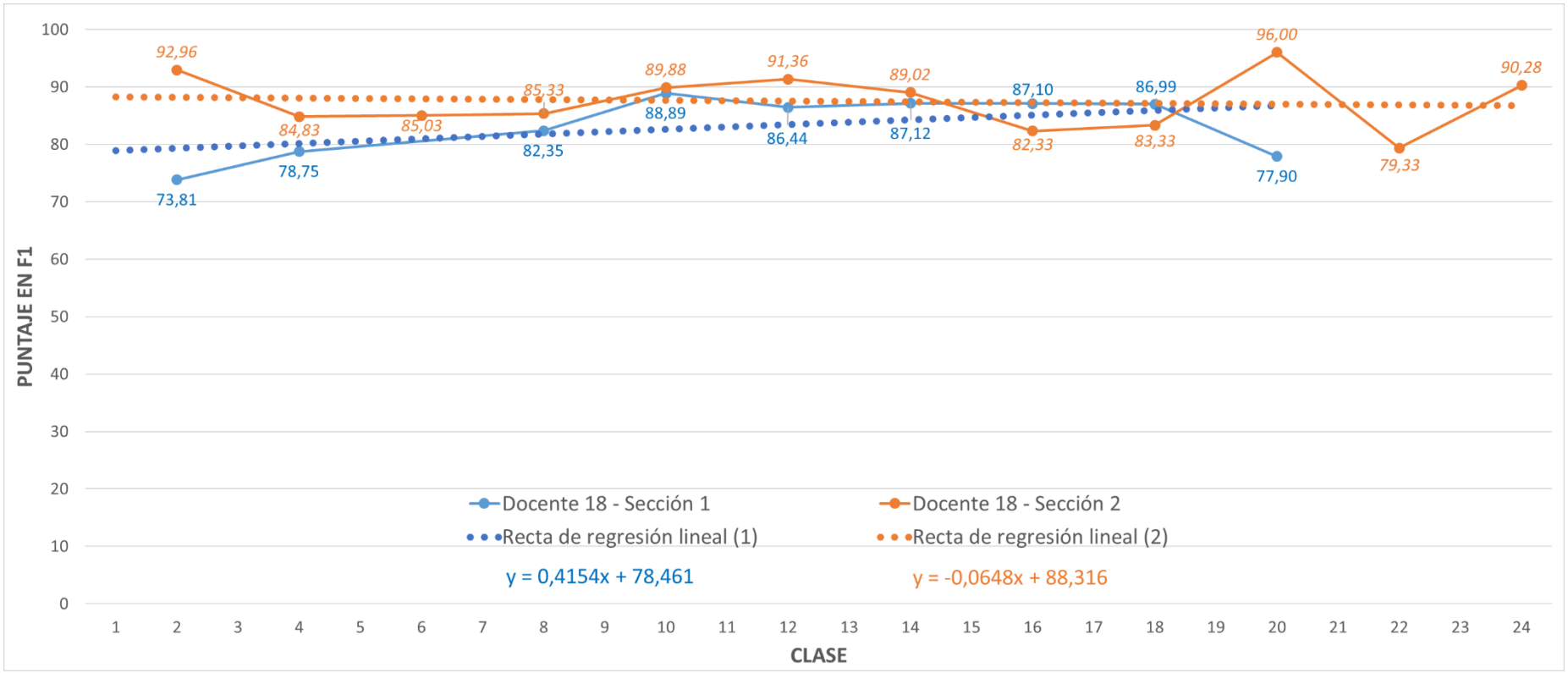
p) F1 - Docente 16 (n=15, $\overline{X}$=71)



q) F1 - Docente 17 (n=8, $\overline{X}$=82,3)

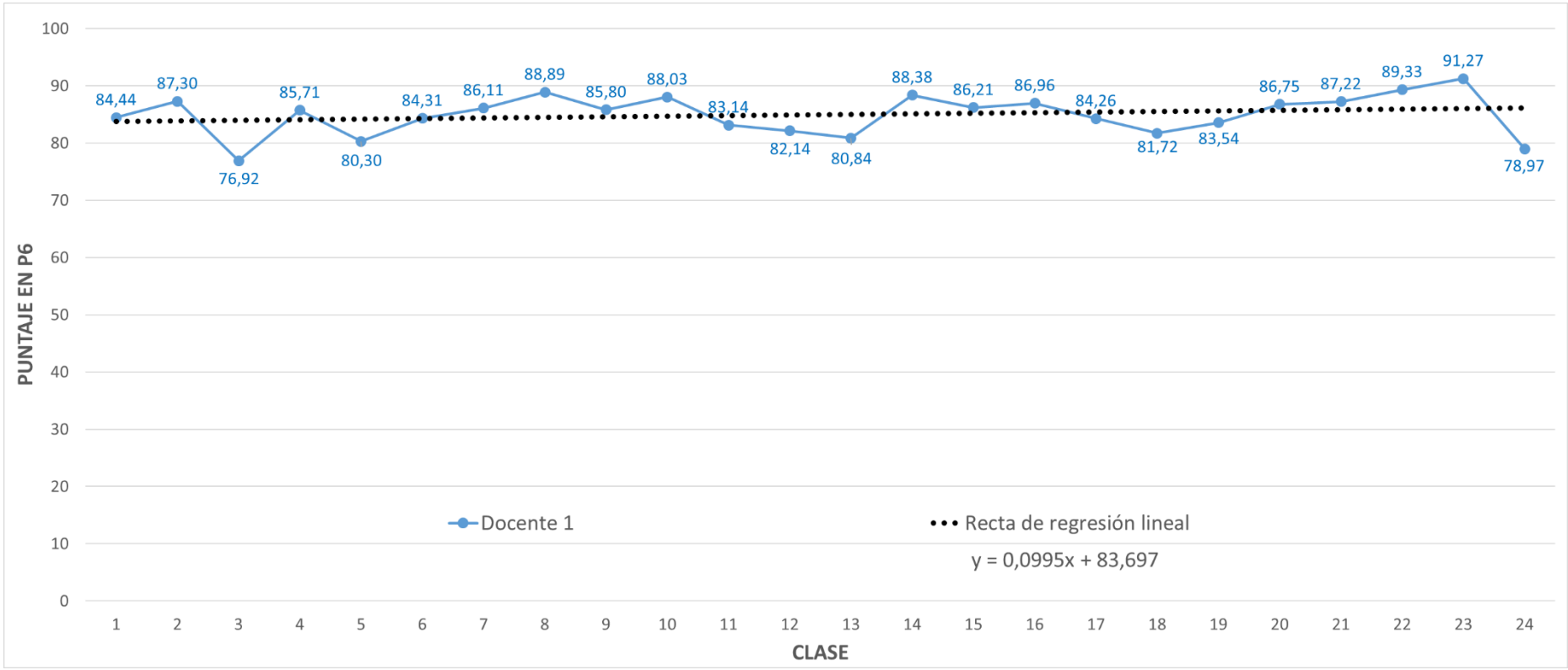


r) F1 - Docente 18 - Sección 1 (n=49, $\overline{X}$=83,3) y Sección 2 (n=36, $\overline{X}$=87,5)

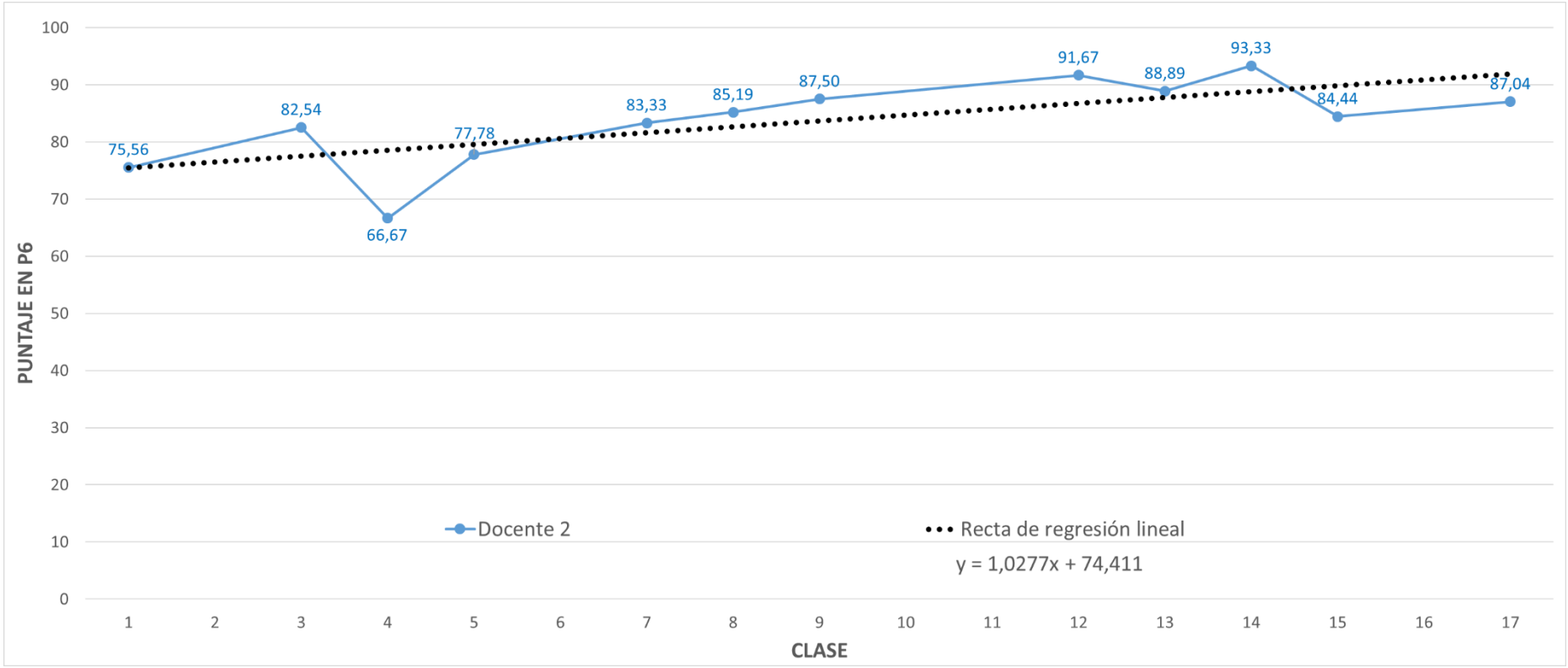


- ***Series de ítem P6: "Calidad general de la clase"***

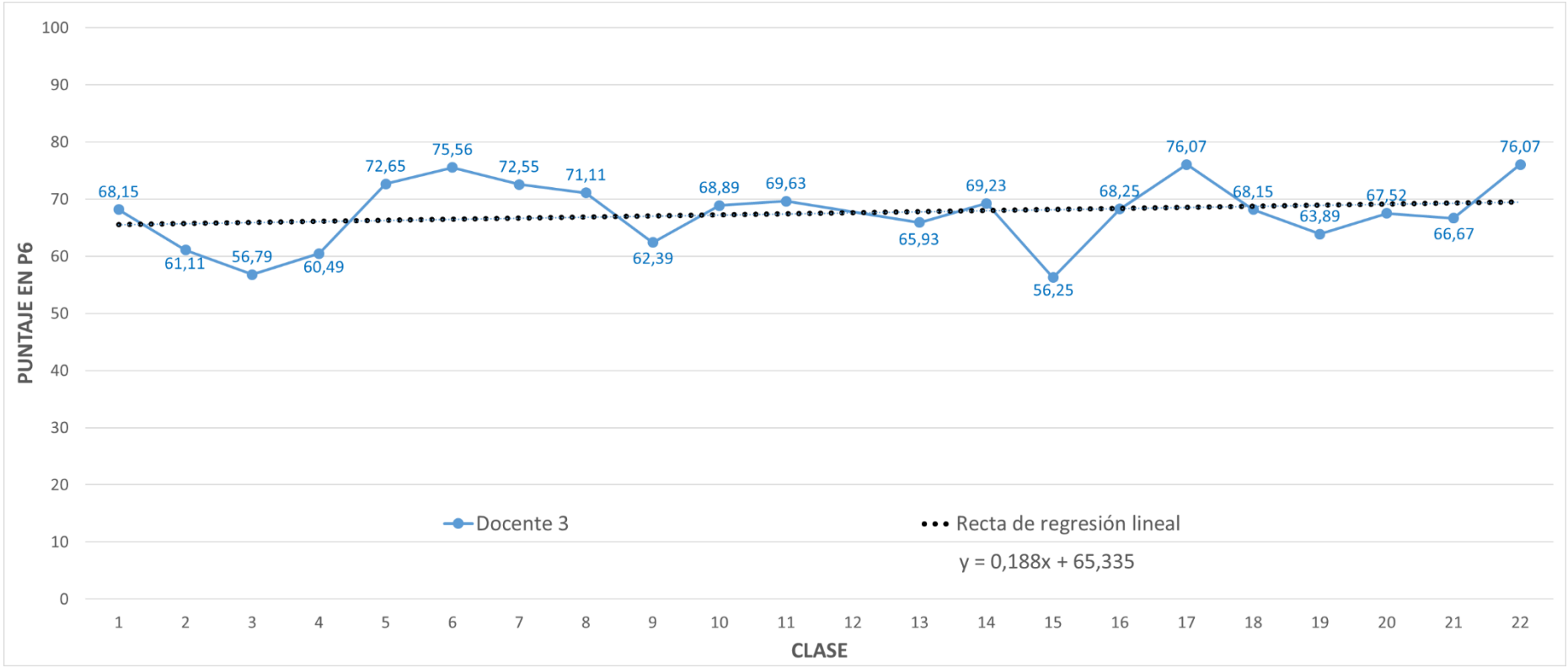
a) P6 - Docente 1 (n=47, $\overline{x}$=84,9)



