Estimació de la magnitud de l'efecte en dissenys de cas únic

Rumen Rumenov Manolov

UNIVERSITAT DE BARCELONA

ESTIMACIÓ DE LA MAGNITUD DE L'EFECTE EN DISSENYS DE CAS ÚNIC

Tesi presentada per

Rumen Manolov

per optar al Grau de Doctor per la Universitat de Barcelona

Director: Dr. Antonio Solanas Pérez

Departament de Metodologia de les Ciències del Comportament
Facultat de Psicologia
Universitat of Barcelona

Programa de doctorat *Mètodes d'Investigació en Psicologia* (2006-2008)

Barcelona, 2010

AGRAÏMENTS

Благодаря на тези, които
ме обичат и ми помагат.

Jag tycker om när jag har dig vid min sida,
tycker om när jag betyder något för dig.

ÍNDEX

1. INTRODUCCIÓ

La present tesi doctoral es centra en l'estimació de la magnitud de l'efecte del tractament en estudis que empren dissenys de cas únic (*N=1*). S'avalua el rendiment de diversos procediments en sèries curtes de dades generades per simulació Monte Carlo. A més de dur a terme aquestes comparacions, les investigacions realitzades modifiquen una tècnica existent i en desenvolupen una de nova, tenint en compte les característiques que solen presentar les dades reals (e.g., autocorrelació). Els següents apartats introdueixen breument els aspectes més importants que defineixen el context de la recerca realitzada, emfasitzant les característiques bàsiques dels dissenys d'*N=1* i els procediments utilitzats per analitzar les dades obtingudes mitjançant aquests.

1.1. Dissenys de cas únic

En aquest apartat només s'esmentaran les característiques més importants dels dissenys de cas únic en relació amb les recerques de la tesi doctoral. Una revisió amb més profunditat es pot trobar en Ato i Vallejo (2007).

Els dissenys de cas únic són estructures longitudinals i seqüencials de recollida de dades que permeten descriure els patrons de conducta d'una unitat que pot ser tant una persona individual com un grup (e.g., una família o una institució) considerat com a una totalitat (Onghena i Edgington, 2005). L'objectiu d'aquests dissenys és avaluar el possible efecte d'una intervenció sistemàtica sobre una conducta específica, la conducta d'interès, en més d'un moment d'observació. La seva utilitat en el context clínic ha estat ressaltada contínuament (Barlow i Hersen, 1973; Blampied, 2000; Horner et al., 2005), i això és visible en contexts tan diversos com, per exemple, educació especial (Mastropieri i Scruggs, 1985), rehabilitació neuropsicològica (Evans, Emslie i Wilson, 1998), problemes del desenvolupament (Reynhout i Carter, 2006), etc. Aquest tipus de dissenys és especialment flexible per adaptar-se al context del client i als objectius del professional (Greenwald, 1976).

Els dissenys d'*N=1* es caracteritzen, a més a més, pel control de la variabilitat de la conducta. Per una banda, degut a la pròpia essència dels dissenys, s'elimina la variabilitat inter-subjecte. Per altra banda, s'avalua la situació inicial mitjançant la fase de línia base que a més d'aquesta funció descriptiva, en té un altre de predictiva: establir els límits dins dels quals oscil·larà la conducta si el tractament no és efectiu; la línia base ha de ser estable per poder atribuir amb major grau de confiança els canvis posteriors a la intervenció realitzada (Kazdin, 2001). En general, la inestabilitat del comportament intra-fase perjudica la comparació entre condicions experimentals (Johnston i Pennypacker, 2008).

Els tipus de control emprats en estudis idiogràfics són diferents dels estudis nomotètics, atès que el moment de mesura cobra un rellevància especial (Mace i Kratochwill, 1986). Elements clau per a la validesa interna en aquests dissenys (Sidman, 1960) són la reversibilitat (retorn de la conducta al retirar el tractament a nivells similars als observats en el patró basal) i la replicació d'efectes (replicació de resultats amb la replicació de l'aplicació d'una determinada intervenció). El terme "validesa interna" reflecteix la idea de la relació funcional entre variable manipulada (i.e., la fase o condició experimental) i la variable mesurada (i.e., la conducta). Per augmentar el control intern s'han de descartar les explicacions alternatives del canvi.

A més de permetre (depenent del disseny concret utilitzat) la inferència de causalitat per a cada estudi individual, les estratègies de cas únic també permeten generalitzar les troballes més enllà del(s) participant(s), context i conductes estudiats. En aquest sentit, les replicacions són bàsiques per aconseguir validesa externa, és a dir , generalització de l'efecte del tractament.

La replicació directa normalment comporta la duplicació de les condicions de l'estudi original amb altres participants; garanteix la generalitat del fenomen en individus diferents a la vegada que incrementa la validesa interna. Les replicacions sistemàtiques varien algun o alguns dels components de l'estudi original per contrastar la generalitat del fenomen en situacions diferents i per esbrinar si els aspectes modificats són crítics per a la relació funcional entre les variables (Sidman, 1960). Una altra manera de contrastar l'efectivitat del tractament en persones i situacions diferents és mitjançant la integració quantitativa d'estudis individuals (i.e., meta-anàlisi); una estratègia que també pot servir per comparar l'efectivitat de tractaments diferents. En aquest sentit la generalitat dels resultats està relacionada amb el grau en que concorden amb la literatura prèvia sobre el tema (Johnston i Pennypacker, 2008).

Existeixen diversos tipus de dissenys que difereixen en el grau de control que permet la seva estructura i que, per tant, difereixen quant a la validesa interna dels resultats obtinguts. Clàssicament s'han establert dues grans categories: els dissenys de reversió i els de no reversió segons si requereixen o no la retirada del tractament. L'estructura més simple és l'AB, seguint la nomenclatura de la tradició clínica (Barlow i Hersen, 1973; Hersen i Barlow, 1976). Aquesta és un reflex del procés natural d'intervenció psicològica en el qual la intervenció comença després d'una avaluació inicial de la situació existent (Rabin, 1981). A més a més, poden ser els més viables considerant les limitacions dels recursos temporals i econòmics en la pràctica psicològica real. Ara bé, la seva capacitat de control és mínima per la qual cosa, metodològicament, una estructura com l'AB no seria suficient per demostrar el control del comportament pel disseny (Wampold i Furlong, 1981a). És recomanable utilitzar estructures que impliquin més de dues fases. En un disseny ABA es podria comprovar si la retirada del tractament en la tercera fase està relacionada amb una reversió de la conducta als nivells de la fase de línia base, tot i que el control seria

també deficitari. En el cas dels dissenys BAB la reintroducció del tractament en l'última condició experimental també hauria de veure's reflectida en el comportament. Aquestes dues estructures tenen limitacions ja que en el primer cas (ABA) no hi ha reintroducció del tractament, la qual cosa no només planteja problemes ètics sinó també metodològics: no es comprova l'eficàcia del tractament amb la replicació dels efectes o, en el segon cas, comencen sense avaluar la situació existent (BAB). Per tant, seria millor que les conductes reversibles es mesuressin seguint un disseny ABAB, única estructura que respon a dissenys veritablement experimentals. Entre les estructures reversibles més complexes es poden esmentar altres tipus de dissenys (de reversió multinivell, multitractament i interactius) però la seva descripció va més enllà dels objectius d'aquesta introducció.

Quan la conducta d'interès no permet o no es aconsellable retornar als nivells obtinguts durant la fase de línia base (e.g., en estudiar processos d'aprenentatge), s'utilitzen els dissenys de no reversió. Partint de l'estructura base AB, és possible replicar-la en diferents contexts, individus o conductes mitjançant els dissenys de línia base múltiple. El control en aquests dissenys rau en la introducció seqüencial del tractament: la relació funcional entre les variables es potenciaria si només s'observen canvis a la conducta concreta intervinguda i no en tots els comportaments (Mace i Kratochwill, 1986). En aquest cas, si s'observen canvis en el comportament següents a la introducció del tractament en cada línia base i aquests canvis es repliquen en altres conductes, situacions o subjectes, es disposaria d'evidència que aquests canvis poden atribuir-se a la intervenció. Les estructures descrites, dissenys de línia base múltiple, proporcionen una validesa interna relativament bona. Tot i que existeixen d'altres dissenys de no reversió (els dissenys de canvi de criteri, els dissenys de tractaments alternants i els dissenys de tractaments simultanis), els de línia base múltiple són en la pràctica els més utilitzats.

La present tesi doctoral es focalitza en dissenys AB, atès que l'estructura més simple de disseny d'*N=1* sembla un punt de partida lògic per obtenir els primers resultats sobre el funcionament de diferents procediments per a la quantificació de la magnitud de l'efecte, remarcant la necessitat que investigacions següents es centrin en estructures més complexes per contrastar la generalitat dels resultats. No obstant, les troballes referents als dissenys AB són aplicables a qualssevulla dues fases adjacents que estiguin incloses en una estructura de disseny de cas únic. Evidentment, quan es treballa amb dissenys de línia base múltiple, es tractaria de quantificar la magnitud de l'efecte tantes vegades com participants, conductes o contexts hi hagi. Addicionalment, per exemple en un disseny ABAB, són possibles tres comparacions entre fases adjacents (i.e., tres càlculs de l'índex de grandària de l'efecte escollit): línia base amb tractament, tractament amb retirada i retirada amb reintroducció del tractament.

1.2. El problema de l'autocorrelació

En els dissenys de cas únic s'observa i es mesura longitudinalment el comportament d'un individu o un grup i, des d'una perspectiva substantiva, s'espera que el comportament d'una unitat experimental en un moment donat estigui relacionat amb el que ha succeït en el passat. Aquest fenomen s'anomena autocorrelació o dependència serial i, a nivell estadístic, es representa com la correlació entre dades de la mateixa sèrie obtinguda al llarg del temps. És imprescindible diferenciar entre autocorrelació positiva i negativa. En el cas de la primera, la conducta pot ser representada per tendències creixents o decreixents. La segona es refereix a una augmentada variabilitat en la taxa de resposta i s'exemplifica mitjançant la ràpida alternança de valors per sobre i per sota del valor esperat.

Més enllà de l'argumentació conceptual, els estudis empírics informen de resultats diferents i extreuen conclusions discrepants. D'una banda, hi ha evidència que la dependència serial en dades conductuals acostuma a ser nul·la i/o no significativa (Huitema, 1985). La revisió de Huitema (1985) ha estat criticada tant des de la vessant metodològica com des de la teòrica. Revisions més recents basades en les mateixes o en noves dades d'*N=1* indiquen que l'autocorrelació és un atribut freqüent en les dades obtingudes mitjançant aquest tipus de dissenys (Busk i Marascuilo, 1988; Matyas i Greenwood, 1991; 1997; Parker, 2006). Esbrinar amb un grau suficient de certesa si les mesures estan correlacionades o no pot ser difícil considerant que la estimació de l'autocorrelació ha demostrat ser problemàtica per l'existència de biaix en els estimadors i la potència estadística insuficient en les proves associades amb aquests estimadors en sèries curtes de dades (Arnau i Bono, 2003; Huitema i McKean, 1991; Matyas i Greenwood, 1991; Solanas, Manolov i Sierra, 2010).

Part de la recerca s'ha centrat en l'avaluació per simulació del rendiment de diverses tècniques estadístiques en presència d'autocorrelació, suposant que aquesta pot ser-hi en les dades obtingudes de la mateixa unitat experimental. Quant a les proves que assumeixen independència de les observacions aquests estudis fan referència a la robustesa de les proves davant de la violació del supòsit. Alguns resultats sobre l'efecte de la dependència serial en les taxes d'error Tipus I seran presentats en el següent apartat que descriu diferents possibilitats d'anàlisi estadística de les dades obtingudes mitjançant dissenys de cas únic.

1.3. Avaluació de l'efectivitat del tractament en dissenys de cas únic

L'objectiu principal d'un investigador aplicat, en realitzar un estudi, és avaluar si la intervenció psicològica ha estat efectiva o no. Hi ha diverses alternatives per obtenir evidències sobre la presència o absència de canvi en la conducta d'interès, però cap d'aquestes alternatives sembla estar exempta de problemes. A continuació es revisen breument les tècniques que han rebut més atenció en la literatura científica, tenint en compte que no és possible comentar tots els procediments existents. L'ordre de presentació es correspon amb l'ordre cronològic de proposta o utilització de les tècniques en el context d'*N=1*.

