DISEÑO Y ANÁLISIS DE DATOS DE DISEÑOS EXPERIMENTALES DE CASO ÚNICO

DESIGN AND DATA ANALYSIS OF SINGLE-CASE EXPERIMENTAL DESIGNS

RUMEN MANOLOV[1]

Cómo referenciar este artículo/How to reference this article:

Manolov, R. (2025). Diseño y análisis de datos de diseños experimentales de caso único [Design and data analysis of single-case experimental designs]. *Acción Psicológica, 22*(1), 7–22. https://doi.org/10.5944/ap.22.1.42833

Resumen

Los diseños experimentales de caso único implican el estudio intensivo de una o pocas unidades (e.g., personas) en diferentes condiciones manipuladas por los investigadores. Algunos diseños conllevan una replicación intra-sujeto (diseño ABAB, diseño de cambio de criterio y diseño de tratamientos alternantes), mientras que el diseño de línea base múltiple suele incluir replicación entre individuos. En ambos casos se dispone de varias oportunidades para demostrar el efecto de la intervención (introduciendo o retirándola) en diferentes momentos del tiempo. Asimismo, es imprescindible la replicación de los resultados en diferentes estudios para poder establecer la generalidad de las conclusiones. En cuanto al análisis de datos, actualmente se dispone de múltiples propuestas sin un consenso sobre cuáles son las opciones más apropiadas. Para favorecer la necesaria justificación de cualquier elección, se ofrece una serie de criterios organizativos que señalan en qué situaciones es más útil cada una de las propuestas comentadas. Asimismo, para acercar a los investigadores aplicados a las opciones analíticas, se comentan las páginas web gratuitas que las implementan. Finalmente, debido a que no es posible discutir con detalle todos los pormenores metodológicos, ni tampoco revisar todas las alternativas analíticas, el lector interesado es dirigido mediante múltiples referencias a las fuentes primarias.

Palabras clave: diseños experimentales de caso único; recomendaciones metodológicas; análisis de datos; software libre.

Abstract

Single-case experimental designs entail the intensive study of one or few entities (e.g., individuals) in different conditions, which are manipulated by the researchers. Some designs include intra-subject replication (ABAB de-

Correspondence address [Dirección para correspondencia]: Rumen Manolov, Facultad de Psicología, Universitat de Barcelona, España.
Email: rrumenov13@ub.edu

ORCID: Rumen Manolov (http://orcid.org/0000-0002-9387-1926)
[1] Universitat de Barcelona.

Recibido: 21 de enero de 2025.
Aceptado: 15 de febrero de 2025.

sign, changing criterion design, and alternating treatments design), whereas the multiple-baseline design usually includes between-subjects replication. For both scenarios, there are several attempts to demonstrate the intervention effect (introducing or withdrawing the intervention) in different moments in time. Moreover, the replication of the results in different studies is necessary for establishing the generality of the conclusions. Regarding data analysis, there are currently multiple proposals, without a consensus regarding which the optimal analytical techniques are. In order to make easier the necessary justification of any data analytical technique chosen, the current text offers a series of organizing principles, which indicate in which situation each of the options is most useful. Furthermore, in order to bring applied researchers closer to the analytical options, the text refers to freely accessible websites implementing them. Finally, given that it is not possible to discuss in detail all methodological aspects, or to review all available data analytical techniques, the interested reader is directed via multiple references to the primary sources.

Keywords: single-case experimental designs; methodological recommendations; data analysis; software.

Diseño y análisis de datos de Diseños Experimentales de Caso Único

Los Diseños Experimentales de Caso Único (DECU) constituyen una metodología de investigación que, en caso de cumplirse determinados requisitos, permite aportar evidencia científica sólida sobre la efectividad de una intervención (Horner et al., 2005). La característica principal de estos diseños es el estudio intensivo y longitudinal de una o más unidades (habitualmente personas, pero también pueden ser grupos considerados en su totalidad). A pesar de su denominación (también se les conoce como diseños de N =1 o diseños intra-sujeto), lo más habitual es que un estudio incluya a más de una persona (e.g., Tanious y Onghena, 2021, reportan que lo más habitual es que haya entre tres y siete personas en un DECU, con media y mediana cercanas a cuatro participantes). En los DECU se toman múltiples medidas obtenidas bajo diferentes condiciones de tratamiento (Tate y Perdices, 2019). Cada unidad se compara consigo misma, siendo uno de sus puntos fuertes la garantía sobre la validez interna de la investigación.

Las condiciones que se comparan suelen ser dos. En primer lugar, se dispone de una línea base que representa la situación habitual (problemática). Posteriormente, se introduce la intervención. También es posible comparar dos intervenciones. La línea base sirve no solo para describir la situación de partida, sino también poder predecir cómo seguiría la conducta de interés en caso de que la intervención no se introduzca o no sea efectiva. Por lo tanto, los DECU implican una comparación entre una predicción o proyección basada en la línea base y la realidad observada durante la intervención.

Objetivo y estructura del texto

El objetivo del artículo es ofrecer una perspectiva general de las características metodológicas de los DECU y de las posibilidades de análisis, tanto visual como cuantitativa. En los apartados siguientes, se describen los diferentes tipos de DECU con ejemplos reales de investigaciones publicadas en diferentes ámbitos. Asimismo, se mencionan los requisitos principales, los retos y las ventajas de los DECU. Posteriormente, se presentan varias alternativas de análisis datos, ofreciendo una clasificación de éstas según diferentes criterios. Dicha clasificación que podría ser útil a la hora de escoger alguna(s) de estas opciones.

Metodología DECU

Tipos principales de Diseños Experimentales de Caso Único

Los DECU se pueden distinguir entre diseños reversibles (en los cuales la intervención se puede retirar; diseño ABAB y diseño de tratamientos alternantes) y diseños irreversibles (diseño de línea base múltiple y diseño de cambio de criterio). Otra posible clasificación es en función de si el diseño permite una comparación entre series (diseño de línea base múltiple) o no (el resto de los dise-

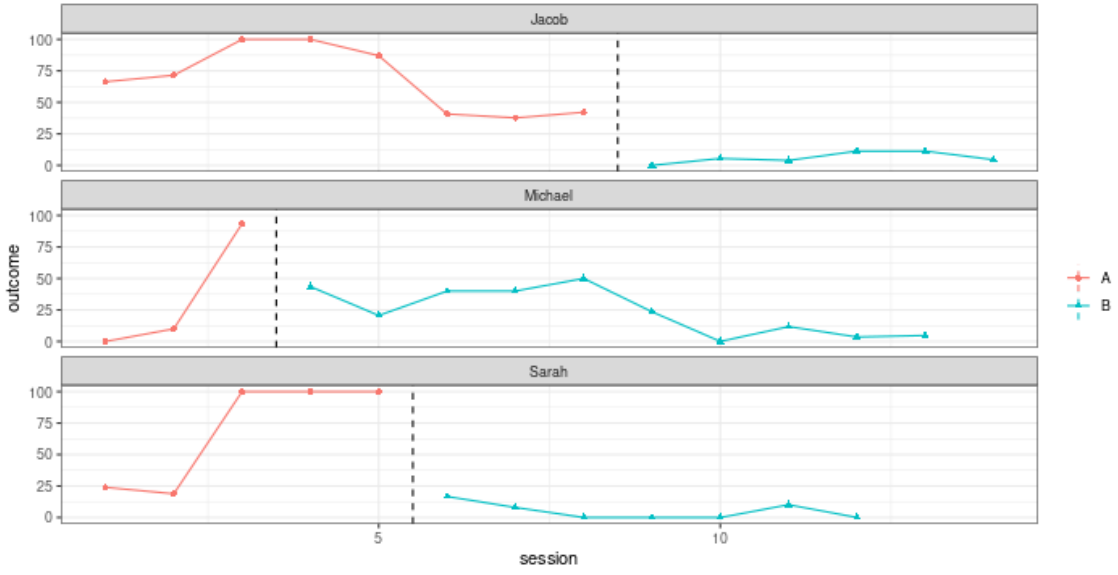
ños). A continuación, se comentan sus características principales.

