La confiabilidad de las observaciones de aula y las encuestas a estudiantes en contextos no experimentales: Evidencia de Argentina*

Alejandro J. Ganimian†
Harvard University/
Universidad de Nueva York

Andrew D. Ho‡
Harvard University

Alejandra Campos Quintero§
Columbia University

11 de mayo de 2026

Resumen

Existe un consenso creciente sobre la necesidad de medir la eficacia de la enseñanza utilizando múltiples instrumentos. Sin embargo, la orientación sobre cómo lograr calificaciones confiables se deriva en gran medida de investigaciones formales en países de ingresos altos. Estudiamos la confiabilidad de observaciones de aula y encuestas a estudiantes realizadas por profesionales en un país de ingresos medios. Ambos instrumentos pueden lograr una confiabilidad relativamente alta (0,6–0,8 en una escala de 0 a 1) cuando se promedian entre evaluadores y ocasiones, pero la confiabilidad de las observaciones varía ampliamente (de 0,4 a 0,8) según la asignación de evaluadores a docentes. Utilizamos teoría de la generalizabilidad para estimar cómo mejora la confiabilidad al aumentar el número de veces que se observa a cada docente o el número de estudiantes encuestados. Recomendamos que los profesionales diseñen sus sistemas de retroalimentación docente a partir de análisis de sus propios datos, en lugar de asumir que los instrumentos y rúbricas generarán puntuaciones con la misma confiabilidad que en contextos de investigación.

Palabras clave: confiabilidad, generalización, eficacia docente, observaciones de aula, encuestas a estudiantes, Argentina.

*Agradecemos los fondos proporcionados por el Banco Interamericano de Desarrollo para este estudio. Agradecemos a Emiliana Vegas, Mariana Alfonso y al equipo de *Enseñá por Argentina*, especialmente a Oscar Ghillione, Fernando Viola, Mariana Albarracín y Laura de Jorge, por hacer posible este estudio. También damos las gracias a Samuel Hansen Freel por su excelente asistencia de investigación.

†Profesor asociado de Educación, Harvard Graduate School of Education. Profesor asociado de Psicología y Economía Aplicada, Steinhardt School of Culture, Education, and Human Development, Universidad de Nueva York. alejandro_ganimian@gse.harvard.edu.

‡Charles William Eliot Profesor de Educación, Harvard Graduate School of Education. andrew_ho@gse.harvard.edu.

§Estudiante doctoral, Teachers College, Columbia University. aac2271@tc.columbia.edu.

1. Introducción

En los últimos dos decenios, los encargados de formular políticas y los profesionales se interesaron cada vez más en medir la eficacia de la enseñanza. Inicialmente, este interés fue motivado en gran medida por la investigación que sugiere que los docentes varían ampliamente en su capacidad para mejorar el logro de sus estudiantes. Varios estudios han encontrado que los estudiantes de algunos docentes obtienen una puntuación más alta en pruebas estandarizadas que las de otros, incluso cuando ambos grupos tienen una demografía similar y comienzan a niveles comparables de progreso (Nye et al., 2004; Rockoff, 2004; Rivkin et al., 2005; Aaronson et al., 2007; Kane et al., 2008; Chetty et al., 2014; Koedel et al., 2015). Esta evidencia dio lugar a esfuerzos para tratar de identificar docentes eficaces para informar sobre contratación, retención, capacitación y remuneración (por ejemplo, Rockoff et al., 2011; Goldhaber et al., 2017; Dee and Wyckoff, 2015).

Esto ha dado lugar a un consenso creciente sobre la necesidad de utilizar múltiples medidas de enseñanza eficaz. Los métodos estadísticos utilizados para estimar la influencia de docentes en los logros estudiantiles han sido criticados por ocasionar a veces resultados imposibles (por ejemplo, una docente que afecta a sus estudiantes ' *anteriores*- puntuaciones de prueba de año; Rothstein, 2010), rendimiento de resultados conflictivos a través de pruebas (Papay, 2011), ignorando otras formas en las que los docentes contribuyen al bienestar de los estudiantes (Blazar, 2018; Kraft, 2019; Jackson, 2020), y descuidando cómo los factores de nivel escolar (por ejemplo, directores, consejeros y compañeros) median la capacidad de docentes para ayudar a los estudiantes (Jackson and Bruegmann, 2009; Johnson et al., 2012; Jackson, 2013; Papay et al., 2020; Mulhern, 2023).

En los últimos años, varios estudios demostraron que otras medidas de calidad de la enseñanza (por ejemplo, observaciones de aula y encuestas a estudiantes) agregan información valiosa no captada por pruebas (Kane and Staiger, 2011, 2012; Kane et al., 2011, 2013, 2014). Muchos de estos estudios ofrecieron asesoramiento práctico sobre cómo administrar esos instrumentos (por ejemplo, debe observarse el número de veces que un docente obtiene calificaciones consistentes de su desempeño; Ho and Kane, 2013). Esta evidencia ha sido citada en el diseño de sistemas de retroalimentación docente en Estados Unidos y en el extranjero.

No está claro, sin embargo, por qué se debe esperar que los datos recopilados con fines de investigación produzcan resultados que sean indicativos de aquellos que se obtengan en entornos de no investigación. En la mayoría de los estudios, docentes voluntarios para participar, las personas que las califican están capacitadas, y existen muchos mecanismos para garantizar la integridad de la información reunida. Por el contrario, los gobiernos y las organizaciones sin fines de lucro tienen que diseñar soluciones que funcionen para todos los docentes y pueden enfrentar limitaciones en la formación o su capacidad para adoptar controles de calidad. Estas diferencias podrían hacer que las métricas no de prueba administradas en

investigación sean mucho más fiables que las que se realizan en la práctica. Si este fuera el caso, los practicantes podrían tomar decisiones sobre cómo diseñar sus sistemas de retroalimentación docente basados en medidas no-test que son más fiables que las que recogen, lo que da lugar a métricas de eficacia docente que son menos fiables de lo que pretendían y realizaban.

Esta pregunta es particularmente apremiante para los profesionales de los países de bajos y medianos ingresos, dado que la gran mayoría de los estudios anteriores en esta literatura se han realizado en los Estados Unidos Tal vez, en los Estados Unidos, las docentes se utilizan más para recibir comentarios y que sus colegas y estudiantes están más acostumbrados a actuar como evaluadores que los del resto del mundo. Estados Unidos también ha invertido más fondos para desarrollar e investigar sistemas de retroalimentación docente. Estas diferencias podrían hacer que las métricas en Estados Unidos sean más fiables que las de otros países.

En el presente estudio, pretendemos aclarar ambas cuestiones evaluando la fiabilidad de las observaciones de aula y encuestas a los estudiantes administrados como parte de un programa de educación (es decir, no con fines de investigación) y comparando nuestros resultados con los de la literatura pertinente. Examinamos la confiabilidad de las observaciones de aula y encuesta a un estudiante de 100 participantes en un camino alternativo hacia la enseñanza en Argentina. Estos docentes fueron marcados en dos puntos de tiempo. La primera fue durante dos semanas de enseñanza práctica, poco después de haber completado un breve curso de formación previa al servicio. El segundo fue durante el año escolar, una vez que comenzaron a enseñar en escuelas de difícil acceso al personal durante dos años. El programa (*Enseñá por Argentina* o ExA) desarrollaron estas medidas aprovechando los instrumentos existentes y administrando las calificaciones exclusivamente con fines de retroalimentación (es decir, no se les asignó ningún interés). Esta configuración nos permite comprender la fiabilidad de las medidas no comprobadas de eficacia docente, ya que son utilizadas frecuentemente por las organizaciones del sector educativo en un entorno poco estudiado. Además, dado que ExA es parte de una red mundial de 60 organizaciones que utilizan medidas y procedimientos similares (enseñanza para todos), vemos nuestro estudio como potencialmente relevante para este grupo más amplio.

Nuestro estudio va más allá de las métricas tradicionales de fiabilidad que cuantifican la consistencia en partituras a través de una fuente de error en un momento (por ejemplo, elementos o evaluadores). Simultáneamente estimamos la contribución de diferentes facetas del error de medición (por ejemplo, dificultad del elemento y rigor del evaluador) y de las interacciones entre estas facetas (por ejemplo, algunos evaluadores son más estrictos en algunos elementos). La principal ventaja de este enfoque es que, al ser más precisos sobre las fuentes de error, también podemos ser más estratégicos acerca de la reducción (por ejemplo, si evaluarr stringency está contribuyendo más al error de medición que la dificultad del elemento, podemos reducir el error más eficientemente aumentando el número de evaluadores en lugar

de elementos). This approach, "teoría de la generalizabilidad ' ' (Lord and Novick, 1968; Nunnally and Bernstein, 1978; Allen and Yen, 1979), se utiliza cada vez más en sistemas de retroalimentación docente en los Estados Unidos. (Hill et al., 2012; Kane and Staiger, 2012; Ho and Kane, 2013). A nuestros conocimientos, no se ha aplicado ampliamente en los países de ingresos bajos o medianos.

Informamos cinco hallazgos principales. En primer lugar, las observaciones de aula realizadas por profesionales pueden alcanzar altos niveles de fiabilidad para hacer ambas *relativo* distinciones (decir qué docentes son más eficaces) y *absoluto* juicios sobre docentes (que dan resultados consistentes para niveles de rendimiento similares) con números actuales de elementos, evaluadores y ocasiones. Durante la práctica clínica, el coeficiente de generalización para un error relativo—una medida de fiabilidad para distinciones relativas que va de 0 (perfectamente poco confiable) a 1 (perfectamente confiable)— fue tan alto como 0,79 en algunos años, y el coeficiente para el error absoluto—una métrica para los juicios absolutos que también va de 0 a 1-reached 0,76. Estas cifras indican que casi 80 % de la variación de las puntuaciones de observación refleja diferencias reales en la eficacia de la enseñanza medida, en comparación con el error de medición, que es alentador. Sobre la base de nuestro examen de estudios anteriores G, las observaciones realizadas para entornos de investigación tienen una media de 0,65 en educación preescolar, 0,33 en educación primaria y 0,51 en educación secundaria, con los valores más altos alcanzando los 0,94 y 0,94 respectivamente (véase la sección siguiente y el Apéndice A).

En segundo lugar, la fiabilidad de estas observaciones varía ampliamente dependiendo de su contexto, de la forma en que los evaluadores se asignan a las lecciones, y del año en que se realizan. Durante el año escolar, los coeficientes de generalización de errores relativos y absolutos alcanzaron niveles inferiores (0,66 y 0,57 respectivamente) que en la práctica clínica (reportada anteriormente). Incluso dentro de la práctica clínica, la confiabilidad variaba según quién actuaba como evaluadores (coaches o pares) y cómo se les asignaba (ya sea que una docente fuera observada por la misma persona o personas diferentes). Las observaciones anotadas por los entrenadores no siempre eran más fiables que las clasificadas por los pares. Más bien, cuando un docente fue observado múltiples veces por la misma persona, los pares fueron más fiables (con coeficientes de error relativo y absoluto de 0.79 y alrededor de 0.75, respectivamente) que los entrenadores (con coeficientes de 0.53-0.55 y 0.38-0.47). Sin embargo, cuando cada lección entregada por un docente fue observada por un par diferente, la fiabilidad fue tanto más baja como más variable de un año a otro (con coeficientes de 0,44 y 0,41 en 2014 y de 0,64 y 0,61 en 2015). Estas cifras indican que, en función de cómo se realizan las observaciones, la variación de las puntuaciones que reflejan las diferencias de eficacia puede ser tan baja como 41 % (casi la mitad de la cifra anterior).

En tercer lugar, es posible mejorar la fiabilidad de las observaciones aumentando el número de veces que las docentes son anotadas. Durante la práctica clínica, observar cada docente tres veces en lugar de dos mejoraría el coeficiente de generalización del error relativo en 5-10 puntos porcentuales (pp.) y el error absoluto en 4-9 pp., dependiendo del tipo evaluador (coaches o pares) y si todas las lecciones entregadas por un docente son observadas por el mismo o un evaluador diferente. Agregar una observación durante el año escolar mejoraría el coeficiente de error relativo en 8 pp. y el de error absoluto en 7-8 pp., dependiendo del año utilizado como referencia. Los aumentos adicionales en el número de observaciones mejorarían la fiabilidad por un pequeño margen.

En cuarto lugar, encuestas a estudiantes administrados por profesionales también pueden alcanzar altos niveles de confiabilidad. Durante la práctica clínica, el coeficiente de generalización del error relativo fue tan alto como 0,76 en algunos años, y el de error absoluto llegó a 0,65. En el año escolar, las cifras correspondientes fueron de 0,62 y 0,63 respectivamente. A pesar de las preocupaciones sobre la fiabilidad de las encuestas a los estudiantes (ver English et al., 2015, para un examen), estos resultados indican que en algún lugar entre 60 y 70 % de variación en las puntuaciones observadas refleja diferencias reales en la eficacia de la enseñanza medida. Notablemente, esta confiabilidad se logró encuestando sólo una muestra de 10 estudiantes cada vez, y varió relativamente poco entre práctica clínica y el año escolar.

Quinto, mejorar la fiabilidad de las encuestas es posible aumentar el número de encuestados con extensiones razonables a las condiciones de administración existentes. Durante la práctica clínica, la encuesta de cinco estudiantes más mejoraría el coeficiente de generalización del error relativo en 9 pp. y el de error absoluto en 8 pp. La adición de cinco estudiantes durante el año escolar mejoraría los coeficientes de error relativo y absoluto en 5-9 pp. y 4-8 pp.

El resto del documento está estructurado como sigue. Sección 2 revisa investigación previa sobre la fiabilidad de las observaciones de aula y encuesta a los estudiantes, mostrando que las medidas recolectadas para fines de investigación muestran niveles de fiabilidad relativamente altos. Sección 3 describe los datos utilizados para este estudio, que se basa en las observaciones de aula y encuesta a un estudiante administrado en dos entornos diferentes a través de dos años de una vía alternativa a la enseñanza en Argentina. Sección 4 explica cómo utilizamos la teoría de la generalización para ser más precisos sobre las fuentes de error de medición en las observaciones y encuestas y más estratégicos sobre cómo reducirlas. Sección 5 presenta nuestras estimaciones de fiabilidad tanto para métricas como para mejorarlas aumentando el número de facetas relevantes de error (por ejemplo, aumentando los evaluadores y/o lecciones).

2. Investigación previa

El estudio de la fiabilidad de las medidas de eficacia docente en general, y de las observaciones de aula y encuestas a estudiantes en particular, ha evolucionado considerablemente en los últimos decenios. Convencionalmente, académicos de medición educativa conciben de la partitura un docente recibe en un procedimiento debido en parte a la eficacia de ese docente y en parte a errores en la medición. Se distingue entre estas partes adoptando múltiples medidas e interpretando similitudes entre las mediciones como indicativas de las primeras y diferencias como indicativas de las últimas. Esta idea está cristalizada en "teoría clásica de prueba"(CTT) y su ecuación $X_i = \tau + \varepsilon_i$, lo que indica que cualquier puntuación observada X_i es igual a una puntuación verdadera τ más el error de ese procedimiento ε_i (Lord and Novick, 1968; Nunnally and Bernstein, 1978; Allen and Yen, 1979). La verdadera puntuación es el promedio de puntajes de larga duración sobre las replicaciones, los errores de medición son desviaciones específicas de replicación de ese promedio, y la fiabilidad es la correlación entre las puntuaciones a través de las replicaciones (la proporción de verdadera a la varianza total de puntuación).