1.3.3. Inspecció visual

En dissenys de cas únic en concret, la tècnica més freqüentment aplicada per prendre decisions respecte a l'efectivitat del tractament ha estat tradicionalment i segueix essent la inspecció visual (Kratochwill i Brody, 1978; Parker, Cryer i Byrns, 2006). Malgrat que va ser proposada per detectar canvis grans en la conducta (Kratochwill i Levin, 1980; Parsonson i Baer, 1986), s'ha demostrat que tant els errors Tipus I com els Tipus II poden ser excessivament probables. Per exemple, Fisch (2001), Matyas i Greenwood (1990) i Normand i Bailey (2006) informen de taxes altes de falses alarmes, mentre que Jones, Weinrott i Vaught (1978) i Ottenbacher (1990) varen trobar taxes altes d'omissió. A més a més, diversos estudis coincideixen en la baixa concordança entre els analistes visuals (e.g., Brossart, Parker, Olsson i Mahadevan, 2006; DeProspero i Cohen, 1979; Ottenbacher, 1990) o baixa consistència al jutjar les mateixes dades (Ximenes, Manolov, Solanas i Quera, 2009). Fins i tot l'experiència en la inspecció visual no sembla garantir un millor rendiment (Knapp, 1983; Richards, Taylor i Ramasamy, 1997), possiblement en relació amb la manca de regles formals de presa de decisions (Wampold i Furlong, 1981b). Entre les limitacions de l'anàlisi visual s'han d'esmentar la manca de guies establertes per a la seva interpretació (Kazdin, 1982) i la impossibilitat d'integrar quantitativament els resultats d'estudis diferents mitjançant meta-anàlisi (Busk i Serlin, 1992).

Una tècnica que guarda relació amb la inspecció visual és el mètode *split-middle* (Miller, 1985; White, 1974), que inclou ajudes visuals per millorar el procés de presa de decisions. Desafortunadament, s'ha demostrat que aquesta tècnica no controla el error Tipus I en presència d'autocorrelació (Crosbie, 1987). Un altre procediment que comporta anàlisi visual i càlculs estadístics simples (i.e., desviacions estàndard al voltant de la mitjana) és el *gràfic de control de processos* o *gràfic de Shewhart* que ha estat discutit com una possible alternativa per analitzar dades conductuals (Callahan i Barisa, 2005; Hantula, 1995; Pfadt, Cohen, Sudhalter, Romanczyk i Wheeler, 1992; Pfadt i Wheeler, 1995). No obstant, aquest tipus d'anàlisi és recomanat només quan

les dades segueixen una distribució normal i no presenten autocorrelació, tendència o valors anòmals (Gottman, 1973).

S'ha de tenir en compte que la inspecció visual i les diferents tècniques estadístiques proporcionen informació diferent. Això junt amb les limitacions de l'anàlisi visual aplicat de forma exclusiva, ha portat a diversos autors a recomanar l'ús conjunt d'aquest tipus d'anàlisi amb procediments quantitatius, atès que no es tracta de tècniques oposades sinó complementàries (e.g., Barlow i Hersen, 2008; Busk i Marascuilo, 1992; Jones et al., 1977).

1.3.2. Tècniques estadístiques clàssiques

L'ús dels dissenys d'*N=1* fora del context experimental ha comportat la disminució del control intern i el subsegüent augment en la variabilitat de les mesures. Un clar exemple són les línies base inestables que, encara que siguin teòricament inacceptables, es poden produir en situacions reals i requereixen l'aplicació de tècniques estadístiques (Kazdin, 1978a). Gran part de la recerca en psicologia utilitza tècniques clàssiques basades en un enfocament nomotètic, com per exemple l'anàlisi de la variància (ANOVA, abreviació del terme *analysis of variance*). S'ha suggerit que aquesta tècnica pot ser aplicada a dissenys de caire idiogràfic si es demostra que les dades són independents (Huitema, 1985; 1988). S'hauria de contrastar doncs la presència o absència d'autocorrelació abans d'utilitzar tècniques paramètriques (Hartmann, 1974; Kazdin, 1978b), cosa que com ja s'ha comentat pot ser difícil amb els procediments estadístics actualment disponibles.

S'ha subratllat que fins i tot nivells baixos i no significatius de dependència serial poden distorsionar substancialment les taxes d'error Tipus I quan s'utilitzen proves estadístiques clàssiques (Busk i Marascuilo, 1988; Sharpley i Alavosius, 1988; Suen, 1987; Suen i Ary, 1987). Les discussions recents segueixen apuntant que les estimacions baixes de l'autocorrelació no garanteixen l'adequació de l'ús de tècniques estadístiques basades en el Model Lineal General per avaluar l'efectivitat del tractament (Ferron, 2002).

D'acord amb aquests comentaris, s'ha demostrat que l'ANOVA és poc apropiat en termes de taxes d'error Tipus I quan les dades no són independents (Scheffé, 1959; Toothaker, Banz, Noble, Camp i Davis, 1983). Per tant, s'ha intentat adaptar aquesta tècnica per poder augmentar la seva utilitat en l'anàlisi de dades d'*N=1*, com per exemple la proposta de Shine i Bower (1971) que consisteix en una modificació en el càlcul de la variància intra-cel·les de l'ANOVA de mesures repetides. No obstant, aquesta proposta és útil només quan les dades es distribueixen normalment i independentment i l'efecte del temps és el mateix per a totes les dades. En aquest cas, s'ha de ressaltar que no només la independència pot ser un supòsit irreal, sinó també

la normalitat, ja que aquesta no sembla ser tan comú en les dades psicològiques (Bradley, 1977; Micceri, 1989).

Una altra possibilitat és combinar les observacions de dissenys ABAB en condicions experimentals (línia base i tractament) per aplicar un ANOVA de mesures repetides, suposant la distribució normal i independent dintre de cada condició (Gentile, Roden i Klein, 1972). Desafortunadament, l'estadístic *F* pot ser massa liberal en casos de dependència serial o massa conservador degut a que la variància de l'error es pot inflar en barrejar fases diferents.

1.3.3. Anàlisi de sèries temporals

Per superar el problema de la dependència serial, s'ha suggerit l'ús de models autoregressius i de mitjanes mòbils (la denominació anglosaxona *autoregressive integrated moving average* - ARIMA; Harrop i Velicer, 1985; Jones, Vaught i Weinrott, 1977; Kratochwill i Levin, 1980; Sharpley i Alavosius, 1988). L'aplicació d'aquesta tècnica requereix els següents passos (Glass, Wilson i Gottman, 1975): a) identificar l'ordre dels paràmetres d'autoregressió, diferenciació i mitjanes mòbils que defineixen el model; b) eliminar la tendència diferenciant les dades; c) estimar els paràmetres autoregressius i de mitjanes mòbils; d) eliminar l'autocorrelació utilitzant els paràmetres estimats en el pas anterior; e) aplicar el Model Lineal General per comprovar l'existència de canvi de nivell o de pendent entre les dades abans i després de la intervenció.

Els inconvenients principals d'aquesta tècnica són la dificultat d'aplicació (i.e., d'identificar el model que s'ajusta millor a les dades empíriques) i la necessitat de sèries llargues de dades. Diversos autors (e.g., Simonton, 1977; Velicer i McDonald, 1984) han suggerit substituir el primer pas per l'aplicació de models assumits a priori, però les evidències no avalen aquesta proposta (Vallejo, 1994). A més, hi ha evidència que l'autocorrelació distorsiona les taxes d'error Tipus I de l'ARIMA en sèries curtes (Greenwood i Matyas, 1990). Dos procediments relacionats amb l'anàlisi de sèries temporals, però superant algunes de les seves limitacions són els anomenats ITSE (Gottman, 1981), provinent del terme *interrupted time series experiments analysis*, i ITSACORR (Crosbie, 1993), nom que combina aquesta mateixa expressió anglosaxona amb la paraula "correlació". No obstant, totes dues tècniques presenten problemes tant a nivell conceptual com a nivell de les estimacions obtingudes (Huitema, 2004; Huitema, McKean i Laraway, 2007).

1.3.4. Altres procediments per prendre decisions estadístiques

Les tècniques incloses en aquest apartat es van proposar fa unes tres dècades per obtenir valors *p* com a indicadors de l'efectivitat del tractament, però el seu ús no ha estat tan extens i tampoc no s'ha revifat en els últims anys. L'estadístic *C* s'ha proposat com a una manera simple d'avaluar l'efectivitat de la intervenció (Tryon, 1982; 1984). Malgrat alguns resultats positius (Arnau i Bono, 1998), el seu rendiment ha demostrat ser inadequat en termes d'error Tipus I i Tipus II (Blumberg, 1984; Crosbie, 1989). Addicionalment, l'estadístic *C* sembla ser més un estimador d'autocorrelació de primer ordre que un estadístic per contrastar la presència d'efectes del tractament (DeCarlo i Tryon, 1993). Una altra tècnica que no ha mostrat rendiment satisfactori amb dades de cas únic és la prova binomial (Crosbie, 1987). També s'han proposat diversos procediments per transformar les dades corregint la dependència serial (e.g., Algina i Swaminathan, 1979; Gorsuch, 1983; Simonton, 1977). Hi ha evidència que la potència estadística d'aquestes tècniques pot ser insuficient per a sèries curtes de dades (Arnau i Bono, 2004).

1.3.5. Proves d'aleatorització

Les proves d'aleatorització constitueixen una manera d'obtenir significació estadística directament des de les dades disponibles i han estat promogudes degut als pocs supòsits distribucionals i la versatilitat per contrastar diferents tipus d'efectes d'intervenció (Busk i Marascuilo, 1992; Edgington i Onghena, 2007). S'ha suggerit que la dependència serial no hauria de suposar un problema per les proves d'aleatorització ja que la significació estadística s'obté directament a partir de les dades (Kratochwill i Levin, 1980) o degut a que la dependència serial té un efecte constant per a totes les divisions de dades possibles (Wampold i Worsham, 1986). Una altra posició és que l'autocorrelació compromet només les taxes d'error Tipus II (i.e., potència estadística o sensibilitat) i no les taxes d'error Tipus I (Edgington, 1980b). Contràriament, Good (1994) considera que l'aplicació d'una prova d'aleatorització requereix que les observacions siguin intercanviables. A més a més, s'ha destacat que la igualtat de variàncies de les dades pertanyents a les diferents condicions experimentals és necessària pel bon funcionament de les proves d'aleatorització (Good, 1994; Gorman i Allison, 1997; Hayes, 1996).

L'aplicació correcta de les proves d'aleatorització requereix que alguna part del disseny sigui aleatoritzada (Edgington, 1980a), cosa que permet controlar les possibles variables de confusió (Onghena i Edgington, 2005), però també restringeix la flexibilitat dels experiments (Kazdin, 1980). Aquesta última raó pot ser la causa del escàs ús de l'assignació aleatòria en contexts aplicats (Ferron i Jones, 2006).

Les troballes inicials suggerien que les taxes d'error Tipus I es corresponen amb els nivells nominals, mentre que la sensibilitat és suficient per detectar tractaments potents i definits per grandàries d'efecte grans (Ferron, Foster-Johnson i Kromrey, 2003; Ferron i Onghena, 1996; Ferron i Sentovich, 2002; Ferron i Ware, 1995). No obstant, estudis recents, seguint un enfocament metodològic diferent, han obtingut resultats discrepants i menys favorables per a les taxes de falses alarmes (Manolov i Solanas, 2009; Manolov, Solanas, Bulté i Onghena, 2010; Sierra, Solanas i Quera, 2005).

1.3.5. Quantificació de la magnitud de l'efecte

S'ha argumentat que les proves estadístiques de significació restringeixen la realització dels estudis aplicats (Michael, 1974) i ignoren la significació clínica a favor de l'estadística (Hugdahl i Öst, 1981). Una aproximació a la importància clínica, educativa o social es pot assolir emprant validació social (e.g., comparació social, autoavaluació) del comportament tractat (Kazdin, 1978a). Una altra manera d'ajudar a un psicòleg a valorar la rellevància del canvi en el client és la quantificació mitjançant els índexs de grandària de l'efecte (Parker i Hagan-Burke, 2007b), encara que aquesta no hauria de substituir les valoracions emeses pel client i el professional en funció dels seus objectius. Es considera que aquests superen les limitacions dels valors *p*, considerats àmpliament com a insuficients per avaluar l'efectivitat d'una intervenció (Cohen 1990; 1994; Kirk, 1996; Rosnow i Rosenthal, 1989; Wilkinson i Task Force on Statistical Inference, 1999). En primer lloc, la grandària de l'efecte no està afectada d'una manera sistemàtica per la grandària de la mostra (Parker i Brossart, 2003). A més a més, la grandària de l'efecte subratlla l'associació entre la variable independent i la dependent, en lloc de centrar-se en la hipòtesi nul·la (Kromrey i Foster-Johnson, 1996). La precisió d'aquesta mesura pot ser avaluada mitjançant la construcció d'intervals de confiança (Kirk, 1996). Addicionalment, la possibilitat de convertir alguns índexs de grandària d'efecte en altres (Friedman, 1982) permet la comparació entre tractaments (Parker i Hagan-Burke, 2007c). Tampoc no s'ha d'oblidar que aquest tipus de mesura permet la realització de meta-anàlisis (Faith, Allison i Gorman, 1997), destacant els models jeràrquics com a una possibilitat versàtil i especialment útil per als dissenys de cas únic (van den Noortgate i Onghena, 2003a; 2003b; 2003c; 2008).