El diseño más habitual es el de línea base múltiple (Tanious y Onghena, 2021), que implica la replicación de la secuencia AB (donde «A» representa la línea base y «B» la fase de intervención) en diferentes personas, conductas de la misma persona o contextos en los que se estudia la misma conducta de la misma persona. Lo fundamental es que la intervención se introduzca de forma escalonada, es decir, en diferentes momentos en el tiempo para las diferentes líneas base. En este tipo de diseños es posible que la línea base empiece en el mismo momento para todos (versión concurrente) o no (no concurrente), véase Slocum et al. (2022) para una discusión de la importancia de esta distinción en cuanto a la validez interna. La principal ventaja es que no es necesario retirar la intervención, lo que demuestra que este DECU es aplicable a situaciones en las que la intervención implica aprendizaje. Se trata de la situación probablemente más habitual en un contexto clínico, sobre todo si se trabaja desde una perspectiva cognitiva o cognitivo-conductual. Desde el punto de vista ético también parece más apropiado no retirar (ni siquiera temporalmente) una intervención que funciona. Un ejemplo de datos recogidos siguiendo un diseño de línea base múltiple puede verse en la Figura 1. Los datos provienen del estudio de Eilers y Hayes (2015) sobre el uso de ejercicios cognitivos para reducir las conductas repetitivas y restrictivas de niños con Trastorno de Espectro Autista. Es posible mejorar el diseño introduciendo aleatorización de diferentes maneras (Levin et al., 2018). Por ejemplo, se puede escoger al azar el momento en el que empieza la intervención para cada participante, entre varios posibles momentos que no se solapen entre participantes. Otra opción es determinar de antemano el momento de intervenir, pero decidir al azar qué participante comienza primero, quién segundo, etc.

El diseño ABAB o diseño de retirada o reversión (Wine et al., 2015) implica una replicación dentro del mismo participante. Se dispone de tres momentos de comparación entre fases adyacentes El diseño es aplicable

Figura 1

Ejemplo de un diseño de línea base múltiple entre personas. Datos descargados de https://osf.io/79dfs*, obtenidos originalmente por Eilers y Hayes (2015). El gráfico se ha obtenido mediante https://jepusto.shinyapps.io/scdhlm.*



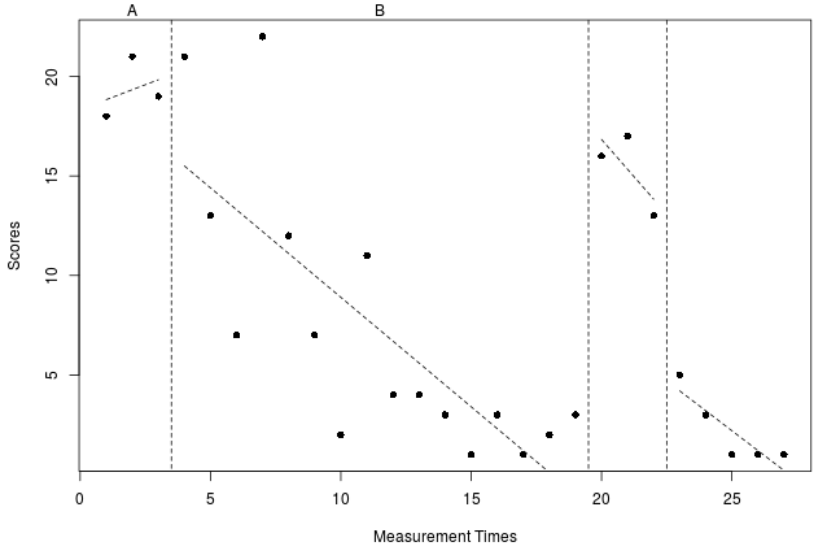
Nota. Los datos rojos son los correspondientes a la línea base (A), mientras que los datos azules son los correspondientes a la fase de intervención (B).

cuando la intervención no provoca un cambio permanente, sino que en ausencia de la intervención es posible volver al nivel inicial de la conducta de interés. En este sentido, la tercera fase es crítica para la inferencia del efecto del tratamiento, que sería posible en caso de que tenga lugar la reversión de la conducta de interés al nivel anterior al tratamiento. Un ejemplo se ofrece en la Figura 2: se trata de datos recogidos por Feeney e Ylvisaker (2006) de un participante que había padecido una lesión cerebral traumática, y a quien se aplicó una intervención cognitivo-conductual para tratar conducta desafiante. Es posible escoger al azar los tres momentos de cambio de fase, entre un listado de posibilidades que respeta un mínimo de longitud de cada fase (Onghena, 1992).

En el diseño de tratamientos alternantes, a diferencia de los dos diseños anteriores, la comparación principal no tiene lugar entre fases. La comparación fundamental se realiza entre las condiciones (habitualmente dos intervenciones diferentes) que están sujetas a una alternancia frecuente. Aparte de esta alternancia (que podría constituir una «fase de comparación»), puede haber -aunque no sea imprescindible- una fase inicial de línea base y una fase final en la que se aplica solo la mejor intervención (Barlow y Hayes, 1979). Dentro de la «fase de comparación», lo habitual es restringir el número de medidas consecutivas dentro de la misma condición a dos. Para que este diseño sea aplicable es necesario que la intervención no tenga efectos duraderos (e.g., intervención farmacológica). Se puede observar un ejemplo en la Figura 3: se trata de datos recogidos por Eilers y Hayes (2005), comparando dos intervenciones diferentes para reducir conductas problemáticas en niños diagnosticados con Trastorno de Espectro Autista. Como se puede apreciar, gráficamente se suelen juntar mediante una línea las medidas pertenecientes a la misma condición y se evalúa la distancia o separación entre estas líneas. Este tipo de inspección visual tiene paralelos en las cuantificaciones propias para los diseños de tratamientos alternantes (Manolov y Onghena, 2018). Asimismo, es posible implementar aleatorización de dife-