Este marco se utiliza a menudo para cuantificar el error de medición de las preguntas (temas) en una prueba. Si todos los elementos están midiendo el mismo constructo, podemos interpretar la expectativa a través de las puntuaciones del artículo como la verdadera puntuación y cualquier desviación de él como error. Por ejemplo, la alfa de Cronbach mide la fiabilidad de la consistencia interna como la proporción de la varianza total de puntuación debido a la variación compartida entre los elementos (Cronbach, 1951). Esta idea también se aplica al error de evaluadores o ocasiones. Si vemos la partitura de cada evaluador (o ocasión) como una replicación, podemos interpretar la correlación de puntuaciones entre evaluadores (o ocasiones) como confiabilidad inter-evaluador (o test-retest).

Una limitación clave de los análisis clásicos de fiabilidad es que no distinguen entre diferentes fuentes de error de medición. Descomponen la varianza observada en la verdadera y no diferenciada varianza de error. Una alternativa es utilizar modelos de efectos aleatorios para analizar la contribución de cada faceta (por ejemplo, elementos o evaluadores) y las interacciones entre ellos (por ejemplo, los evaluadores son más estrictos en algunos elementos). This approach, “teoría de la generalizabilidad ’ ’ (Cronbach et al., 1972; Brennan, 2001), nos permite describir la varianza de error con más precisión y ser más estratégico sobre la reducción al aumentar las replicaciones sobre las facetas que añaden el más ruido. Por ejemplo, si el evaluador stringency contribuye más al error que la dificultad del elemento, aumentar el número de evaluadores reducirá el error por un margen mayor que aumentar el número de elementos.

En los últimos decenios, los estudios de G(eneralizabilidad) se convirtieron en un método cada vez más prominente para examinar la fiabilidad de las medidas de no prueba de la

eficacia docente en la educación K-12. Estos estudios ofrecieron orientación práctica sobre cómo diseñar sistemas docente-feedback para producir resultados fiables. Tal vez más famoso, el estudio de Medidas de Enseñanza Eficaz (MET), que comparó la fiabilidad de cuatro protocolos de observación de aulas ampliamente utilizados en cinco distritos escolares en los EE.UU., concluyó que "para lograr confiabilidad en el barrio de 0,65... tuvimos que marcar cuatro lecciones diferentes, cada una con un evaluador diferente ' ' (Kane and Staiger, 2012) y posteriormente se determinaron múltiples enfoques de observaciones fiables (MET Project, 2013). Muchos encargados de formular políticas y profesionales de todo el mundo se han basado en estas directrices al diseñar sus propios sistemas (por ejemplo, Pouezevara et al., 2016; Cruz-Aguayo et al., 2020).

Gran parte de lo que sabemos sobre la fiabilidad de las métricas alternativas de eficacia docente, sin embargo, se deriva de un conjunto relativamente pequeño de medidas y contextos. Buscamos estudios G de observaciones de aula y encuestamos a estudiantes de educación preescolar a secundaria en países de ingresos bajos/medios y de altos ingresos. No encontramos ningún estudio de encuestas a estudiantes, pero encontramos 12 estudios de observaciones de aula (ver Tabla A.1 en el ApéndiceA). La mayoría se centra en tres instrumentos: el sistema de evaluación de aulas (CLASE, Mashburn et al., 2008; Pianta et al., 2008; Hamre et al., 2013); Marco para la Enseñanza (FfT, Danielson, 2011); y la calidad matemática de la instrucción (MQI, Hill et al., 2011, 2012). Se establecieron tres cuartos en los Estados Unidos y todos ellos estaban en países de ingresos altos. Estos patrones plantean dudas sobre la validez externa de la orientación de estos estudios G.

Las observaciones de aula en estos estudios se llevaron a cabo con fines de investigación y incorporan varios mecanismos de garantía de calidad que probablemente mejoren su fiabilidad, tales como: formación evaluador, evaluación, certificación y práctica adicional (por ejemplo, entrenamiento profundo, entrenamiento de uno a uno, observaciones emparejadas y calibración de grupo; Jerald, 2012); codificación maestro (en los que los expertos discutan y concuerdan en las puntuaciones correctas y puntuan racionales; McClellan, 2013), y un motor de validación (incluyendo una biblioteca de vídeo en línea, una rúbrica de puntuación, comparaciones con otras métricas e informes automatizados; MET Project, 2010), entre otros (por ejemplo, pilotando el protocolo de observación; Joe et al., 2013). Si las observaciones realizadas por los profesionales, con menos de estos mecanismos, pueden alcanzar niveles similares de fiabilidad, sigue siendo una cuestión abierta.

3. Datos

3.1. Contexto

Realizamos nuestro estudio en Argentina, un país de ingresos medio superior con altos niveles de matriculación en la escuela primaria y secundaria, pero resultados de aprendizaje más bajos que sus vecinos. El ingreso per cápita de Argentina (USD 13,730) es comparable al de China, México, Rusia y Turquía (World Bank, 2024a), pero ha sufrido recientemente varias crisis económicas y políticas que la distinguen tanto de estos países como de la mayoría de sus vecinos sudamericanos. Según los últimos datos, 4 de cada 10 personas viven por debajo del umbral de pobreza (World Bank, 2024c). Más de 99 % niños y jóvenes matriculados en la escuela primaria y secundaria, pero sólo 90 % lo hacen en la escuela secundaria superior y sólo 70 % graduado de la secundaria (World Bank, 2024b). Incluso entre los que llegan al último año de secundaria, 43 % puntuación en los niveles más bajos de la evaluación nacional en idioma y 82 % lo hacen en matemáticas (Ganimian and Mesalles, 2025). La proporción de niños de 15 años en los niveles más bajos de pruebas globales es mayor: 55 % en idioma y 73 % en matemáticas (OECD, 2023). Además, los estudiantes más pobres son 21 y 42 puntos porcentuales más propensos a anotar en estos niveles en lectura y matemáticas que sus compañeros más ricos, respectivamente.

Nos centramos en la Provincia de Buenos Aires (PBA), el sistema escolar subnacional más grande del país. En Argentina, las provincias (como Estados Unidos) se encargan de proporcionar enseñanza preescolar a la enseñanza terciaria y al gobierno federal para proporcionar educación superior y asistencia técnica y financiera a las provincias (*Ley de Educación Nacional*, 2006). PBA atiende a 4,3 millones de estudiantes: 654,958 en educación preescolar, 1,7 millones en educación primaria, 1,7 millones en educación secundaria y 260.082 en nivel terciario (MdCH, 2024). El PBA es representativo del país en su conjunto, con un ingreso familiar medio de 117.278 ARS por mes, que es casi idéntico al promedio nacional. También es comparable en la desigualdad de ingresos, con un coeficiente Gini ligeramente inferior al promedio nacional (INDEC, 2024). Sus resultados de aprendizaje reflejan esta realidad económica: sus calificaciones en la evaluación nacional se asemejan estrechamente a las de la provincia promedio en el país (Ganimian and Mesalles, 2025).

Obtuvimos los datos de nuestro estudio a partir de *Enseñá por Argentina* (ExA), una organización sin fines de lucro que recluta graduados universitarios para enseñar en escuelas de difícil acceso al personal durante dos años. Para 2024, 15 años después de su fundación, ExA había colocado 400 docentes sirviendo a 130.000 estudiantes en siete provincias (la Provincia y Ciudad de Buenos Aires, Chaco, Mendoza, Neuquén, Salta y Santa Fe). Además, sigue procesos similares para formar y desarrollar sus docentes como otras 60 organizaciones de todo el mundo que forman la red de Teach for All. Consideramos que nuestro estudio es

relevante para este grupo más amplio y para otras organizaciones que utilizan instrumentos y procedimientos comparables.

3.2. Procedimiento

En este estudio examinamos la fiabilidad de dos medidas de eficacia docente (observaciones de aula y encuestas a estudiantes) desarrolladas y administradas por ExA para fines de retroalimentación. ExA proporciona a los docentes informes sobre ambas medidas para ayudarles a mejorar su instrucción. En 2014 y 2015, ExA administró estas medidas justo después de que se contratara a docentes, durante su instituto de formación de verano (una formación previa al servicio de cuatro semanas, que concluye con dos semanas de enseñanza práctica) y durante el año escolar, una vez que los docentes ya estaban en el aula. Nos referimos al anterior proceso como "práctica clínicaz al último como el .ªño escolar". Todos los nuevos docentes participaron en la práctica clínica sólo en el año en que fueron contratados (por ejemplo, si un docente fue contratado en 2014, sólo participaron en la práctica clínica en 2014) y los docentes nuevos y existentes enseñados durante el año escolar durante dos años (por ejemplo, el conjunto de datos del año escolar 2014 incluye a los docentes contratados en 2013 [segundo año] y 2014 [primer año]).

3.2.1. Práctica clínica

Durante la práctica clínica, cada docente enseñó a un grupo de estudiantes voluntarios durante dos semanas y fue observado en dos lecciones, con un evaluador calificando cada lección en seis dominios. En los estudios G, esta configuración de docentes, lecciones, evaluadores y dominios se denota como un diseño dominio-por-lección-en-docente, o $d \times (l : t)$. En este diseño, los dominios se cruzan con docentes y lecciones (como indica el signo $\times$) porque todos los docentes fueron calificados con el mismo protocolo de observación de aula (ver sección 3.4.1) en todas las lecciones. Las lecciones están anidadas dentro de docentes (como indica el signo :) porque cada docente enseñó lecciones diferentes (por ejemplo, docente A enseñó matemática de 5.º grado; docente B enseñó lengua de 6.º grado). No se puede estimar el efecto del evaluador en la confiabilidad porque sólo había un evaluador por lección, por lo que no podemos saber cómo otro evaluador habría calificado las mismas lecciones. Algunos docentes fueron calificados por el mismo entrenador en ambas lecciones, otros por el mismo par y otros por un par diferente. Los entrenadores (pero no los pares) observaron a múltiples docentes, por lo que realizamos estudios separados para cada entrenador y reportamos el resultado promedio entre entrenadores para cada año. Esta configuración nos permite comparar la confiabilidad de estos tres enfoques para asignar evaluadores, que pueden ser de interés para profesionales que buscan equilibrar experiencia y disponibilidad de evaluadores.

Los estudiantes de cada docente también fueron encuestados en la última lección de práctica clínica sobre siete dominios. En este caso, los estudiantes actúan como evaluadores. En la notación de la Teoría G, este arreglo está representado como un dominio por valorador-en-docente o $d \times (r : t)$ diseño. Dominios se cruzan con evaluadores y docentes porque todos los docentes fueron marcados en la misma encuesta (ver sección 3.4.2). Los evaluadores están anidados porque cada docente enseñó un grupo diferente de estudiantes (por ejemplo, docente A fue calificado por estudiantes 1-10, mientras que docente B por estudiantes 11-20). Muestramos al azar 10 encuestas a estudiantes por docente para mantener constante el número de evaluadores. El efecto de la dificultad de la lección sobre la confiabilidad no puede ser estimado porque los estudiantes fueron encuestados sólo una vez, por lo que no podemos saber cómo los mismos estudiantes habrían calificado su docente en una lección diferente.

3.2.2. Año escolar

Durante el año escolar, docentes enseñan en múltiples escuelas, grados y asignaturas durante 11 meses. Cada docente fue marcado en dos ocasiones por un evaluador sobre los mismos dominios que en la práctica clínica. Nos referimos a las ocasiones aquí porque estas observaciones ocurrieron en diferentes momentos, a diferencia de las lecciones en práctica clínica, que tuvo lugar en estrecha sucesión. Esta es una docente-por-dominio-por-ocasión o $t \times d \times o$ diseño. Dominios se cruzan con todo por las mismas razones que antes. Las ocasiones también se cruzan porque todos los docentes fueron observados a mediados y finales del año. En 2014, ambas observaciones tuvieron lugar en la misma escuela, grado, sección, y sujetas a mantenerlas comparables; en 2015 se realizaron en diferentes clases para ser más completas. No se puede estimar el efecto del evaluador en la fiabilidad porque sólo hubo un evaluador por ocasión. Cada evaluador observó múltiples docentes, por lo que realizamos un estudio separado para cada evaluador y reportamos el resultado promedio entre evaluadores.

Los estudiantes de cada docente fueron encuestados dos veces usando la misma herramienta de práctica clínica. Estos son dominio-por-evaluador-en-docente o $d \times (r : t)$ diseños por ocasión. Dominios se cruzan con todo por las mismas razones que antes. Los evaluadores están anidados dentro de docentes porque cada docente tiene un conjunto diferente de estudiantes. Al igual que en la práctica clínica, mostramos al azar 10 estudiantes por ocasión para mantener constante el número de evaluadores. Realizamos un análisis separado por ocasión, en lugar de cruzar las ocasiones con todo lo demás, porque las encuestas eran anónimas, por lo que no podemos asegurar que los 10 estudiantes que mostramos en ambas ocasiones sean los mismos estudiantes. Además, en 2014, ExA revisó el mismo grupo de estudiantes en ambas ocasiones, pero en 2015 realizó encuestas de diferentes clases. Por lo tanto, es poco probable que los evaluadores sean cruzados con las ocasiones en 2014 y definitivamente no se cruzan con las ocasiones en 2015.

3.3. Sampling

Nuestro marco de muestreo incluye 100 docentes únicos que participaron en ExA en 2014 y 2015: 23 iniciaron el programa antes de 2014 y permanecieron, 32 comenzaron en 2014, y 45 comenzaron en 2015. Tenemos datos sobre los dos últimos cohortes para la práctica clínica y el año escolar, pero sólo vemos la primera cohorte durante el año escolar porque completó la práctica clínica antes de nuestro estudio.

Nuestras muestras para cada análisis no incluyen todas las docentes en una cohorte dada. Algunos docentes fueron observados menos veces que el resto, por lo que los dejamos para asegurar que todos los docentes tengan suficientes datos para estimar los componentes pertinentes de las diferencias. Durante la práctica clínica, algunos docentes fueron observados por el mismo entrenador o par en las lecciones y otros por un par diferente por lección (ver sección 3.2.1). Analizamos cada grupo por separado. Algunos docentes se incluyen en múltiples análisis, pero ninguno contribuye más de una vez al mismo análisis. Cuadro1 muestra el número de docentes, lecciones o ocasiones, evaluadores y dominios, y el diseño para cada análisis.