La importància de les mesures de grandària de l'efecte en dissenys de cas únic es mostra en la quantitat d'estudis recents que comparen tècniques diferents (e.g., Brossart et al., 2006; Campbell, 2004; Parker i Brossart, 2003; Parker i Hagan-Burke, 2007a) o il·lustren la seva aplicació (e.g., Olive i Smith, 2005; Parker i Brossart, 2006; Parker i Hagan-Burke, 2007b). La tendència a promoure l'ús de tractaments psicològics amb una base empírica sòlida ha arribat fins als dissenys d'*N=1* (Jenson,

Clark, Kircher i Kristjansson, 2007; Schlosser i Sigafoos, 2008; Shadish, Rindskopf i Hedges, 2008) i queda reflectida en el fet que la revista *Evidence-Based Communication Assessment and Intervention* va dedicar un número especial a aquest tema durant l'any 2008 comptant amb la col·laboració dels autors més destacats de l'àmbit.

Deixant de banda les diferències de mitjanes estandarditzades que van ser proposades en el marc dels dissenys de comparació de grups, els índexs de grandària de l'efecte en *N=1* es basen principalment en l'anàlisi de la regressió o en la inspecció visual. Tres dels procediments basats en la regressió (*Trend analysis* de Gorsuch, 1983; la *d* de White, Rusch, Kazdin i Hartmann, 1989, i el model d'Allison i Gorman, 1993) es comenten amb detall en el primer estudi inclòs en la present tesi doctoral, mentre que el model de Center, Skiba i Casey (1985-1986) no es va estudiar degut a que la proposta d'Allison i Gorman (1993) en representa una millora. Segons una revisió bibliogràfica recent (Schlosser, Lee i Wendt, 2008) la tècnica més freqüentment utilitzada per quantificar la grandària de l'efecte i en meta-anàlisi de dissenys de cas únic és el percentatge de no solapament entre les dades (PND – *Percent of nonoverlapping data*; Scruggs, Mastropieri i Casto, 1987). Com a millores d'aquesta tècnica s'han proposat el percentatge de valors més grans que la mediana (PEM – *Percentage of data points exceeding the median*; Ma, 2006) i el percentatge de no solapament entre totes les dades (PAND - *Percentage of all nonoverlapping data*; Parker, Hagan-Burke i Vannest, 2007). Les tres característiques d'aquestes tres tècniques relacionades amb l'anàlisi visual es presenten en els articles que composen la present tesi doctoral.

Les tècniques d'estimació de la grandària de l'efecte presenten certs avantatges en comparació amb les tècniques que es centren en la significació estadística, però tampoc no són perfectes. En primer lloc, s'ha d'emfasitzar que segons el procediment emprat, les conclusions sobre la força de la relació entre les variables estudiades poden ser diferents (McGrath i Meyer, 2006; Parker et al., 2005). En segon lloc, hi ha una manca de criteris per interpretar les grandàries d'efecte de dissenys de cas únic (Parker i Brossart, 2006), excepte en el cas del PND (Scruggs i Mastropieri, 1998).

2. RECERCA REALITZADA EN EL MARC DE LA TESI DOCTORAL

2.1. Problema de recerca i objectius

La present tesi neix de la necessitat d'avaluar els procediments més comuns proposats per estimar la magnitud de l'efecte en dissenys de cas únic. Algunes d'aquestes tècniques ja han estat estudiades i comparades entre sí, però gairebé sempre en el context de dades empíriques. Per tant, la recerca duta a terme durant el període de la tesi doctoral pretén complementar la literatura científica existent aplicant les tècniques a dades generades amb paràmetres coneguts en una gran varietat de condicions experimentals. Això és important sobretot per a les tècniques desenvolupades en els darrers anys (i.e., PEM i PAND) i per a les que es proposen dins del marc de la tesi doctoral.

Per aconseguir informació valuosa sobre el rendiment de les tècniques és necessari trobar una manera adequada de comparar procediments que tenen un fonament diferent i que comporten estimacions de la magnitud de l'efecte en termes diversos (e.g., R^2 en les tècniques basades en l'anàlisi de la regressió, percentatges en el cas de PND i la seva modificació, PEM i PAND). Conseqüentment, un objectiu complementari era establir uns criteris de comparació adients.

Una de les finalitats dels estudis metodològics relacionats amb els dissenys de cas únic és trobar procediments d'estimació de la magnitud de l'efecte que representin les dades obtingudes de dissenys de cas únic d'una manera correcta i que siguin útils per als investigadors aplicats. La recerca realitzada representa un pas en l'assoliment d'aquest objectiu ambiciós i es centra en cercar tècniques que potenciïn la sinèrgia entre l'anàlisi visual i la quantificació. D'aquesta manera, a la inspecció visual freqüentment emprada pels professionals s'afegiria una tècnica per sistematitzar la presa de decisions i es possibilitaria la documentació dels resultats dels estudis aplicats.

2.2. Mètode d'investigació seguit

Una part important de tots els articles són els mètodes Monte Carlo degut a la seva utilitat per obtenir estimacions quan la distribució mostral de l'estadístic (e.g., de l'índex de grandària de l'efecte) és desconeguda, però es coneix la distribució de la variable aleatòria en la població (Noreen, 1989). La simulació Monte Carlo també s'empra per estudiar el rendiment de tècniques analítiques els supòsits de les quals (e.g., dependència entre les observacions) són violats (Serlin, 2000). En els quatre estudis presentats, la quantitat d'iteracions utilitzada garanteix estimacions precises, segons els criteris presentats en Robey i Barcikowski (1992). Finalment, l'efecte dels valors inicials (les llavors del procés de simulació) anòmals i la dependència entre

sèries de dades consecutives van ser controlats eliminant els 20 números anteriors a les dades de cada sèrie, seguint les indicacions de Greenwood i Matyas (1990) i Huitema, McKean i McKnight (1999), respectivament.

Un propòsit complementari va consistir en promoure una ampliació en el mètode de generació de dades que representin mesures d'*N=1*. Les recerques de caire metodològic en dissenys de cas únic solen simular dades generades per un procés autoregressiu de primer ordre i que segueixen una distribució normal unitaria. No obstant, les dades reals no són necessàriament normals (Bradley, 1977; Micceri, 1989) i la dependència serial pot ser deguda a altres tipus de processos (Harrop i Velicer, 1985). Les investigacions més recents de la tesi inclouen també un procés de mitjanes mòbils, MA(1), com a fonament de la generació de les dades. A més, quant a la distribució dels errors, s'afegeix una distribució exponencial negativa, degut a la seva marcada asimetria, i una distribució uniforme, atès que presenta una curtosi menor en comparació amb la normal.

Quant a la realització de les comparacions, va ser necessari definir quin és el rendiment desitjable de les tècniques. En aquest sentit, és important que les estimacions de la magnitud de l'efecte representin correctament les característiques existents en les dades (i.e., dels paràmetres de simulació). Un dels criteris era la discriminació entre presència i absència d'efecte de la intervenció i aquesta discriminació s'expressa en raons entre les estimacions obtingudes en tots dos casos. Quan més gran sigui la distància entre les estimacions, en termes relatius, millor. També mitjançant raons es va avaluar la distorsió introduïda per l'autocorrelació i la tendència en les dades. En aquests casos, el rendiment apropiat seria marcat per unes estimacions semblants tant en presència com en absència d'aquests factors de confusió.

2.3. Progressió de la recerca

El primer estudi es va centrar en comparar índexs amb fonaments diferents: tres tècniques que utilitzen l'anàlisi de la regressió controlant la relació entre la conducta d'interès i el temps, dues tècniques basades en diferències de mitjanes estandarditzades i una tècnica fonamentada en la inspecció visual. Els resultats suggereixen que els procediments més sofisticats i que semblen més apropiats des del punt de vista conceptual no són necessàriament millors. El bon rendiment del PND, combinat amb una de les seves grans limitacions – l'afectació per la presència valors anòmals – va motivar el segon estudi. Aquest es va centrar en el PND i en dues tècniques (PEM i PAND) que pretenen millorar-lo, especialment la influència de valors extrems. Per a l'elecció de tècniques a estudiar també es va considerar que la facilitat de càlcul pot estar relacionada amb l'atractivitat dels índexs per als investigadors aplicats. Les troballes varen indicar que les tècniques ideades per

perfeccionar el PND no aconseguien el seu propòsit i es mostraven inferiors respecte a la discriminació entre presència i absència d'efecte del tractament. Per tant, el focus del tercer estudi també va ser el PND – l'índex visual més conegut i més emprat en l'àmbit d'*N=1* – amb l'objectiu d'intentar superar un dels seus desavantatges: les distorsions provocades per la tendència general en les dades (Schlosser, Lee i Wendt, 2008). La correcció proposada es va anomenar *Percentage of nonoverlapping corrected data* (PNCD) i no implica anàlisi de la regressió com els procediments existents. Els resultats indiquen que el PNCD assoleix controlar efectivament la tendència, atès que aquesta exerceix una influència pràcticament nul·la sobre les estimacions de la magnitud de l'efecte. El quart estudi va sorgir de la necessitat de desenvolupar un procediment que no només fos insensible a la tendència, sinó que a més a més estimés d'una manera precisa el canvi que es produeix en el comportament amb la finalització d'una condició experimental i la iniciació de la següent. Es va considerar rellevant la idea de quantificar separadament el canvi de nivell i el canvi de pendent (Beretvas i Chung, 2008a). El procediment proposat es va anomenar temptativament SLC, un acrònim de *slope and level change*. Aquesta tècnica, de càlcul relativament fàcil, es va avaluar mitjançant mètodes Monte Carlo indicant manca de biaix en l'estimació i un bon rendiment en termes d'error estàndard en condicions amb autocorrelació positiva.

En síntesi, els estudis inclosos en la present tesi doctoral representen una progressió lògica des d'anàlisis comparatius d'índexs amb diferents fonaments fins a comparacions entre tècniques estretament relacionades, seguit per la introducció de modificacions en una tècnica existent i arribant a proposar un nou procediment d'estimació de la magnitud de l'efecte de la intervenció. Aquesta línia de treball que inclou tant les propostes d'altres autors com els desenvolupaments del grup de recerca del que forma part el doctorand, correspon a la idea d'apostar per índexs relativament simples tant a nivell de càlcul com d'interpretació i que poden ser combinats amb la inspecció de gràfiques que representin les dades conductuals. En el cas de les propostes realitzades, es va considerar necessari facilitar l'aplicació dels índex mitjançant codis de programació que es poden emprar fàcilment en software de distribució lliure. Malgrat els progressos, la línia de recerca s'hauria de continuar debatent i millorant les tècniques que realment són utilitzades (o que són potencialment útils) pels professionals.

3. ESTUDIS REALITZATS

La tesi doctoral "Quantificació de la magnitud de l'efecte en dissenys de cas únic" es presenta com a compendi d'estudis publicats com a articles científics. A continuació s'especifiquen les referències dels articles i la informació més rellevant sobre les revistes corresponents.

Estudi 1:

Manolov, R. i Solanas, A. (2008). Comparing N = 1 effect size indices in presence of autocorrelation. *Behavior Modification, 32,* 860-875.

Informació sobre la revista: Factor d'impacte 2008 en ISI Journal Citation Reports: 1.559 – Posició 42 de les 88 revistes de l'àrea *Psychology, Clinical*.

Estudi 2:

Manolov, R., Solanas, A. i Leiva, D. (2010). Comparing "visual" effect size indices for single-case designs. *Methodology, 6,* 49-58.

Informació sobre la revista: Indexada a les bases de dades PsycINFO, PSYNDEX i ERIH.

Estudi 3:

Manolov, R. i Solanas, A. (2009). Percentage of nonoverlapping corrected data. *Behavior Research Methods, 41*, 1262-1271.

Informació sobre la revista: Factor d'impacte 2008 en ISI Journal Citation Reports: 1.737 – Posició 3 de les 11 revistes de l'àrea *Psychology, Mathematical*; Posició 34 de les 71 revistes de l'àrea *Psychology, Experimental*.

Estudi 4:

Solanas, A., Manolov, R. i Onghena, P. (2010). Estimating slope and level change in N=1 designs. *Behavior Modification, 34,* 195-218.

Informació sobre la revista: Factor d'impacte 2008 en ISI Journal Citation Reports: 1.559 – Posició 42 de les 88 revistes de l'àrea *Psychology, Clinical*.