Figura 2

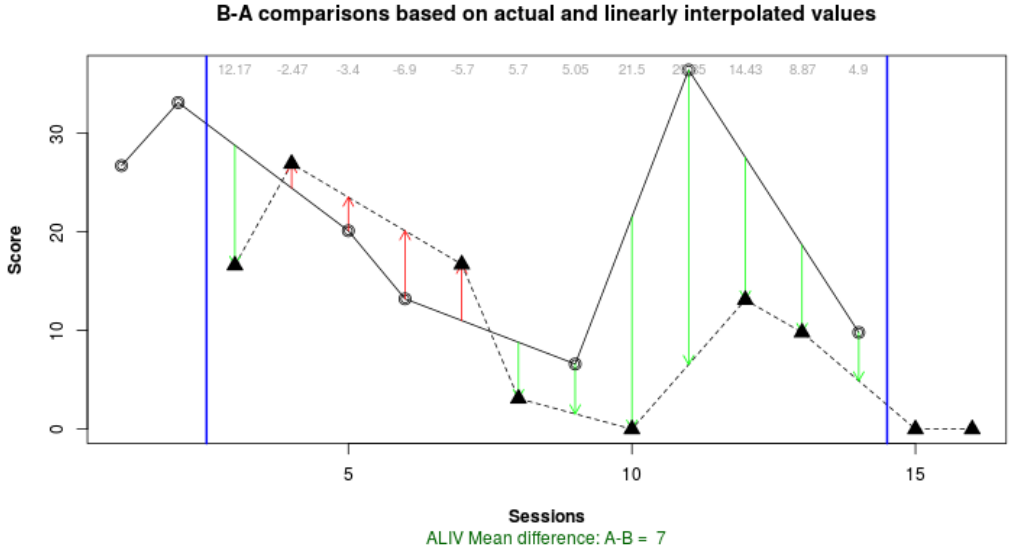
Ejemplo de un diseño ABAB. Datos descargados de https://osf.io/79dfs*, originalmente obtenidos por Feeney e Ylvisaker (2006). El gráfico se ha obtenido mediante https://tamalkd.shinyapps.io/scda/*



Nota. El primer recuadro representa la línea base inicial. El segundo recuadro es la primera introducción de la intervención. El tercer recuadro es la retirada de la intervención (i.e., la vuelta a la línea base). El cuarto recuadro es la reintroducción de la intervención.

Figura 3

Ejemplo de un diseño de línea base múltiple entre personas. Datos descargados de https://osf.io/kaphj, obtenidos originalmente por Eilers y Hayes (2015). El gráfico se ha obtenido mediante https://manolov.shinyapps.io/ATDesign.



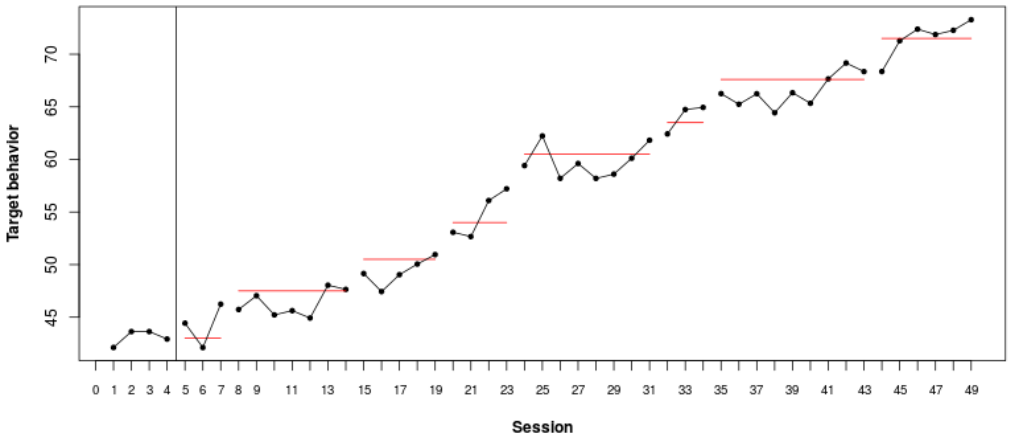
Nota. Las líneas verticales verdes muestran comparaciones entre condiciones que favorecen a la condición "B" (valores inferiores de la conducta indeseable). Las líneas verticales rojas muestran comparaciones que favorecen a la condición "A". A la izquierda de la primera línea azul y a la derecha de la segunda línea azul hay datos para los cuales es imposible comparar las líneas que juntan los puntos de las dos condiciones.

rentes maneras. Por ejemplo, una posible aleatorización implicaría seleccionar al azar qué condición tiene lugar en cada momento de medida, respetando la limitación de un número máximo (de dos, habitualmente) medidas consecutivas en la misma condición (Onghena y Edgington, 1994).

El diseño de cambio de criterio implica la consecución gradual de la meta final mediante el establecimiento, habitualmente de forma conjunta entre investigador y participante, de criterios cada vez más restrictivos (Hartmann y Hall, 1976). En concreto, en un contexto de modificación de conducta, con la consecución de cada criterio intermedio, el participante recibe reforzamiento. Posibles aplicaciones de este tipo de DECU incluirían ir asumiendo los diferentes pasos necesarios en una tarea compleja (e.g., que un niño aprenda a atarse los cordones) o conseguir niveles cada vez más adecuados de una conducta de interés (e.g., mejorar el cumplimiento de una intervención médica). Se ha resaltado importancia de variar la longitud de las subfases de intervención, la magnitud del cambio de criterio y también introducir retrocesos a criterios previos menos exigentes (Klein et al., 2017). Un ejemplo puede verse en la Figura 4, que corresponde a datos recogidos por Facon et al. (2008), quienes utilizaron procedimientos de aprendizaje operante para tratar un mutismo selectivo severo en un niño con discapacidad intelectual, intentando aumentar progresivamente los decibelios de sus verbalizaciones. Entre los desarrollos para los diseños de cambio de criterio hay que destacar el diseño con rango aceptable para la conducta dentro de cada subfase (McDougall, 2005) y la posibilidad de introducir aleatorización, por ejemplo, escogiendo al azar la longitud de las subfases de la intervención (Onghena et al., 2019).

Figura 4

Ejemplo de un diseño de cambio de criterio. Datos descargados de https://osf.io/kcjnu*, obtenidos originalmente por Facon et al. (2008). El gráfico se ha obtenido mediante https://manolov.shinyapps.io/ChangingCriterion.*



Nota. El primer recuadro representa los datos de la línea base. El segundo recuadro incluye las cuatro subfases de la fase de intervención: los criterios se marcan mediante las líneas horizontales rojas, mientras que las líneas horizontales azules son los rangos aceptables para cada subfase.

Recomendaciones metodológicas: rúbricas

Se dispone de varios instrumentos para valorar el rigor metodológico o guías para sugerir cómo proceder en estudios DECU (e.g., Perdices et al., 2023; What Works Clearinghouse, 2022). Se trata de componentes de diseño que potencien la validez interna, es decir, el grado de confianza en una relación causal entre la intervención y la conducta de interés (e.g., número ocasiones para demostrar el efecto de la intervención, número de medidas, fidelidad de implementación, aleatorización). En cuanto a la validez externa o capacidad de generalizar, se requiere informar de detalles suficientes sobre el participante, la intervención y el contexto para poder valorar en qué tipo de situaciones se puede considerar que una intervención es efectiva.