3.4. Medidas

3.4.1. Observaciones de clase

ExA desarrolló su protocolo de observación de aulas basado en cinco medidas creadas y administradas en los Estados Unidos: el sistema de evaluación de aulas (CLASE, Mashburn et al., 2008; Pianta et al., 2008; Hamre et al., 2013); Marco para la Enseñanza (FfT, Danielson, 2011); Enseñanza como Liderazgo (TAL, Farr, 2010); el Protocolo para la observación de la enseñanza de las artes lingüísticas (PLATO, Grossman et al., 2013, 2015); y la calidad matemática de la instrucción (MQI, Hill et al., 2011, 2012). Cubrió seis dominios: presentar el contenido claramente, controlar la comprensión, gestionar el comportamiento de los estudiantes, implementar procedimientos de clase, crear un entorno propicio para el aprendizaje y desarrollar un sentido de posibilidad. Cada dominio fue marcado basado en cinco a siete elementos en una escala 1 (prenovicio) a 5 (exemplario). Cada artículo incluyó una breve descripción para cada puntuación posible para ayudar a los evaluadores con su selección. Incluimos histogramas de las partituras de nivel académico y docente, gráficos de barras de las clasificaciones de dominio, y tablas con correlaciones entre ellos en Apéndice A. Describimos los dominios y proporcionamos elementos de ejemplo traducidos para el protocolo, y enlace al protocolo completo, en Apéndice B.

3.4.2. Encuestas de estudiantes

ExA tradujo la encuesta Tripod (Ferguson, 2010, 2012). La encuesta cubre siete dominios: atención (atendiendo a las necesidades de los estudiantes), conferir (aprendizaje de estudiantes

en conversaciones), cautivar (aprendizaje de estudiantes), aclarar (controlar la comprensión de los estudiantes), consolidar (ayudar a los estudiantes a integrar conceptos), desafiar (tener altos estándares para los estudiantes), y controlar (manejar el comportamiento de los estudiantes). Cada dominio fue marcado basado en dos a siete artículos en una escala de 1 ('nunca"') a 5 ('siempre'). Las distribuciones de los puntajes evaluadores, docentes y de dominio están en Apéndice A, y las descripciones de dominios y elementos de ejemplo en Apéndice B.

4. Análisis

4.1. Estudios de generalización

Estimamos la fiabilidad de las observaciones de aula y encuestas a un estudiante durante práctica clínica y el año escolar realizando estudios G. En todos los estudios, concibemos de la puntuación observada X_i que un docente recibe en replicación i como compuesto de una puntuación del universo τ (es decir, promedio de larga duración sobre las replicaciones) y *múltiple* facetas de error (por ejemplo, desviaciones de τ debido a las diferencias en la dificultad de dominio o la corrección evaluador). En cada estudio, descomponemos la varianza observada-score en la varianza universal-score (es decir, diferencias reales en la eficacia) y diferentes tipos de varianza de error (es decir, diferencias debido a facetas de error e interacciones entre ellos).

Como se discutió en las secciones 3.2.1 y 3.2.2, el diseño del estudio, o la forma en que los docentes fueron asignados a dominios, lecciones u ocasiones y evaluadores, difiere entre contextos y años. Cada uno de estos diseños nos permite distinguir entre diferentes fuentes de varianza de error. A continuación, explicamos cómo analizamos los datos de cada diseño utilizando modelos de efectos aleatorios.

4.1.1. El $d \times (l : t)$ y $d \times (r : t)$ diseños

Como se explica en secciones 3.2.1- 3.2.2, observaciones de aula durante práctica clínica seguir a $d \times (l : t)$ diseño y encuestas estudiantes a durante práctica clínica y el año escolar siguen a $d \times (r : t)$ diseño. En ambos, todos los docentes están marcados en los mismos dominios, pero cada docente se enfrenta a diferentes lecciones o evaluadores. Estos diseños nos permiten distinguir entre cinco fuentes de diferencia:

$$X_{dl:t} = \mu + \nu_t + \nu_d + \nu_{l:t} + \nu_{dt} + \nu_{dl:t,e} \tag{1}$$

o

$$X_{dr:t} = \mu + \nu_t + \nu_d + \nu_{r:t} + \nu_{dt} + \nu_{dr:t,e}, \tag{2}$$

Donde $X_{dl:t}$ o $X_{dr:t}$ es la puntuación observada para docente t sobre el dominio d, evaluado en la lección l o por evaluador r; μ es la gran media (es decir, la puntuación media en todos los docentes, dominios y lecciones o evaluadores); ν_t es el efecto docente (es decir, cuánto docente t difiere en su desempeño); ν_d es el efecto del dominio (es decir, cuánto dominio d difiere en su dificultad); $\nu_{l:t}$ o $\nu_{r:t}$ son el efecto evaluación o evaluador (es decir, cuánto lección l difiere en su dificultad o evaluador r en su rigor), anidado dentro de docentes; ν_{dt} es el efecto dominio-por-docente (es decir, cuánto dominio d difiere en su dificultad para docente t); y $\nu_{dl:t,e}$ o $\nu_{dr:t,e}$ es el efecto dominio-por-lección o dominio-por-evaluador (es decir, cuánto dominio d difiere en su dificultad para la lección l o evaluador r), anidado dentro de docentes y confundido con variación residual. Los parámetros de interés no son estos efectos aleatorios, sino sus diferencias, que se calculan directamente mediante una probabilidad máxima restringida.

En estos diseños, podemos estimar la varianza relativa de error $\hat{\sigma}_\delta^2$ (es decir, variación de las puntuaciones de las facetas de error que afectan a la posición relativa o clasificación de docentes) utilizando las fórmulas:

$$\hat{\sigma}_\delta^2 = \frac{\hat{\sigma}_{l:t}^2}{n_l} + \frac{\hat{\sigma}_{dt}^2}{n_d} + \frac{\hat{\sigma}_{dl:t,e}^2}{n_d n_l}, \tag{3}$$

o

$$\hat{\sigma}_\delta^2 = \frac{\hat{\sigma}_{r:t}^2}{n_r} + \frac{\hat{\sigma}_{dt}^2}{n_d} + \frac{\hat{\sigma}_{dr:t,e}^2}{n_d n_r}, \tag{4}$$

Donde $\hat{\sigma}_{l:t}^2$ y $\hat{\sigma}_{r:t}^2$ son las diferencias estimadas de las lecciones y evaluadores, anidadas dentro de docentes; $\hat{\sigma}_{dt}^2$ es la diferencia de la interacción de dominio por centro; $\hat{\sigma}_{dl:t,e}^2$ o $\hat{\sigma}_{dr:t,e}^2$ es la diferencia de la interacción entre dominios y lecciones o evaluadores, anidado dentro de docentes y confundido con error residual; y n_d, n_l, y n_r son los números de dominios, lecciones y evaluadores.

También podemos estimar la varianza absoluta de error $\hat{\sigma}_\Delta^2$ (es decir, la variación de las puntuaciones de las facetas de error que afectan no sólo la clasificación, sino también las ubicaciones de docentes en la escala de puntuación) como:

$$\hat{\sigma}_\Delta^2 = \frac{\hat{\sigma}_d^2}{n_d} + \hat{\sigma}_\delta^2, \tag{5}$$

Donde $\hat{\sigma}_d^2$ es la diferencia de dominio estimada y todo lo demás es como arriba.

Podemos utilizar nuestras estimaciones de la varianza de error relativa y absoluta para obtener coeficientes de generalización para errores relativos y absolutos $\mathbb{E}\hat{\rho}^2$ y Φ. Estos son similares a los coeficientes de confiabilidad de CTT como el alfa de Cronbach, pero son más generales porque tienen en cuenta la variabilidad de error derivada de múltiples facetas de error y de interacciones entre ellos. Definen la fiabilidad como la parte de la varianza total

explicada por la varianza de puntuación del universo:

$$\mathbb{E}\hat{\rho}^2 = \frac{\hat{\sigma}_t^2}{\hat{\sigma}_t^2 + \hat{\sigma}_\delta^2} \tag{6}$$

y

$$\hat{\Phi} = \frac{\hat{\sigma}_t^2}{\hat{\sigma}_t^2 + \hat{\sigma}_\Delta^2} \tag{7}$$

Donde $\hat{\sigma}_t^2$ es la variación del universo-score estimada y todo lo demás es como arriba. Estas fórmulas son siempre las mismas independientemente del diseño del estudio, por lo que no las repetimos a continuación.

4.1.2. El $t \times d \times o$ diseño

Como se explica en la sección 3.2.2, durante el año escolar observaciones de aula seguir a $t \times d \times o$ diseño. En este diseño, todos los docentes están marcados en los mismos dominios y ocasiones. Este diseño nos permite descomponer las puntuaciones observadas en siete fuentes de varianza:

$$X_{tdo} = \mu + \nu_t + \nu_d + \nu_o + \nu_{dt} + \nu_{to} + \nu_{do} + \nu_{tdo,e}, \tag{8}$$

Donde X_{tdo} es la puntuación observada para docente t sobre el dominio d y ocasión o; μ es el gran medio; ν_t es el efecto docente; ν_d es el efecto de dominio; ν_o es el efecto de la ocasión (es decir, cuánto ocasión o difiere en su dificultad); ν_{dt} es el efecto dominio-por-docente; ν_{to} es el efecto docente-por-ocasión (es decir, cuánto docente t difiere en su actuación en ocasión r); ν_{do} es el efecto de dominio por ocasión (es decir, cuánto dominio d difiere en su dificultad en la ocasión o); y $\nu_{tdo,e}$ es el efecto docente-por-dominio-por-ocasión (es decir, cuánto docente t difiere en su desempeño en el dominio d y ocasión o), confundido con error residual.

Podemos estimar la varianza relativa de error como:

$$\hat{\sigma}_\delta^2 = \frac{\hat{\sigma}_{dt}^2}{n_d} + \frac{\hat{\sigma}_{to}^2}{n_o} + \frac{\hat{\sigma}_{tdo,e}^2}{n_d n_o}, \tag{9}$$

Donde $\hat{\sigma}_{to}^2$ y $\hat{\sigma}_{tdo,e}^2$ son las diferencias estimadas para las interacciones docente-por-ocasión y docente-por-dominio-por-ocasión; n_d y n_o son los números de dominios y ocasiones; y todo lo demás es como arriba. También podemos estimar la varianza absoluta de error como:

$$\hat{\sigma}_\Delta^2 = \frac{\hat{\sigma}_d^2}{n_d} + \frac{\hat{\sigma}_o^2}{n_o} + \frac{\hat{\sigma}_{do}^2}{n_d n_o} + \hat{\sigma}_\delta^2, \tag{10}$$

Donde $\hat{\sigma}_d^2$, $\hat{\sigma}_o^2$, y $\hat{\sigma}_{do}^2$ son las diferencias estimadas para los dominios, las ocasiones y la interacción dominio- by-ocasión; y todo lo demás es como arriba.

4.2. Estudios de decisión

A continuación, identificamos el enfoque óptimo para aumentar la confiabilidad de las observaciones de aula y encuestas a los estudiantes usando estudios de D(ecision). En cada estudio D, tomamos los coeficientes de generalización del error relativo y absoluto de un diseño de estudio, que capturan la confiabilidad de estos instrumentos en las condiciones actuales, y calculamos cómo cambiarían si promediamos más evaluadores y lecciones o ocasiones en cada procedimiento de medición. Como se explica en la sección 4.1, estos coeficientes se derivan de las estimaciones de la varianza relativa y absoluta de errores basadas en los componentes de varianza de cada estudio G. El cálculo de estas diferencias incluye el número de réplicas para cada faceta de error en cada diseño. Al permitir que algunos de estos números varían, podemos anticipar su impacto esperado en la confiabilidad.

4.2.1. El $d \times (l : t)$ y $d \times (r : t)$ diseños

Como ecuaciones (5)-(7) mostrar, en estos diseños, relativa y absoluta variabilidad de errores y sus coeficientes de generalización dependen en parte del número de dominios y lecciones o evaluadores. Por lo tanto, si aumentamos alguno de ellos, la varianza de error disminuiría y aumentaría la confiabilidad. Esto tiene sentido intuitivo: si los docentes están marcados en más dominios o lecciones o por más evaluadores, sus puntajes deben ser más fiables (porque estamos aumentando el número de réplicas). Asumimos que el protocolo de observación y la encuesta tienen fuertes justificaciones teóricas y estimamos cómo aumentar el número de lecciones o evaluadores afectaría su fiabilidad. Dejaremos que el número de lecciones o evaluadores variará en el cálculo de la varianza relativa de error:

$$\hat{\sigma}_\delta^2 = \frac{\hat{\sigma}_{l:t}^2}{n_l} + \frac{\hat{\sigma}_{dt}^2}{n_d} + \frac{\hat{\sigma}_{dl:t,e}^2}{n_d n_l'}, \tag{11}$$

o

$$\hat{\sigma}_\delta^2 = \frac{\hat{\sigma}_{r:t}^2}{n_r} + \frac{\hat{\sigma}_{dt}^2}{n_d} + \frac{\hat{\sigma}_{dr:t,e}^2}{n_d n_r'}, \tag{12}$$

y también en el cálculo de la varianza absoluta de error:

$$\hat{\sigma}_\Delta^2 = \frac{\hat{\sigma}_d^2}{n_d} + \hat{\sigma}_\delta^2, \tag{13}$$

Donde n_l' y n_r' son el número de lecciones y evaluadores que se permiten variar y todo lo demás es como arriba. Si aumentamos estos números, la varianza de error disminuiría (porque están en el denominador de ambos conjuntos de expresiones) y los coeficientes de generalización aumentarían (porque la variabilidad de error en sus denominadores; ver ecuaciones [6] y [7]).

4.2.2. El $t \times d \times o$ diseño

Como ecuaciones (9)-(10) mostrar, en este diseño, relativa y absoluta variabilidad de errores y sus coeficientes de generalización dependen en parte del número de dominios y ocasiones. Si nuevamente mantenemos el número de dominios constantes en observaciones y encuestas, podemos permitir que el número de ocasiones cambie para estimar cómo aumentarlos afectaría la variabilidad relativa del error:

$$\hat{\sigma}_{\delta}^2 = \frac{\hat{\sigma}_{dt}^2}{n_d} + \frac{\hat{\sigma}_{to}^2}{n'_o} + \frac{\hat{\sigma}_{tdo,e}^2}{n_d n'_o}, \tag{14}$$

y varianza absoluta de error:

$$\hat{\sigma}_{\Delta}^2 = \frac{\hat{\sigma}_{d}^2}{n_d} + \frac{\hat{\sigma}_{o}^2}{n'_o} + \frac{\hat{\sigma}_{do}^2}{n_d n'_o} + \hat{\sigma}_{\delta}^2, \tag{15}$$

Donde n'_o es el número variable de ocasiones y todo lo demás es como arriba.

5. Resultados

5.1. Observaciones de clase

Las observaciones de clase en este entorno pueden alcanzar altos niveles de fiabilidad para hacer tanto distinción relativa como juicios absolutos sobre docentes. Como tabla2 muestra que los coeficientes de error relativo (en la tercera fila desde abajo) oscilaron entre 0,53 y 0,79 y los de error absoluto (en la segunda fila a última) oscilaron entre 0,38 y 0,76. El hecho de que el primero sea ligeramente más grande que el segundo no debe ser sorprendente, ya que el error relativo incluye fuentes de variabilidad que sólo cambian la posición relativa de los docentes, mientras que el error absoluto incluye aquellos que cambian sus posiciones relativas y absolutas (como sección 4.1 muestra, el cálculo de la varianza absoluta de error incluye, y por lo tanto siempre es igual o mayor que, error relativo). Los valores más altos en ambos conjuntos de coeficientes indican que estas observaciones podrían utilizarse para identificar cuáles son las docentes más necesitadas de apoyo y asignar a todos los docentes puntuación por debajo de un umbral a una intervención (por ejemplo, recursos, capacitación, coaching).