Estudi 1©

Comparing N = 1 effect size indices in presence of autocorrelation

Rumen Manolov i Antonio Solanas

Departament de Metodologia de les Ciències del Comportament

Facultat de Psicologia

Universitat de Barcelona

Resum. L'article es centra en sis índexs de mesura de grandària de l'efecte degut a la seva importància per a les meta-anàlisis com a eina que potencia la generalització dels resultats d'estudis individuals. Els procediments estudiats tenen fonaments estadístics diferents: tres es basen en l'anàlisi de la regressió, dos en diferències de mitjanes estandarditzades i un en la inspecció visual. El context de comparació el constitueixen sèries de dades generades mitjançant simulació Monte Carlo que permet establir tant la presència com l'absència d'efectes del tractament (canvi de nivell i canvi de pendent) com de factors de confusió com la tendència general. Addicionalment, es va estudiar la rellevància de la dependència serial entre les dades, així com la variació del rendiment de les tècniques en augmentar la quantitat de mesures disponibles. Els resultats suggereixen que les tècniques que distingeixen millor entre presència i absència d'efecte de la intervenció són el percentatge de no solapament entre les dades i les diferències de mitjanes estandarditzades, índexs que, a més a més, es veuen menys afectats per l'autocorrelació. A banda de la pitjor discriminació entre patrons de dades, una de les tècniques basades en la regressió va mostrar estimacions molt baixes de la grandària de l'efecte, mentre que les dues restants van produir estimacions excessivament altes fins i tot per a tractaments inefectius.

Behavior Modification
Volume 32 Number 6
November 2008 860-875

10.1177/0145445508318866
http://bmo.sagepub.com
hosted at
http://online.sagepub.com

Comparing $N = 1$ Effect Size Indices in Presence of Autocorrelation

Rumen Manolov
Antonio Solanas
University of Barcelona

Generalization from single-case designs can be achieved by replicating individual studies across different experimental units and settings. When replications are available, their findings can be summarized using effect size measurements and integrated through meta-analyses. Several procedures are available for quantifying the magnitude of treatment effect in $N = 1$ designs, and some of them are studied in this article. Monte Carlo simulations were used to generate different data patterns (trend, level change, and slope change). The experimental conditions simulated were defined by the degrees of serial dependence and phase length. Out of all the effect size indices studied, the percentage of nonoverlapping data and standardized mean difference proved to be less affected by autocorrelation and to perform better for shorter data series. The regression-based procedures proposed specifically for single-case designs did not differentiate between data patterns as well as did simpler indices.

Keywords: *single case, AB designs, effect size, autocorrelation*

$N = 1$ designs have been criticized because of their problematic statistical generalizations. A possible solution to this problem consists in replication across participants and settings to establish the generality of the treatment effects. The quantitative integration of these replications can be accomplished by means of meta-analysis. A prior step to integration is summarizing the evidence from each study, a stage in which effect sizes are of maximum relevance. The measurements of the magnitude of effect have gained importance as they overcome *p* values' limitations (Cohen, 1990, 1994; Kirk, 1996; Rosnow & Rosenthal, 1989; Wilkinson & the Task Force on Statistical Inference, 1999). Effect size is an objective measurement of

Authors' Note: Please address correspondence to Rumen Manolov, Passeig de la Vall d'Hebron 171, Barcelona 08035, Spain; rrumenov13@ub.edu.

the strength of the intervention and provides clinical and social researchers with more useful information than the significance level. In contrast to the latter, effect sizes are not systematically affected by sample size (Parker & Brossart, 2003) and focus on the strength of association between the independent and dependent variables instead of centering on the null hypothesis (Kromrey & Foster-Johnson, 1996). Moreover, effect size allows comparison of treatments and is useful for documenting results for posterior meta-analysis and power analysis (Parker & Hagan-Burke, 2007). Another advantage is the possibility of constructing confidence intervals for the effect size (Kirk, 1996).

One of the peculiarities of single-case designs is that they generally include few measurement times (Huitema, 1985). However, several surveys (e.g., Busk & Marascuilo, 1988; Matyas & Greenwood, 1990, 1996; Parker, 2006) have reported that autocorrelation is a common feature of $N = 1$ designs. It has been claimed that even low and statistically nonsignificant levels of autocorrelation can have critical influence on the analytical techniques used (Busk & Marascuilo, 1988; Sharpley & Alavosius, 1988; Suen, 1987; Suen & Ary, 1987). Moreover, empirical findings have suggested that autocorrelation affects a great variety of statistical techniques such as analysis of variance (Toothaker, Banz, Noble, Camp, & Davis, 1983), the binomial test and the split-middle method (Crosbie, 1987), randomization tests (Gorman & Allison, 1996; Sierra, Solanas, & Quera, 2005), and visual analysis (Jones, Weinrott, & Vaught, 1978; Matyas & Greenwood, 1990).

The typical phase length and the likely presence of serial dependence have influenced the lack of consensus about the optimal effect size measurement in single-case research. The most frequent formulas, such as standardized mean differences (e.g., Cohen's *d,* Hedges's *g,* and Glass's Δ) and correlations (e.g., η^2, ω^2, and R^2), have been conceptualized and developed for group designs and focus solely on the average level in the control and treatment conditions. There have also been proposed indices destined specifically for $N = 1$ designs, such as the percentage of nonoverlapping data (PND) or the regression indices (Allison & Gorman, 1993; Center, Skiba, & Casey, 1985-1986; Gorsuch, 1983; White, Rusch, Kazdin, & Hartmann, 1989). PND, as its name suggests, centers on a criterion frequently used in visual inspection, which is still the most commonly applied single-case data analysis technique (Parker, Cryer, & Byrns, 2006). The regression procedures take into account mean levels and the possible slope changes between conditions and also control for

trends not associated with the intervention. The comparison between studies is enhanced by the possibility of converting one type of index into another (Friedman, 1982).

Each of the indices mentioned has its drawbacks: deficient performance in the presence of outliers and trend, ignoring all phase A data points but one (PND); no accounting for changes in slope (Gorsuch's [1983] trend analysis and White et al.'s [1989] *d*); conservativeness, attainment of more than one magnitude of effect index, and impossibility of obtaining a negative *d* (Center, Skiba, & Casey's [1985-1986] procedure); possibility of producing unreliable estimates of trend because of short baseline and overestimation of effect size (Allison & Gorman's [1993] procedure). Regarding the limitations of the latter, which appears to be the conceptually most appropriate one, too-large effect sizes may potentially affect interpretability (Campbell, 2004). With respect to that, Scruggs and Mastropieri (1998) pointed out that an effect size of $d = 3.0$ implies that percentile 50 of the treatment phase corresponds to percentile 99.9 of the baseline phase, making greater values of *d* practically useless. Finally, applied researchers have to keep in mind that when the parametric assumptions of regression-based procedures are not met, the correctness of the effect sizes calculated is not guaranteed. We performed a small revision of scientific literature and found that PND seems to be used more frequently (e.g., Bellini, Peters, Benner, & Hopf, 2007; Mathur, Kavale, Quinn, Forness, & Rutherfod, 1998; Scruggs & Mastropieri, 1994; Scruggs, Mastropieri, Forness, & Kavale, 1988) than regression-based methods (Allison, Faith, & Franklin, 1995; Skiba, Casey, & Center, 1986), probably because of the latter's relatively greater complexity.

The objective of the present investigation was to assess the performance of six proposed measures of effect sizes for AB designs in the presence of different degrees of autocorrelation. The comparison between the indices was done in terms of R^2 (except for PND) because this indicator ranges from 0 to 1 and is easily interpreted as the variance of the dependent variable explained by the change in phase. Because estimating autocorrelation from real data, and testing it for significance, may be problematic (Huitema & McKean, 1991; Matyas & Greenwood, 1990), we decided to test the effect size procedures with data constructed with known parameters (i.e., serial dependence, trend, level change, and slope change), a method that has already been applied in single-case effect size studies (Parker & Brossart, 2003). Another aim was to evaluate the influence of series length, as suggested by Campbell (2004).

Method

Design Selection

We studies two-phase AB designs with different total (N) and phase length (n_A and n_B). Short series were chosen as they are more feasible in applied settings: (a) $N = 10$ and $n_A = n_B = 5$; (b) $N = 15$, $n_A = 5$, and $n_B = 10$; (c) $N = 15$, $n_A = 7$, and $n_B = 8$; (d) $N = 20$, $n_A = 5$, and $n_B = 15$l (e) $N = 20$ and $n_A = n_B = 10$; and (f) $N = 30$ and $n_A = n_B = 15$.

Data Generation

The data for the abovementioned series lengths were generated according to an expression that allows specification of level and slope changes and trend. The statistical model was the same as in previous investigations (e.g., Huitema & McKean, 2000, 2007):

$$y_t = \beta_0 + \beta_1 * T_t + \beta_2 * D_t + \beta_3 * SC_t + \varepsilon_t,$$

where y_t is the value of the dependent variable at moment t; β_0 is the intercept; β_1, β_2, and β_3 are the partial correlation coefficients; T_t is the value of the time variable at moment t (takes values from 1 to N); D_t is the dummy variable for level change (0 for phase A and 1 for phase B); SC_t is the value of the slope change variable, with $SC_t = [T_t - (n_A + 1)] * D_t$ taking 0 for phase A and values from 0 to $(n_B - 1)$ for phase B; and ε_t is the error term. The error term (ε_t) was generated following a first-order autoregressive model: $\varepsilon_t = \varphi_1 * \varepsilon_{t-1} + u_t$. The values of serial dependence (φ_1) ranged from –0.9 to 0.9 in steps of 0.1. The u_t term represents white noise at moment t, and $\varepsilon_1 = u_1$.

The value of the intercept parameter β_0 was set to zero as it does not affect effect size calculation. However, our goal was to guarantee suitable comparisons between experimental conditions. Therefore, it was important that the two types of effects (i.e., level change associated with parameter β_2 and slope change associated with β_3) and trend (extraneous variable associated with parameter β_1) produce comparable mean differences between phase B and phase A. First, two criteria were chosen: (a) for *series length,* the shortest design, $n_A = n_B = 5$, was chosen to explore whether longer series imply better effects detection, and (b) for the *partial correlation coefficient,* level change (β_2) was selected as it remains constant throughout the whole intervention phase. As the u_t term was generated following $\mathbf{N}(0,1)$, the phase A values approximate zero ($y_{Ai} \approx 0$). Being present, a level change of

β_2, $y_{Bi} = y_{Ai} + \beta_2 = 0 + \beta_2 = \beta_2$. $\beta_2 = 0.3$ was chosen as it proved to avoid floor and ceiling effects (i.e., R^2 not approaching 0 or 1, respectively). The change in slope produces $(n_B - 1)$ increments, and it was necessary to find a β_3 value so that the median phase B point would be equal to β_2, which will make the phase B mean also equal to β_2. As $\overline{y}_B - \overline{y}_A = \beta_2$, a β_3 value implying the same mean difference can be calculated as

$$\beta_3 = \frac{\beta_2}{\frac{n_B - 1}{2}},$$

which for $\beta_2 = 0.3$ leads to

$$\beta_3 = \frac{0.3}{\frac{5-1}{2}} = \frac{0.6}{4} = 0.15.$$

As trend involves increments from the first observation, the accomplishment of the $\overline{y}_B - \overline{y}_A = \beta_2$ criterion required meeting the following equality $y_{Bi} - y_{Ai} = \beta_2$. The needed β_1 value can be found as

$$\beta_1 = \frac{\beta_2}{\frac{n_A + n_B}{2}},$$

which for $\beta_2 = 0.3$ leads to

$$\beta_1 = \frac{0.3}{\frac{5+5}{2}} = \frac{0.6}{10} = 0.06.$$

We could verify that the β_1 and β_3 values are appropriate for producing β_2 mean differences even for the most extreme levels of serial dependence (−0.9 and 0.9) whenever $n_A = n_B$. In total, there were eight data patterns studied, defined by the presence and combination of trend, level change, and slope change (i.e., β_1, β_2, and β_3 being equal to or different from zero).

It is likely that for series with high negative autocorrelation, unstable baselines will be obtained. Therefore, we used a large number of iterations to ensure that the indices' comparison did not depend on few clinically improbable data sets.

The 50 numbers previous to each simulated data series were eliminated to reduce artificial effects (Greenwood & Matyas, 1990) and to avoid dependence between successive data series (Huitema, McKean, & McKnight, 1999).

Analysis

We calculated the effect size for each experimental condition using the following indices.

Percentage of nonoverlapping data:

1. Calculate the number of phase B data points that exceed the highest data point in phase A. Simulating increases in behavior with the introduction of treatment ensures that this step is appropriate.
2. Divide the value obtained in Step 1 by the number of observations in Phase B and multiply by 100 to convert the proportion into a percentage.