Ventajas

En los diseños clásicos de comparación de grupos, la aleatorización en la asignación de las personas los grupos experimentales permite asumir la equivalencia inicial de los grupos y cualquier diferencia posterior se atribuye al efecto del tratamiento. No obstante, este tipo de diseños suele implicar criterios de inclusión estrictos para que se puedan formar grupos homogéneos que suelen ser puros en cuanto a la problemática de interés (i.e., sin comorbilidades). Esto limita la posibilidad de generalizar los resultados a individuos con características menos típicas. Asimismo, inferir de un resultado general a un individuo puede resultar en una falacia ecológica, puesto que el promedio no representa necesariamente a ninguna persona en concreto. En contraste, los DECU tratan directamente con la unidad sobre la cual se quiere extraer una conclusión (i.e., el individuo) y permiten el estudio de poblaciones muy heterogéneas (e.g., personas diagnosticas con Trastorno del Espectro Autista).

Otra ventaja de los DECU es la posibilidad de estudiar el proceso de cambio, gracias a las múltiples medidas de las que se dispone. Asimismo, el estudio intensivo de la persona permite explicar posibles valores anómalos que puedan obtenerse en algún momento determinado.

Finalmente, la estructura básica de los DECU mimetiza la práctica profesional habitual: una fase de evaluación inicial seguida de una intervención. Por lo tanto, los profesionales pueden ejercer a la vez de investigadores y aprovechar su experiencia, publicándola y aportando datos para establecer la base científica de las intervenciones.

Retos

La limitación fundamental relacionada con los DECU es la generalización o validez externa. La manera más segura de dar pasos hacia la generalización es la replicación (Tate y Perdices, 2019). Solo añadiendo más casos y atendiendo a sus características se puede generalizar, de forma empírica y gracias a la lógica, a otras personas de características parecidas.

Un segundo reto es conseguir que el diseño sea lo más riguroso posible para potenciar la validez interna, incorporando los componentes recomendados en los instrumentos que evalúan el rigor metodológico. Por una parte, es fundamental que haya varias demostraciones del cambio en la conducta de interés concurrente con el cambio en la condición experimental, en diferentes momentos del tiempo, para descartar la «historia» (i.e., ocurrencia de eventos externos a la intervención) como posible razón de dicho cambio. Por otra parte, la aleatorización (e.g., selección al azar del momento en que cambiar de condición experimental en un diseño ABAB o del orden en que intervenir los participantes en un diseño de línea base múltiple) se ha resaltado como un elemento fundamental para la validez interna (e.g., Jacobs, 2019).

Finalmente, un tercer reto es el análisis de datos. Se dispone de múltiples opciones analíticas sin un consenso claro respecto a cuál escoger, sobre todo en relación con la posible presencia de dependencia serial, tendencia a la mejora espontánea, o la variedad de tipos de DECU. Además, la autocorrelación (o dependencia serial entre las medidas obtenidas longitudinalmente de la misma unidad) dificulta la aplicación de pruebas inferenciales clásicas. El presente texto pretende presentar una estructura para que los investigadores aplicados puedan realizar una elección y presentar una justificación con una base sólida.

Análisis de Datos DECU

Clasificación de las alternativas de análisis

Formativo o Sumativo

Una primera distinción es entre el análisis *formativo* y *sumativo* (Ledford et al., 2019). El primero forma parte de la experimentación guiada por los datos, utilizada para determinar cuándo cambiar las condiciones (i.e., decidir mientras los datos aún se están recogiendo). En cambio, el análisis sumativo sirve para documentar y comunicar el grado de efectividad de la intervención, una vez que todos los datos ya estén recogidos. El análisis formativo se lleva a cabo principalmente mediante la inspección visual de la representación gráfica de los datos (e.g., Byun et al., 2017). Las secciones siguientes se refieren a organizar opciones analíticas para análisis *sumativo*.

Objetivo de la evaluación de los datos

Uno de los objetivos de la evaluación de los datos es establecer una relación funcional o causal, es decir, valorar si se puede inferir que los cambios en la conducta objeto se deben al efecto de tratamiento. Se compara visualmente el patrón de datos esperado, según el diseño, y el obtenido. Para el mismo objetivo se podría utilizar el *p*-valor de una prueba de aleatorización, que serviría para una inferencia causal tentativa (que no una inferencia poblacional basada en supuestos y modelos; Manolov y Onghena, 2018). La idea de la prueba de aleatorización es que se escoge de antemano un estadístico de prueba, y este estadístico se calcula para todas las divisiones de datos posibles (i.e., todas las aleatorizaciones o maneras de asignar los momentos de medida a diferentes condiciones). De esta manera, el *p*-valor se obtiene directamente a partir de los datos, sin necesidad de asumir una distribución (e.g., normal) para el estadístico de prueba o para los datos (Heyvaert y Onghena, 2014).

Un segundo objetivo podría ser, en algunos casos, comparar los datos a un resultado final deseable, como cuando se utilizan criterios de maestría y niveles preestablecidos de rendimiento (McDougale et al., 2015). En este

sentido, es posible cuantificar el grado en qué se ha conseguido el objetivo, comparando el nivel deseado con el nivel conseguido (Ferron et al., 2020).

Un tercer objetivo es la cuantificación de la magnitud de la diferencia mediante la obtención tamaño del efecto, aunque puede presentar retos interpretativos, considerando que los criterios a seguir deberían ser específicos de cada ámbito de investigación (Vannest y Sallese, 2021). Las secciones siguientes se refieren a la cuantificación del tamaño de la diferencia entre condiciones.

Finalmente, se puede considerar una comparación más generalizada entre la situación antes y después del tratamiento a través del índice de cambio fiable (Estrada et al., 2019), en caso de utilizarse medidas con propiedades psicométricas conocidas.

Escala de medida y unidad de medida

En cuanto a la *escala de medida* de la variable dependiente, si ésta es ordinal, se pueden utilizar índices de no solapamiento (Parker et al., 2011). Cuando la escala de medida es de intervalo o razón, se pueden calcular diferencias en medias y tendencias. En estos casos, el investigador puede seleccionar las *unidades de medida* deseadas para la cuantificación resumen: porcentajes (e.g., el logaritmo de la razón de respuestas se puede convertir a un cambio porcentual; Pustejovsky, 2018), estandarizadas (e.g., la diferencia de medias estandarizada entre casos: Shadish et al., 2014; modelos multinivel tras estandarizar los datos), o no estandarizadas (e.g., modelos multinivel con los datos originales).

Intra-individual o entre individuos

Otro criterio para escoger el enfoque analítico es el nivel de análisis. Si el foco es obtener cuantificaciones separadas para cada individuo, se pueden utilizar medidas intra-individuales como la diferencia de media estandarizada (Busk y Serlin, 1992) y los índices de no solapamiento (Parker et al., 2011). Si el objetivo es obtener una única cuantificación general para varios individuos, se puede usar la diferencia de medias estandarizada entre individuos (Shadish et al., 2014) o modelos multinivel (Ferron et al., 2009, 2010).