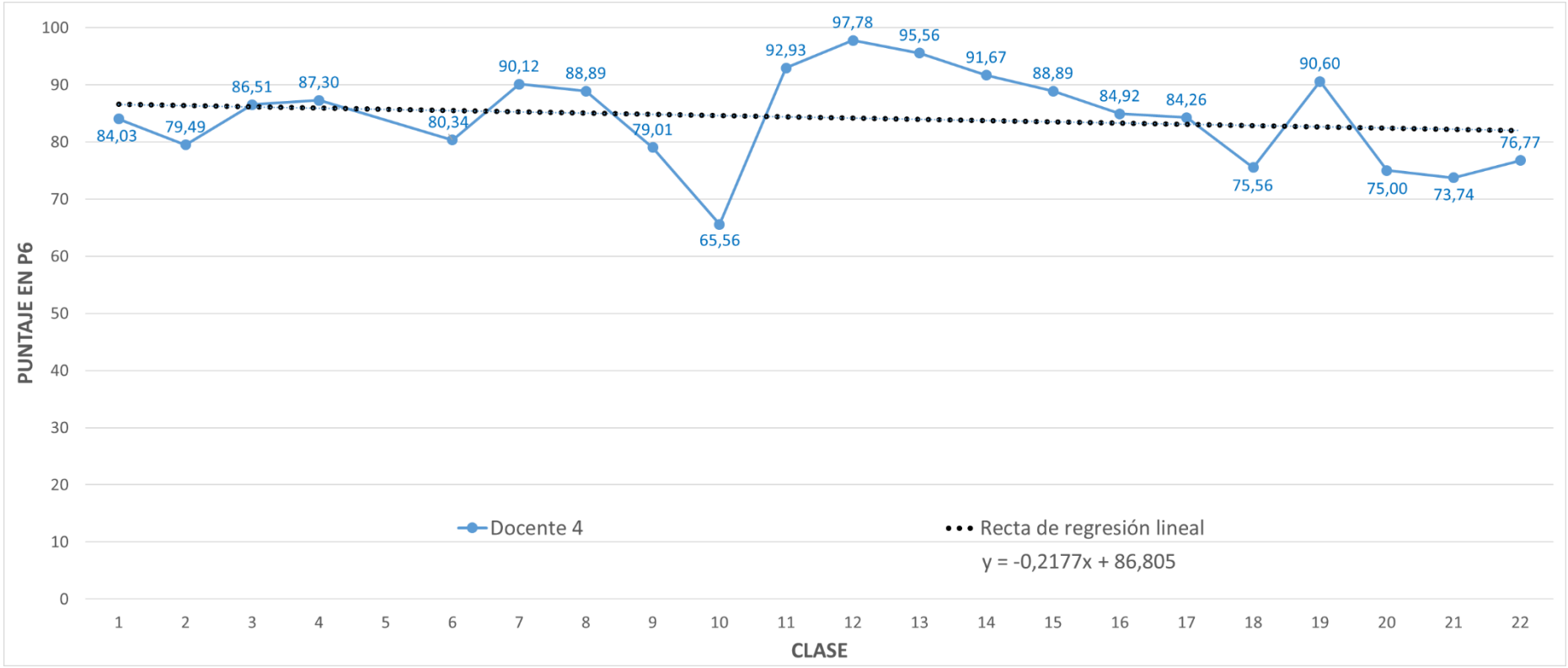
b) P6 - Docente 2 (n=8, $\overline{X}$=83,7)



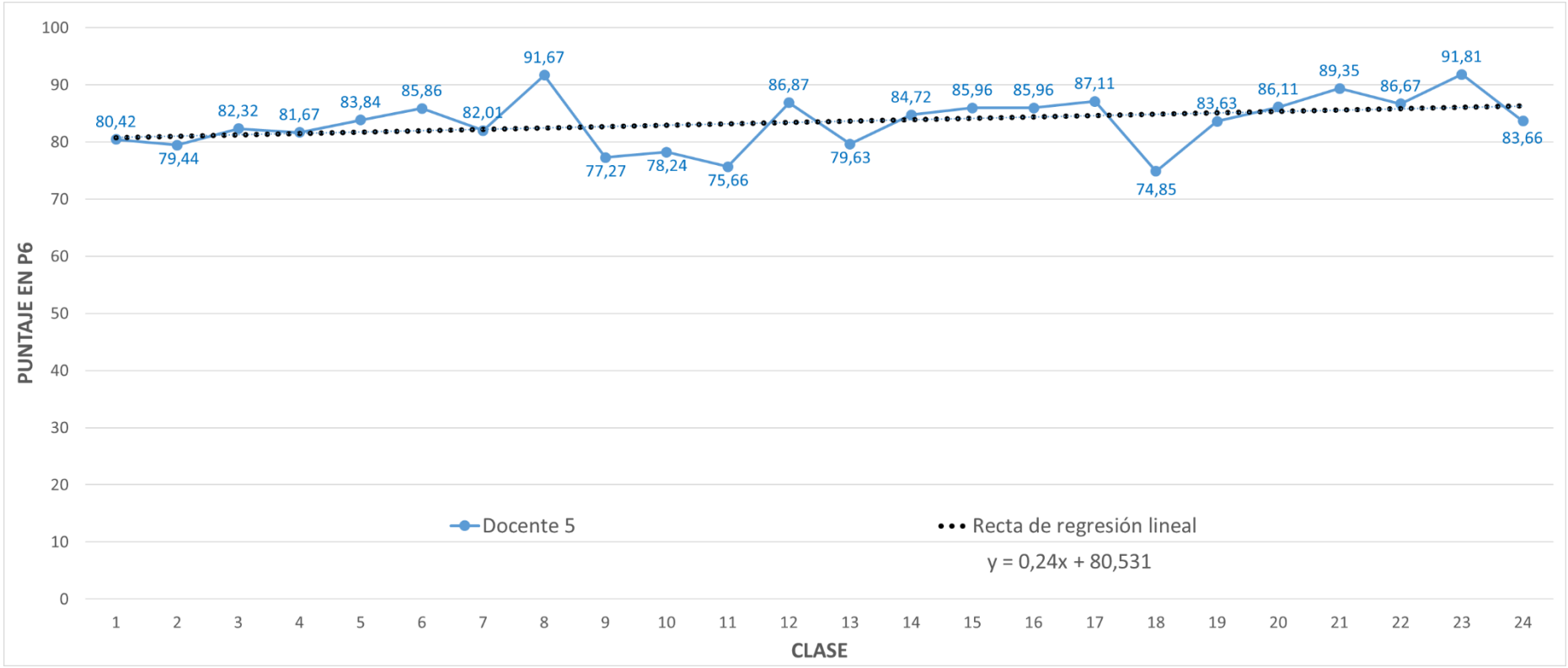
c) P6 - Docente 3 (n=22, $\overline{X}$=67,5)



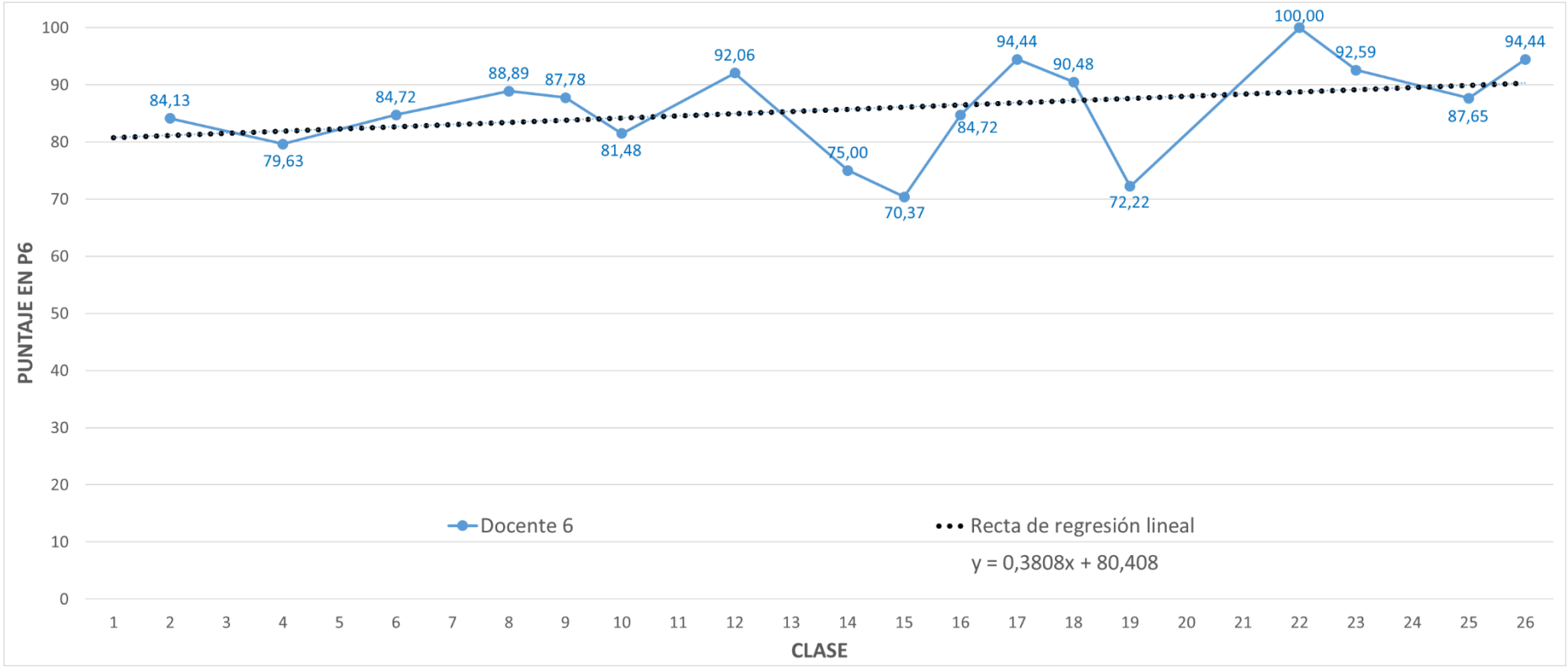
d) P6 - Docente 4 (n=17, $\overline{X}$=84,2)



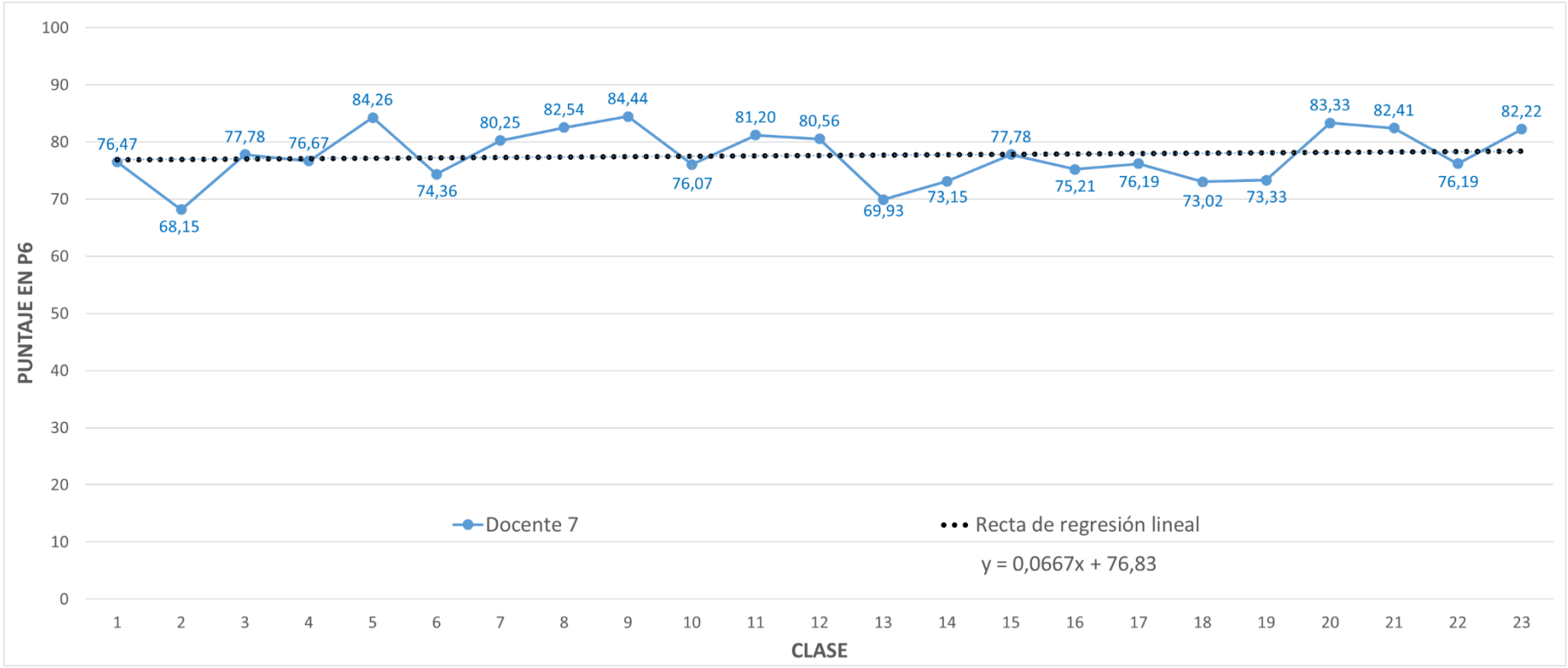
e) P6 - Docente 5 (n=28, $\overline{X}$=83,5)



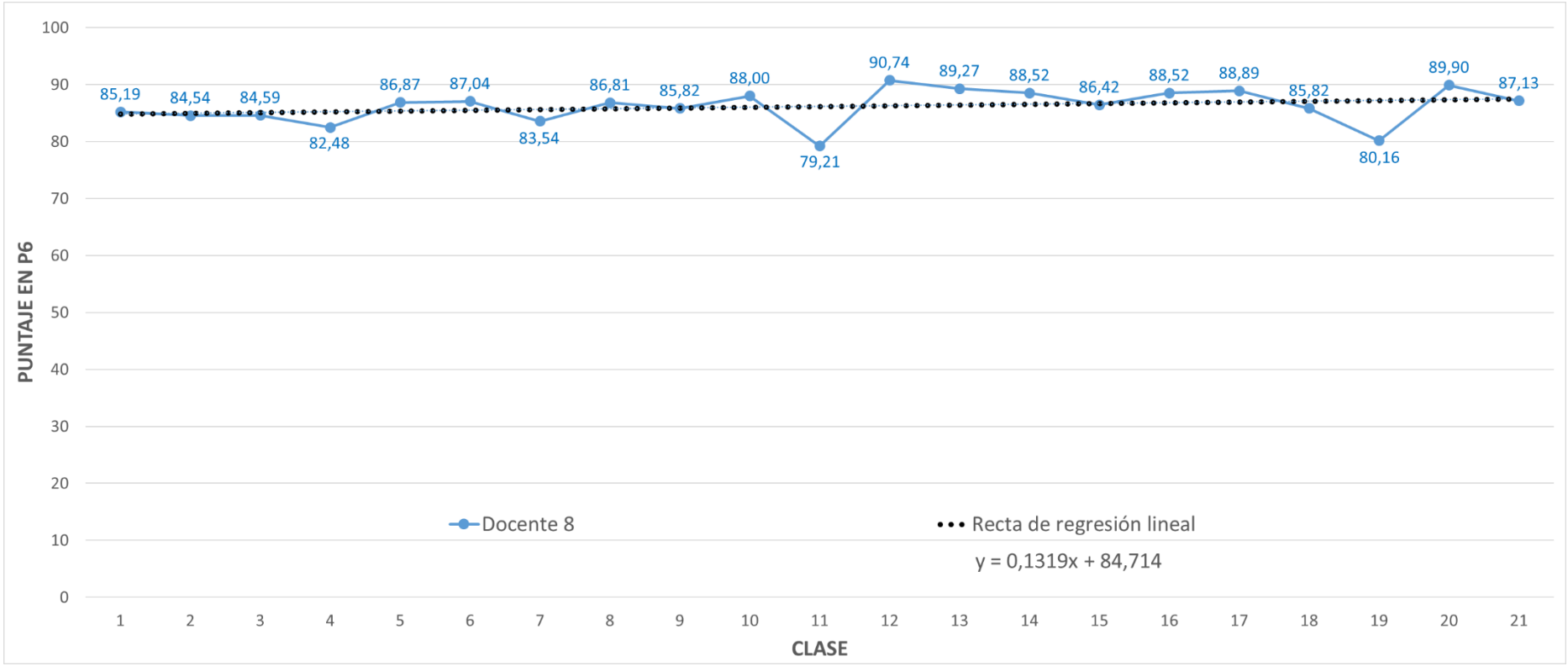
f) P6 - Docente 6 (n=14, $\overline{X}$=85,9)