Cohen's *d*:

1. Obtain the difference between the means of both phases: $\overline{y}_B - \overline{y}_A$.
2. Calculate the standard deviation of each phase.
3. Divide the value obtained in Step 1 by the phase A standard deviation or by the pooled standard deviation (obtaining d_A and d_{AB}, respectively).
4. Convert *d* to R^2, using $R^2 = \frac{d^2}{d^2 + 4}$

Gorsuch's (1983) trend analysis:

1. Calculate a simple linear regression using time ($T = 1, 2, \ldots, n$) as a predictor variable and the original dependent variable: $\boldsymbol{Y = a + b_t * T + u_t}$.
2. Calculate a simple linear regression using the treatment variable ($X = 0$ for phase A and $X = 1$ for phase B) as a predictor and the residual of the Step 1 regression as a dependent variable: $\textbf{residual}(\boldsymbol{Y}) = \boldsymbol{a + b_x * X + u_t}$.
3. Calculate R^2 as the sum of squares explained by the Step 2 model divided by the total sum of squares.

White et al.'s *d* (1989, using the correction in Faith, Allison, & Gorman, 1996):

1. Calculate a simple linear regression using phase A data and the time variable as predictor.
2. Use the Step 1 regression coefficients (intercept and slope) to obtain the predicted value of the dependent variable for the last day of phase B; this value is called $\hat{y}_A$.
3. Calculate a simple linear regression using phase B data and the time variable as predictor.
4. Use the Step 3 regression coefficients (intercept and slope) to obtain the predicted value of the dependent variable for the last day of phase B; this value is called $\hat{y}_B$
5. Calculate the difference $\hat{y}_B - \hat{y}_A$, which represents the numerator in White et al.'s (1989) formula.
6. Calculate the pooled standard deviation of phases A and B.

7. Calculate the Pearson product–moment correlation coefficient between the dependent variable and the time variable.
8. Calculate d through the expression

$$d = \frac{\hat{y}_B - \hat{y}_A}{\sqrt{(1 - r^2) * \sqrt{(s_A^2 + s_B^2)/2}}}.$$

9. Convert d to R^2.

Allison and Gorman (1993):

1. Calculate a simple linear regression using phase A data and the time variable as predictor: $\boldsymbol{Y_A = b_0 + b_1 * T_A + e}$.
2. Calculate the predicted values for Y and the residuals for both phases.
3. Calculate zero-order correlations between the treatment variable X ($X = 0$ for phase A and $X = 1$ for phase B) and residual (Y), on one hand, and between $X * T$ and residual (Y), on the other. If both correlations share the same sign, then proceed with Step 4. Otherwise, go to Step 6.
4. Calculate a multiple linear regression with the treatment variable X and $X * T$ as predictors: $\boldsymbol{\text{residual}(Y) = b_0 + b_1 * X + b_3 * X * T + e}$.
5. Obtain the adjusted R^2 for the Step 4 equation.
6. In case the zero-order correlations associated with level and slope have different signs, it is only necessary to estimate the effect of the treatment variable X through a simple linear regression, as the change in slope will attenuate this effect. Obtain the adjusted R^2.

Simulation

The specific steps that were implemented in the Fortran programs (one for each of the six series lengths) were the following:

1. Systematically select each of the 19 degrees of serial dependence.
2. Systematically select the (β_1, β_2, and β_3) parameters for data generation: $2^3 = 8$ data patterns – autoregressive model; trend; level change; slope change; trend and level change; trend and slope change; level and slope change; and trend, level, and slope change.
3. One hundred thousand iterations of Steps 4 through 17.
4. Generate an array with $50 + N$ data following a normal distribution with mean zero and unitary standard deviation by means of NAG*fl90* mathematical-statistical libraries (specifically external subroutines *nag_rand_seed_set* and *nag_rand_normal*).
5. Eliminate the first 50 numbers.
6. Assign the following N numbers to array u_t.

7. Establish $\varepsilon_1 = u_1$.
8. Obtain the array of ε_t using the equation $\varepsilon_t = \varphi_1 * \varepsilon_{t-1}$.
9. Obtain the time array $T_t = 1, 2, \ldots, N$.
10. Obtain the dummy treatment variable array D_t, where $D_t = 0$ for phase A and $D_t = 1$ for phase B.
11. Obtain the slope change array according to Huitema and McKean's (2007) expression: $SC_t = [T_t - (n_A + 1)] * D_t$ used for data generation.
12. Obtain the slope change array $T_t * D_t$ according to Allison and Gorman's (1993) procedure used in the effect size computation.
13. Obtain the y_t array containing measurements (i.e., dependent variable) following Huitema and McKean's (2007) model: $y_t = \beta_0 + \beta_1 * T_t + \beta_2 * D_t + \beta_3 * SC_t + \varepsilon_t$.
14. Calculate the PND.
15. Calculate effect size according to the two versions of Cohen's d (d_A and d_{AB}). Convert d to R^2.
16. Calculate effect size (R^2) according to Gorsuch's (1983) trend analysis.
17. Calculate White et al.'s (1989) d and convert to R^2.
18. Calculate effect size (adjusted R^2) according to Allison and Gorman's (1993) procedure. NAG*fl90* libraries external subroutine *nag_mult_lin_reg* was used to obtain the multiple regression coefficients.
19. Average the obtained R^2 from the 100,000 replications of each experimental condition.

During program elaboration, the appropriate performance of the programs was verified through comparisons with the output of statistical packages and with the examples presented in Faith et al. (1996).

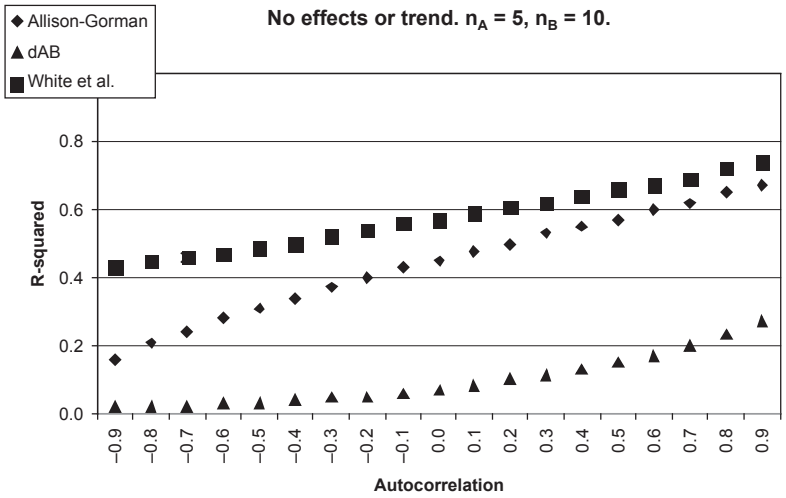
Results

Because of the low magnitude of effect estimates produced by Gorsuch's (1983) trend analysis, we do not comment on this procedure in the following sections. The values, ranging from 0.01 to 0.06 for all experimental conditions and concurring with Parker and Brossart's (2003) results, show the influence of autocorrelation and the zero sensitivity to the differential data patterns.

Autocorrelation Effect

To explore the effect produced by the presence of serial dependence in data, we constructed figures crossing each of the six effect size indices with the eight data patterns. In each of these 6 * 8 = 48 figures, degree of autocorrelation is placed on the abscissa and the index value (R^2 or percentage)

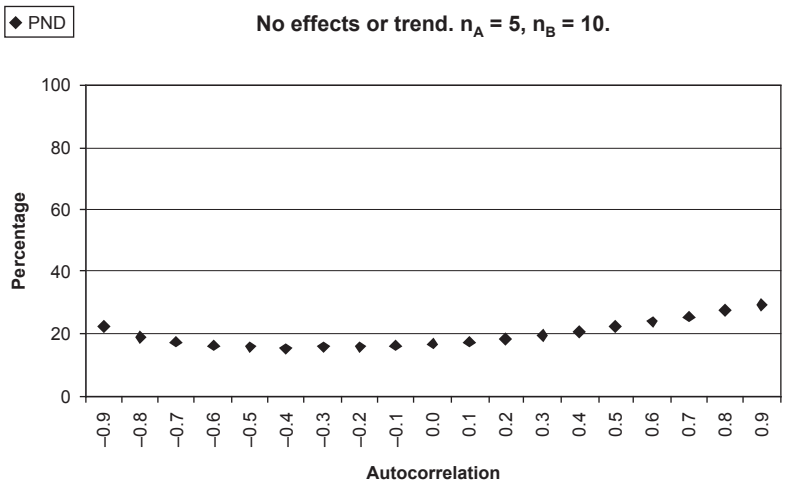
Figure 1
Autocorrelation Effect on Different Effect Size Measures



on the ordinate, superimposing the different phase lengths. Visual inspection for simpler data patterns (i.e., when none or only one type of effect is present) showed that negative serial dependence is associated with lower R^2 values, whereas positive serial dependence correlates with higher effect size estimates. There appears to be an approximately linear relation between φ_1 and R^2. Figure 1 compares several techniques and illustrates the fact that for Cohen's *d* we observed a greater increment in R^2 for positive autocorrelation ($0.0 \leq \varphi_1 \leq 0.9$) than for negative autocorrelation ($-0.9 \leq \varphi_1 \leq 0.0$). As Figure 2 shows, for PND there is a nonlinear relation between autocorrelation and the effect size measurement, which in this case, because of the peculiarities of the index, is the percentage itself rather than an R^2.

Comparing differences in R^2 between high negative ($\varphi_1 = -0.9$) and zero autocorrelation, on one hand, and high positive ($\varphi_1 = 0.9$) and zero autocorrelation, on the other, it appears that White et al.'s (1989) *d* and Allison and Gorman's (1993) procedure are the most affected ones, whereas Cohen's *d* and PND are less sensitive to serial dependence. When the data pattern is more complex (i.e., including different types of effect and/or trend), the effect of autocorrelation becomes curvilinear and the R^2 variation diminishes for all indices.

Figure 2
Autocorrelation Effect on the Effect Size Calculated Through the Percentage of Nonoverlapping Data

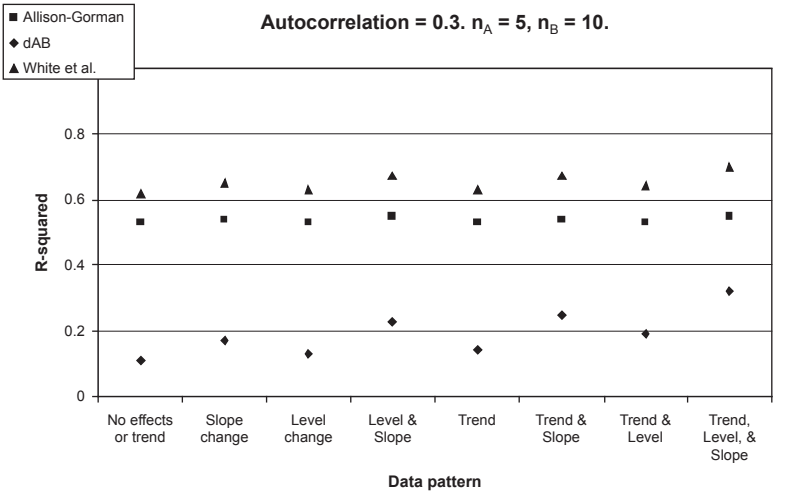


Effect of Data Pattern

We carried out the exploration of data patterns' detection by constructing graphs combining the six procedures (PND, Cohen's [1990, 1994] d_A and d_{AB}, Gorsuch's [1983] trend analysis, White et al.'s [1989] d, and Allison & Gorman's [1993] procedure) for computing the magnitude of effect with the six series lengths. In each of these 6 * 6 = 36 graphs, we put data patterns in the abscissa and the effect size index (R^2 or percentage) in the ordinate, superimposing several autocorrelation levels. The ideal pattern of effect detection would be represented by greater effect sizes for combined level and slope change, followed by second greater values for each of those effects separately and smaller values for data with no effect. A perfect index would not be affected by general trend not related to treatment's introduction. Therefore, greater discrepancy in R^2 or percentage between effects of interest and the remaining conditions meant better differentiation and indicated a more desirable performance.