Una clasificación parecida procede del tipo de diseño. Por ejemplo, un diseño ABAB implica una comparación intra-serie, mientras que un diseño de línea base múltiple permite tanto la comparación intra-serie, como entre-series (ver Ferron et al., 2014). En este último tipo de diseño, la comparación entre-series está ligada al inicio concurrente de la fase de línea base (Christ, 2007).

Técnicas analíticas y su implementación en software gratuito

Inspección Visual

En cuanto a la inspección visual, elemento analítico que suele estar presente siempre, los desarrollos se pueden organizar en seis ámbitos. Primero, se han hecho recomendaciones sobre las características deseables de los gráficos como representaciones visuales (Dart y Radley, 2018). Segundo, se han listado aspectos de los datos a considerar (Ledford et al., 2019; Maggin et al., 2018): nivel, tendencia, variabilidad, inmediatez, solapamiento y consistencia. Concretamente, dentro de cada fase, se puede valorar el nivel, la tendencia y la variabilidad. Complementariamente, a la hora de comparar fases adyacentes, se pueden identificar cambios de nivel o cambios de tendencia, además de valorar si dichos cambios son inmediatos o demorados. Otro tipo de comparación se refiere al grado en que las diferentes fases incluyen valores parecidos (i.e., grado de solapamiento). Finalmente, al considerar varias ocasiones de demostración de efecto, se puede valorar la consistencia entre fases parecidas y la consistencia del efecto (Manolov y Tanious, 2022). Tercero, se ha propuesto usar medianas, líneas de tendencia y de variabilidad y cuantificaciones de no solapamiento que acompañen a la valoración visual de estos aspectos (Lane y Gast, 2014). Cuarto, se han propuesto protocolos que sistematicen los diferentes pasos del análisis visual, aunque sin necesariamente acudir a cuantificaciones para cada uno de los aspectos de los datos. (Wolfe et al., 2019). Quinto, se han propuesto ayudas visuales en forma de líneas de tendencia central y variabilidad superpuestas (e.g., Fisher et al., 2003). Sexto, se han propuesto gráficos para un análisis conjunto de varias comparaciones entre condiciones (Manolov, Tanious et al., 2022).

Opciones intra-individuales

En la presente sección las diferentes opciones analíticas intra-individuales se organizarán según el aspecto focal de los datos: nivel (media o mediana), tendencia, variabilidad, solapamiento, inmediatez y consistencia.

Los índices de *no solapamiento* (comparados en Parker, Vannest y Davis, 2011) son un grupo de cuantificaciones que se centran en la información ordinal contenida en los datos. Específicamente, comparan datos de diferentes condiciones en cuanto a cuál de ellos es superior, sin tener en cuenta la distancia (i.e., cuán superior). Algunos índices resumen los datos de la fase de línea base mediante su mejor dato (el índice con acrónimo PND; Scruggs et al., 1987) o la mediana (el índice con acrónimo PEM; Ma, 2016), mientras que otros utilizan todos los datos sin resumirlos (el índice con acrónimo NAP; Parker & Vannest, 2009). A su vez, algunos índices no tienen en cuenta una posible tendencia hacia la mejora espontánea durante la fase de línea base (PND, PEM, NAP), mientras que otros sí controlan este tipo de tendencia (los índices Tau de Parker, Vannest, Davis y Sauber 2011). En cuanto al software, la página web https://jepusto.shinyapps.io/SCD-effect-sizes proporciona explicaciones y fórmulas, además de las cuantificaciones.

Al centrarse en el *nivel*, las cuantificaciones propuestas han sido la diferencia de medias estandarizada (Busk y Serlin, 1992) y el logaritmo de la razón de respuestas (Pustejovsky, 2018). Para diferencias de medias estandarizadas intra-individuales y una cuantificación en términos de porcentaje se puede utilizar https://jepusto.shinyapps.io/SCD-effect-sizes.

En cuanto a las opciones que incorporan la *tendencia*, se trata de propuestas basadas en modelos de regresión, por ejemplo: (a) cuantificación separada del cambio de pendiente y del cambio de nivel para la primera ocasión de la fase de intervención, conocidos en inglés como modelos "piecewise" (Center et al., 1985; Moeyaert et al., 2014); y (b) cuantificación conjunta a través del promedio de las diferencias entre la tendencia proyectada de la fase de línea base y la tendencia ajustada en la fase de intervención (Swaminathan et al., 2014). En cuanto al software, el análisis de un único nivel siguiendo un modelo "piecewise" (i.e., una comparación A-B) se puede llevar a mediante http://34.251.13.245/MultiSCED/ (Declercq et al., 2020) y https://manolov.shinyapps.io/Regression/, que también incorpora el modelo de Swaminathan et al. (2014).

En cuanto a la *inmediatez* del cambio, la propuesta inicial fue comparar la media de las tres últimas medidas de la fase de línea base con la media de las tres primeras medidas de la fase de intervención, aplicable mediante https://manolov.shinyapps.io/Overlap/. Una propuesta más reciente permite identificar el momento (o los momentos) más probable(s) en que se produjo el mayor cambio y, por lo tanto, valorar si dicho momento coincide con el momento de cambio de fase (en inglés, *Bayesian Unknown Change Point Model*, Natesan y Hedges, 2017). El código de R para esta opción está disponible en https://github.com/prathiba-stat/BUCP. Otra opción es explorar diferentes latencias, teniendo en cuenta si el tipo de efecto esperado es abrupto o progresivo (Manolov y Onghena, 2022).

Finalmente, nótese que las pruebas de aleatorización (Heyvaert y Onghena, 2014) permiten seleccionar como estadístico de prueba una cuantificación centrada en el nivel (e.g., una diferencia de medias), en la tendencia (e.g., diferencia de pendientes), en la variabilidad (e.g., razón de varianzas) o en el solapamiento. Mediante dichas pruebas se asocia un *p*-valor a estas cuantificaciones para representar probabilidad de obtener una diferencia tan grande o mayor en ausencia de efecto de la intervención.

Opciones entre individuos

La recomendación realizada por What Works Clearinghouse (2022), pero no compartida por todos (e.g., Kratochwill et al., 2021), es utilizar una diferencia de medias estandarizada que permita obtener una única cuantificación para varios participantes. Se trata de dos procedimientos diferentes. Uno no tiene la tendencia en cuenta (utilizando estimación por el método de los momentos; Shadish et al., 2014), mientras que el otro sí modela la tendencia (utilizando estimación por máxima verosimilitud restringida; Pustejovsky et al. 2014). La estandarización se consigue teniendo en cuenta tanto variabilidad intra-individual, como entre individuos, para obtener una cuanti-

ficación comparable a la que se obtiene en diseños de comparación de grupos. La autocorrelación se modela a la hora de obtener el intervalo de confianza alrededor de la estimación puntual.

Un objetivo similar se consigue mediante los modelos multinivel de dos niveles (Ferron et al., 2009), que también permiten obtener cuantificaciones separadas para cada individuo a través de estimaciones Bayesianas empíricas (Ferron et al., 2010).