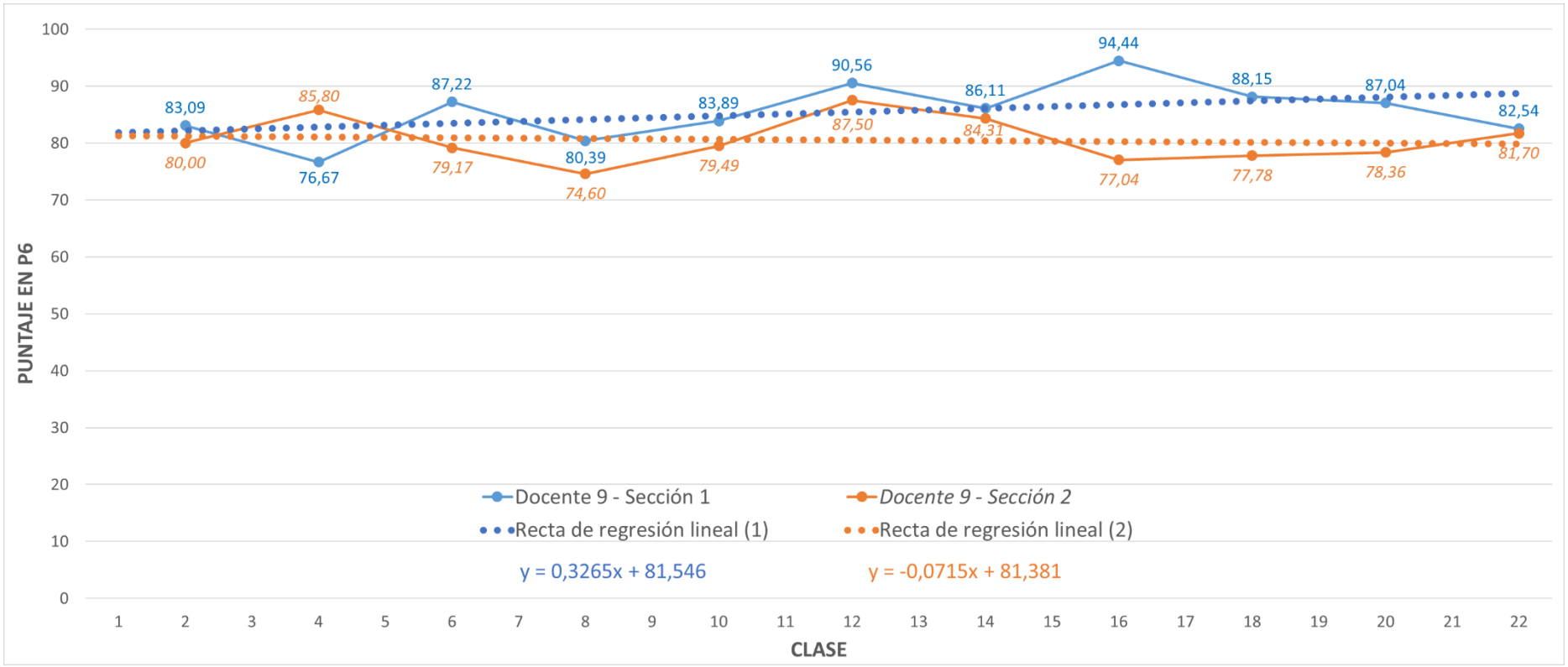
g) P6 - Docente 7 (n=18, $\overline{X}$=77,6)



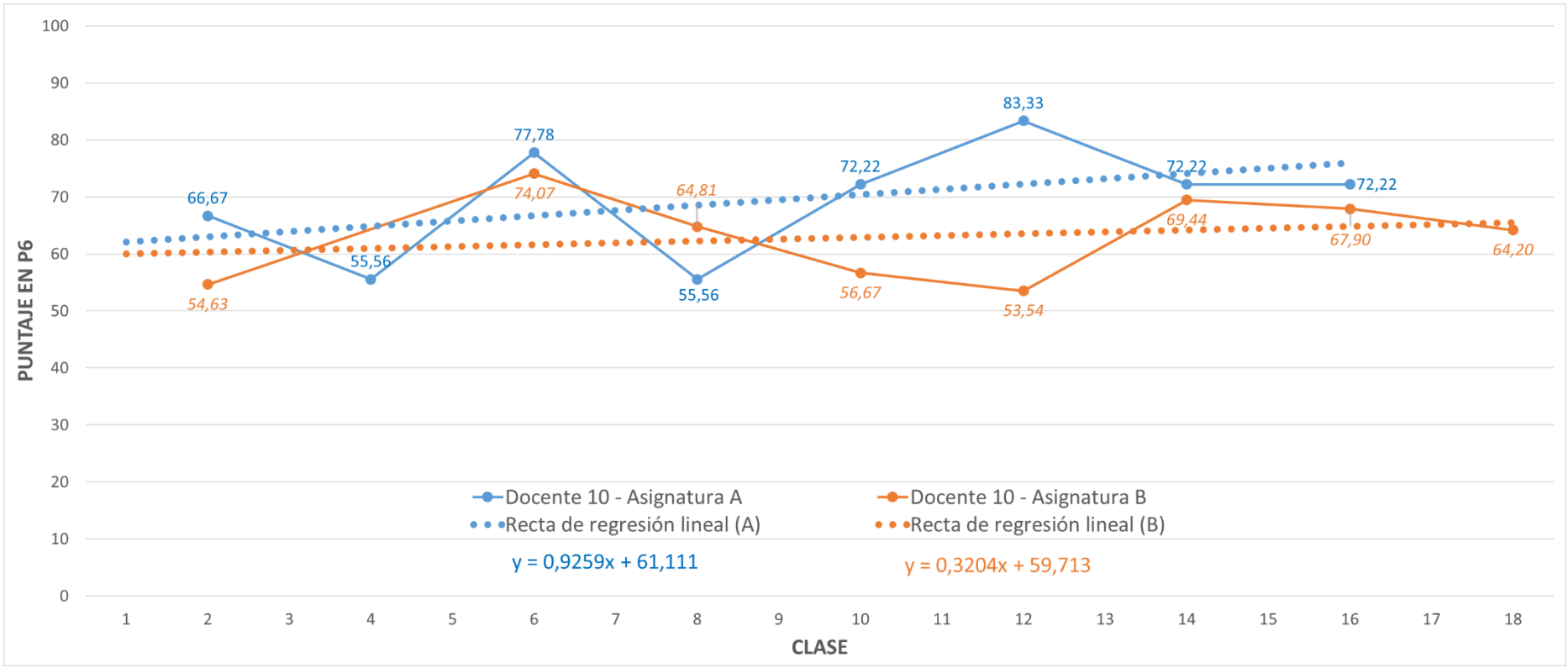
h) P6 - Docente 8 (n=43, $\overline{X}$=86,2)



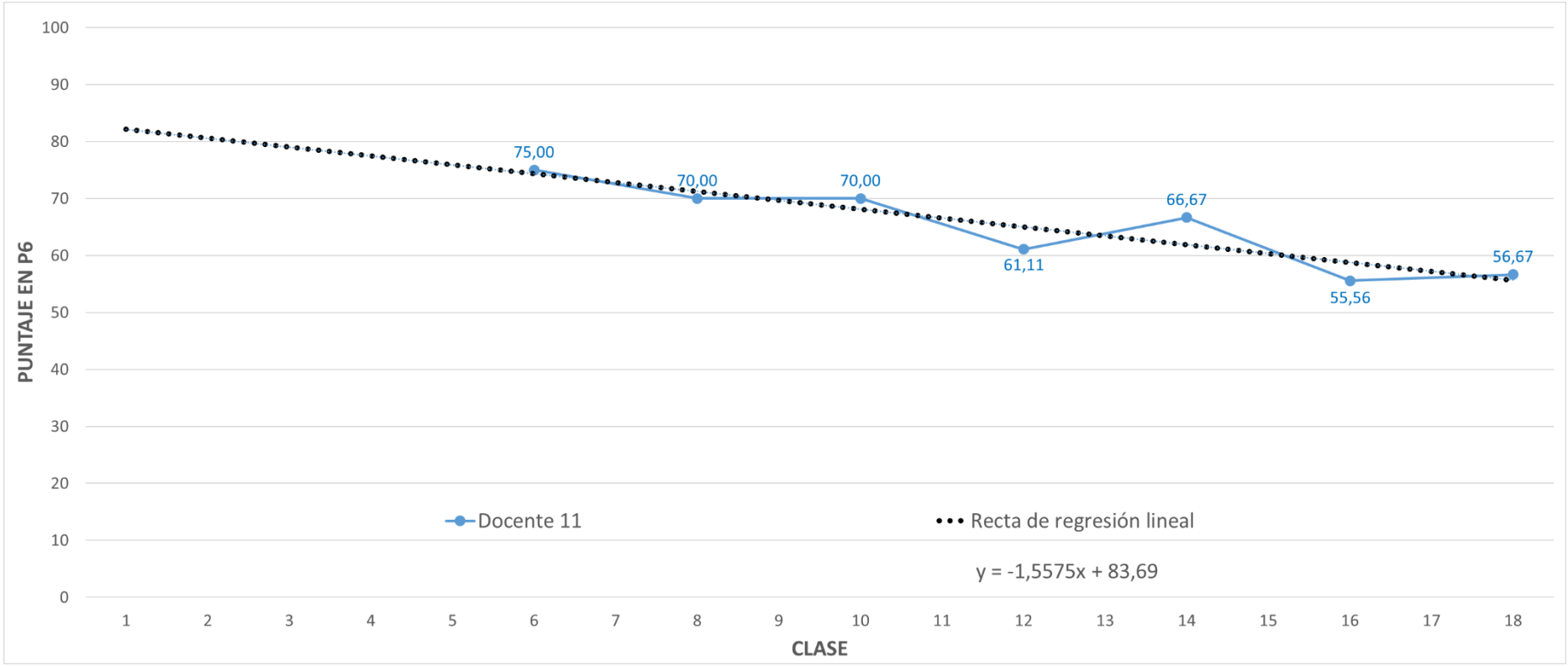
i) P6 - Docente 9 - Sección 1 (n=28, $\overline{X}$=85,5) y Sección 2 (n=33, $\overline{X}$=80,5)



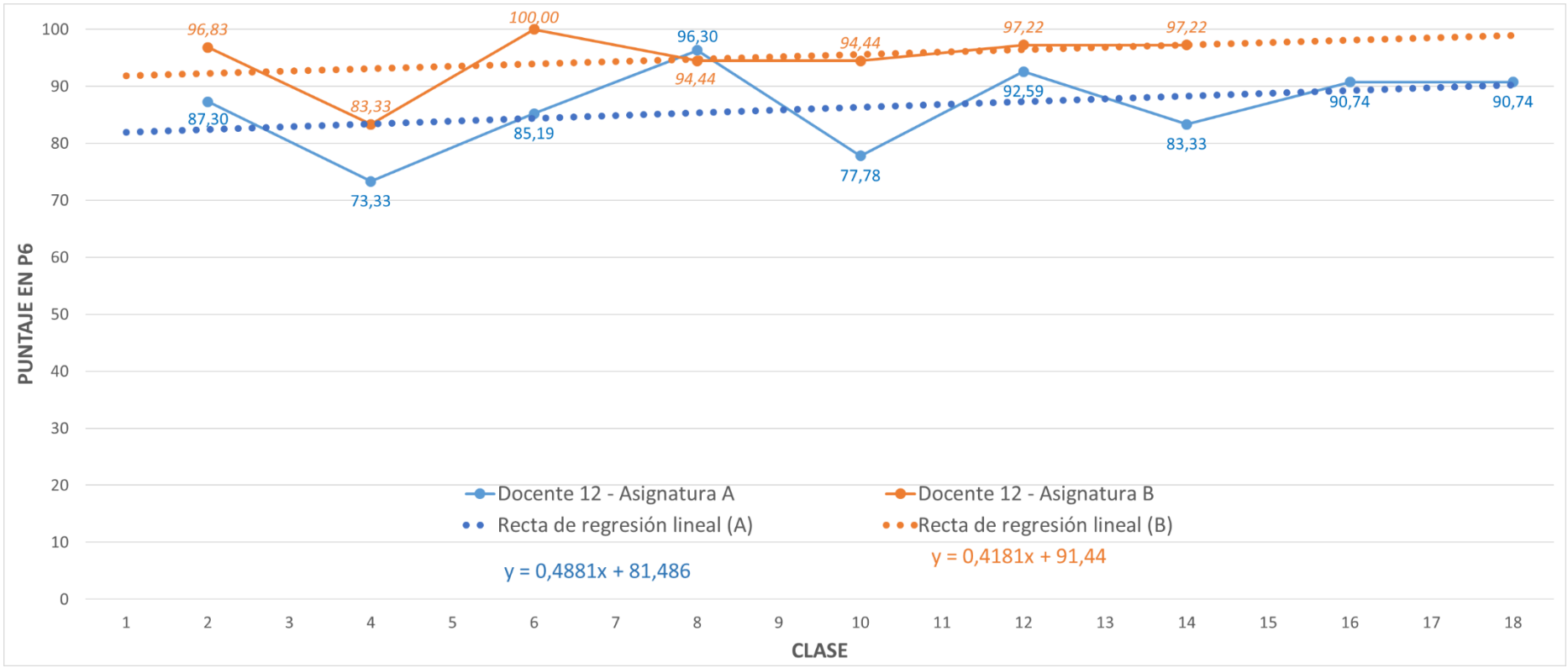
j) P6 - Docente 10 - Asignatura A (n=3, $\overline{X}$=69,4) y Asignatura B (n=16, $\overline{X}$=63,2)



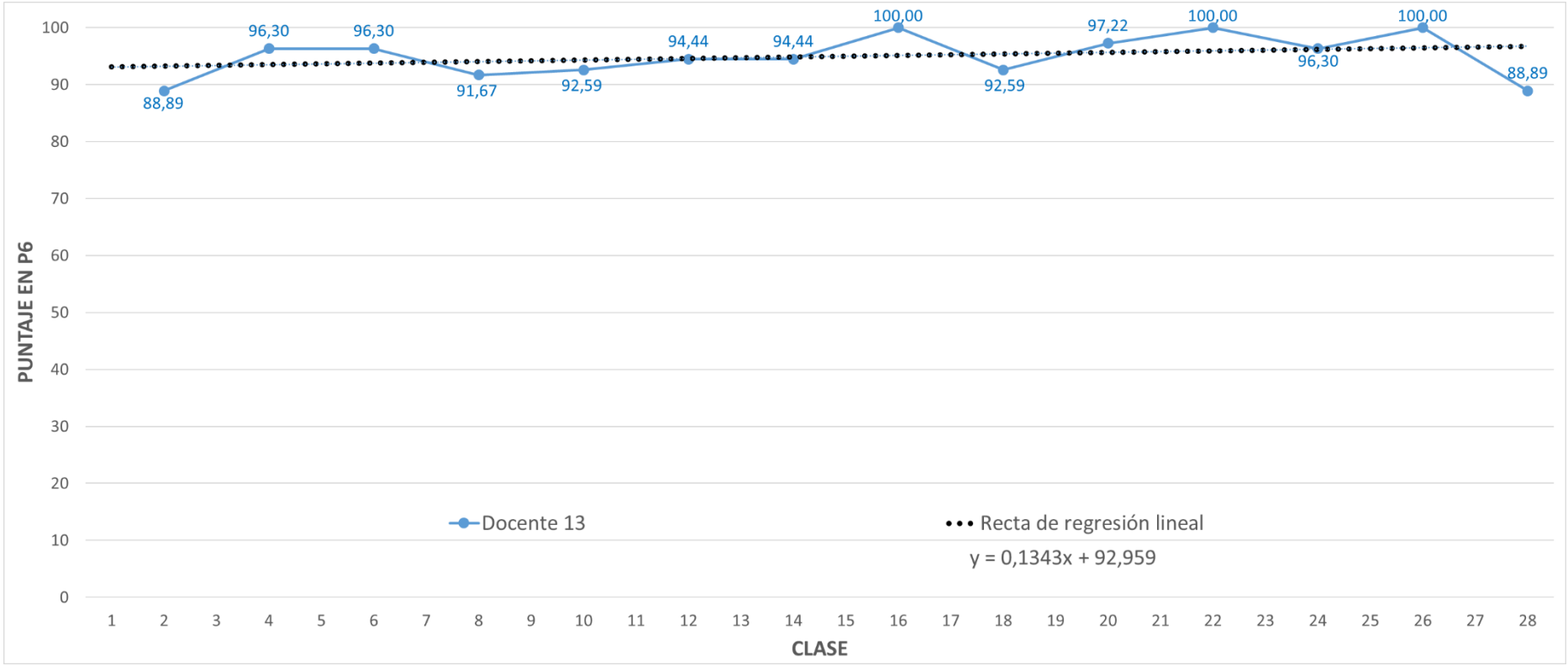
k) P6 - Docente 11 (n=13, $\overline{X}$=65)