The visual inspection carried out following those criteria suggests that the regression-based indices differentiate data patterns only for long and balanced series ($n_A = n_B = 10$ or 15) and also produce greater R^2. d_A and d_{AB}

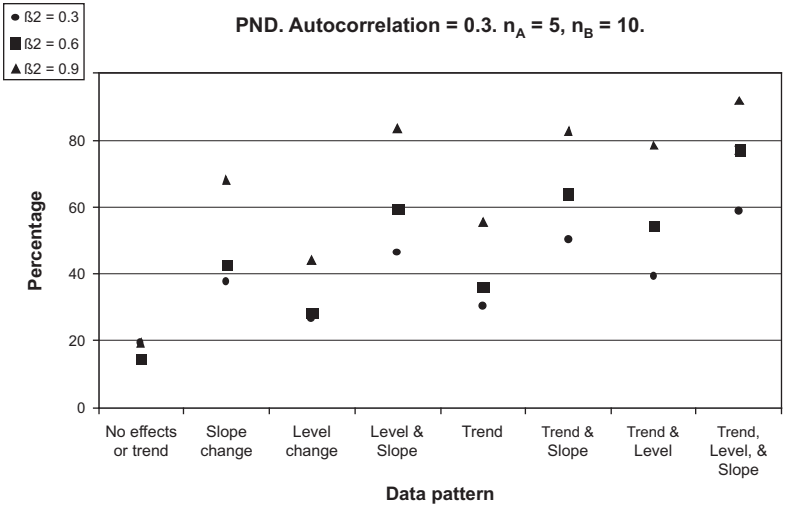
Figure 3
Effect Sizes Calculated for Different Data Patterns Through Two Regression-Based Indices and One Standardized Mean Difference Index



differentiate more than White et al.'s (1989) and Allison and Gorman's (1993) indices, d_{AB} being the index that produces lower estimates of the magnitude of effect. PND proved to be the measurement that most detected the differences between patterns even for short series ($n_A = n_B = 5$). A common problem of PND and the standardized mean differences is that they produce greater effect sizes in the presence of trend (extraneous variable) than in the presence of level change (intervention effect). As expected, complex patterns are associated with greater effect sizes for all indices.

As shown in Figure 3, Cohen's *d*s are more sensitive to differential patterns. Nevertheless, the effect size values obtained through d_A and d_{AB} are smaller than the ones obtained via the regression-based procedures. Thus, the former indices have a lower probability of producing great effect sizes in the absence of effects, a finding that becomes more evident in longer series. Figure 4 illustrates the higher differentiation between patterns accomplished by PND—the index that seemed to better approximate the ideal pattern described above. The figures show examples for $\varphi_1 = 0.3$, as it represents a level of serial dependence likely to be found in behavioral data (Parker, 2006), but the abovementioned tendencies were found for all φ_1 values simulated.

Figure 4
Effect Sizes Calculated for Different Data Patterns by Means of the Percentage of Nonoverlapping Data (PND)



Series Length Effect

Analysis of the results revealed that incrementing series length leads to a higher differentiation between the data patterns. This, however, does not imply obtaining greater R^2. Actually, we found that simple patterns (containing only one type of effect) produced higher estimations for $n_A = n_B =$ 5 and $n_A = 5$, $n_B = 15$ than for $n_A = n_B = 10$ and 15. Consistent with the data simulation method, we obtained greater effect sizes for the (incremental) change in slope than for the (constant) change in level. As mentioned earlier, for the regression-based indices the values of n_A and n_B (and the relation between those) are relevant as they affect pattern distinction.

Discussion

The purpose of the present study was to explore performance of different effect size indices applied to data with known parameters. In applied settings, it is common to have only few behavioral measurements that can be

sequentially related. Therefore, the most useful indices to summarize magnitude of the treatment effect will be the ones sensitive to effects in short data series and less affected by serial dependence. Out of the indices studied, the ones that performed better in the aforementioned terms were PND and standardized mean differences (d_A and d_{AB}). Other advantages of these indices are calculus easiness and the fact that they are more widely known (especially *d*) in comparison to regression-based procedures—a feature that might make them more attractive to applied researchers with lower degrees of expertise in statistics. These indices better differentiate between the distinct data patterns and appear to have lower probability of false alarms in absence of treatment effects, but their results are distorted by trend. Hence, visual inspection can be used to detect trend and outliers before deciding whether *d* and PND are appropriate effect size measures. A modification in the latter index will enable its application in cases in which reduction rather than increment in the behavior of interest is expected. Recent proposals related to the PND are the percentage of data points exceeding the mean (Ma, 2006) and the percentage of all nonoverlapping data (Parker, Hagan-Burke, & Vannest, 2007), and their properties require further research.

It was surprising to find that the more sophisticated indices conceptualized for single-case designs (i.e., taking into account trend, level, and slope change) performed worse than simpler and theoretically less appropriate strategies. Thus, future investigation is necessary to improve regression-based indices. Meanwhile, the use of simpler indices in $N = 1$ designs can be recommended whenever complementary information about trend is also taken into consideration. A possible source for additional information is visual analysis, which can enhance the choice of an appropriate effect size index and validate the results obtained by it (Parker et al., 2006).

Among limitations of the study, we have to mention that only AB designs were studied because of their applicability in nonreversal behaviors. Nevertheless, the results presented here can also be useful for multiple-baseline designs for which there can be an effect size computed for each baseline (Busse, Kratochwill, & Elliott, 1995).

It has to be commented that values of β_1, β_2, and β_3 were not extracted from a previously published investigation because of the lack of indication in scientific literature. Apart from the β values discussed, we also tried $\beta_2 = 0.6$ and $\beta_2 = 0.9$, varying the β_1 and β_3 values according to the formulas presented. Very similar results were obtained, and as expected, all procedures showed greater discrimination between patterns (Figure 4 shows an example for one of the best-performing indices). Nevertheless, future studies may continue exploring the optimal values of β_1, β_2, and β_3 for simulating different

magnitudes of different data patterns. Another possible line of research is the application of the effect size indices to more-phased designs (e.g., ABAB), which are more suitable for controlling extraneous variables. In such a study, it would be interesting to explore the variations in effect size as a function of how it was calculated: (a) from phases A_1 and B_1; (b) from phases A_1 and B_2; (c) using the means of both A and both B phases; or (d) calculating an effect size for each change in phase.

References

Allison, D. B., Faith, M. S., & Franklin, R. (1995). Antecedent exercise in the treatment of disruptive behavior: A review and meta-analysis. *Clinical Psychology: Science and Practice, 2*, 279-303.

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy, 31*, 621-631.

Bellini, S., Peters, J. K., Benner, L., & Hopf, A. (2007). A meta-analysis of school-based social skills interventions for children with autism spectrum disorders. *Remedial and Special Education, 28*, 153-162.

Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment, 10*, 229-242.

Busse, R. T., Kratochwill, T. R., & Elliott, S. N. (1995). Meta-analysis for single-case consultation outcomes: Applications to research and practice. *Journal of School Psychology, 33*, 269-285.

Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification, 28*, 234-246.

Center, B. A., Skiba, R. J., & Casey, A. (1985-1986). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education, 19*, 387-400.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Crosbie, J. (1987). The inability of the binomial test to control Type I error with single-subject data. *Behavioral Assessment, 9*, 141-150.

Faith, M. S., Allison, D. B., & Gorman, D. B. (1996). Meta-analysis of single-case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 245-277). Mahwah, NJ: Lawrence Erlbaum.

Friedman, H. (1982). Simplified determinations of statistical power, magnitude of effect and research sample sizes. *Educational and Psychological Measurement, 42*, 521-526.

Gorman, B. S., & Allison, D. B. (1996). Statistical alternatives for single-case designs. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 159-214). Mahwah, NJ: Lawrence Erlbaum.

Gorsuch, R. L. (1983). Three methods for analyzing limited time-series (N of 1) data. *Behavioral Assessment, 5*, 141-154.

Greenwood, K. M., & Matyas, T. A. (1990). Problems with application of interrupted time series analysis for brief single-subject data. *Behavioral Assessment, 12*, 355-370.

Huitema, B. E. (1985). Autocorrelation in behavior analysis: A myth. *Behavioral Assessment, 7*, 107-118.

Huitema, B. E., & McKean, J. W. (1991). Autocorrelation estimation and inference with small samples. *Psychological Bulletin, 110*, 291-304.

Huitema, B. E., & McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement, 60*, 38-58.

Huitema, B. E., & McKean, J. W. (2007). An improved portmanteau test for autocorrelated errors in interrupted time-series regression models. *Behavior Research Methods, 39*, 343-349.

Huitema, B. E., McKean, J. W., & McKnight, S. (1999). Autocorrelation effects on least-squares intervention analysis of short time series. *Educational and Psychological Measurement, 59*, 767-786.

Jones, R. R., Weinrott, M. R., & Vaught, R. S. (1978). Effects of serial dependence on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis, 11,* 277-283.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746-759.

Kromrey, J. D., & Foster-Johnson, L. (1996). Determining the efficacy of intervention: The use of effect sizes for data analysis in single-subject research. *Journal of Experimental Education, 65*, 73-93.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject research: Percentage of data points exceeding the median. *Behavior Modification, 30*, 598-617.

Mathur, S. R., Kavale, K. A., Quinn, M. M., Forness, S. R., & Rutherford, R. B., Jr. (1998). Social skills interventions with students with emotional and behavioral problems: A quantitative synthesis of single-subject research. *Behavioral Disorders, 23*, 193-201,

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis for single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23*, 341-351.

Matyas, T. A., & Greenwood, K. M. (1996). Serial dependency in single-case time series. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 215-243). Mahwah, NJ: Lawrence Erlbaum.

Parker, R. I. (2006). Increased reliability for single-case research results: Is bootstrap the answer? *Behavior Therapy, 37*, 326-338.

Parker, R. I., & Brossart, D. F. (2003). Evaluating single-case research data: A comparison of seven statistical methods. *Behavior Therapy, 34*, 189-211.

Parker, R. I., Cryer, J., & Byrns, G. (2006). Controlling baseline trend in single-case research. *School Psychology Quarterly, 21*, 418-443.

Parker, R. I., & Hagan-Burke, S. (2007). Useful effect size interpretations for single case research. *Behavior Therapy, 38*, 95-105.

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data: An alternative to PND. *Journal of Special Education, 40*, 194-204.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.

Scruggs, T. E., & Mastropieri, M. A. (1994). The utility of the PND statistic: A reply to Allison and Gorman. *Behaviour Research and Therapy, 32*, 879-883.

Scruggs, T. E., & Mastropieri, M. A. (1998). Summarizing single-subject research: Issues and applications. *Behavior Modification, 22*, 221-242.

Scruggs, T. E., Mastropieri, M. A., Forness, S. R., & Kavale, K. A. (1988). Early language intervention: A quantitative synthesis of single-subject research. *Journal of Special Education, 20*, 259-283.

Sharpley, C. F., & Alavosius, M. P. (1988). Autocorrelation in behavior data: An alternative perspective. *Behavior Assessment, 10*, 243-251.

Sierra, V., Solanas, A., & Quera, V. (2005). Randomization tests for systematic single-case designs are not always appropriate. *Journal of Experimental Education, 73*, 140-160.

Skiba, R. J., Casey, A., & Center, B. A. (1986). Nonaversive procedures in the classroom behavior problems. *Journal of Special Education, 19*, 459-481.

Suen, H. K. (1987). On the epistemology of autocorrelation in applied behavior analysis. *Behavioral Assessment, 9,* 113-124.

Suen, H. K., & Ary, D. (1987). Autocorrelation in behavior analysis: Myth or reality? *Behavioral Assessment, 9*, 150-130.

Toothaker, L. E., Banz, M., Noble, C., Camp, J., & Davis, D. (1983). N = 1 designs: The failure of ANOVA-based tests. *Journal of Educational Statistics, 4*, 289-309.

White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta-analysis in individual subject research. *Behavioral Assessment, 11*, 281-296.

Wilkinson, L., & The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 694-704.

Rumen Manolov is a PhD student at the Faculty of Psychology at the University of Barcelona. His investigation is focused on the analysis of data obtained from single-case designs.

Antonio Solanas, PhD, is full professor at the Faculty of Psychology at the University of Barcelona. He is also the director of the Department of Behavioral Sciences Methods. His main research interests include single-case designs analysis, social reciprocity measurement, and multivariate data analysis methods.

Estudi 2©

Comparing "visual" effect size indices for single-case designs

Rumen Manolov, Antonio Solanas i David Leiva

Departament de Metodologia de les Ciències del Comportament
Facultat de Psicologia
Universitat de Barcelona

Resum. Aquest article es centra en l'índex (percentatge de no solapament entre les dades; PND) que va mostrar millor rendiment en l'Estudi 1 inclòs en la present tesi doctoral. Degut a certes limitacions presentes en PND, s'han proposat dues tècniques relacionades (PEM – percentatge de dades superiors a la mediana i PAND – percentatge de no solapament entre totes les dades) que pretenen assolir un rendiment més òptim en presència de tendència en les dades. L'Estudi 2 té l'objectiu de comprovar si aquestes tècniques representen una millora, tal i com suggereixen els seus autors. La comparació es duu a terme en el context de dades amb característiques conegudes (presència o absència d'efecte i de tendència, nivell d'autocorrelació). Els resultats indiquen que PND és l'índex que discrimina millor entre els diferents patrons de dades, és a dir, mostra una distància més gran entre les estimacions de la grandària de l'efecte en els casos de tractament efectiu i no efectiu. PEM és una tècnica menys afectada per la dependència serial, específicament en absència d'efecte de la intervenció. S'ha de remarcar que fins i tot tècniques molt semblants comporten estimacions diferents de la magnitud de l'efecte del tractament.