En cuanto al software, para las dos versiones de la diferencia de medidas estandarizada entre casos se puede utilizar https://jepusto.shinyapps.io/scdhlm. La potencia se puede calcular mediante https://abkpowercalculator.shinyapps.io/ABkpowercalculator/.

La web http://34.251.13.245/MultiSCED/ (Declercq et al., 2020) permite llevar a cabo un análisis de dos niveles. Otra web, https://manolov.shinyapps.io/SeveralAB/, ofrece las estimaciones Bayesianas empíricas de los efectos individuales y permite modelar la autocorrelación y varianza residual heterogénea y además de ofrecer los valores de los criterios informacionales AIC y BIC que permiten la comparación entre modelos.

Una manera diferente de combinar resultados de varios casos es valorar si un efecto puede considerar exitosamente replicado (Manolov, Tanious et al., 2022) gracias a una definición a priori del nivel deseado tras la intervención y del mínimo cambio deseable. Esta opción está implementada en https://manolov.shinyapps.io/Brinley/.

Opciones entre estudios

Debido a la importancia de la replicación para generalizar conclusiones en el contexto DECU, el metaanálisis de resultados de diferentes estudios sobre la misma problemática y con la misma intervención es necesario. Recientemente, se han distinguido dos enfoques (Declercq et al., 2022): combinar tamaños del efecto (dos etapas) y combinar datos directamente (una etapa).

En cuanto al software, http://34.251.13.245/MultiSCED/ (Declercq et al., 2020) permite implementar un modelo de tres niveles (i.e., una etapa), mientras que https://manolov.shinyapps.io/Change permite un metaanálisis de dos etapas combinando tamaños del efecto.

Discusión

Recomendaciones

Planificar

Antes de recoger y analizar datos, hay que asegurarse que el estudio puede aportar evidencia científica sólida: siguiendo las recomendaciones metodológicas (e.g., Perdices et al., 2023; What Works Clearinghouse, 2022). Asimismo, hay que valorar si con los recursos disponibles o factibles (participantes y número de momentos de medida), las técnicas analíticas funcionarían adecuadamente en términos de ausencia de sesgo, eficiencia, tasa de error Tipo I y potencia estadística.

En cuanto a la planificación, en caso de que haya aleatorización en el diseño y de que se utilice una prueba de aleatorización para inferencia causal tentativa, es necesario comprobar si el número de aleatorizaciones posibles (según el diseño) permite obtener un *p*-valor igual o inferior al alfa nominal (habitualmente 0.05). El *p*-valor no puede ser más pequeño que 1 dividido entre el número de aleatorizaciones. El cálculo del número de aleatorizaciones puede obtenerse para diferentes DECU a través de la web https://tamalkd.shinyapps.io/scda.

Informar

Se recomienda seguir las guías de publicación elaboradas por un conjunto de expertos en los DECU (Tate et al., 2016). Asimismo, una justificación es necesaria para la elección de la técnica. Esta justificación puede basarse en varios criterios: (a) el problema de investigación y la posibilidad de obtener información útil; (b) el patrón de datos esperado (e.g., la presencia de mejora espontánea durante la fase de línea base; (c) la adecuación de las propiedades estadísticas (ausencia de sesgo, mayor eficiencia, mayor potencia estadística); (d) facilidad de interpretación, más allá de meramente reportar valores. Consideramos que hay

dos justificaciones que no deberían serlo: (a) facilidad de cálculo y (b) tradición (e.g., publicaciones previas).

Aparte de valorar si un efecto es visualmente claro, si es grande o estadísticamente significativo, se ha recomendado valorar la validez social (Snodgrass et al., 2023). Se trata de una aproximación a la significación práctica: el funcionamiento del individuo después de la intervención, el mantenimiento del efecto en el tiempo, la posibilidad de que la intervención se implemente por agentes típicos y con los recursos disponibles en la práctica profesional, etc.

Limitaciones y aportaciones

El objetivo de este artículo es ofrecer una amplia panorámica de las principales características de los DECU y sus diferentes tipologías. Esto se ha hecho previamente en inglés (e.g., Ledford et al., 2019; Maggin et al., 2018) y también se dispone de texto en castellano (e.g., Bono y Arnau, 2014), aunque sin incluir los últimos desarrollos a nivel de análisis de datos. Adicionalmente, también se quería ofrecer una estructura organizativa y una panorámica de las diferentes técnicas de análisis de datos disponibles. También se dispone de textos *en inglés* sobre esta temática (e.g., Maggin et al., 2019; Manolov, Moeyaert et al., 2022), pero los criterios organizativos que aquí se presentan, para guiar a la hora de escoger cómo analizar los datos, son más completos, al atender al nivel de análisis deseado, al tipo de análisis que se desea realizar, a la escala de medida de la variable de interés, y a la necesidad (o no) de considerar una tendencia en los datos.

Otra aportación del presente texto es el listado organizado de software gratuito disponible. Lo mencionado en el texto se complementa por un proyecto de *Open Science Framework* https://osf.io/t6ws6., donde se dispone de ejemplos de la manera en la que los datos han de organizarse para cada una de las webs creadas con R y Shiny.

En cuanto a las limitaciones, las restricciones de longitud del texto han impedido profundizar en los detalles técnicos de los procedimientos. Para obtener información más detallada, se invita al lector a consultar la siguiente lista de bibliografía relevante sobre DECU: https://osf.io/u9g2r.

Referencias

Barlow, D. H. y Hayes S. C. (1979). Alternating treatments design: One strategy for comparing the effects of two treatments in a single subject. *Journal of Applied Behavior Analysis, 12*(2), 199–210. https://doi.org/10.1901/jaba.1979.12-199

Bono, R. y Arnau, J. (2014). *Diseños experimentales de caso único en ciencias sociales y de la salud* [Single-case Experimental Designs in Social and health sciences]. Síntesis.

Busk, P. L. y Serlin, R. C. (1992). Meta-analysis for single-case research. En T. R. Kratochwill y J. R. Levin (Eds.), *Single-case Research Designs and Analysis: New Directions for Psychology and Education* (pp. 187−212). Lawrence Erlbaum.

Byun, T. M., Hitchcock, E. R. y Ferron, J. (2017). Masked visual analysis: Minimizing Type I error in visually guided single-case design for communication disorders. *Journal of Speech, Language, and Hearing Research, 60*(6), 1455−1466. https://doi.org/10.1044/2017_JSLHR-S-16-0344

Center, B. A., Skiba, R. J. y Casey, A. (1985). A Methodology for the Quantitative Synthesis of intra-Subject Design Research. *The Journal of Special Education, 19*(4), 387−400. https://doi.org/10.1177/002246698501900404

Christ, T. J. (2007). Experimental Control and Threats to Internal Validity of Concurrent and Nonconcurrent Multiple Baseline Designs. *Psychology in the Schools, 44*(5), 451–459. https://doi.org/10.1002/pits.20237

Dart, E. H. y Radley, K. C. (2018). Toward a standard assembly of linear graphs. *School Psychology Quarterly, 33*(3), 350−355. https://doi.org/10.1037/spq0000269

Declercq, L., Cools, W., Beretvas, S. N., Moeyaert, M., Ferron, J. M. y Van den Noortgate, W. (2020).