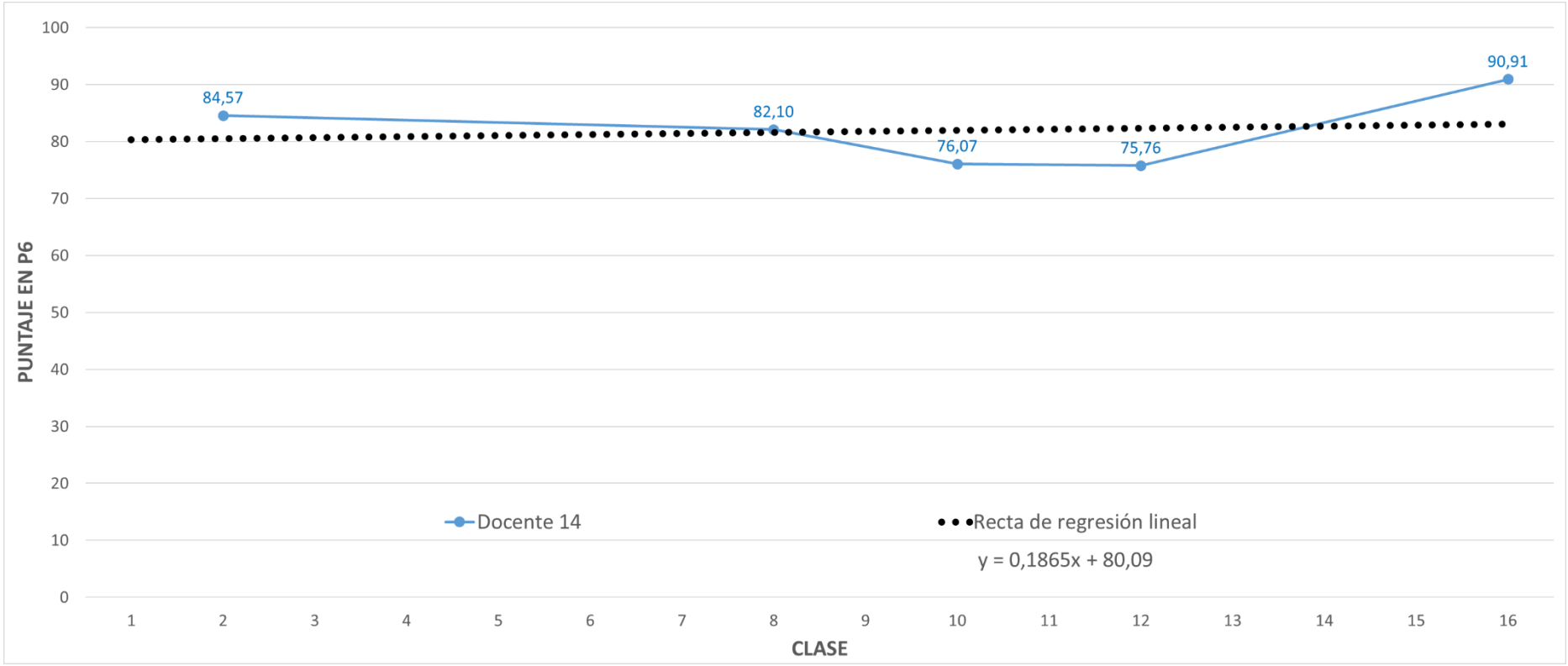
l) P6 - Docente 12 - Asignatura A (n=9, $\overline{X}$=86,4) y Asignatura B (n=7, $\overline{X}$=94,8)



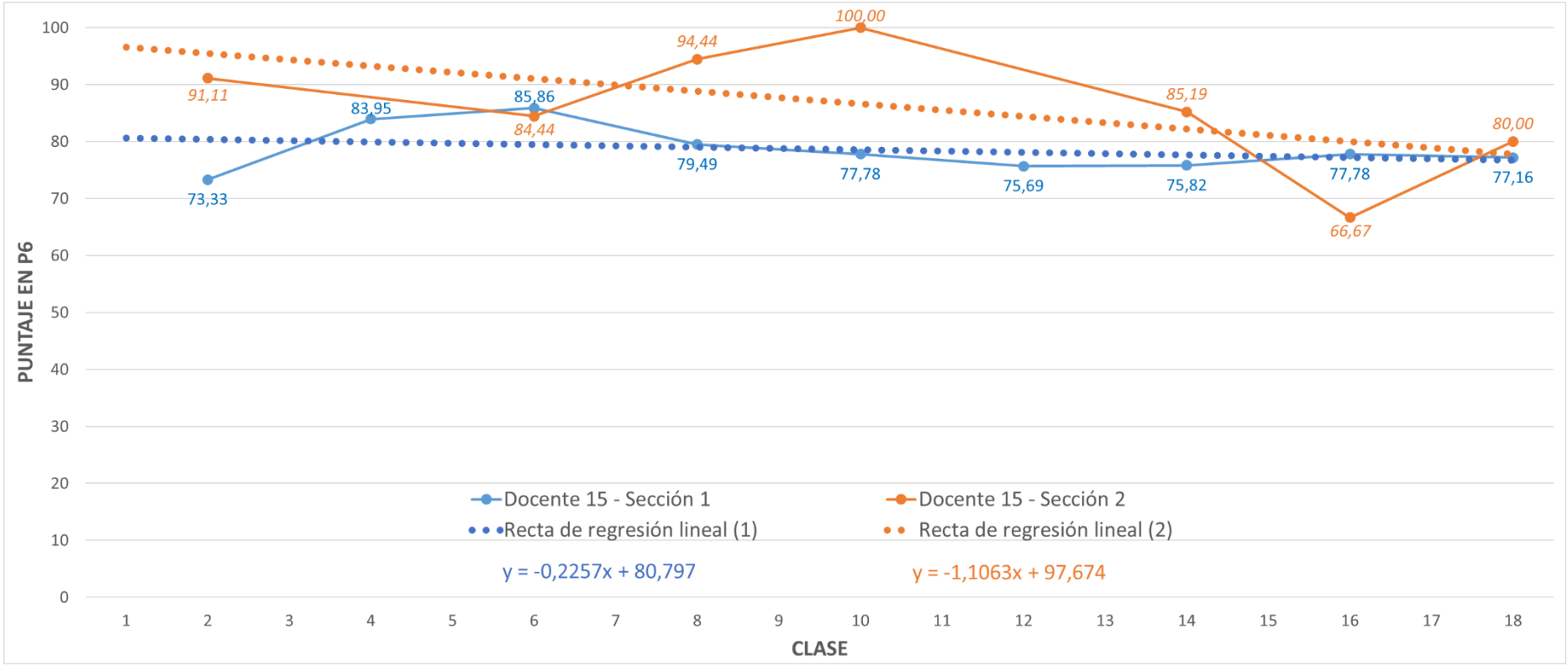
m) P6 - Docente 13 (n=4, $\overline{X}$=95)



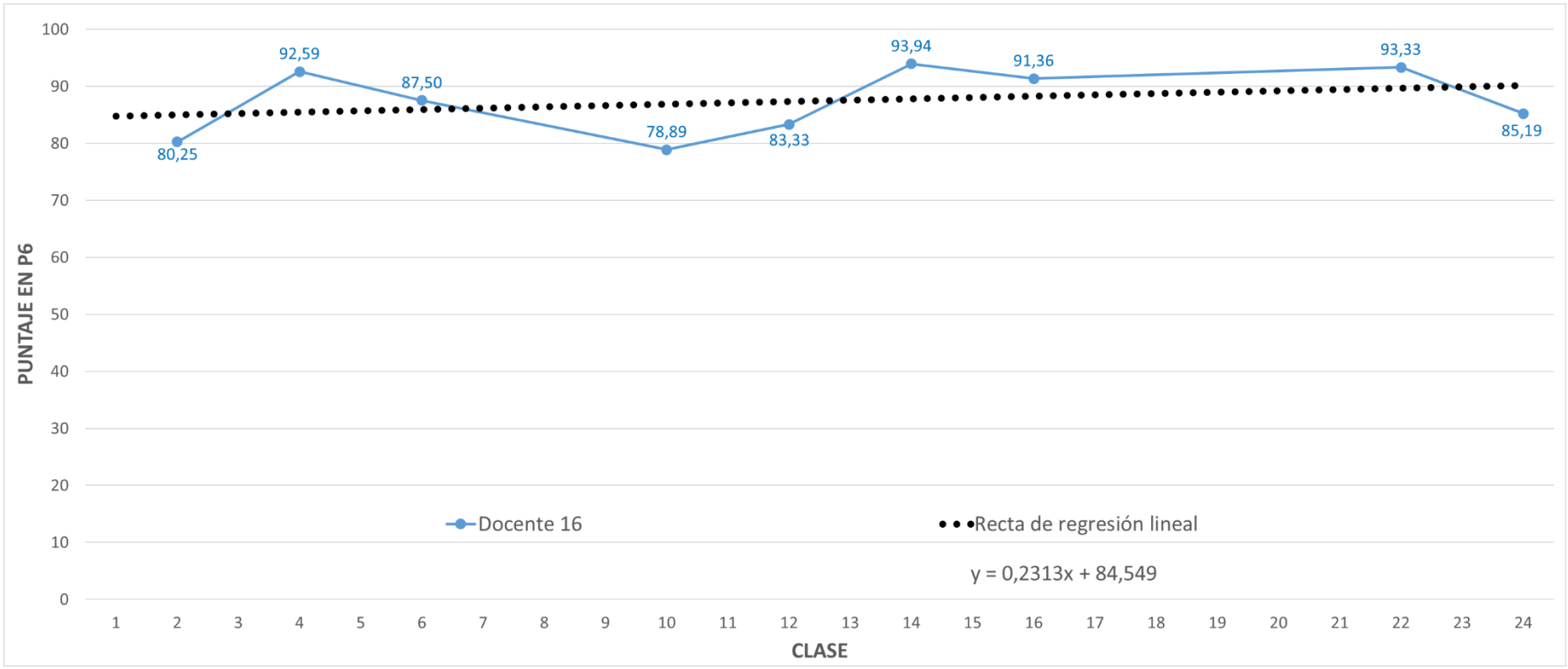
n) P6 - Docente 14 (n=25, $\overline{X}$=81,9)



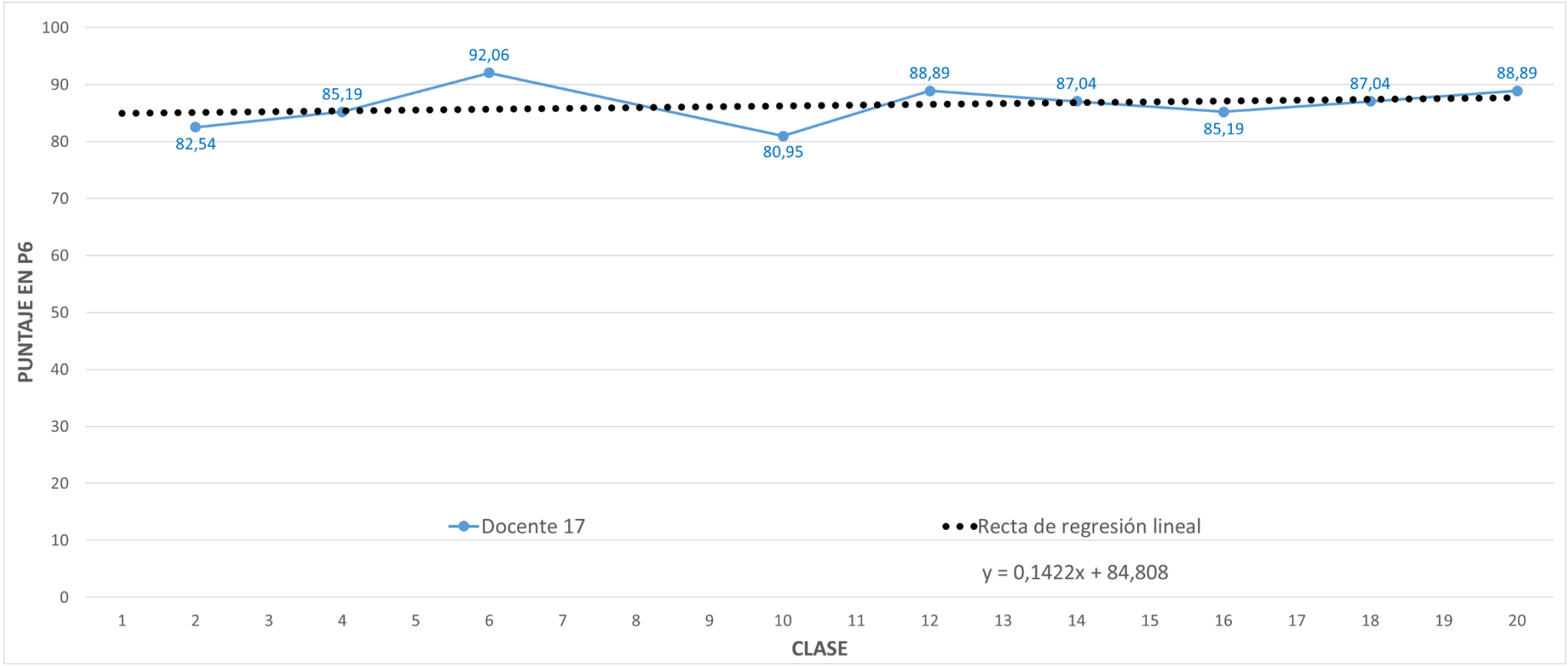
o) P6 - Docente 15 - Sección 1 (n=20, $\overline{X}$=78,5) y Sección 2 (n=15, $\overline{X}$=86)