Original Article

Comparing "Visual" Effect Size Indices for Single-Case Designs

Rumen Manolov, Antonio Solanas, and David Leiva

Department of Behavioral Sciences Methods, University of Barcelona, Spain

Abstract. Effect size indices are indispensable for carrying out meta-analyses and can also be seen as an alternative for making decisions about the effectiveness of a treatment in an individual applied study. The desirable features of the procedures for quantifying the magnitude of intervention effect include educational/clinical meaningfulness, calculus easiness, insensitivity to autocorrelation, low false alarm, and low miss rates. Three effect size indices related to visual analysis are compared according to the aforementioned criteria. The comparison is made by means of data sets with known parameters: degree of serial dependence, presence or absence of general trend, and changes in level and/or in slope. The percent of nonoverlapping data showed the highest discrimination between data sets with and without intervention effect. In cases when autocorrelation or trend is present, the percentage of data points exceeding the median may be a better option to quantify the effectiveness of a psychological treatment.

Keywords: single-case, AB designs, effect size, autocorrelation

Single-case designs present problems for both data analysis of the specific study and quantitative integration of different studies. Replicating across subjects and settings in order to obtain evidence on the strength of the intervention is useful only when there are summary measures available to be used in meta-analyses.

The difficulties in single-case designs' analysis are related to the scarce number of observations usually available (Huitema, 1985) and to the serial dependence between the measurements obtained from the same experimental unit (Busk & Marascuilo, 1988; Matyas & Greenwood, 1991, 1997; Parker, 2006). Whether being statistically significant or not, autocorrelation has been alleged to affect the analytical techniques employed (Busk & Marascuilo, 1988; Sharpley & Alavosius, 1988; Suen, 1987; Suen & Ary, 1987). Scientific evidence points out that serial dependence alters the performance of procedures as diverse as ANOVA (Toothaker, Banz, Noble, Camp, & Davis, 1983), the split-middle method (Crosbie, 1987), and randomization tests (Gorman & Allison, 1997; Sierra, Solanas, & Quera, 2005). On the other hand, for determining the effectiveness of a treatment in an individual study it is not sufficient to obtain a p value, due to the disadvantages of this indicator (Cohen, 1990, 1994; Kirk, 1996; Rosnow & Rosenthal, 1989; Wilkinson & The Task Force on Statistical Inference, 1999). Clinical, educational, and social researchers need more meaningful information than the one provided by the statistical significance. Visual analysis, as an alternative, is more subjective and does not allow quantification. Moreover, it has been found to be distorted by the presence of serial dependence (Jones, Weinrott, & Vaught, 1978; Matyas & Greenwood, 1990). An objective measurement that can be used to quantify the relationship between the treatment and the behavior of interest is effect size.

In contrast with p values, effect size indices are useful for documenting results for posterior meta-analysis and power analysis (Parker & Hagan-Burke, 2007b). Among the advantages of effect size, the following have been stated: (a) it is not systematically affected by sample size (Parker & Brossart, 2003); (b) it uses on the strength of association between the independent and the dependent variables, instead of centering on the null hypothesis (Kromrey & Foster-Johnson, 1996); (c) it allows treatments' comparison (Parker & Hagan-Burke, 2007b); and (d) it is possible to construct confidence intervals about the effect size (Kirk, 1996).

The most widely known effect size indices based on standardized mean differences (e.g., Cohen's d; Hedges' g; and Glass' Δ) and measurements of association (e.g., η^2; ω^2; and R^2) were not developed for single-case designs but rather for designs involving groups' comparison and, thus, focus only on the average levels of behavior in the different conditions. Nonetheless, there are also procedures conceptualized for $N = 1$ designs – some of them based on regression analysis and others closely related to visual analysis. It is possible to convert some effect size indices into others (Friedman, 1982), allowing the comparison between meta-analyses using different measures. The bibliographic search we performed suggests that visually based indices are applied more often (e.g., Bellini, Peters, Benner, & Hopf, 2007; Mathur, Kavale, Quinn, Forness, & Rutherfod, 1998; Scruggs & Mastropieri, 1994; Scruggs, Mastropieri, Forness, & Kavale, 1988) than regression-based methods (Allison, Faith, & Franklin, 1995; Skiba, Casey, & Center, 1986) in meta-analyses. This could be due to the advantages of visual indices, such as calculus easiness and increased interpretability from clinical and educational perspective.

DOI: 10.1027/1614-2241/a000006

Regression-Based Effect Size Indices

The regression-based procedures incorporate predictor variables in order to model changes in level and in slope and also try to control for extraneous variables such as trends. The following procedures are some of the most studied ones in scientific literature:

1) Gorsuch's (1983) trend analysis includes time as covariate and eliminates its influence prior to testing for level change.
2) White, Rusch, Kazdin, and Hartmann's (1989) *d*, taking into consideration the correction presented in Faith, Allison, and Gorman (1997), compares two predicted values – the last treatment phase point according to baseline phase regression equation with the last treatment phase point as predicted by the treatment phase regression equation. The model also takes into account the possible relation between time and the measured behavior.
3) Center, Skiba, and Casey's (1985–1986) model, in contrast with the abovementioned procedures, can account for both changes in level and slope, while controlling for the presence of trend. Among the limitations of this procedure have been stated the attainment of more than one magnitude of effect index and the impossibility to obtain a negative *d*.
4) Allison and Gorman's (1993) model pretends to improve the previous technique, estimating trend solely from the baseline phase and allowing the correspondence between the type of treatment effect (i.e., reducing or increasing the behavior of interest) and the sign of the effect size index (negative or positive, respectively). A shortcoming of the model is the possible effect size overestimation.

Common drawbacks of the regression-based procedures are the parametric assumptions, while there is also evidence that despite their conceptual appropriateness those models do not perform as well as simpler indices (Manolov & Solanas, 2008).

Visual Effect Size Indices

These effect size indices are based on a criterion employed in visual analysis in order to decide the effectiveness of a treatment – the amount of overlap between the data points pertaining to baseline and treatment phases. Their attractiveness to applied researchers is related to calculation easiness and to the fact that visual inspection is still the most commonly applied single-case data analysis technique (Parker, Cryer, & Byrns, 2006). Some of the procedures proposed for using in psychological studies are:

1) Scruggs, Mastropieri, and Casto's (1987) percent of nonoverlapping data (PND). PND is based on the proportion of treatment phase measurements greater than the highest baseline phase data point. It has been criticized for ignoring all phase A data points except for one, a reason for which the following two indices were proposed.
2) Ma's (2006) percentage of data points exceeding the median (PEM). PEM was proposed to correct some of the potential drawbacks of PND, like the sensitivity to floor or ceiling effects, while maintaining its advantages. As its name suggests, this index computes the percentage of treatment measurements greater than the baseline phase median.
3) Parker, Hagan-Burke, and Vannest's (2007) percentage of all nonoverlapping data (PAND). PAND was introduced as an alternative to PND for larger data sets. It takes into account all data points and counts the minimum number of measurements that need to be removed in order to obtain series with no overlap. The ratio between the remaining data points and series' length is the basis of the index. The authors also suggest that the index can be converted into a *Phi* effect size index or an improvement rate difference.

The objective of the present study was to extend the scientific literature (e.g., Parker & Hagan-Burke, 2007a) assessing the performance of the three measures of effect sizes for AB designs in presence of different degrees of autocorrelation. We aimed to explore which index discriminates better between the distinct data patterns, while an additional purpose was to evaluate the influence of series' length, following Campbell's (2004) suggestions. As the estimation and hypothesis testing of serial dependence from real data can be problematic (Huitema & McKean, 1991; Matyas & Greenwood, 1991), we decided to test the effect size procedures with data constructed with known parameters (i.e., serial dependence, trend, level change, and slope change), a method that has already been applied in single-case effect size studies (Manolov & Solanas, 2008; Parker & Brossart, 2003).

Method

Design Selection

The study focused on AB designs with several series' lengths (N) and phase lengths (n_A and n_B), short enough to be feasible in applied settings where the temporal cost has to be taken into consideration. We chose the following values in order to cover a range of possible "short series:"

a) $N = 10$; $n_A = n_B = 5$.
b) $N = 15$; $n_A = 5$; $n_B = 10$.
c) $N = 15$; $n_A = 7$; $n_B = 8$.
d) $N = 20$; $n_A = 5$; $n_B = 15$.
e) $N = 20$, $n_A = n_B = 10$.
f) $N = 30$, $n_A = n_B = 15$.

Data Generation

For each series' length we generated data sets with different patterns, defined by the presence or absence of general trend, level change and/or in slope. The statistical model used was suggested by Huitema and McKean (2000, 2007):

$y_t = \beta_0 + \beta_1 \times T_t + \beta_2 \times D_t + \beta_3 \times SC_t + \varepsilon_t$, where:
y_t: the value of the dependent variable at moment t;
β_0: intercept;
β_1: coefficient associated with general trend;
β_2: coefficient associated with level change;
β_3: coefficient associated with slope change;
T_t: value of the time variable at moment t (takes values from 1 to N);
D_t: dummy variable for level change. For Phase A it was set to 0 and for Phase B to 1;
SC_t: value of the slope change variable, computed as $[T_t - (n_A + 1)] \times D_t$, so that it is equal to 0 for phase A, and takes values from 0 to $(n_B - 1)$ for phase B; and
ε_t: error term.

The error term (ε_t) was generated following a first-order autoregressive model: $\varepsilon_t = \varphi_1 \times \varepsilon_{t-1} + u_t$. The values of serial dependence (φ_1) ranged from −0.9 to 0.9 in steps of 0.1. The u_t term represents white noise at moment t generated following N (0, 1) and $\varepsilon_1 = u_1$.

The value of the intercept parameter β_0 was set to zero as it does not affect effect size calculation. In order to ensure the adequacy of the comparison between experimental conditions, we chose the values of β_1, β_2, and β_3 so that they produce comparable mean differences between the two phases. We chose to set first the β_2 parameter, as the level change is maintained constant throughout the whole intervention phase. Afterwards, we set the values of β_1 and β_3 leading to the same difference $\bar{y}_B - \bar{y}_A$. Those steps were initially carried out for the shortest series (i.e., $n_A = n_B = 5$) in order to explore if longer series imply better discrimination of data patterns. We tested several values for β_2 (from 0.1 to 0.6 in steps of 0.1) for all experimental conditions seeking its most appropriate value. We found that for $\beta_2 = 0.1$ the values of PND were all too low, while for $\beta_2 = 0.6$ PEM was close to reaching its maximum value. To avoid the floor and ceiling effects (see Figure 1), which make impossible data pattern discrimination, we decided to set β_2 to 0.3.

The use of $\beta_2 \neq 0$ implies that $\bar{y}_B - \bar{y}_A = \beta_2$ if the other parameters are set to zero. The value of β_3 that leads to the same mean difference can be found through the following expression:

$$\beta_3 = \frac{\beta_2}{\frac{n_B - 1}{2}}, \text{ which for } \beta_2 = 0.3 \text{ leads to } \beta_3 = \frac{0.3}{\frac{5-1}{2}} = \frac{0.6}{4} = 0.15,$$

while the appropriate value of β_1 is obtained as

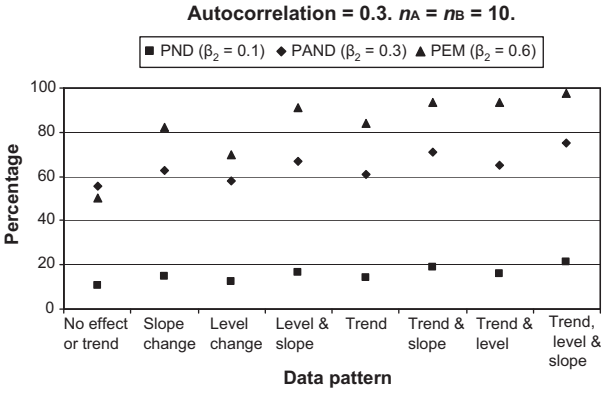


Figure 1. Influence of the simulation parameters β on the effect size indices.

$$\beta_1 = \frac{\beta_2}{\frac{n_A + n_B}{2}}, \text{ which for } \beta_2 = 0.3 \text{ leads to } \beta_1 = \frac{0.3}{\frac{5+5}{2}} = \frac{0.6}{10} = 0.06.$$

We could verify that the β_1 and β_3 values are appropriate for producing mean differences equal to the value of β_2 even for the most extreme levels of serial dependence (−0.9 and 0.9), whenever $n_A = n_B = 5$. In total there were eight data patterns studied, defined by the presence and combination of trend, level change, and slope change (i.e., β_1, β_2, and β_3 being equal to or different from zero).

Finally, in order to guarantee suitable simulated data, the 50 values previous to each simulated data series were eliminated in order to reduce artificial effects (Greenwood & Matyas, 1990) and to avoid dependence between successive data series (Huitema, McKean, & McKnight, 1999).