MultiSCED: A Tool for (Meta-)Analyzing Single-Case Experimental Data with Multilevel Modeling. *Behavior Research Methods, 52*(1), 177–192. https://doi.org/10.3758/s13428-019-01216-2

Declercq, L., Jamshidi, L., Fernández Castilla, B., Moeyaert, M., Beretvas, S. N., Ferron, J. M. y Van den Noortgate, W. (2022). Multilevel Meta-Analysis of Individual Participant Data of Single-Case Experimental Designs: One-stage versus Two-Stage Methods. *Multivariate Behavioral Research, 57*(2–3), 298–317. https://doi.org/10.1080/00273171.2020.1822148

Eilers, H. J. y Hayes, S. C. (2015). Exposure and Response Prevention Therapy with Cognitive Defusion Exercises to Reduce Repetitive and Restrictive Behaviors Displayed by Children with Autism Spectrum Disorder. *Research in Autism Spectrum Disorders, 19*, 18–31. https://doi.org/10.1016/j.rasd.2014.12.014

Estrada, E., Ferrer, E. y Pardo, A. (2019). Statistics for Evaluating Pre-Post Change: Relation between Change in the Distribution Center and Change in the Individual Scores. *Frontiers in Psychology, 9*, Artículo 2696. https://doi.org/10.3389/fpsyg.2018.02696

Facon, B., Sahiri, S. y Riviere, V. (2008). A Controlled Single-Case Treatment of Severe Long-Term Selective Mutism in a Child with Mental Retardation. *Behavior Therapy, 39*(4), 313–321. https://doi.org/10.1016/j.beth.2007.09.004

Feeney, T. y Ylvisaker, M. (2006). Context-Sensitive Cognitive-Behavioural Supports for Young Children with TBI: A Replication Study. *Brain Injury, 20*(6), 629–645. https://doi.org/10.1080/02699050600744194

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G. y Hibbard, S. T. (2009). Making Treatment Effect Inferences from Multiple-Baseline Data: The Utility of Multilevel Modeling Approaches. *Behavior Research Methods, 41*(2), 372–384. https://doi.org/10.3758/BRM.41.2.372

Ferron, J. M., Farmer, J. L. y Owens, C. M. (2010). Estimating Individual Treatment Effects from Multiple-Baseline Data: A Monte Carlo study for multilevel-modeling approaches. *Behavior Research Methods, 42*(4), 930–943. https://doi.org/10.3758/BRM.42.4.930

Ferron, J. M., Goldstein, H., Olszewski, A. y Rohrer, L. (2020). Indexing Effects in Single-Case Experimental Designs by Estimating the Percent of Goal Obtained. *Evidence-Based Communication Assessment and Intervention, 14*(1–2), 6–27. https://doi.org/10.1080/17489539.2020.1732024

Ferron, J. M., Moeyaert, M., Van den Noortgate, W. y Beretvas, S. N. (2014). Estimating Causal Effects from Multiple-Baseline Studies: Implications for Design and Analysis. *Psychological Methods, 19*(4), 493–510. https://doi.org/10.1037/a0037038

Fisher, W. W., Kelley, M. E. y Lomas, J. E. (2003). Visual Aids and Structured Criteria for Improving Visual Inspection and Interpretation of single-Case Designs. Journal of Applied Behavior Analysis, 36(3), 387–406. https://doi.org/10.1901/jaba.2003.36-387

Hartmann, D. P. y Hall, R. V. (1976). The Changing Criterion Design. *Journal of Applied Behavior Analysis, 9*(4), 527–532. https://doi.org/10.1901/jaba.1976.9-527

Heyvaert, M. y Onghena, P. (2014). Analysis of Single-Case Data: Randomisation Tests for Measures of Effect Size. *Neuropsychological Rehabilitation, 24*(3–4), 507–527. https://doi.org/10.1080/09602011.2013.818564

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S. y Wolery, M. (2005). The Use of Single-Subject Research to Identify Evidence-Based Practice in Special Education. *Exceptional Children, 71*(2),

165−179. https://doi.org/10.1177/001440290507100203

Jacobs, K. W. (2019). Replicability and Randomization Test Logic in Behavior Analysis. *Journal of the Experimental Analysis of Behavior, 111*(2), 329–341. https://doi.org/10.1002/jeab.501

Kratochwill, T. R., Horner, R. H., Levin, J. R., Machalicek, W., Ferron, J. y Johnson, A. (2021). Single-case Design Standards: An Update and Proposed Upgrades. *Journal of School Psychology, 89*, 91–105. https://doi.org/10.1016/j.jsp.2021.10.006

Klein, L. A., Houlihan, D., Vincent, J. L. y Panahon, C. J. (2017). Best Practices in Utilizing the Changing Criterion Design. *Behavior Analysis in Practice, 10*(1), 52–61. https://doi.org/10.1007/s40617-014-0036-x

Lane, J. D. y Gast, D. L. (2014). Visual Analysis in Single Case Experimental Design Studies: Brief Review and Guidelines. *Neuropsychological Rehabilitation, 24*(3–4), 445–463. https://doi.org/10.1080/09602011.2013.815636

Ledford, J. R., Barton, E. E., Severini, K. E. y Zimmerman, K. N. (2019). A Primer on Single-Case Research Designs: Contemporary Use and Analysis. *American Journal on Intellectual and Developmental Disabilities, 124*(1), 35–56. https://doi.org/10.1352/1944-7558-124.1.35

Levin, J. R., Ferron, J. M. y Gafurov, B. S. (2018). Comparison of Randomization-Test Procedures for Single-Case Multiple-Baseline Designs. *Developmental Neurorehabilitation, 21*(5), 290–311. https://doi.org/10.1080/17518423.2016.1197708

Ma, H. H. (2006). An Alternative Method for Quantitative Synthesis of Single-Subject Research: Percentage of Data Points Exceeding the Median. *Behavior Modification, 30*(5), 598–617. https://doi.org/10.1177/0145445504272974

Maggin, D. M., Cook, B. G. y Cook, L. (2018). Using Single-Case Research Designs to Examine the Effects of Interventions in Special Education. *Learning Disabilities Research & Practice, 33*(4), 182–191. https://doi.org/10.1111/ldrp.12184

Maggin, D. M., Cook, B. G. y Cook, L. (2019). Making Sense of Single-Case Design Effect Sizes. *Learning Disabilities Research & Practice, 34*(3), 124–132. https://doi.org/10.1111/ldrp.12204

Manolov, R., Moeyaert, M. y Fingerhut, J. (2022). A Priori Justification for Effect Measures in Single-Case Experimental Designs. *Perspectives on Behavior Science, 45*(1), 156–189. https://doi.org/10.1007/s40614-021-00282-2

Manolov, R. y Onghena, P. (2018). Analyzing Data from Single-Case Alternating Treatments Designs. *Psychological Methods, 23*(3), 480–504. https://doi.org/10.1037/met0000133

Manolov, R. y Onghena, P. (2022). Defining and Assessing Immediacy in Single Case Experimental Designs. *Journal of the Experimental Analysis of Behavior, 118*(3), 462−492. https://doi.org/10.1002/JEAB.799