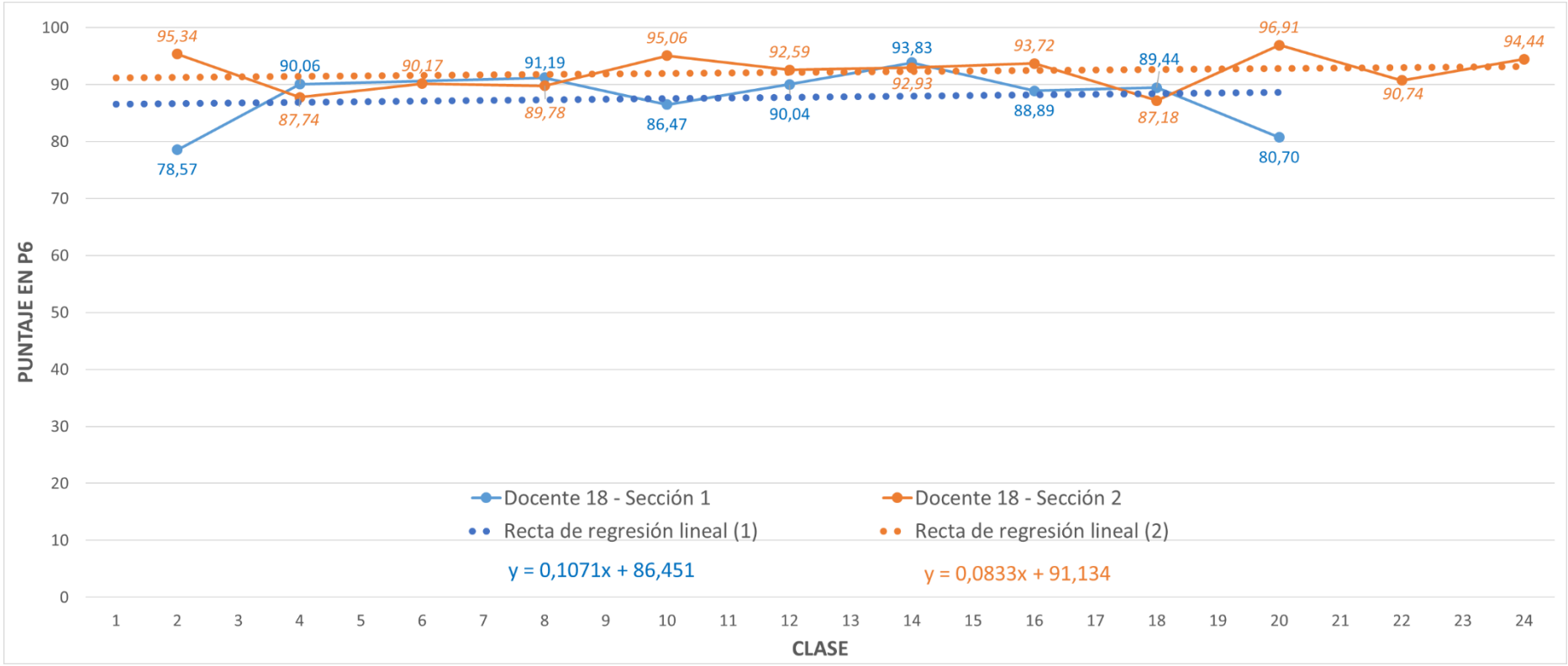
p) P6 - Docente 16 (n=15, $\overline{X}$=87,4)



q) P6 - Docente 17 (n=8, $\overline{X}$=86,4)

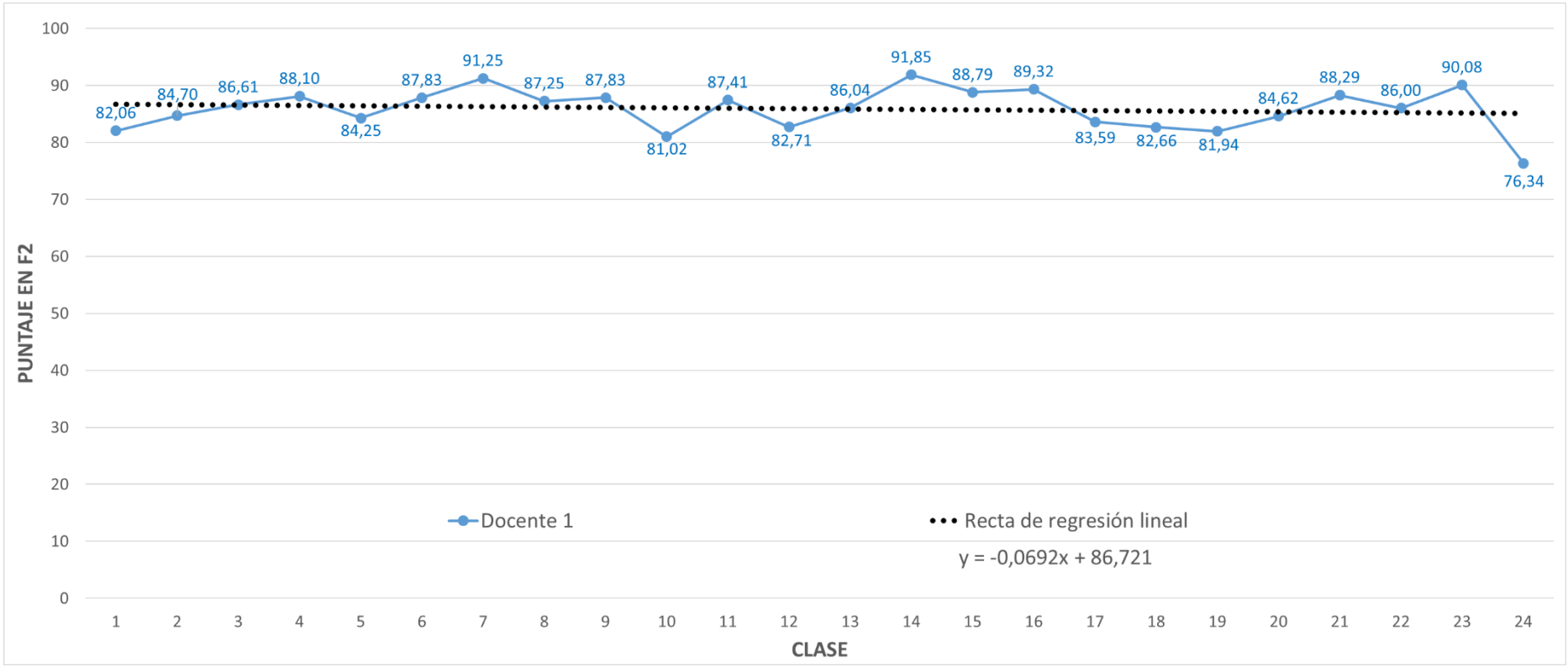


r) P6 - Docente 18 - Sección 1 (n=49, $\overline{X}$=87,7) y Sección 2 (n=36, $\overline{X}$=92,2)

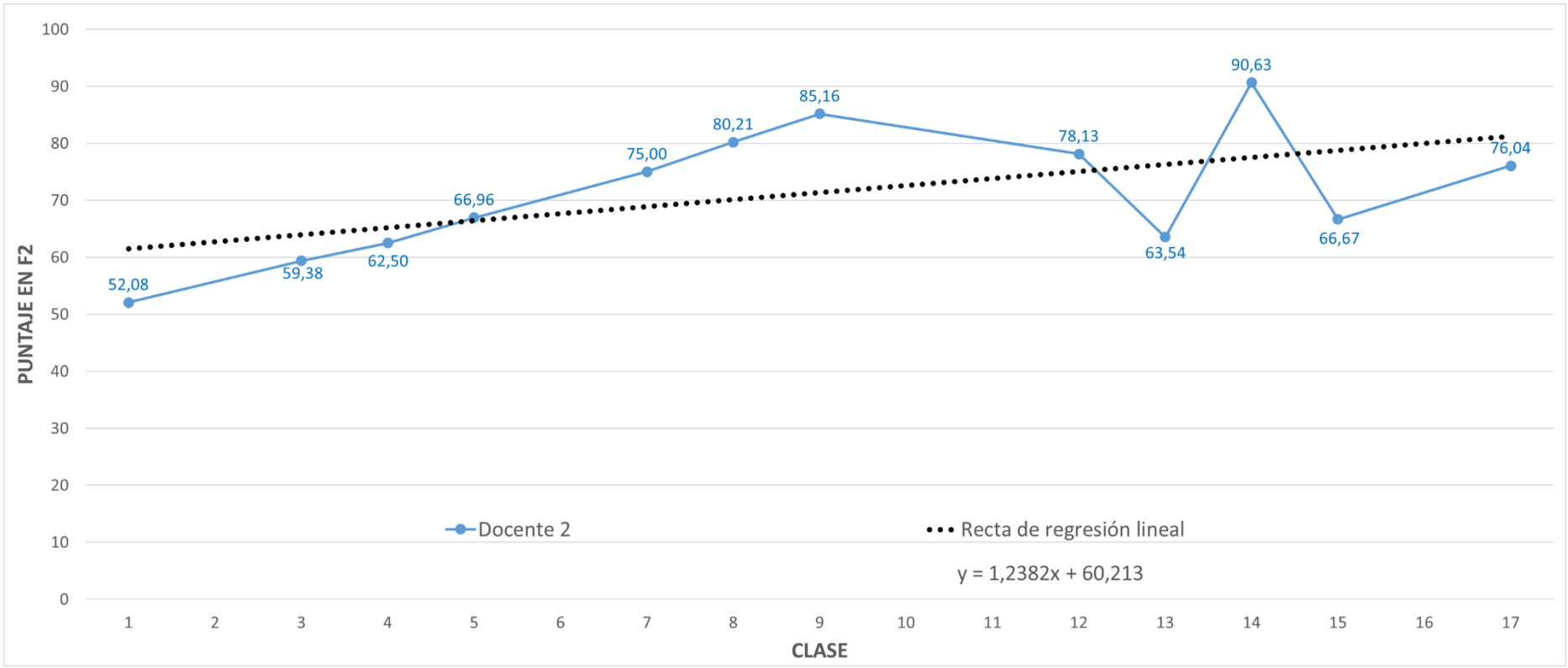


- ***Series de dimensión F2: "Capacidad de la clase para motivar la participación"***

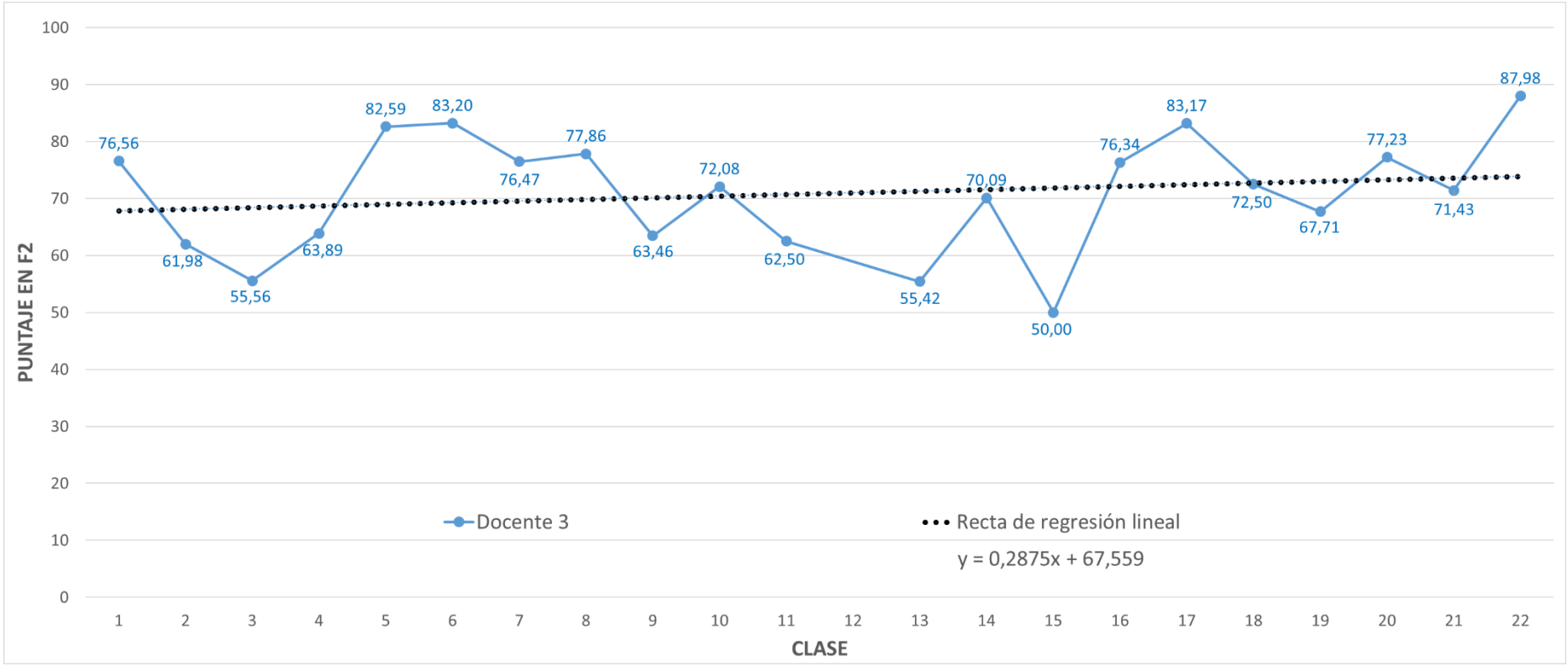
a) F2 - Docente 1 (n=47, $\overline{X}$=85,9)



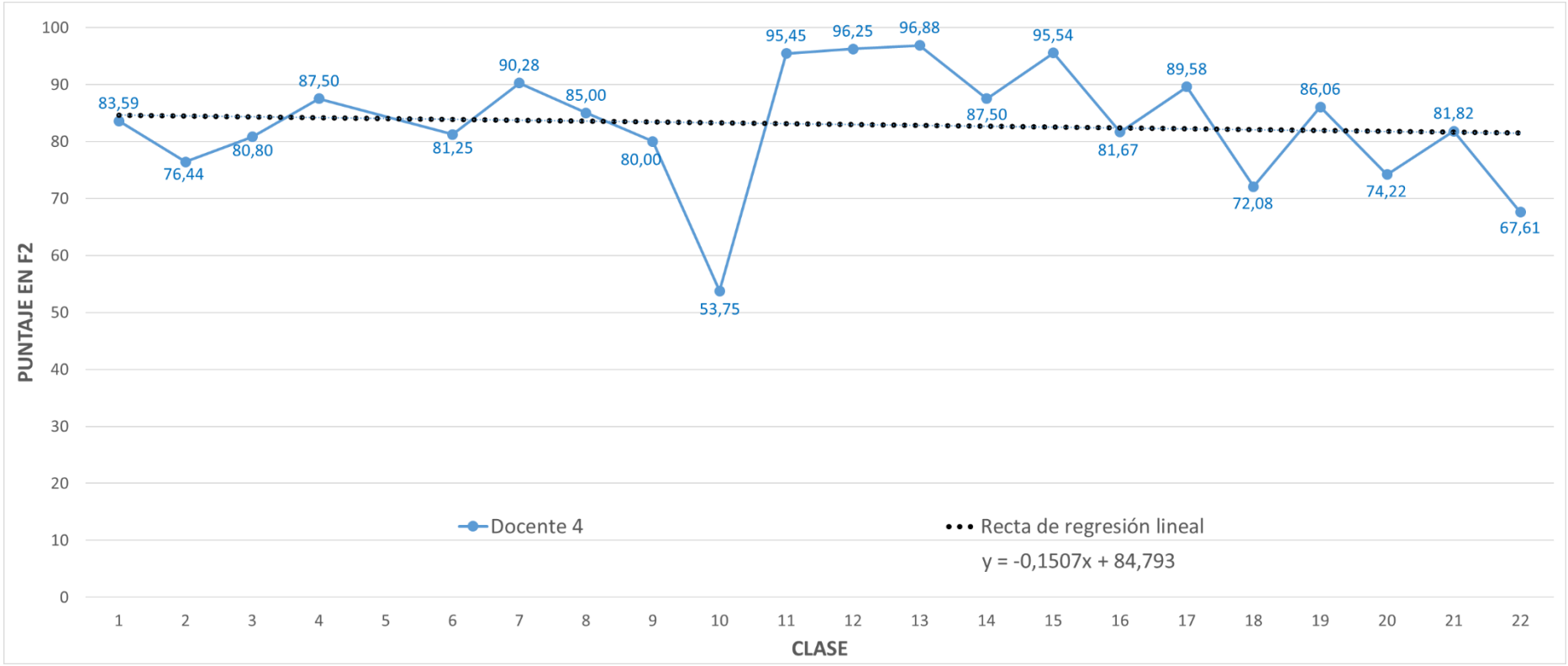
b) F2 - Docente 2 (n=8, $\overline{X}$=71,4)