En promedio, las observaciones realizadas durante la práctica clínica tenían niveles de confiabilidad similares a los de la escuela del año. El coeficiente de generalización media para el error relativo de práctica clínica fue de 0.62 y el del año escolar fue de 0.64. El coeficiente medio para el error absoluto de práctica clínica fue 0.56 y el del año escolar fue 0.47. Sin embargo, la fiabilidad de las observaciones de prácticas clínicas varía más que la de las observaciones de los años escolares. Los coeficientes de error relativo y absoluto en la práctica clínica oscilaron

entre 0,44 y 0,79 y de 0,38 a 0,76. Los correspondientes al año escolar oscilaron entre 0,62 y 0,6 y entre 0,44 y 0,51. Estos resultados sugieren que un contexto no era inherentemente más propicio para la fiabilidad y que hay otros factores que explican la variabilidad durante la práctica clínica.

Un factor que podría explicar las similitudes práctica clínica y el año escolar y la variabilidad en confiabilidad durante la práctica clínica es la forma en que se asignaron los evaluadores. Si comparamos las observaciones en las que el mismo entrenador anotó ambas lecciones en práctica clínica y el año escolar, ambos conjuntos de observaciones tenían una fiabilidad similar. Los coeficientes de generalización para errores relativos oscilaron entre 0,53 y 0,55 durante la práctica clínica y entre 0,62 y 0,66 durante el año escolar, y los errores absolutos oscilaron entre 0,38 y 0,47 durante la práctica clínica y entre 0,43 y 0,57 durante el año escolar. Además, en la práctica clínica, las observaciones en las que el mismo evaluador (coach o peer) anotó ambas lecciones entregadas por un docente tenían mayor fiabilidad que aquellas en las que una persona diferente anotó cada lección. Los coeficientes de error relativo para los primeros variaron de 0,53 a 0,79 en el primer caso y de 0,44 a 0,64 en el segundo caso, y los de error absoluto variaron de 0,38 a 0,76 en el primer caso y de 0,41 a 0,61 en el segundo. Estos resultados indican que la asignación evaluador importa más que el contexto en que se realizan observaciones o incluso quién actúa como evaluador.

Los cuadros que presentan los resultados de las descomposiciones de las diferencias de puntuación suelen incluir también columnas que indican el porcentaje de la varianza total que representa cada componente de la varianza. Es importante recordar, sin embargo, que los componentes de las diferencias son variaciones estimadas de las distribuciones de los puntajes más elementales (por ejemplo, en los $t \times d \times o$ diseño, X_{tdo} o la puntuación observada para docente t sobre el dominio d y ocasión o), no las puntuaciones promedio que usamos (por ejemplo, en el mismo diseño, $\bar{X}_t$ o la puntuación media para docente t a través de dominios y ocasiones). Para describir la importancia de una fuente de error en términos de su impacto en la fiabilidad para las puntuaciones que utilizamos más comúnmente, reportamos los resultados de nuestros estudios D.

El aumento de la confiabilidad de los juicios relativos de las observaciones parece factible tanto en la práctica clínica como en el año escolar. Como muestra la figura 1, el coeficiente de generalización del error relativo en las observaciones de práctica clínica está entre 0,44 y 0,79 cuando cada docente es calificado dos veces, independientemente del tipo de evaluador (ver la coordenada del eje y de las líneas azules en 2 lecciones en los paneles A-F). Agregar una lección mejoraría considerablemente este coeficiente en 5-10 pp. (ver la coordenada del eje y de las mismas líneas en 3 lecciones). Los aumentos adicionales en el número de lecciones sólo mejorarían marginalmente la confiabilidad en 3-7 pp., a pesar de hacer tales observaciones más complejas logísticamente (nótese que las pendientes de estas líneas son cada vez más planas

después de 3 lecciones). El coeficiente de error relativo en las observaciones del año escolar es de alrededor de 0,6 cuando cada docente es calificado dos veces (ver paneles G-H). Añadir una ocasión mejoraría este coeficiente en 8 pp., pero los aumentos posteriores en el número de ocasiones lo mejorarían sólo en 5 pp. adicionales.

Agregar lecciones u ocasiones tendría un impacto ligeramente menor en la confiabilidad de los juicios absolutos de las observaciones. Como muestra la figura 1, el coeficiente de generalización del error absoluto está entre 0,38 y 0,76 durante la práctica clínica cuando un docente es calificado dos veces por el mismo entrenador o por un par diferente en cada lección (ver las coordenadas del eje y de las líneas rojas en 2 lecciones en los paneles A-B y E-F). Agregar una lección elevaría este coeficiente en 4-9 pp. (ver las coordenadas del eje y de las mismas líneas en 4 lecciones en los paneles A, E-F). Otros aumentos en el número de lecciones lograrían mejoras más pequeñas en la confiabilidad, de 2-6 pp. El patrón es similar para el año escolar. El coeficiente de error absoluto está entre 0,43 y 0,57 cuando cada docente es calificado dos veces (ver paneles G-H). Añadir una ocasión mejoraría este coeficiente en 7-8 pp., pero los aumentos posteriores lo mejorarían sólo en 4-5 pp. adicionales.

5.2. Encuestas de estudiantes

Las encuestas de estudiantes también pueden alcanzar altos niveles de fiabilidad. Como tabla3 muestra que los coeficientes de error relativo oscilaron entre 0,5 y 0,76 y los de error absoluto entre 0,37 y 0,65. Como en el caso de las observaciones de aula, los valores más altos en ambos conjuntos de coeficientes indican que las encuestas pueden ayudar a hacer distinciones relativas entre y juicios absolutos sobre docentes basados en sólo 10 estudiantes por docente.

Hubo relativamente poca variación en la fiabilidad de las encuestas en contextos y años. El coeficiente de generalización media para el error relativo de práctica clínica fue de 0,6 y el del año escolar fue de 0,63. El coeficiente medio para el error absoluto de práctica clínica fue 0.58 y el del año escolar fue 0.51. Estos resultados sugieren que la fiabilidad de las encuestas es estable en contextos y asignaciones de evaluadores. Una advertencia importante, sin embargo, es que la fiabilidad fue más baja en el año 2015 escolar, cuando ExA cambió de encuestar a los mismos estudiantes dos veces para encuestar diferentes grupos de estudiantes (ver sección 3.2.2). Más ampliamente, ExA hizo pocos cambios en los diseños de estudio para encuestas a estudiantes, por lo que la aparente estabilidad en las estimaciones de fiabilidad puede ser en parte una función de sólo dos diseños que se comparan.

Aumentar el número de evaluadores mejoraría la confiabilidad de los juicios relativos. Como muestra la figura 2, el coeficiente de generalización del error relativo en las encuestas de práctica clínica está entre 0,56 y 0,63 con 10 estudiantes (ver las coordenadas del eje y de las líneas azules en 10 estudiantes en los paneles A-B). Añadir 5 estudiantes mejoraría la

confiabilidad en 9 pp. (ver las coordenadas del eje y de estas líneas en 15 estudiantes), pero añadir 5 más sólo la mejoraría de 7 a 9 pp. (ver las coordenadas del eje y en 20 estudiantes). El impacto de añadir evaluadores en el año escolar es ligeramente inferior. El coeficiente de error relativo está entre 0,35 y 0,64 con 10 estudiantes. Añadir 5 estudiantes lo aumentaría en 5-9 pp., y añadir 5 estudiantes más lo haría sólo en 3-5 pp.

Añadir evaluadores tendría un impacto similar en la confiabilidad de los juicios absolutos de las encuestas. Como muestra la figura 2, el coeficiente de generalización del error absoluto está entre 0,53 y 0,62 para los estudios de práctica clínica con 10 estudiantes (ver las coordenadas del eje y de las líneas rojas en 10 estudiantes en los paneles A-B). Añadir 5 estudiantes mejoraría la confiabilidad en 8 pp. (ver las coordenadas del eje y en 15 estudiantes), pero añadir 5 más sólo la mejoraría en 5 pp. (ver las coordenadas del eje y en 20 estudiantes). Una vez más, añadir evaluadores tendría un menor impacto en la confiabilidad durante el año escolar. El coeficiente de error absoluto está entre 0,37 y 0,65 con 10 estudiantes. Añadir 5 estudiantes lo aumentaría en 4-8 pp., pero 5 más sólo lo harían en 2-5 pp.

6. Debate

En este documento, presentamos uno de los primeros estudios G de dos medidas no-pruebas de eficacia docente en un país de ingresos medianos: observaciones de aula y encuestas a estudiantes. Nuestra motivación era doble. En primer lugar, estudios anteriores G se basaron en los datos recogidos para la investigación, con varios mecanismos de garantía de calidad existentes, por lo que evaluamos si sus resultados son indicativos de la fiabilidad de los instrumentos administrados para la práctica. En segundo lugar, estudios anteriores G se centraron en un pequeño conjunto de medidas y contextos, por lo que se evaluó si son representativos de las realidades de los instrumentos menos establecidos administrados en los CMI. Obtuvimos datos de una educación sin ánimo de lucro y examinamos la confiabilidad de sus métricas.

Encontramos que tanto las observaciones de aula como las encuestas a los estudiantes administrados en la práctica pueden alcanzar altos niveles de fiabilidad. Creemos que este hallazgo es importante porque demuestra que los practicantes no siempre necesitan adoptar costosos mecanismos de garantía de calidad para producir calificaciones fiables de docentes. También encontramos, sin embargo, que la fiabilidad de las observaciones de aula variaba ampliamente dependiendo de cómo se asignan los evaluadores. Vemos esto como una buena razón para que los practicantes realicen sus propios estudios sobre la fiabilidad de sus instrumentos, en lugar de depender de nuestras estimaciones para decidir cómo diseñar sus sistemas de retroalimentación docente. Para apoyarlos en este esfuerzo, hemos explicado en gran detalle cómo comprender el diseño de cada procedimiento de medición y cómo analizar

los datos que cada uno produce. También hemos hecho disponibles los conjuntos de datos y el código de nuestros análisis con este documento.

Como ilustramos utilizando nuestros propios datos, los estudios G pueden ser útiles para comprender no sólo la fiabilidad actual de un procedimiento de medición, sino también el enfoque óptimo para mejorarlo. Mostramos que simplemente añadiendo una lección o ocasión en observaciones de aula o cinco evaluadores en encuestas un estudiante logró mejoras significativas en la fiabilidad. Igualmente importante, también demostramos la disminución del rendimiento marginal de nuevas expansiones en el número de lecciones, ocasiones o evaluadores. Consideramos que este enfoque es particularmente útil para los profesionales. En primer lugar, les permite evaluar la compensación entre las posibles mejoras en la fiabilidad de los cambios a sus sistemas existentes contra su costo. En segundo lugar, les permite utilizar sus recursos de la manera más eficiente posible, cambiando sólo lo necesario para lograr medidas fiables.

Al utilizar los estudios G para diseñar sus sistemas de retroalimentación docente, es importante que los profesionales observen que las recomendaciones de los estudios D se estiman con una imprecisión considerable. Un estudio reciente utilizó la estimación bayesiana para reanalizar datos de un estudio G en el campo médico y encontró que el número mínimo de evaluadores necesarios para alcanzar niveles adecuados de fiabilidad era mayor y más variable que el estudio había implicado (Himmelsbach and Gilbert, 2025). Por lo tanto, no recomendamos que los practicantes lleven a cabo un estudio único de G y D y asuman que sus reliquias son precisas ni que se aplican a todas las administraciones posteriores de sus instrumentos. Como lo ilustra nuestro análisis, puede haber una variación significativa año a año en la fiabilidad de las medidas no de prueba de eficacia docente administradas por los profesionales. En cambio, alentamos a los practicantes a examinar periódicamente la fiabilidad de sus medidas y hacer ajustes según sea necesario. Nuestros resultados sugieren que este enfoque sería preferible suponer que las recomendaciones de los entornos de investigación formal se generalicen por sí mismas.

Nos asociamos con este camino alternativo a la enseñanza porque es miembro de una red internacional que utiliza prácticas similares para capacitar y proporcionar retroalimentación a sus docentes. Por lo tanto, anticipamos que los procesos y percepciones de nuestro estudio serán relevantes para estos otros programas, impactando a miles de docentes cada año y a los estudiantes que sirven. Sin embargo, para responder adecuadamente a la pregunta de si las observaciones de aula y encuestas a los estudiantes administrados por profesionales pueden producir resultados fiables, debe haber más análisis de los datos recogidos por los gobiernos y las organizaciones sin fines de lucro, especialmente en los centros de enseñanza preescolar. Vemos este cambio como similar al que ha tenido lugar en la literatura de evaluación del impacto, que ha pasado de evaluar programas en pequeña escala dirigidos por organizaciones

altamente capaces en un conjunto estrecho de contextos para tratar de comprender el efecto de las iniciativas cuando se implementan a escala.

Referencias

Aaronson, D., L.˜Barrow, and W.˜Sander (2007). Teachers and student achievement in the chicago public high schools. *Journal of Labor Economics˜25*(1), 95–135.

Allen, M.˜J. and W.˜M. Yen (1979). Introduction to measurement theory. Prospect Heights, IL: Waveland Press.

Blazar, D. (2018). Validating teacher effects on students' attitudes and behaviors: Evidence from random assignment of teachers to students. *Education Finance and Policy˜13*(3), 281–309.

Brennan, R.˜L. (2001). *Generalizability Theory*. New York: Springer-Verlag.

Chetty, R., J.˜N. Friedman, and J.˜E. Rockoff (2014). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review˜104*(9), 2593–2632.

Cronbach, L.˜J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika˜16*(3), 297–334.

Cronbach, L.˜J., G.˜C. Gleser, H.˜Nanda, and N.˜Rajaratnam (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Cruz-Aguayo, Y., D.˜Hincapié, and C.˜Rodríguez (2020). Testing our teachers: Keys to a successful teacher evaluation system.

Danielson, C. (2011). *Enhancing professional practice: A framework for teaching*. Association for Supervision and Curriculum Development (ASCD).

Dee, T.˜S. and J.˜Wyckoff (2015). Incentives, selection, and teacher performance: Evidence from impact. *Journal of Policy Analysis and Management˜34*(2), 267–297.

English, D., J.˜Burniske, D.˜Meibaum, and L.˜Lachlan-Haché (2015). Uncommon measures: Student surveys and their use in measuring teaching effectiveness. Washington, DC: American Institutes for Research (AIR).

Farr, S. (2010). *Teaching as leadership: The highly effective teacher's guide to closing the achievement gap*. San Francisco, CA: Teach for America.

Ferguson, R.˜F. (2010). Student perceptions of teaching effectiveness. Boston, MA: The National Center for Teacher Effectiveness and the Achievement Gap Initiative.