Manolov, R. y Tanious, R. (2022). Assessing Consistency in Single-Case Data Features using Modified Brinley Plots. *Behavior Modification, 46*(3), 581–627. https://doi.org/10.1177/0145445520982969

Manolov, R., Tanious, R. y Fernández-Castilla, B. (2022). A Proposal for the Assessment of Replication of Effects in Single-Case Experimental Designs. *Journal of Applied Behavior Analysis, 55*(3), 997–1024. https://doi.org/10.1002/jaba.923

McDougale, C. B., Richling, S. M., Longino, E. B. y O'Rourke, S. A. (2020). Mastery Criteria and Maintenance: A Descriptive Analysis of Applied Research Procedures. *Behavior Analysis in*

Practice, 13(2), 402–410. https://doi.org/10.1007/s40617-019-00365-2

McDougall, D. (2005). The Range-Bound Changing Criterion Design. *Behavioral Interventions, 20*(2), 129–137. https://doi.org/10.1002/bin.189

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S. N. y Van den Noortgate, W. (2014). The Influence of the Design Matrix on Treatment Effect Estimates in the Quantitative Analyses of Single-Case Experimental Designs Research. *Behavior Modification, 38*(5), 665–704. https://doi.org/10.1177/0145445514535243

Natesan, P. y Hedges, L. V. (2017). Bayesian Unknown Change-Point Models to Investigate Immediacy in Single Case Designs. P*sychological Methods, 22*(4), 743–759. https://doi.org/10.1037/met0000134

Onghena, P. (1992). Randomization Tests for Extensions and Variations of ABAB single-Case Experimental Designs: A Rejoinder. *Behavioral Assessment, 14*(2), 153–171.

Onghena, P. y Edgington, E. S. (1994). Randomization Tests for Restricted Alternating Treatments Designs. *Behaviour Research and Therapy, 32*(7), 783–786. https://doi.org/10.1016/0005-7967(94)90036-1

Onghena, P., Tanious, R., De, T. K. y Michiels, B. (2019). Randomization Tests for Changing Criterion Designs. *Behaviour Research and Therapy, 117*(6), 18–27. https://doi.org/10.1016/j.brat.2019.01.005

Parker, R. I. y Vannest, K. J. (2009). An Improved Effect Size for Single-Case Research: Nonoverlap of all Pairs. *Behavior Therapy, 40*(4), 357−367. https://doi.org/10.1016/j.beth.2008.10.006

Parker, R. I., Vannest, K. J. y Davis, J. L. (2011). Effect Size in Single-Case Research: A Review of Nine Nonoverlap Techniques. *Behavior Modification, 35*(4), 303–322. https://doi.org/10.1177/0145445511399147

Parker, R. I., Vannest, K. J., Davis, J. L. y Sauber, S. B. (2011). Combining Nonoverlap and Trend for Single-Case Research: Tau-U. *Behavior Therapy, 42*(2), 284−299. https://doi.org/10.1016/j.beth.2010.08.006

Perdices, M., Tate, R. L. y Rosenkoetter, U. (2023). An Algorithm to Evaluate Methodological Rigor and Risk of Bias in Single-Case Studies. *Behavior Modification, 47*(6), 1482–1509. https://doi.org/10.1177/0145445519863035

Pustejovsky, J. E. (2018). Using Response Ratios for Meta-Analyzing Single-Case Designs with Behavioral Outcomes. *Journal of School Psychology, 68*(6), 99−112. https://doi.org/10.1016/j.jsp.2018.02.003

Pustejovsky, J. E., Hedges, L. V. y Shadish, W. R. (2014). Design-Comparable Effect Sizes in Multiple Baseline Designs: A General Modeling Framework. *Journal of Educational and Behavioral Statistics, 39*(5), 368–393. https://doi.org/10.3102/1076998614547577

Scruggs, T. E., Mastropieri, M. A. y Casto, G. (1987). The Quantitative Synthesis of Single-Subject Research: Methodology and Validation. *Remedial and Special Education, 8*(2), 24–33. https://doi.org/10.1177/074193258700800206

Shadish, W. R., Hedges, L. V. y Pustejovsky, J. E. (2014). Analysis and Meta-Analysis of Single-Case Designs with a standardized Mean Difference Statistic: A Primer and Applications. *Journal of School Psychology, 52*(2), 123–147. https://doi.org/10.1016/j.jsp.2013.11.005

Slocum, T. A., Pinkelman, S. E., Joslyn, P. R. y Nichols, B. (2022). Threats to Internal Validity in Multiple-baseline Design Variations. *Perspectives on Behavior Science, 45*(3), 619−638. https://doi.org/10.1007/s40614-022-00326-1

Snodgrass, M., Cook, B. G. y Cook, L. (2023). Considering Social Validity in Special Education Research. *Learning Disabilities Research & Practice, 38*(4), 311–319. https://doi.org/10.1111/ldrp.12326

Swaminathan, H., Rogers, H. J., Horner, R., Sugai, G. y Smolkowski, K. (2014). Regression Models for the Analysis of Single Case Designs. *Neuropsychological Rehabilitation, 24*(3–4), 554−571. https://doi.org/10.1080/09602011.2014.887586

Tanious, R. y Onghena, P. (2021). A Systematic Review of Applied Single-Case Research Published between 2016 and 2018: Study Designs, Randomization, Data Aspects, and Data Analysis. *Behavior Research Methods, 53*(4), 1371–1384. https://doi.org/10.3758/s13428-020-01502-4

Tate, R. L. y Perdices, M. (2019). *Single-case Experimental Designs for Clinical Research and Neurorehabilitation Settings: Planning, Conduct, Analysis, and Reporting*. Routledge.

Tate, R. L., Perdices, M., Rosenkoetter, U., McDonald, S., Togher, L., Shadish, W., Horner, R., Kratochwill, T., Barlow, D. H., Kazdin, A. E., Sampson, M., Shamseer, L. y Vohra, S. (2016). The Single-Case Reporting Guideline in Behavioural Interventions (SCRIBE) 2016: Explanation and elaboration. *Archives of Scientific Psychology, 4*(1), 10–31. https://doi.org/10.1037/arc0000027

Vannest, K. J. y Sallese, M. R. (2021). Benchmarking Effect Sizes in Single-Case Experimental Designs. *Evidence-Based Communication Assessment and Intervention, 15*(3), 142–165. https://doi.org/10.1080/17489539.2021.1886412

What Works Clearinghouse. (2022). *Procedures and Standards Handbook, Version 5.0. U.S.* Department of Education, Institute of Education Sciences. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/Final_WWC-HandbookVer5.0-0-508.pdf

Wine, B., Freeman, T. R. y King, A. (2015). Withdrawal versus Reversal: A Necessary Distinction? *Behavioral Interventions, 30*(1), 87–93. https://doi.org/10.1002/bin.1399

Wolfe, K., Barton, E. E. y Meadan, H. (2019). Systematic Protocols for the Visual Analysis of Single-Case Research Data. *Behavior Analysis in Practice, 12*(2), 491–502. https://doi.org/10.1007/s40617-019-00336-7