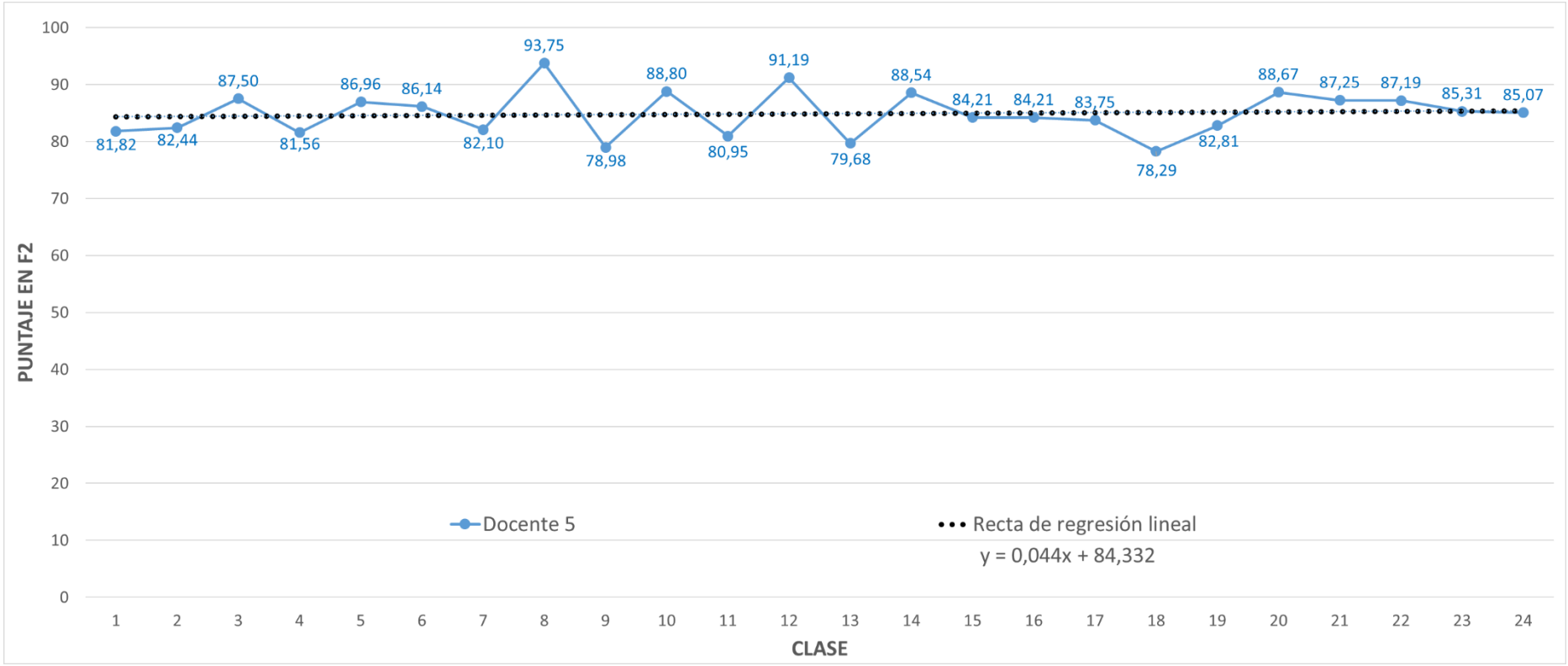
c) F2 - Docente 3 (n=22, $\overline{X}$=70,9)



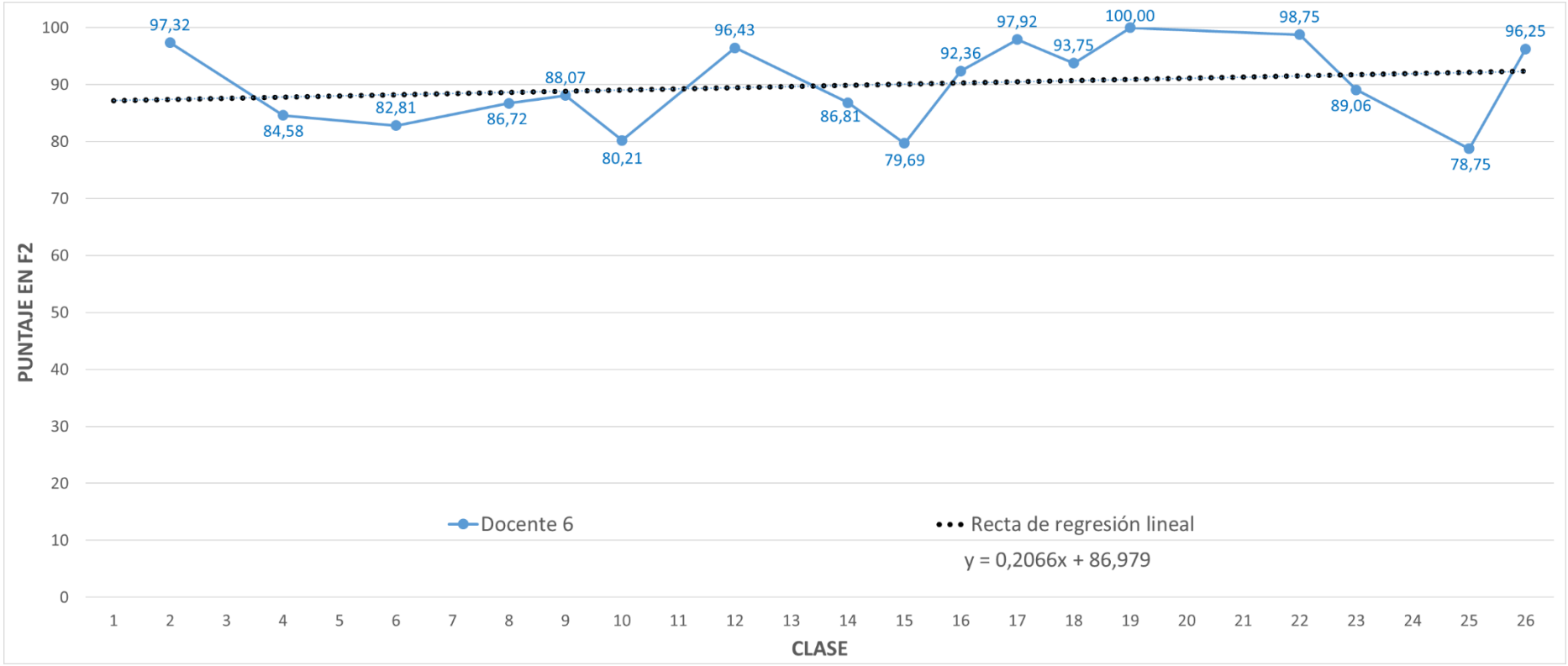
d) F2 - Docente 4 (n=17, $\overline{X}$=83)



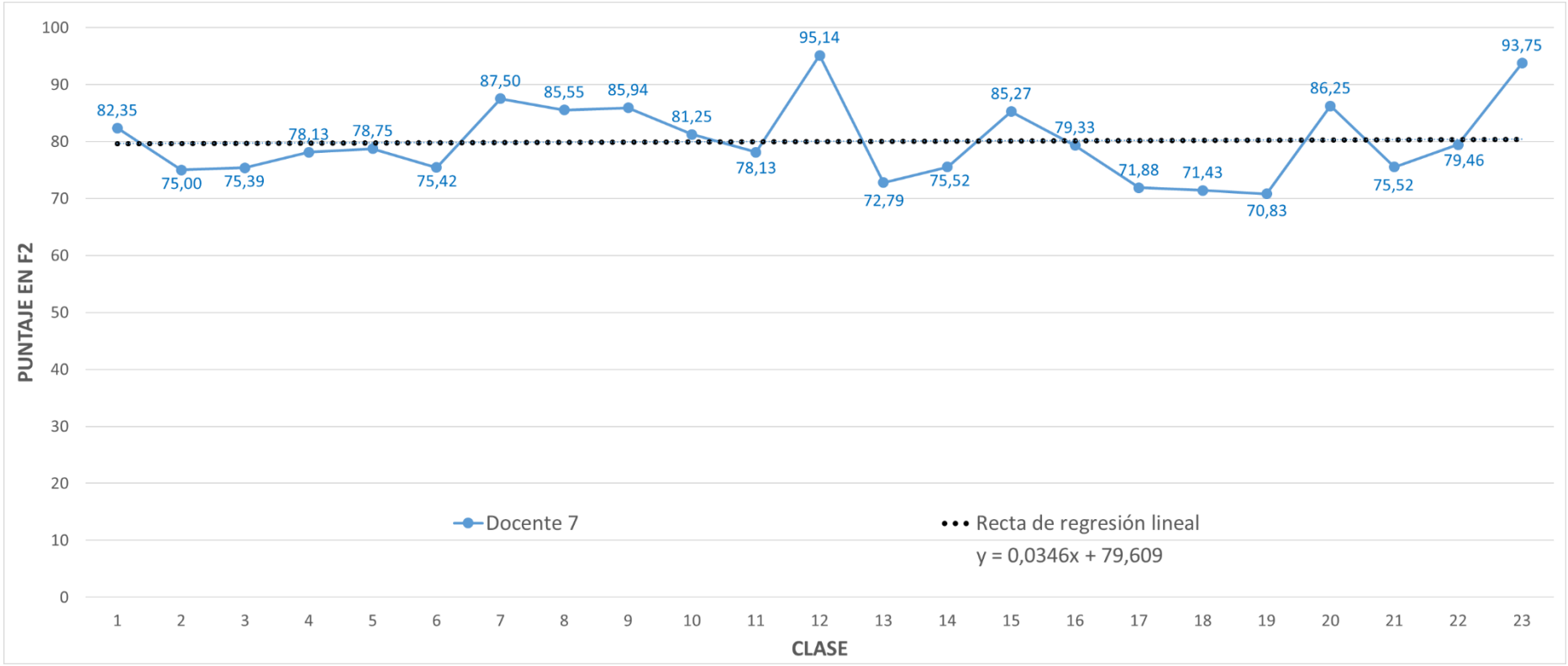
e) F2 - Docente 5 (n=28, $\overline{X}$=84,9)



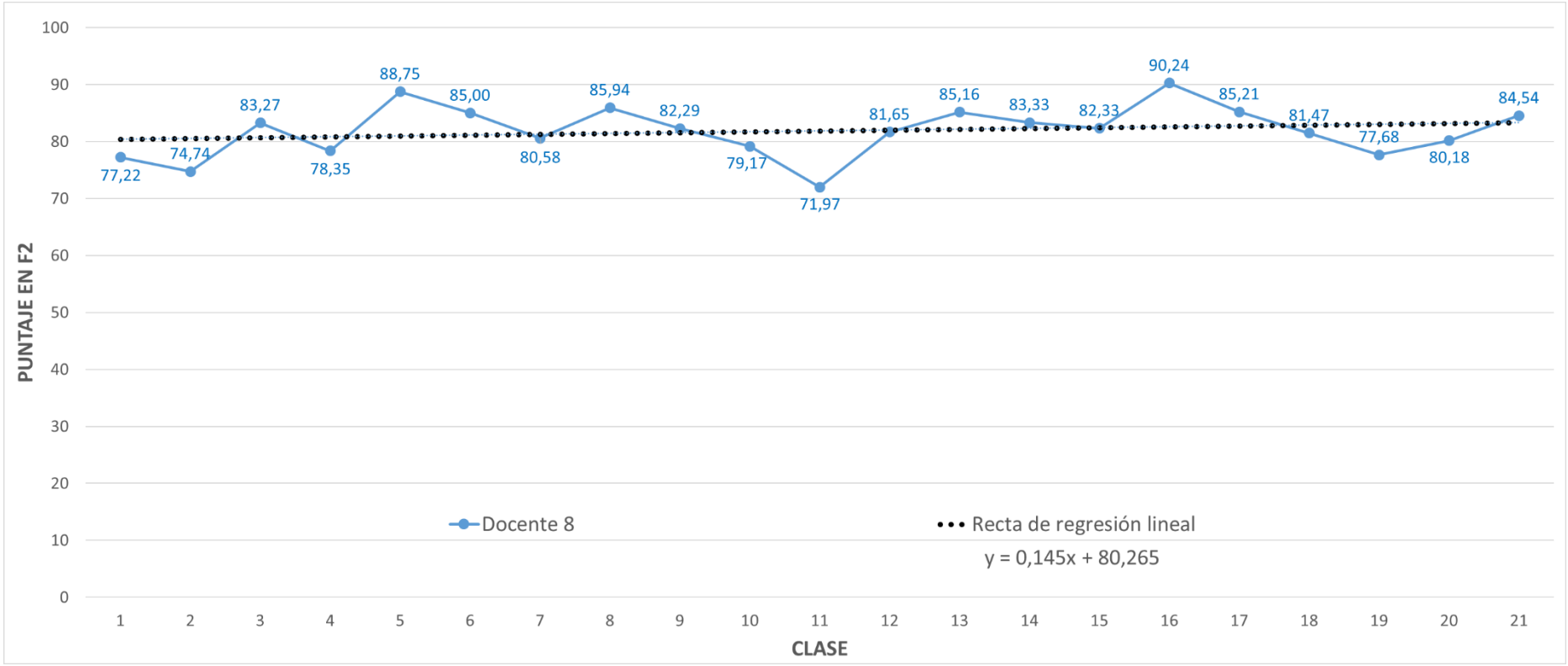
f) F2 - Docente 6 (n=14, $\overline{X}$=90)