Ferguson, R.˜F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan˜94*(3), 24–28.

Gage, N.˜A., H.˜Han, A.˜S. MacSuga-Gage, D.˜Prykanowski, and A.˜Harvey (2018). *A generalizability study of a direct observation screening tool of teachers' classroom management skills*, Volume Emerging research and issues in behavioral disabilities, pp. 29–50. Emerald Publishing.

Ganimian, A.˜J. and V.˜Mesalles (2025). ¿Qué aprendimos de Aprender? Informe sobre el desempeño de las 24 jurisdicciones argentinas en las evaluaciones nacionales. Ciudad Autónoma de Buenos Aires, Argentina: Educar 2050 and Argentinos por la Educación.

Goldhaber, D., C.˜Grout, and N.˜Huntington-Klein (2017). Screen twice, cut once: Assessing the predictive validity of applicant selection tools. *Education Finance and Policy˜12*(2), 197–223.

Grossman, P., J.˜Cohen, and L.˜Brown (2015). *Understanding instructional quality in English language arts: Variations in PLATO scores by content and context*, pp. 303–331. Wiley Online Library.

Grossman, P., S.˜Loeb, J.˜Cohen, and J.˜Wyckoff (2013). Measure for measure: The relationship between measures of instructional practice in middle school english language arts and teachers' value-added scores. *American Journal of Education˜119*(3), 445–470 0195–6744.

Hamre, B.˜K., R.˜C. Pianta, J.˜T. Downer, J.˜DeCoster, A.˜J. Mashburn, S.˜M. Jones, J.˜L. Brown, E.˜Cappella, M.˜Atkins, S.˜E. Rivers, M.˜Atkins, S.˜E. Rivers, M.˜A. Brackett, and A.˜Hamagami (2013). Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms. *The elementary school journal˜113*(4), 461–487.

Hill, H.˜C., C.˜Y. Charalambous, and M.˜A. Kraft (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher˜41*(2), 56–64.

Hill, H.˜C., L.˜Kapitula, and K.˜Umland (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal˜48*(3), 794–831.

Himmelsbach, Z. and J.˜Gilbert (2025). The case for bayesian estimation of the d-study. Presentation at the Measurement Lab. Cambridge, MA: Harvard Graduate School of Education (HGSE).

Ho, A.˜D. and T.˜J. Kane (2013). The reliability of classroom observations by school personnel. Seattle, WA: Bill and Melinda Gates Foundation.

INDEC (2024). Encuesta Permanente de Hogares (EPH) total urbano. Evolución de la distribución del ingreso. Tercer trimestre de 2023. (Trabajo e ingresos, Vol. 8, no. 2). Buenos Aires, Argentina: Instituto Nacional de Estadística y Censos (INDEC).

Jackson, C.˜K. (2013). Match quality, worker productivity, and worker mobility: Direct evidence from teachers. *Review of Economics and Statistics˜95*(4), 1096–1116.

Jackson, C.˜K. (2020). What do test scores miss? The importance of teacher effects on non-test-score outcomes. *Journal of Political Economy˜126*(5), 2072–2107.

Jackson, C.˜K. and E.˜Bruegmann (2009). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics˜1*(4), 85–108.

Jerald, C. (2012). Ensuring accurate feedback from observations. *Perspectives on Practive*. Seattle, WA: Bill and Melinda Gates Foundation.

Joe, J.˜N., C.˜M. Tocci, S.˜L. Holtzman, and J.˜C. Williams (2013). Foundations of observation: Considerations for developing a classroom observation system that helps districts achieve consistent and accurate scores. *Policy and Practice Brief*. Seattle, WA: Bill and Melinda Gates Foundation.

Johnson, S.˜M., M.˜A. Kraft, and J.˜P. Papay (2012). How context matters in high-need schools: The effects of teachers' working conditions on their professional satisfaction and their students' achievement. *Teachers College Record˜114*(10), 1–39.

Kane, T.˜J., K.˜Kerr, and R.˜C. Pianta (2014). *Designing teacher evaluation systems: New guidance from the measures of effective teaching project*. John Wiley & Sons.

Kane, T.˜J., D.˜F. McCaffrey, T.˜Miller, and D.˜O. Staiger (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. *Measures of Effective Teaching Project*. Seattle, WA: Bill and Melinda Gates Foundation.

Kane, T.˜J., J.˜E. Rockoff, and D.˜O. Staiger (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review˜27*(6), 615–631.

Kane, T.˜J. and D.˜O. Staiger (2011). Learning about teaching: Initial findings from the measures of effective teaching project. Bill and Melinda Gates Foundation. Seattle, WA.

Kane, T.˜J. and D.˜O. Staiger (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Seattle, WA: Bill and Melinda Gates Foundation.

Kane, T.˜J., E.˜S. Taylor, J.˜H. Tyler, and A.˜L. Wooten (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources˜46*(3), 587–613.

Koedel, C., K.˜Mihaly, and J.˜E. Rockoff (2015). Value-added modeling: A review. *Economics of Education Review*.

Kraft, M.˜A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources˜54*(1), 1–36.

Ley de Educación Nacional (2006). Ley de Educación Nacional No. 26.206. Buenos Aires, Argentina: Honorable Congreso de la Nación Argentina.

Lord, F.˜M. and M.˜R. Novick (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Mantzicopoulos, P., B.˜F. French, and H.˜Patrick (2018). The mathematical quality of instruction (mqi) in kindergarten: An evaluation of the stability of the mqi using generalizability theory. *Early Education and Development˜29*(6), 893–908.

Mantzicopoulos, P., B.˜F. French, H.˜Patrick, J.˜S. Watson, and I.˜Ahn (2018). The stability of kindergarten teachers' effectiveness: A generalizability study comparing the framework for teaching and the classroom assessment scoring system. *Educational Assessment˜23*(1), 24–46.

Mashburn, A.˜J., J.˜T. Downer, S.˜E. Rivers, M.˜A. Brackett, and A.˜Martinez (2014). Improving the power of an efficacy study of a social and emotional learning program: Application of generalizability theory to the measurement of classroom-level outcomes. *Prevention science˜15*, 146–155.

Mashburn, A.˜J., J.˜P. Meyer, J.˜P. Allen, and R.˜C. Pianta (2014). The effect of observation length and presentation order on the reliability and validity of an observational measure of teaching quality. *Educational and Psychological Measurement˜74*(3), 400–422.

Mashburn, A.˜J., R.˜C. Pianta, B.˜K. Hamre, J.˜T. Downer, O.˜A. Barbarin, D.˜Bryant, M.˜Burchinal, D.˜M. Early, and C.˜Howes (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child development˜79*(3), 732–749.

McClellan, C. (2013). What it looks like: Master coding videos for observer training and assessment. *Policy and Practice Brief*. Seattle, WA: Bill and Melinda Gates Foundation.

MdCH (2024). Anuario estadístico 2023. Buenos Aires, Argentina: Secretaría de Educación, Ministerio de Capital Humano.

MET Project (2010). Validation engine for observational protocols. Seattle, WA: Bill and Melinda Gates Foundation.

MET Project (2013). Ensuring fair and reliable measures of effective teaching: Culminating findings from the met project's three-year study. *Policy and Practice Brief*. Seattle, WA: Bill and Melinda Gates Foundation.

Meyer, J.˜P., A.˜H. Cash, and A.˜Mashburn (2011). Occasions and the reliability of classroom observations: Alternative conceptualizations and methods of analysis. *Educational Assessment˜16*(4), 227–243.

Mulhern, C. (2023). Beyond teachers: Estimating individual school counselors' effects on educational attainment. *American Economic Review˜113*(11), 2846–2893.

Nunnally, J.˜C. and I.˜H. Bernstein (1978). Psychometric theory (2nd edition). New York: McGraw-Hill.

Nye, B., S.˜Konstantopoulos, and L.˜V. Hedges (2004). How large are teacher effects? *Educational evaluation and policy analysis˜26*(3), 237–257.

OECD (2023). PISA 2022 results (Vol. I): The state of learning and equity in education. Paris, France: Organization for Economic Cooperation and Development (OECD).

Papay, J.˜P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal˜48*(1), 163–193.

Papay, J.˜P., E.˜S. Taylor, J.˜H. Tyler, and M.˜E. Laski (2020). Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data. *American Economic Journal: Economic Policy˜12*(1), 359–388.

Patrick, H., B.˜F. French, and P.˜Mantzicopoulos (2020). The reliability of framework for teaching scores in kindergarten. *Journal of Psychoeducational Assessment˜38*(7), 831–845.

Pianta, R.˜C., K.˜M. La˜Paro, B.˜K. Hamre, and P.˜H. (2008). Classroom assessment scoring system (class) manual: K-3. Baltimore, MD: Paul Brookes Publishing Co.

Pouezevara, S., A.˜Pflepsen, L.˜Nordstrum, S.˜King, and A.˜Gove (2016). Measures of quality through classroom observation for the sustainable development goals: Lessons from low- and middle-income countries. Paris, France: United Nations Educational, Scientific, and Cultural Organization (UNESCO).

Praetorius, A.-K., C.˜Pauli, K.˜Reusser, K.˜Rakoczy, and E.˜Klieme (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and instruction˜31*, 2–12.

Rivkin, S.˜G., E.˜A. Hanushek, and J.˜F. Kain (2005). Teachers, schools, and academic achievement. *Econometrica*, 417–458.

Rockoff, J.˜E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, 247–252.

Rockoff, J.˜E., B.˜A. Jacob, T.˜J. Kane, and D.˜O. Staiger (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy˜6*(1), 43–74.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics˜125*(1), 175–214.

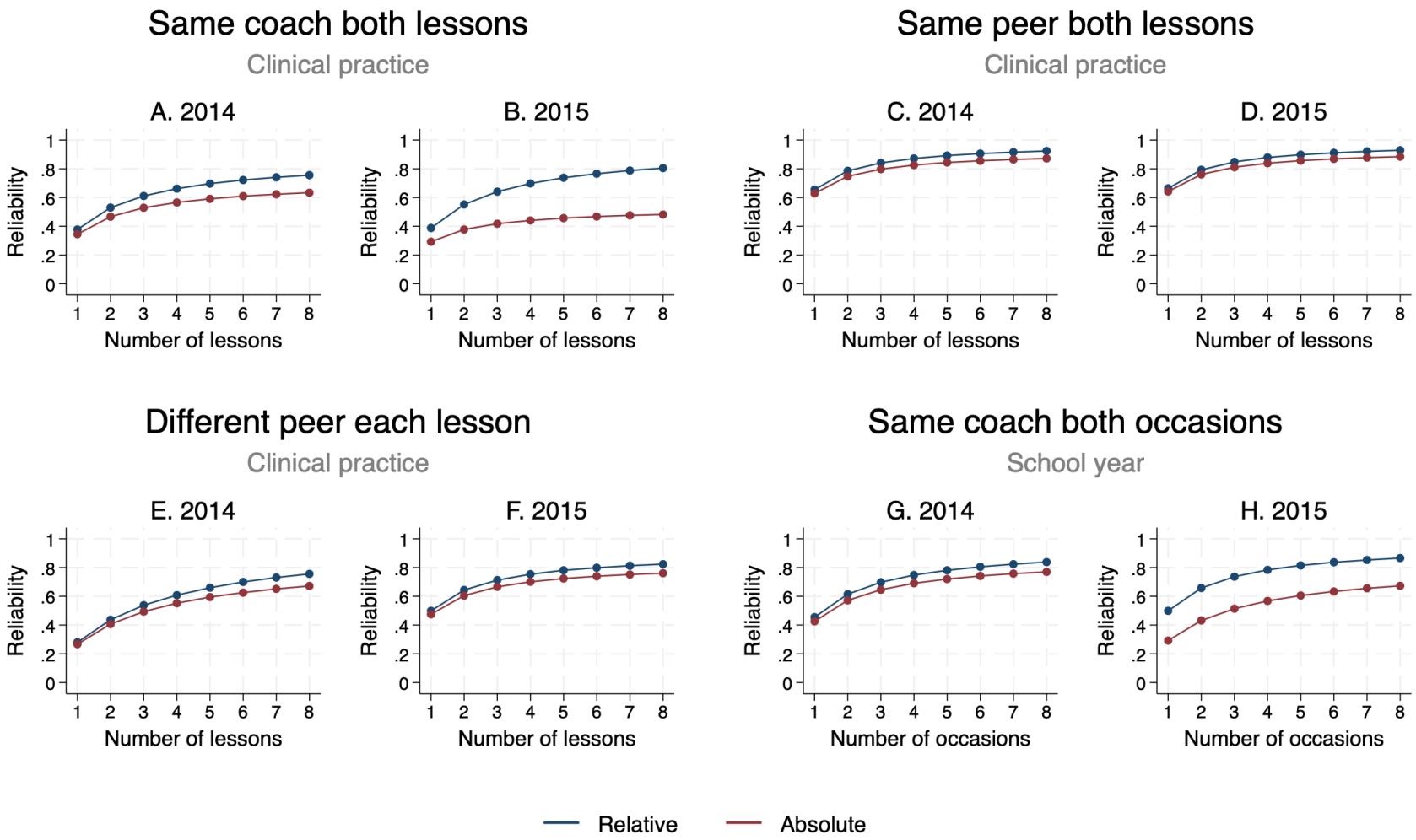
van˜der Lans, R.˜M., W.˜J. C.˜M. van˜de Grift, K.˜van Veen, and M.˜Fokkens-Bruinsma (2016). Once is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. *Studies in Educational Evaluation˜50*, 88–95.

World Bank (2024a). DataBank - Education Statistics. Retrieved from `https://databank.worldbank.org/`.

World Bank (2024b). Education statistics (Edstats). `https://datatopics.worldbank.org/education/` Retrieved: June 8, 2023.

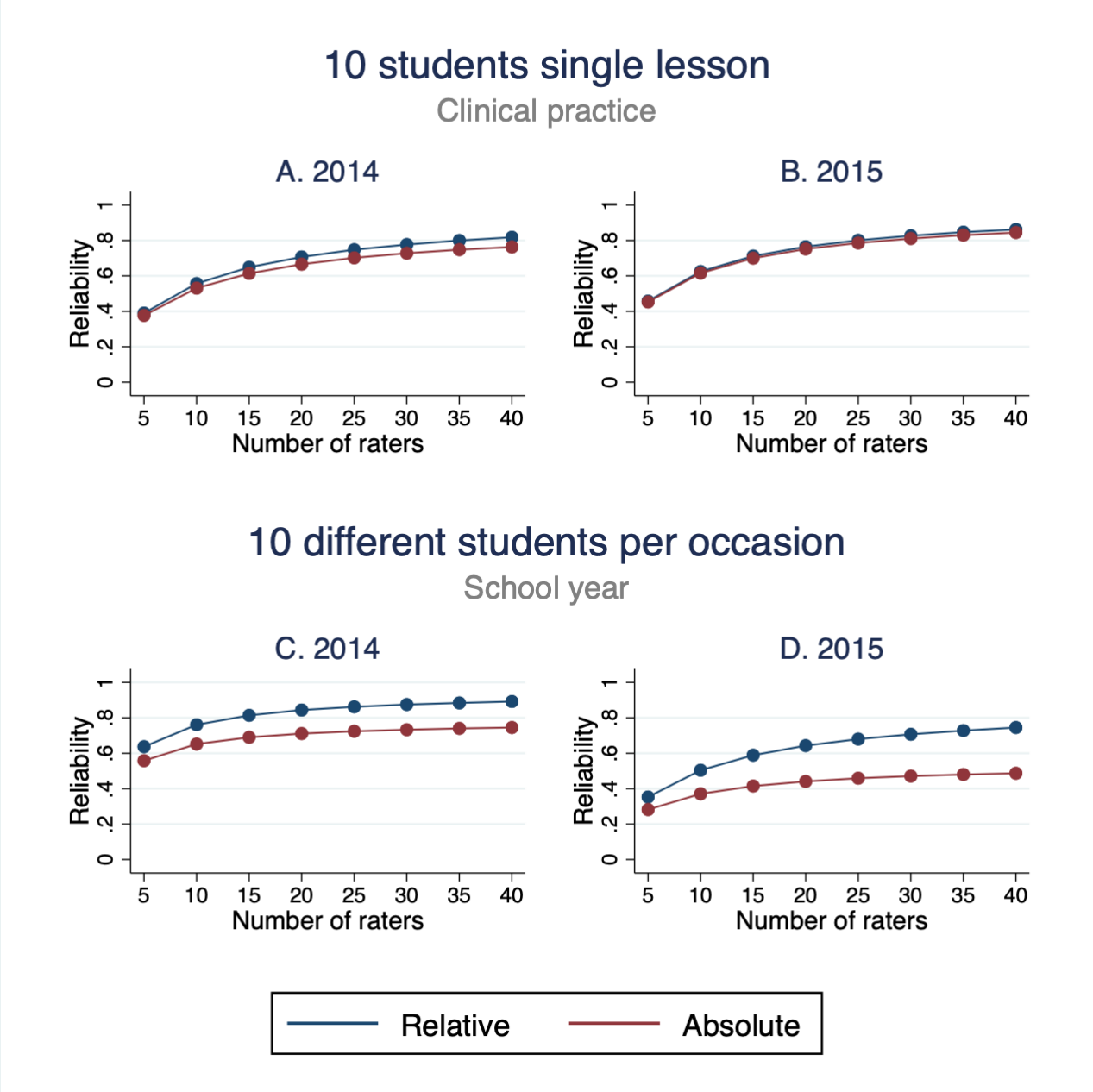
World Bank (2024c). Poverty and inequality platform: Argentina country profile. Washington, DC: The World Bank. `https://pip.worldbank.org/country-profiles/ARG`.

Figura 1: Confiabilidad de las observaciones de aula en diferentes números de lecciones, práctica clínica y año escolar, 2014 y 2015



Notas: Esta figura muestra cómo cambiaría la confiabilidad de las observaciones de aula al aumentar el número de lecciones. Incluye todos los diseños de la tabla 2. La línea azul se refiere a la confiabilidad de la posición relativa de docentes y la roja a la de las puntuaciones absolutas de docentes.

Figura 2: Fiabilidad de encuestas a estudiantes en diferentes números de evaluadores, práctica clínica y año escolar, 2014 y 2015



Notas: Esta figura muestra cómo cambiaría la confiabilidad de las encuestas a estudiantes al aumentar el número de evaluadores. Incluye todos los diseños de la tabla 3. La línea azul se refiere a la confiabilidad de la posición relativa de docentes y la roja a la de las puntuaciones absolutas de docentes.

Cuadro 1: Muestras analíticas de datos, 2014 y 2015

Contexto	Año	Docentes	Lecciones/ ocasiones	Evaluadores por lección/ocasión	Dominios	Diseño del estudio	Media	DE
A. Observaciones de clase								
Práctica clínica	2014	30	2	Mismo entrenador ambas lecciones	6	$7[d \times (l : t)]$	2.48	0.73
	2015	37	2	Mismo entrenador ambas lecciones	6	$8[d \times (l : t)]$	2.60	0.77
	2014	25	2	Mismo par ambas lecciones	6	$d \times (l : t)$	2.95	0.78
	2015	43	2	Mismo par ambas lecciones	6	$d \times (l : t)$	3.59	0.73
	2014	25	2	Diferente par por lección	6	$d \times (l : t)$	2.92	0.70
	2015	20	2	Diferente par por lección	6	$d \times (l : t)$	3.49	0.73
Año escolar	2014	48	2	Mismo entrenador ambas ocasiones	6	$3(t \times d \times o)$	3.01	0.77
	2015	35	2	Mismo entrenador ambas ocasiones	6	$4(t \times d \times o)$	2.96	0.72
B. Encuestas de estudiantes								
Práctica clínica	2014	23	1	10 estudiantes	7	$d \times (r : t)$	4.47	0.75
	2015	31	1	10 estudiantes	7	$d \times (r : t)$	4.44	0.88
Año escolar	2014	33	2	10 estudiantes diferentes por ocasión	7	$2[d \times (r : t)]$	3.75	0.94
	2015	28	2	10 estudiantes diferentes por ocasión	7	$2[d \times (r : t)]$	3.86	0.88

Notas: Esta tabla enumera el número de docentes, lecciones o ocasiones, evaluadores por lección, dominios y diseño de estudio de las observaciones de aula y encuestas a estudiantes durante práctica clínica y el año escolar 2014 y 2015. También muestra la desviación media y estándar de las puntuaciones .[el]ementales"(es decir, en el nivel docente-por-lección-por-evaluador-por-item).

Cuadro 2: Variación de las puntuaciones de dominio en las observaciones de aula, práctica clínica y año escolar, 2014 y 2015

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Práctica clínica						Año escolar	
	El mismo entrenador en ambas lecciones		El mismo par en ambas lecciones		Diferentes pares por lección		El mismo entrenador en ambas ocasiones	
	2014	2015	2014	2015	2014	2015	2014	2015
Componente de varianza	Var.	Var.	Var.	Var.	Var.	Var.	Var.	Var.
Docente	.075	.049	.258	.226	.093	.131	.11	.076
Dominio	.117	.245	.1	.075	.093	.079	.065	.08
Lección : Docente	.089	.046	.097	.084	.201	.087		
Ocasión							0	.09
Dominio × Docente	.06	.015	.03	.024	0	.079	.033	.015
Ocasión × Docente							.088	.032
Dominio × Ocasión							.032	.026
Residual	.145	.172	.203	.188	.232	.188	.231	.251
SD del efecto docente	.274	.221	.508	.475	.305	.362	.332	.276
EEM de una observación	.258	.2	.265	.248	.346	.269	.262	.199
Confiabilidad de una sola observación								
Posición relativa de docentes	.53	.55	.79	.79	.44	.64	.62	.66
Puntajes absolutos de docentes	.47	.38	.75	.75	.41	.61	.57	.43
Número de docentes	30	37	25	43	25	20	48	35

Notas: Esta tabla muestra la varianza en las notas de las aulas por contexto y año. Todas las columnas muestran los componentes de la varianza. La desviación estándar del efecto docente es la raíz cuadrada de la varianza del universo-score. El error estándar de medición de una sola observación es la raíz cuadrada de la varianza relativa de error. Los componentes estimados como negativos se fijaron en cero. Los componentes dejados en blanco para un diseño no se estimaron para ese diseño.

Cuadro 3: Variación de las puntuaciones de dominio en encuestas a estudiantes, práctica clínica y año escolar, 2014 y 2015

	(1)	(2)	(3)	(4)
	Práctica clínica		Año escolar	
	Los mismos 10 estudiantes		10 estudiantes diferentes por ocasión	
	2014	2015	2014	2015
Componente de varianza	Var.	Var.	Var.	Var.
Docente	.025	.116	.113	.04
Dominio	.017	.012	.166	.156
Evaluador : Docente	.223	.458	.232	.218
Dominio × Docente	.014	.009	.036	.034
Residual	.265	.246	.409	.37
SD del efecto docente	.158	.341	.336	.2
EEM de una observación	.168	.225	.185	.179
Confiabilidad de una sola observación				
Posición relativa de docentes	.47	.7	.77	.56
Puntajes absolutos de docentes	.45	.69	.66	.42
Número de docentes	23	31	33	25

Notas: La tabla muestra la varianza en el dominio marca en cuatro administraciones de encuestas a estudiantes: dos bajo práctica clínica y dos durante el año escolar, en 2014 y 2015. Todas las columnas muestran los componentes de la varianza para el objeto de medición (varia de punto) y cada faceta de error. La desviación estándar del efecto docente es dada por la raíz cuadrada de la verdadera diferencia de puntuación, y corresponde a la distribución de puntuaciones medias por docente. El error estándar de medición de una sola encuesta se da por la raíz cuadrada de la varianza relativa de error. El componente de varianzas estimado como negativo se ha fijado en cero. Componente de varianzas dejado en blanco para un diseño no se estimó para ese diseño.

Apéndice A Otras cifras y cuadros

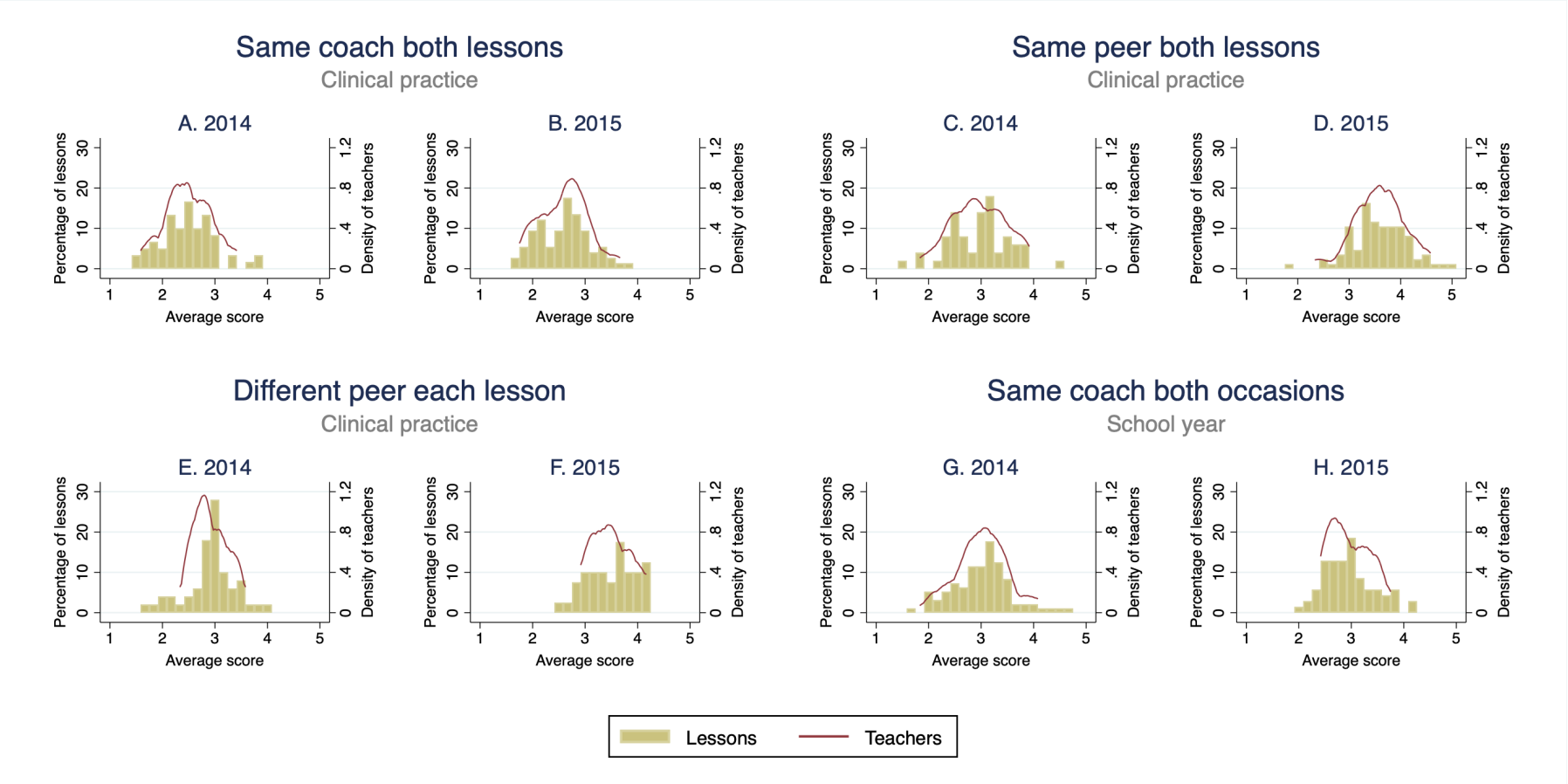
Cuadro A.1: Estudios de generalización de las observaciones de aula

Estudio	Contexto	Instrumento	Docentes	Obs. por docente	Media puntuación	SD del docente efecto	SEM de una sola obs.	Confiabilidad de una sola obs.
A. Preprimaria								
Mantzicopoulos et al. (2018)	Midwestern U.S.	CLASE K-3-EMSUP	10	4	4.75/7	0.43	0.23	0.78
		CLASE K-3-CLORG			5.19/7	0.23	0.24	0.47
		CLASE K-3-INSUP			3.04/7	0.41	0.34	0.61
		FfT Medio ambiente de clase			2.36/4	0.26	0.18	0.68
		FfT Instrucciones de clase			1.87/4	0.19	0.22	0.44
Mantzicopoulos et al. (2018)	Midwestern U.S.	MQI-R	20	5		0.17	0.10	0.80
		MQI-WWSM				0.13	0.10	0.60
		MQI-EI				0.00	0.10	0.10
		MQI-CCASP				0.09	0.10	0.58
		MQI-CWCM				0.13	0.14	0.82
		Toda la lección				0.29	0.14	0.70
Patrick et al. (2020)	Indiana, IN	FfT Reading	20	10	2.47/4	0.37	0.09	0.94
		FfT Math			2.37/4	0.35	0.10	0.93
B. Primaria								
Meyer et al. (2011)	Sudeste de Estados Unidos.	CLASE-EMSUP	118	4	5.37/7	0.52	0.39	0.64
		CLASE-INSUP			2.88/7	0.22	0.43	0.20
		CLASS-CLORG			5.19/7	0.36	0.41	0.43
Gage et al. (2018)	Sudeste de Estados Unidos.	CMS Alabado	11		0.28/1	0.13	0.08	0.19
		CMS BSP			0.61/1	0.45	0.25	0.45
		CMS OTR			2.57/7	0.32	0.42	0.20
		CMS PE			0.26/1	0.10	0.05	0.19
C. Enseñanza secundaria								
Hill et al. (2012)	Southwestern U.S.	MQI-R	24	1				0.45
		MQI-EI						0.37
		MQI-SPMMR						0.46
Kane and Staiger (2012)	Charlotte-Mecklenburg, NC;	FfT	1333	4		0.29	0.38	0.37
	Dallas, TX; Denver, CO;	CLASE						0.31
	Hillsborough Co., FL;	PLATO						0.34
	Nueva York, NY;	MQI						0.14
	Memphis, TN	UTOP	1000					0.30
Mashburn et al. (2014)	Sudeste de Estados Unidos.	CLASE-EMSUP	47	3	4.11/7	0.46	0.32	0.67
		CLASE-INSUP			3.21/7	0.43	0.39	0.54