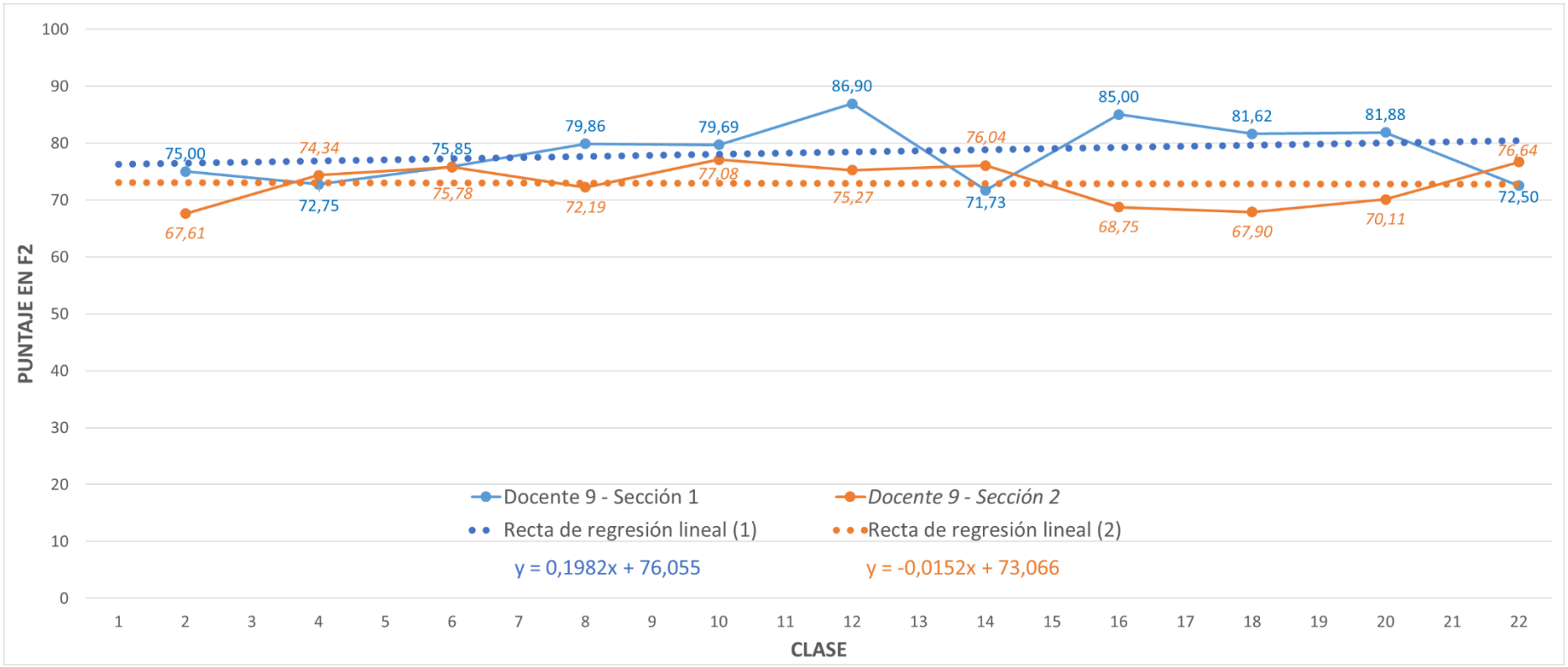
g) F2 - Docente 7 (n=18, $\overline{X}$=80)



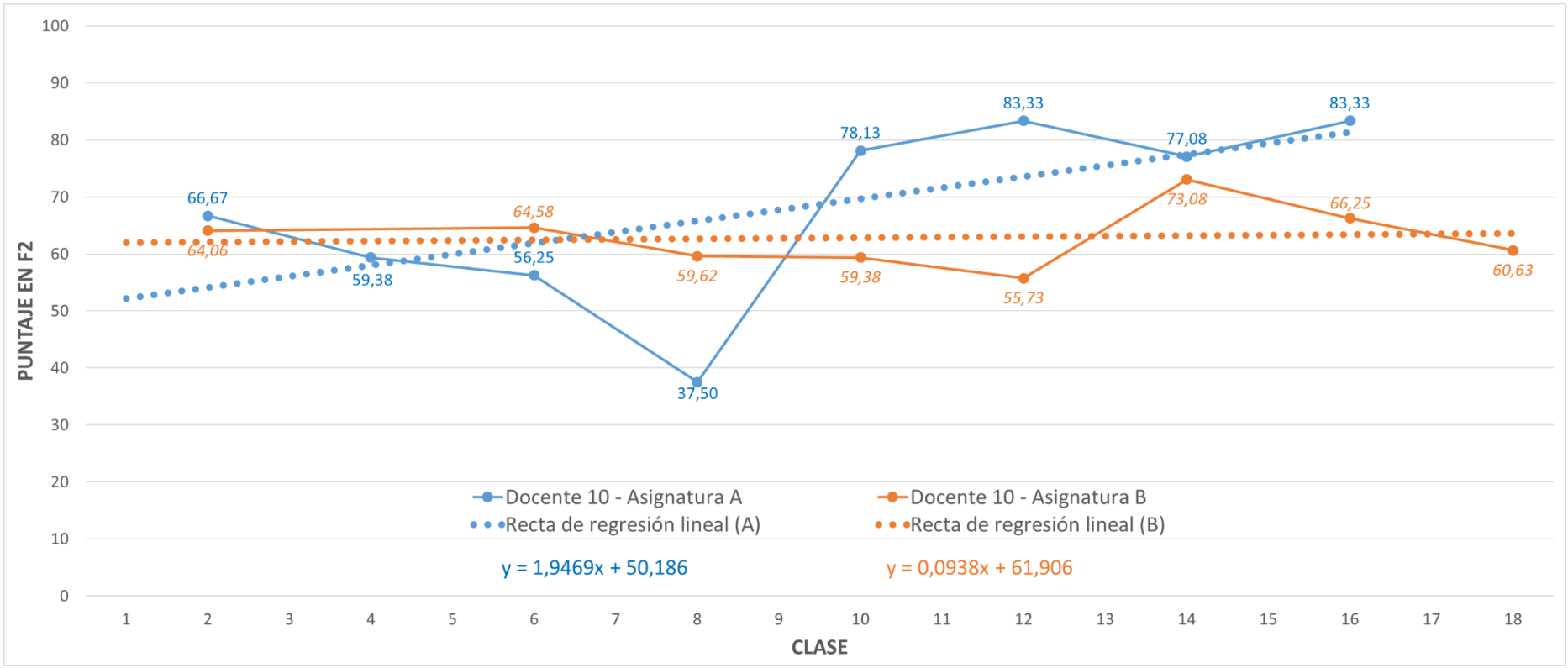
h) F2 - Docente 8 (n=43, $\overline{X}$=81,9)



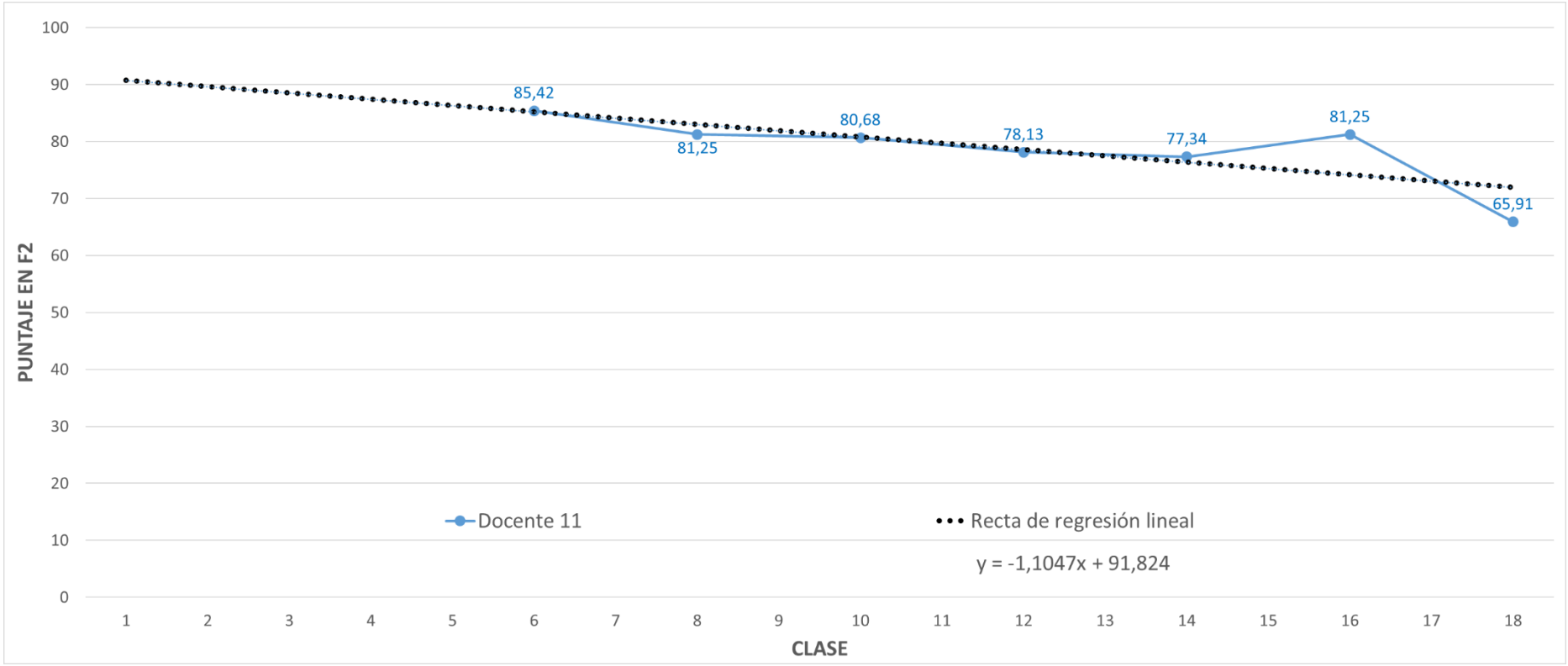
i) F2 - Docente 9 - Sección 1 (n=28, $\overline{X}$=78,4) y Sección 2 (n=33, $\overline{X}$=72,9)



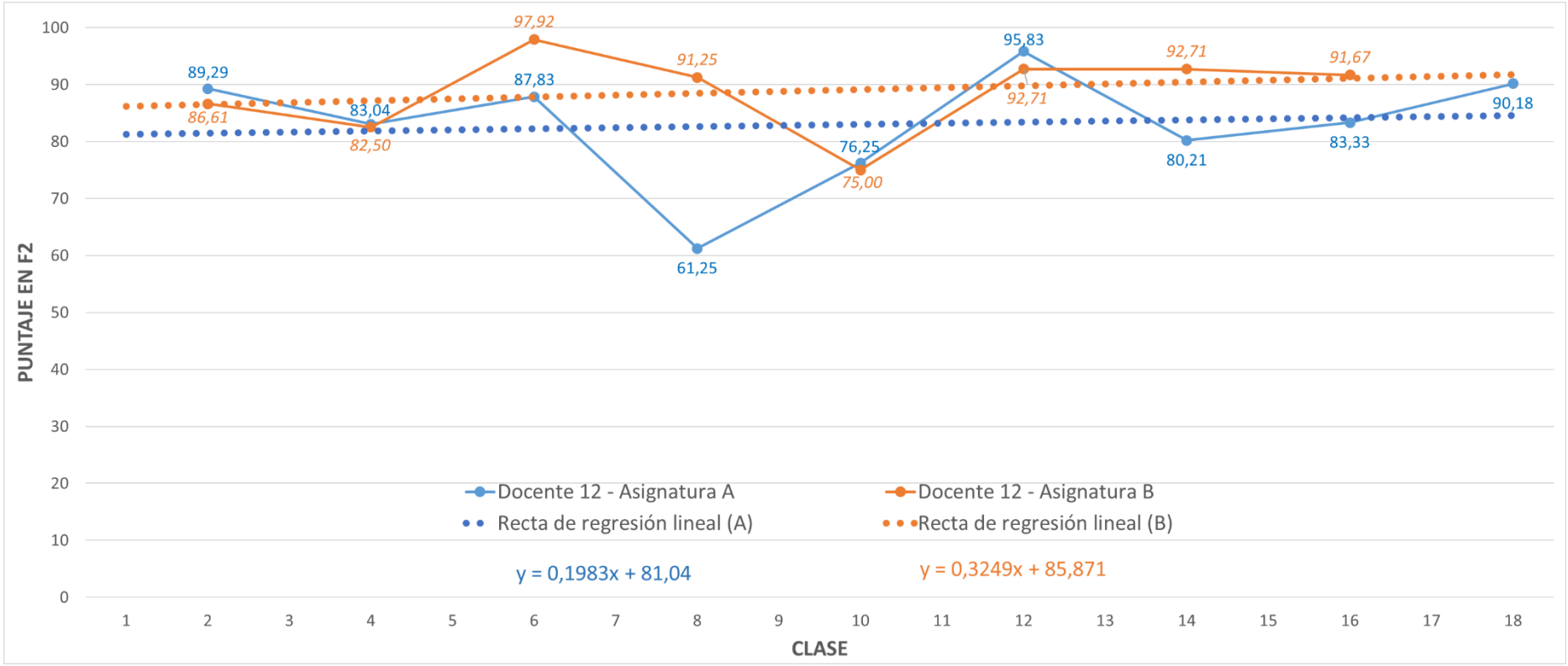
j) F2 - Docente 10 - Asignatura A (n=3, $\overline{X}$=67,7) y Asignatura B (n=16, $\overline{X}$=62,9)



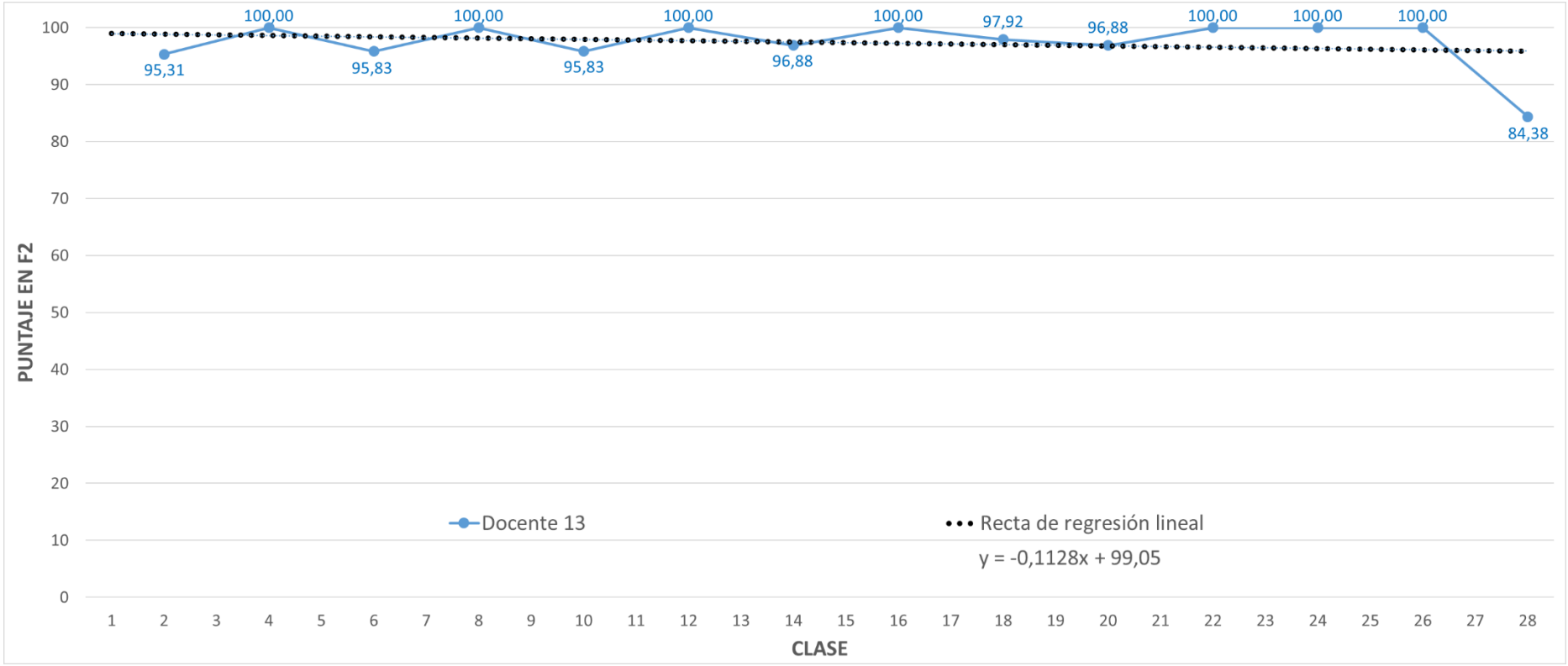
k) F2 - Docente 11 (n=13, $\overline{X}$=78,6)