		CLASS-CLORG			5.18/7	0.65	0.34	0.78
Mashburn et al. (2014)	Brooklyn y Queens, NY	CLASE-EMSUP	48	6		0.48		0.48
		CLASE-INSUP				0.49		0.51
		CLASS-CLORG				0.56		0.44
Praetorius et al. (2014)	Alemania y Suiza	CLASE Gestión de las aulas	38	5	3.64/7	0.10		0.92
		Apoyo al aprendizaje personal			2.60/7	0.00		0.94
		Activación cognitiva			1.93/7	0.02		0.63
D. Múltiples niveles								
Ho and Kane (2013)	Hillsborough Co., FL	FfT	67	46	2.58/4	0.27	0.34	0.39
van der Lans et al. (2016)	Países Bajos	ICALT3	69	3		1.14		0.51

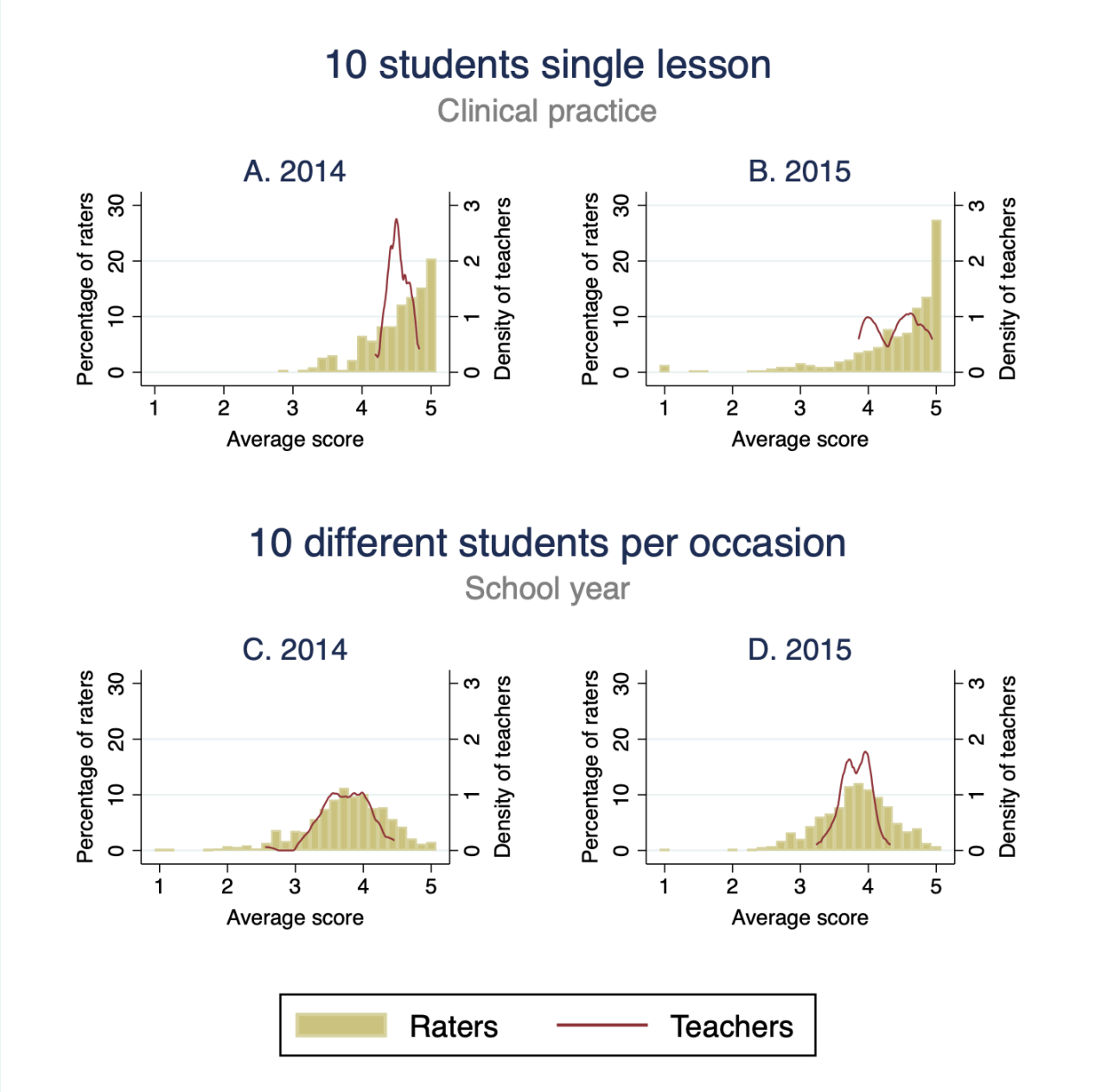
Notas: Esta tabla ofrece una visión general de los estudios anteriores sobre la fiabilidad de las observaciones de aula y encuesta a los estudiantes. La desviación estándar (SD) del efecto docente es la raíz cuadrada de la varianza de verdadero núcleo. El error estándar de medición (SEM) de una sola observación es la raíz cuadrada de la varianza relativa-error. Las células que quedan en blanco se refieren a valores no reportados. CLASS representa el sistema de evaluación de aulas. EMSUP, INSUP y CLORG son sus dominios: Apoyo emocional, apoyo instructivo y organización aula. MQI representa la calidad matemática de la instrucción. R, EI, CCASP, WWSM, CWCM y SPMMR son sus dominios: Richness, Errores e Imprecisión, Prácticas Estudiantiles Básicas Comunes, Trabajando con Estudiante y Matemáticas, el trabajo de aula se conecta a las Matemáticas y la participación de los estudiantes en la creación y la razón. FfT significa Marco para la enseñanza, PLATO para el Protocolo para la Observación de la Enseñanza de las Artes Lingüísticas, UTOP para el Protocolo de Observación de UTeach, ICALT3 para el Análisis Comparativo Internacional del Aprendizaje y la Enseñanza. NSSE representa la Encuesta Nacional de Participación Estudiantil. Mantzicopoulos et al. (2018) reporta un coeficiente de fiabilidad para cinco observaciones. In Ho and Kane (2013), cada docente fue observado en promedio 46 observaciones por docente por diferentes observadores y lecciones.

Figura A.1: Distribución de las puntuaciones medias de nivel de instrucción/ocasión en las observaciones de aula (2014 y 2015)



Notas: Esta cifra muestra la distribución de niveles de lección o ocasión (histograma) y niveles docentes (planeta de canal) puntuación media en las observaciones de aula de ExA docentes durante la práctica clínica y el año escolar 2014 y 2015.

Figura A.2: Distribución de las puntuaciones medias de nivel evaluador y docente en encuestas a estudiantes (2014 y 2015)



Notas: Esta cifra muestra la distribución del nivel evaluador (histograma) y las puntuaciones medias de nivel docente en encuestas a estudiantes de ExA docentes durante la práctica clínica y el año escolar en 2014 y 2015.

Figura A.3: Distribución de las puntuaciones de dominio sobre las observaciones de aula (2014 y 2015)

Notas: Esta figura muestra la distribución de puntajes de dominio sobre las observaciones de aula de ExA docentes durante la práctica clínica y el año escolar en 2014 y 2015. Los seis dominios son: presentar el contenido claramente, controlar la comprensión, gestionar el comportamiento de los estudiantes, implementar procedimientos de clase, crear un entorno de aprendizaje y desarrollar un sentido de posibilidad (ver sección 3.4.1).

Figura A.4: Distribución de puntajes de dominio en encuestas a estudiantes (2015)

Notas: Esta cifra muestra la distribución de puntajes de dominio en encuestas a estudiantes de ExA docentes durante la práctica clínica y el año escolar en 2014 y 2015. Los siete dominios son: cuidado, conferir, cautivar, aclarar, consolidar, desafiar y controlar (ver sección 3.4.2).

Cuadro A.2: Correlación entre las puntuaciones de dominio en observaciones de aula (2014)

	Práctica clínica						Año escolar					
	Presentación del contenido claramente	Comprobación de comprensión	Gestión del comportamiento estudiantil	Implementación de procedimientos de clase	Creación de entornos de aprendizaje	Desarrollo del sentido de la posibilidad	Presentación del contenido claramente	Comprobación de comprensión	Gestión del comportamiento estudiantil	Implementación de procedimientos de clase	Creación de entornos de aprendizaje	Desarrollo del sentido de la posibilidad
A. Práctica clínica												
Presentación del contenido claramente	1.00											
Comprobación de comprensión	0.79***	1.00										
Gestión del comportamiento estudiantil	0.52***	0.39**	1.00									
Implementación de procedimientos de clase	0.52***	0.40**	0.59***	1.00								
Creación de entornos de aprendizaje	0.54***	0.38*	0.57***	0.67***	1.00							
Desarrollo del sentido de la posibilidad	0.69***	0.60***	0.42**	0.49**	0.65***	1.00						
B. Año escolar												
Presentación del contenido claramente	0.18	0.30	0.02	0.23	0.21	0.28	1.00					
Comprobación de comprensión	0.22	0.32	0.10	0.00	0.19	0.22	0.21	1.00				
Gestión del comportamiento estudiantil	0.27	0.39*	-0.03	0.12	0.24	0.35*	0.53***	0.50***	1.00			
Implementación de procedimientos de clase	0.36*	0.32	-0.15	-0.01	0.06	0.25	0.25	0.58***	0.67***	1.00		
Creación de entornos de aprendizaje	0.25	0.35*	0.02	0.12	0.21	0.40**	0.28	0.68***	0.53***	0.64***	1.00	
Desarrollo del sentido de la posibilidad	0.08	0.15	0.11	0.13	0.21	0.25	0.35*	0.47**	0.44**	0.50***	0.73***	1.00

Notas: Esta tabla muestra los coeficientes de correlación entre las puntuaciones de dominio sobre observaciones de aula de ExA docentes durante práctica clínica y el año escolar en 2014. Esta tabla incluye sólo a los docentes con ambos conjuntos de puntajes. * significativo a 10 %; ** significativo a 5 %*** significativa a 1 %.

Cuadro A.3: Correlación entre las puntuaciones de dominio en observaciones de aula (2015)

	Práctica clínica						Año escolar					
	Presentación del contenido claramente	Comprobación de comprensión	Gestión del comportamiento estudiantil	Implementación de procedimientos de clase	Creación de entornos de aprendizaje	Desarrollo del sentido de la posibilidad	Presentación del contenido claramente	Comprobación de comprensión	Gestión del comportamiento estudiantil	Implementación de procedimientos de clase	Creación de entornos de aprendizaje	Desarrollo del sentido de la posibilidad
A. Práctica clínica												
Presentación del contenido claramente	1.00											
Comprobación de comprensión	0.57**	1.00										
Gestión del comportamiento estudiantil	0.65***	0.65***	1.00									
Implementación de procedimientos de clase	0.27	0.13	0.64***	1.00								
Creación de entornos de aprendizaje	0.60***	0.66***	0.63***	0.39*	1.00							
Desarrollo del sentido de la posibilidad	0.61***	0.57**	0.74***	0.65***	0.72***	1.00						
B. Año escolar												
Presentación del contenido claramente	0.01	-0.14	0.09	0.20	-0.07	-0.02	1.00					
Comprobación de comprensión	-0.28	-0.01	-0.20	-0.33	-0.25	-0.54**	0.21	1.00				
Gestión del comportamiento estudiantil	0.21	0.39	0.25	-0.06	0.25	-0.10	0.50**	0.57**	1.00			
Implementación de procedimientos de clase	-0.28	-0.18	-0.33	-0.49**	-0.38	-0.50**	0.25	0.75***	0.34	1.00		
Creación de entornos de aprendizaje	0.15	0.13	0.27	0.33	0.14	-0.08	0.48**	0.59***	0.70***	0.18	1.00	
Desarrollo del sentido de la posibilidad	-0.11	0.07	-0.29	-0.47**	0.01	-0.38	0.16	0.68***	0.55**	0.58***	0.36	1.00

Notas: Esta tabla muestra los coeficientes de correlación entre las puntuaciones de dominio sobre observaciones de aula de ExA docentes durante práctica clínica y el año escolar en 2015. Esta tabla incluye sólo a los docentes con ambos conjuntos de puntajes. * significativo a 10 %; ** significativo a 5 %*** significativa a 1 %.

Cuadro A.4: Correlación entre puntuaciones de dominio en encuestas a estudiantes (2014)

	Práctica clínica							Año escolar						
	Atención	Conferir	Cautivar	Aclarar	Consolidación	Desafío	Control	Atención	Conferir	Cautivar	Aclarar	Consolidación	Desafío	Control
A. Práctica clínica														
Atención	1.00													
Conferir	0.72***	1.00												
Cautivar	0.71***	0.88***	1.00											
Aclarar	0.45	0.17	0.29	1.00										
Consolidación	0.21	-0.05	0.04	-0.06	1.00									
Desafío	0.64**	0.69***	0.73***	0.10	0.38	1.00								
Control	0.40	0.60**	0.69***	-0.08	0.00	0.59**	1.00							
B. Año escolar														
Atención	0.39	0.46	0.70***	0.22	0.08	0.48*	0.58**	1.00						
Conferir	0.39	0.51*	0.72***	0.38	-0.21	0.45	0.66**	0.75***	1.00					
Cautivar	0.28	0.59**	0.73***	0.21	-0.08	0.41	0.65**	0.84***	0.82***	1.00				
Aclarar	0.44	0.66**	0.82***	0.11	0.10	0.56**	0.66**	0.92***	0.77***	0.91***	1.00			
Consolidación	0.46	0.57**	0.66**	0.49*	-0.12	0.43	0.33	0.64**	0.74***	0.79***	0.68**	1.00		
Desafío	0.17	0.44	0.66**	0.19	-0.30	0.13	0.63**	0.78***	0.76***	0.88***	0.79***	0.60**	1.00	
Control	0.37	0.64**	0.74***	-0.00	0.16	0.61**	0.69***	0.88***	0.66**	0.90***	0.96***	0.61**	0.71***	1.00

Notas: Esta tabla muestra los coeficientes de correlación entre las puntuaciones de dominio en encuestas a estudiantes de ExA docentes durante práctica clínica y el año escolar en 2014. Esta tabla incluye sólo a los docentes con ambos conjuntos de puntajes. * significativo a 10 %; ** significativo a 5 %*** significativa a 1 %.

Cuadro A.5: Correlación entre puntuaciones de dominio en encuestas a estudiantes (2015)

	Práctica clínica							Año escolar						
	Atención	Conferir	Cautivar	Aclarar	Consolidación	Desafío	Control	Atención	Conferir	Cautivar	Aclarar	Consolidación	Desafío	Control
A. Práctica clínica														
Atención	1.00													
Conferir	0.73	1.00												
Cautivar	0.93***	0.78*	1.00											
Aclarar	0.74*	0.56	0.88**	1.00										
Consolidación	0.77*	0.49	0.90**	0.96***	1.00									
Desafío	0.81*	0.51	0.88**	0.96***	0.94***	1.00								
Control	0.92**	0.59	0.93***	0.90**	0.91**	0.97***	1.00							
B. Año escolar														
Atención	-0.12	-0.14	-0.25	-0.56	-0.37	-0.59	-0.43	1.00						
Conferir	-0.22	-0.08	-0.39	-0.78*	-0.69	-0.67	-0.48	0.66	1.00					
Cautivar	0.02	-0.12	-0.16	-0.45	-0.27	-0.47	-0.30	0.97***	0.57	1.00				
Aclarar	0.36	0.45	0.19	-0.15	-0.08	-0.21	-0.04	0.75*	0.44	0.79*	1.00			
Consolidación	0.03	-0.23	-0.08	-0.30	-0.07	-0.32	-0.19	0.91**	0.39	0.94***	0.60	1.00		
Desafío	-0.23	-0.16	-0.30	-0.56	-0.39	-0.64	-0.51	0.98***	0.59	0.93***	0.70	0.89**	1.00	
Control	-0.71	-0.36	-0.56	-0.53	-0.46	-0.71	-0.76*	0.53	0.22	0.39	0.13	0.43	0.68	1.00

Notas: Esta tabla muestra los coeficientes de correlación entre las puntuaciones de dominio en encuestas a estudiantes de ExA docentes durante práctica clínica y el año escolar en 2015. Esta tabla incluye sólo a los docentes con ambos conjuntos de puntajes. * significativo a 10 %; ** significativo a 5 %*** significativa a 1 %.

Apéndice B Instrumentos

B.1 Observación aula

ExA desarrolló su protocolo de observación de aulas basado en medidas previas y la utilizó para proporcionar retroalimentación a sus docentes durante la práctica clínica y el año escolar. Cubrió seis dominios: presentar el contenido claramente, controlar la comprensión, gestionar el comportamiento de los estudiantes, implementar procedimientos de clase, crear un entorno propicio para el aprendizaje y desarrollar un sentido de posibilidad. Cada dominio incluía cinco a siete elementos. Cada artículo fue marcado de 1 ('pre-novice") a 5 ("exemplary"). Cada puntuación de los elementos posibles contenía una breve descripción para ayudar a los evaluadores a elegir entre ellos. El protocolo se puede acceder a: `https://bit.ly/3NhS7nv`.

La presentación del contenido incluía claramente siete elementos: a) ¿el maestro docente es el material?; b) ¿anuncian qué estudiantes aprenderán al comienzo de la clase?; c) ¿Usan el lenguaje corporal adecuado?; d) ¿Su explicación sigue una estructura clara?; e) ¿hacen un uso efectivo de ayudas visuales?; f) ¿mantienen un ritmo adecuado?; g) terminan la clase revisando conceptos clave o lecciones aprendidas? Por ejemplo, para el punto a), las descripciones fueron: 1) pre-novicio: no, cometen errores de contenido en su explicación y respuestas a preguntas de los estudiantes; 2) novicio: no, su presentación es correcta pero demasiado elemental y no pueden responder preguntas básicas; 3) intermedio: más o menos, su presentación es correcta pero no pueden responder preguntas avanzadas; 4) avanzado: sí, su presentación es correcta y completa y pueden responder a la mayoría de preguntas; 5) ejemplar: sí, su presentación es correcta, completa, y pueden responder a todas las preguntas.

Comprobación para la comprensión incluye siete artículos: a) ¿Hacen preguntas a los estudiantes docentes para comprobar su comprensión?; b) ¿Las preguntas abarcan una amplia gama de habilidades?; c) ¿todos los estudiantes participan en las preguntas?; d) ¿Ofrecen los comentarios docentes sobre las respuestas de los estudiantes?; e) alientan a los estudiantes a hablar entre sí?; f) responden a respuestas incorrectas ayudando a los estudiantes a mejorar sus respuestas?; y g) ¿consiguieron que no son claros? Por ejemplo, para el punto a), las descripciones fueron: 1) prenovicio: no, el docente habla durante toda la lección; 2) novicio: no, la clase incluye una conferencia y una actividad, pero no hay interacciones entre estudiantes y doctores; 3) intermedio: más o menos, el docente hace sólo unas pocas preguntas; 4) avanzado: sí, hacen preguntas en varios momentos de la lección; y 5) ejemplar: sí, incorporan preguntas a lo largo de la lección.

Administrar el comportamiento de los estudiantes incluyó siete artículos: a) ¿el docente establece reglas para el comportamiento?; b) ¿se aplican estas reglas constantemente?; c) minimizan el tiempo gastado en temas disciplinarios?; d) ¿Hay recompensas y consecuencias cuando los estudiantes siguen las reglas?; e) ¿son tales recompensas y consecuencias acordes a

las reglas que se aplican?; f) ¿son los docentes respetuosos con los estudiantes cuando aplican las reglas?; y g) determinan dónde deben estar los estudiantes Por ejemplo, para el punto a), las descripciones fueron: 1) pre-novicio: no, no hay signos en el aula y el docente nunca se alude a las reglas; 2) novicio: no, hay signos en el aula, pero el docente nunca se refiere a ellos; 3) intermedio: más o menos, hay signos, pero el docente se refiere a ellos selectivamente; 4) avanzado: sí, hay signos y el docente se refiere a ellos consistentemente; 5) ejemplar: sí, hay signos y el docente y los estudiantes se refieren consistentemente.

La implementación de los procedimientos de clase incluye cinco elementos: a) ¿tienen las rutinas establecidas para los procedimientos de clase?; b) implementan estas rutinas consistentemente?; c) minimizan el tiempo dedicado a los procedimientos de clase?; d) ¿Hay consecuencias claras para el incumplimiento de las rutinas establecidas?; y e) ¿tiene el docente un sistema para abordar circunstancias excepcionales? Por ejemplo, para el punto a), las descripciones fueron: 1) pre-novicio: no, no hay signos en el aula y el docente nunca se alude a las rutinas; 2) novicio: no, hay signos en el aula, pero el docente nunca se refiere a ellos; 3) intermedio: más o menos, hay signos en el aula, pero el docente se refiere selectivamente a ellos; 4) avanzado: sí, hay signos en el aula y el docente se refiere consistentemente a ellos; 5) ejemplar: sí, y hay signos de enseñanza.

Crear un entorno propicio para el aprendizaje incluye seis elementos: a) ¿es el docente respetuoso con los estudiantes?; b) ¿se aseguran que los estudiantes se respeten mutuamente?; c) se aseguran de que los estudiantes se sientan cómodos para hacer preguntas?; d) ¿se aseguran de que los estudiantes se sientan cómodos para compartir errores en el trabajo doméstico o en el aula?; e) ¿las reglas facilitan un entorno de aprendizaje?; f) ¿Los docente transmiten los objetivos de aprendizaje? Por ejemplo, para el punto a), las descripciones fueron: 1) pre-novicio: no, son hostiles y ofensivos hacia los estudiantes; 2) novicio: no, no son hostiles/ofensivos, pero hacen comentarios en mal gusto; 3) intermedio: más o menos, no son hostiles/ofensivos y sus comentarios no son de mal gusto, pero tratan a los estudiantes desigualmente; 4) avanzado: sí, son respetuosos y tratan a todos los estudiantes por igual; 5) ejemplar: sí, son respetuosos y tratan a todos los estudiantes genuinas

Desarrollar un sentido de posibilidad incluye siete elementos: a) ¿reconoce la fuerza y las mejoras de los estudiantes?; b) ¿muestran los procedimientos apropiados para resolver problemas en actividades, deberes o evaluaciones?; c) aconsejan a los estudiantes sobre cómo estudiar?; d) ¿proporcionan actividades modelo, deberes o evaluaciones?; e) transmiten la relevancia del contenido que se está enseñando?; f) ¿conocen la importancia de hacer bien en la escuela? Por ejemplo, para el punto a), las descripciones fueron: 1) pre-novicio: no, no felicitan a los estudiantes por realizar bien en actividades, deberes o evaluaciones; 2) novicio: no, ellos elogian a los estudiantes en general, pero no indican qué estudiantes hicieron bien o mejoraron; 3) intermedio: más o menos, comentan sobre el rendimiento de los estudiantes en términos

generales; 4) avanzado: sí, comentan sobre el desempeño o las mejoras de los estudiantes individuales; y 5) ejemplar: sí, comentan sobre el rendimiento o las mejoras específicas.

B.2 Encuestas de estudiantes

ExA adaptó y tradujo la encuesta Tripod (Ferguson, 2010, 2012) para proporcionar información a los docentes durante la práctica clínica y el año escolar. Cubrió siete dominios: cuidado (atención a las necesidades de los estudiantes), conferir (aprendizaje de los estudiantes en conversaciones), cautivar (interés de los estudiantes), aclarar (verificación de la comprensión de los estudiantes), consolidar (ayudar a los estudiantes a integrar conceptos), desafiar (establecer altos estándares para los estudiantes) y controlar (manejar el comportamiento de los estudiantes). Cada ítem fue calificado de 1 ("nunca") a 5 ("siempre"). Las encuestas son: `https://bit.ly/4dvIwV6` (primaria) y `https://bit.ly/3BGRNMw` (secundaria).

Cuidado incluye seis artículos: a) Me gusta la forma en que mi docente me trata cuando necesito ayuda; b) mi docente me hace sentir que realmente se preocupa por mí; c) si estoy triste o enojado, mi docente me ayuda a sentirme mejor; d) mi docente me anima a hacer mi mejor esfuerzo; e) mi docente sabe si algo me molesta; y f) mi docente nos da tiempo para explicar nuestras ideas.

Conferir incluye siete ítems: a) cuando nos están enseñando, mi docente nos pregunta si entendemos; b) mi docente hace preguntas para estar seguros de que estamos siguiendo lo que están diciendo; c) mi docente verifica que entendamos lo que nos está enseñando; d) mi docente nos dice qué estamos aprendiendo y por qué; e) mi docente quiere que compartamos nuestros pensamientos; f) los estudiantes hablan y comparten sus ideas sobre el trabajo de clase;

Cautivar incluye dos ítems: a) el trabajo escolar es interesante; y b) la tarea me ayuda a aprender.

Aclarar incluyó siete ítems: a) mi docente explica las cosas de manera muy ordenada; b) en esta clase, aprendemos a corregir nuestros errores; c) mi docente explica claramente las cosas difíciles; d) mi docente tiene varias buenas maneras de explicar cada tema que cubrimos en esta clase; e) esta clase es ordenada—todo tiene un lugar y las cosas son fáciles de encontrar; y f) si no entiendo algo, mi docente lo explica de otra manera.

Consolidar incluyó dos elementos: a) mi docente toma el tiempo para resumir lo que aprendemos cada día; y b) cuando mi docente marca mi trabajo, escriben en mis papeles para ayudarme a entender.

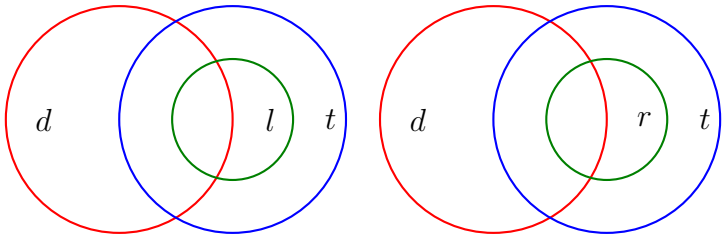
Desafiar incluyó dos ítems: a) mi docente nos empuja a pensar con profundidad sobre las cosas que hacemos; b) en esta clase, mi docente no acepta menos que nuestro máximo esfuerzo.

El control incluyó tres elementos: a) mis compañeros de clase se comportan de la manera en que mi docente quiere; b) nuestra clase se mantiene ocupada y no pierde tiempo; y c) todo el mundo sabe lo que deben hacer y aprender en esta clase.

Apéndice C Diseños de estudio

C.1 El $d \times (l : t)$ y $d \times (r : t)$ diseños

En la teoría G, la $d \times (l : t)$ y $d \times (r : t)$ diseños están representados por diagramas Venn como sigue:



$d \times (l : t)$ $d \times (r : t)$

Esta es una representación gráfica de los diseños de estudio descritos en las secciones 3.2.1 y 3.2.2. En ambos casos, el círculo para d se intersecta con todo lo demás para indicar que los dominios se cruzan con docentes y lecciones (en el primer caso) o evaluadores (en el segundo caso). Los círculos para l y r están dentro del círculo de t para indicar que lecciones y evaluadores están anidados dentro de docentes.

Prácticamente, lo que esto significa es que nuestros conjuntos de datos para cada estudio son los siguientes:

Cuadro C.1: Serie de sesiones de datos $d \times (l : t)$ diseño

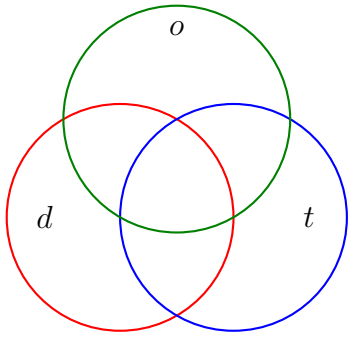
	Docente 1		Docente 2		Docente 3	
	Lección 1	Lección 2	Lección 3	Lección 4	Lección 5	Lección 6
Dominio 1	X	X	X	X	X	X
Dominio 2	X	X	X	X	X	X
Dominio 3	X	X	X	X	X	X
Dominio 4	X	X	X	X	X	X
Dominio 5	X	X	X	X	X	X
Dominio 6	X	X	X	X	X	X

C.2 El $t \times d \times o$ diseños

En la teoría G, la $t \times d \times o$ diseño está representado por el siguiente diagrama Venn:

Cuadro C.2: Serie de sesiones de datos $d \times (r : t)$ diseño

	Docente 1		Docente 2		Docente 3	
	Evaluador 1	Evaluador 2	Evaluador 3	Evaluador 4	Evaluador 5	Evaluador 6
Dominio 1	X	X	X	X	X	X
Dominio 2	X	X	X	X	X	X
Dominio 3	X	X	X	X	X	X
Dominio 4	X	X	X	X	X	X
Dominio 5	X	X	X	X	X	X
Dominio 6	X	X	X	X	X	X



En este caso, los círculos para t, d, y o intersección entre sí para indicar que este es un diseño completamente cruzado: todos los docentes son marcados todos los dominios y las ocasiones.

Esto significa que nuestros conjuntos de datos para este estudio son los siguientes:

Cuadro C.3: Serie de sesiones de datos $d \times (l : t)$ diseño

	Docente 1		Docente 2		Docente 3	
	Ocasión 1	Ocasión 2	Ocasión 1	Ocasión 2	Ocasión 1	Ocasión 2
Dominio 1	X	X	X	X	X	X
Dominio 2	X	X	X	X	X	X
Dominio 3	X	X	X	X	X	X
Dominio 4	X	X	X	X	X	X
Dominio 5	X	X	X	X	X	X
Dominio 6	X	X	X	X	X	X