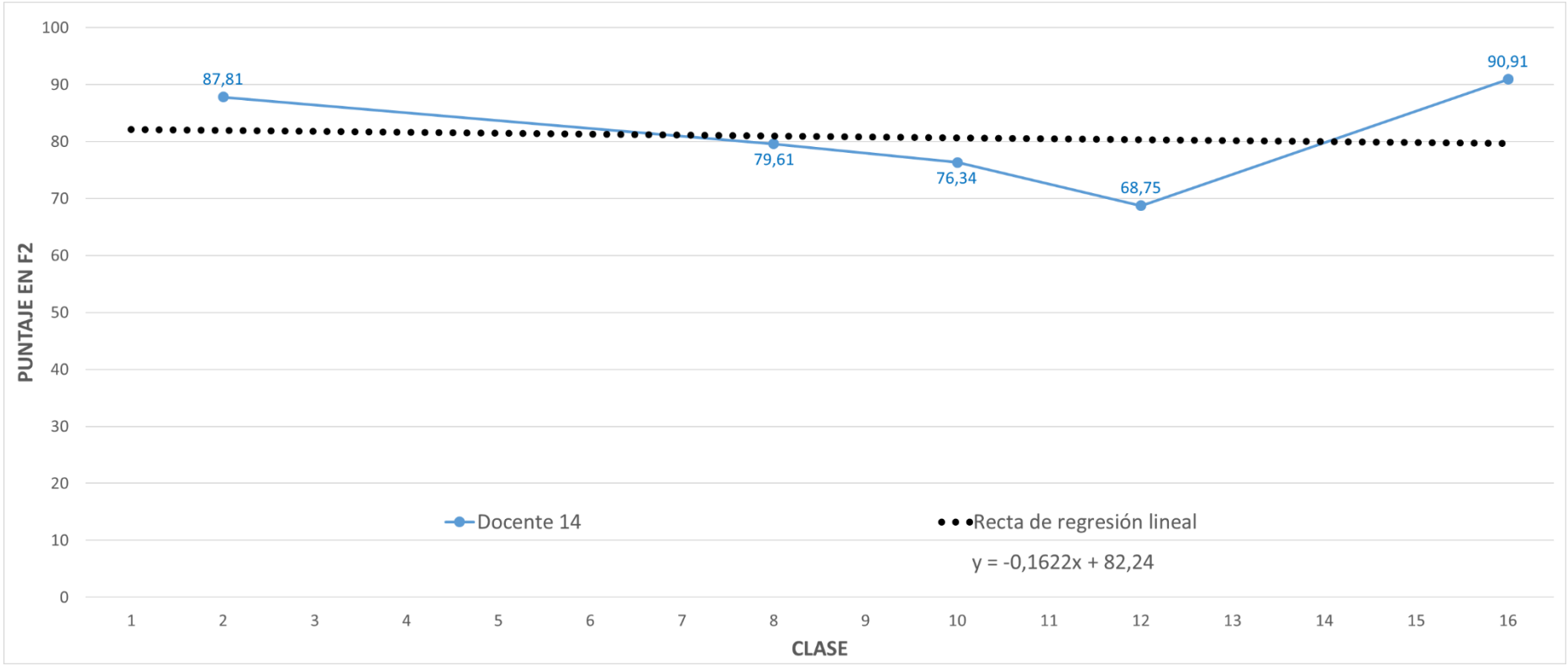
l) F2 - Docente 12 - Asignatura A (n=9, $\overline{X}$=83) y Asignatura B (n=7, $\overline{X}$=88,8)



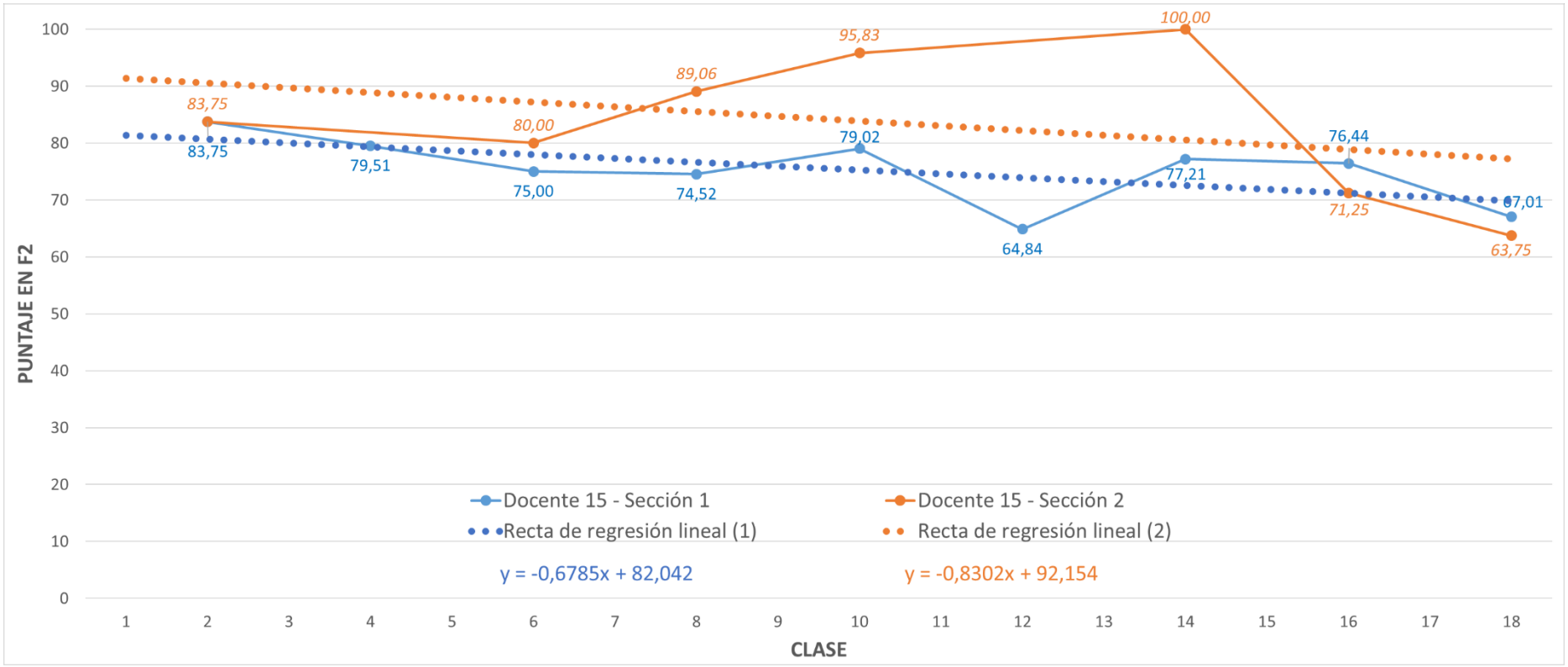
m) F2 - Docente 13 (n=4, $\overline{X}$=97,4)



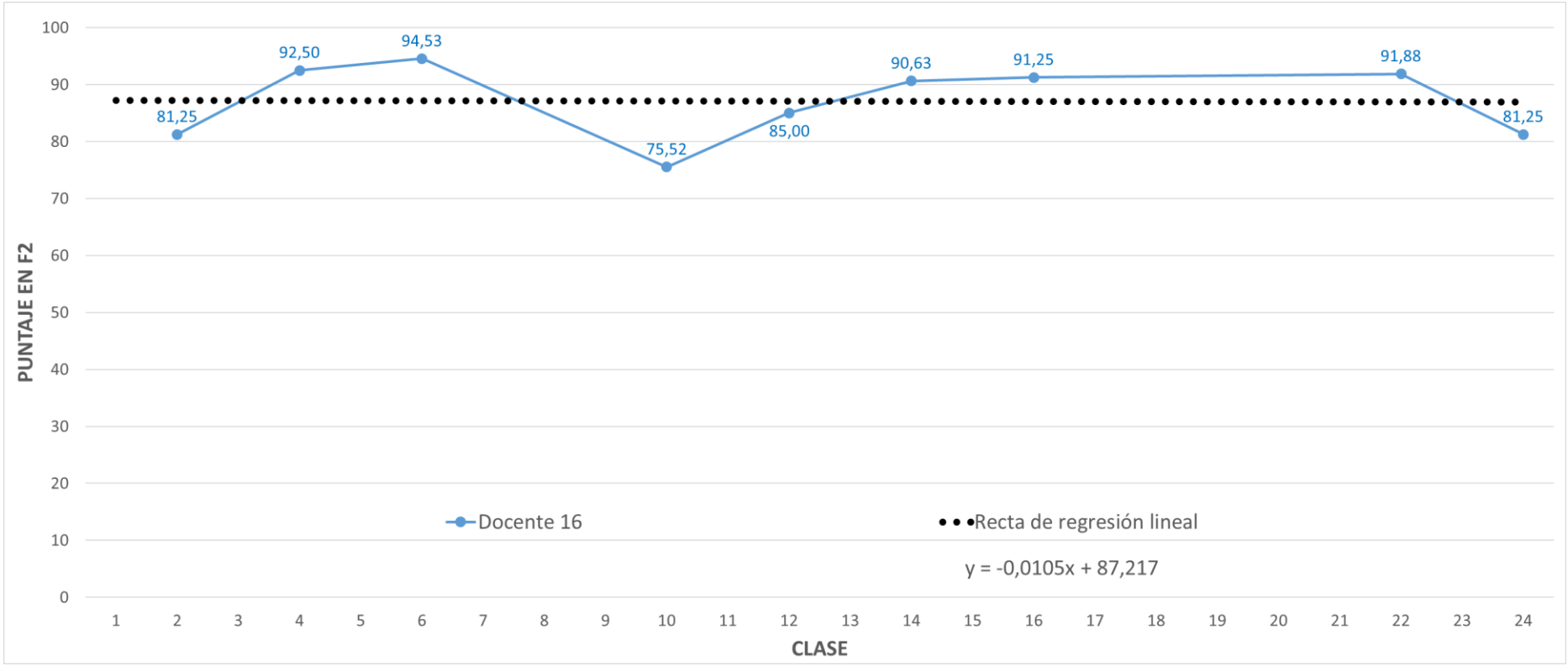
n) F2 - Docente 14 (n=25, $\overline{X}$=80,7)



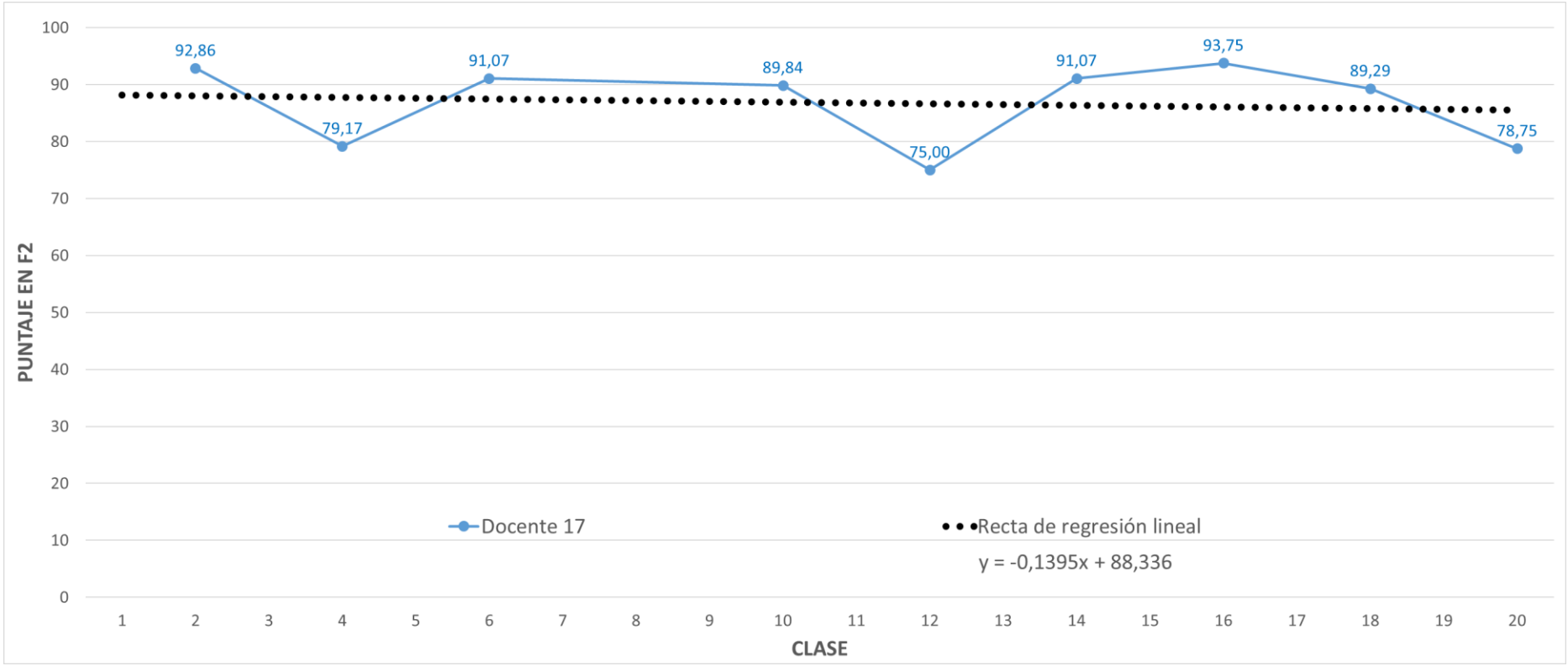
o) F2 - Docente 15 - Sección 1 (n=20, $\overline{X}$=75,3) y Sección 2 (n=15, $\overline{X}$=83,4)



p) F2 - Docente 16 (n=15, $\overline{X}$=87,1)



q) F2 - Docente 17 (n=8, $\overline{X}$=86,8)



r) F2 - Docente 18 - Sección 1 (n=49, $\overline{X}$=81) y Sección 2 (n=36, $\overline{X}$=86,2)