Analysis

Prior to presenting in detail the steps needed to compute the three effect size indices included in the present study, an example of a fictitious data set is presented. Consider a psychological study applying the Parent-Child Interaction Therapy (for an in-depth description see Borrego, Anhalt, Terao, Vargas, & Urquiza, 2006) in which the number of praises a parent directs to a child is registered 5 days prior to treatment introduction and 5 days during intervention. The data gathered using the AB design structure (4, 5, 3, 6, and 3 praises during baseline and 7, 5, 8, 9, and 7 praises during treatment phase) can be represented graphically as shown in Figure 2. In the following section, each of the procedures is applied to the data set presented in order to illustrate their calculus.

We calculated the effect size for each experimental condition using the following indices:

PND

1) Identify the highest measurement in Phase A. In the example it is 6 praises corresponding to baseline day 4.

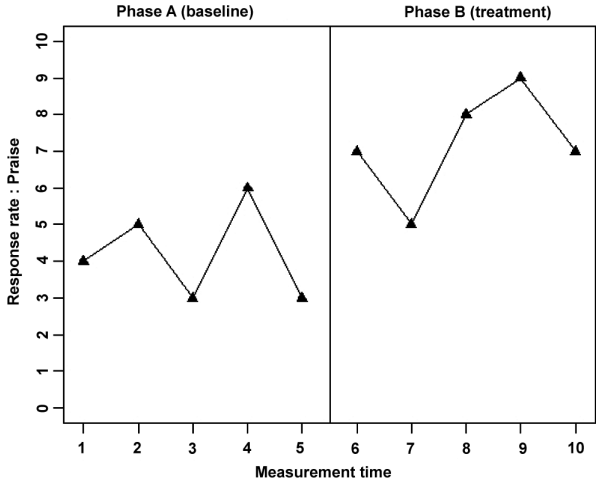


Figure 2. A fictitious example of an AB data series with $n_A = n_B = 5$.

2) Calculate the number of Phase B data points that exceed the value identified in the previous step. The measurements corresponding to days 6, 7, 9, and 10 are > 6, so there are four values exceeding Phase A's highest value.
3) Divide the value obtained in Step 2 by the number of observations in Phase B. The number of Phase B observations is 5 and the result of the division is 4/5 = 0.8.
4) Multiply the value obtained in Step 3 by 100 in order to convert the proportion into a percentage. The percentage obtained for the example is 0.8 × 100 = 80%.

PAND

1) Identify the highest measurement in Phase A. As obtained above this value is 6.
2) Calculate the minimal number of data points to be eliminated in order to have no interphase overlap. If the measurement corresponding to day 7 (i.e., 5 praises) is eliminated, then Phases A and B would not overlap – all Phase B data points would be greater than the Phase A measurements.
3) Divide the value obtained in Step 2 by the total number of observations. A single value to be eliminated means that the correct division is 1/10 = 0.1.
4) Multiply the value obtained in Step 3 by 100. The value obtained is 0.1 × 100 = 10%.
5) Subtract the value obtained in Step 4 from 100. The percentage of all nonoverlapping data is equal to 100 – 10 = 90%.

Percentage of Data PEM

1) Calculate the median of Phase A. In the example, the sorted baseline measurements are 3, 3, 4, 5, and 6 and, therefore, the Phase A median is equal to 4.
2) Calculate the number of Phase B data points that exceed the value identified in the previous step. All data points from the treatment phase are > 4, so the value obtained is 5 (equal to n_B).
3) Divide the value obtained in Step 2 by the number of observations in Phase B. The division to be made is 5/5 = 1.
4) Multiply the value obtained in Step 3 by 100 in order to convert the proportion into a percentage. In the example presented, the percentage of data PEM obtained is, thus, 1 × 100 = 100%.

Simulation

The specific steps that were implemented in the Fortran programs (one for each of the six series' lengths) were the following ones:

1) Systematic selection of each of the 19 degrees of serial dependence.
2) Systematic selection of the (β_1, β_2, and β_3) parameters for data generation, leading to eight different data patterns – autoregressive model (i.e., no effect or trend); trend; level change; slope change; trend and level change; trend and slope change; level and slope change; and trend, level, and slope change.
3) 100,000 iterations of Steps 4–15.
4) Generate an array with 50 + N data following a normal distribution with mean zero and unitary standard deviation by means of NAG*fl90* mathematical-statistical libraries (specifically external subroutines *nag_rand_seed_set* and *nag_rand_normal*).
5) Eliminate the first 50 numbers.
6) Assign the following N numbers to array u_t.
7) Establish $\varepsilon_1 = u_1$.
8) Obtain the array of ε_t using the equation $\varepsilon_t = \varphi_1 \times \varepsilon_{t-1}$.
9) Obtain the time array $T_t = 1, 2, \ldots, N$.
10) Obtain the dummy treatment variable array D_t, where $D_t = 0$ for Phase A and $D_t = 1$ for Phase B.
11) Obtain the slope change array according to Huitema and McKean's (2007) expression: $SC_t = [T_t - (n_A + 1)] \times D_t$ used for data generation.
12) Obtain the y_t array containing measurements (i.e., dependent variable) following Huitema and McKean's (2007) model: $y_t = \beta_0 + \beta_1 \times T_t + \beta_2 \times D_t + \beta_3 \times SC_t + \varepsilon_t$.
13) Calculate PND.
14) Calculate PAND.
15) Calculate PEM.
16) Average the obtained percentages from the 100,000 replications of each experimental condition.

Results

This section is organized according to the objectives of the study: To explore the effect of autocorrelation, to compare

Table 1. Distortion due to autocorrelation when no trend or effect is present in data – the values represent the ratio $\varphi_1 \neq 0/\varphi_1 = 0$

		Series' length					
φ	Effect size indices	5 + 5	5 + 10	7 + 8	5 + 15	10 + 10	15 + 15
−0.9	PND	1.32	1.36	1.46	1.38	1.60	1.78
	PAND	1.05	1.09	1.06	1.13	1.05	1.05
	PEM	1.00	1.00	1.00	1.00	1.00	1.00
−0.6	PND	0.97	0.97	0.99	0.97	1.02	1.06
	PAND	1.00	0.99	1.00	0.99	1.00	1.00
	PEM	1.00	1.00	1.00	1.00	1.00	1.00
−0.3	PND	0.93	0.93	0.93	0.93	0.94	0.97
	PAND	0.99	0.98	0.99	0.98	1.00	1.00
	PEM	1.00	1.00	1.00	1.00	1.00	1.00
0.3	PND	1.16	1.17	1.17	1.16	1.17	1.18
	PAND	1.02	1.04	1.02	1.05	1.01	1.01
	PEM	1.00	1.00	1.00	1.00	1.00	1.00
0.6	PND	1.38	1.44	1.50	1.47	1.58	1.64
	PAND	1.05	1.11	1.06	1.16	1.05	1.04
	PEM	1.00	1.00	1.00	1.00	1.00	1.00
0.9	PND	1.61	1.76	1.94	1.83	2.27	2.86
	PAND	1.09	1.19	1.12	1.28	1.11	1.11
	PEM	1.00	1.00	1.00	1.00	0.99	1.00

data pattern discrimination, and to assess the importance of series' length.

Autocorrelation Effect

In order to quantify the degree to which autocorrelation introduces distortion in the effect size estimates, we divided the estimates obtained for $\varphi_1 \neq 0$ by the one obtained for $\varphi_1 = 0$. We performed those calculi for the case of no effect or trend simulated to avoid confounding variables. If the ratio obtained is equal to 1, then there is no influence of serial dependence. Ratios < 1 imply an underestimation of the effect size associated with autocorrelation, while values > 1 entail overestimation. As Table 1 shows, PEM yields practically the same values regardless of the degree of serial dependence. For PND and PAND, greater negative or positive autocorrelation is generally associated with higher effect size estimates, being PND the more affected of the two indices. Figure 2 shows an example of those findings.

When there was treatment effect simulated in data, PEM proved to be sensitive to the presence of autocorrelation – positive as well as negative serial dependence leads to lower effect size estimates (see Figure 3 for an example). For PND and PAND, the type of relationship between autocorrelation and effect size depends on the type of effect in data. When the intervention involves a level change, positive and negative φ_1 overestimate effect size. When the treatment effect is expressed as slope change, it would be underestimated if either PND or PAND are used. Figure 4 is an illustration of these tendencies.

Data Pattern Discrimination

The comparison of data pattern discrimination was carried out by constructing graphs combining the three procedures for computing the magnitude of effect with the six series' lengths. In each of these $3 \times 6 = 18$ graphs we put data patterns in the abscissa and the effect size index (i.e., percentage) in the ordinate, superimposing several autocorrelation levels.

We consider that an effect size index should detect (i.e., yield highest effect size estimates) powerful treatments, like the ones represented by changes in slope and in level in the same direction. The indices would also have to respond with high estimates to the occasions when either a level change or a slope change is present. On the other hand, when the intervention is not effective the effect size index ought to yield low (ideally zero) percentages. Additionally, a perfect index would not be sensitive to a general trend, which has no relation to the introduction of a psychological treatment.

The visual inspection carried out following those criteria suggests that PND and PEM approximate the ideal discrimination pattern. Nonetheless, there is one relevant discrepancy between those two indices due to the essence of their calculus – PND yields smaller effect size estimates than PEM. PAND seems to be more deficient, as it yields more similar estimates for data sets with and without treatment effects. An example of those findings can be seen in Figure 5, which is constructed for $\varphi_1 = 0.3$, as it represents a level of serial dependence likely to be found in behavioral data (Parker, 2006), although the abovementioned tendencies are common to all φ_1 values studied. All of the indices

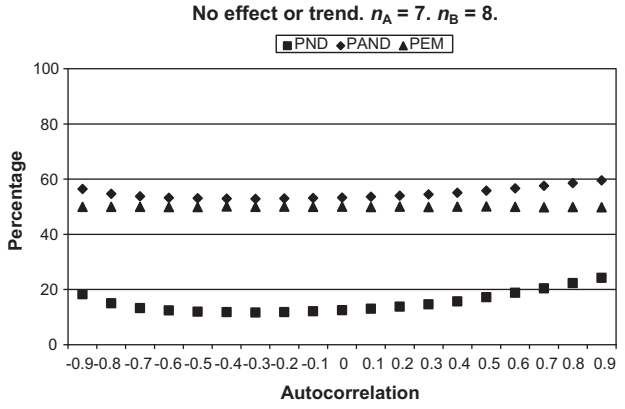


Figure 3. Autocorrelation effect on the effect size indices when no effect or trend is present in data.

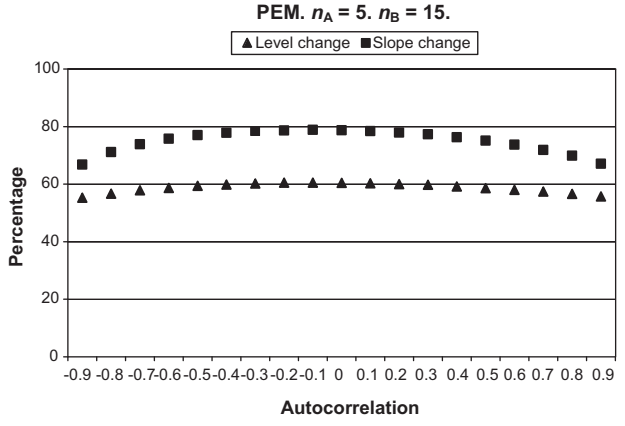


Figure 4. Autocorrelation effect on PEM when treatment effects are present in data.

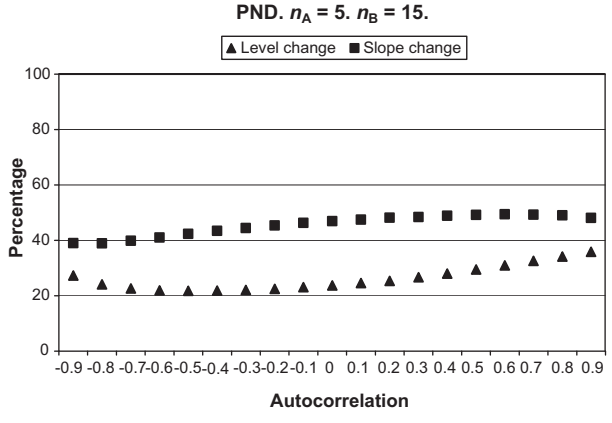


Figure 5. Autocorrelation effect on PND when treatment effects are present in data.

tested share a common drawback – they are affected by the presence of trend in data which leads to overestimating effect size. As expected, complex patterns are associated with greater effect size estimates for all indices.

Complementing the analyses performed, we divided the effect size estimates for series with effect and/or trend present by the estimate for data with no effect or trend simulated. These calculi were carried out for each of the three indices and for all series' lengths. Ratios equal to 1 suggest that there are the same estimates obtained in presence and in absence of effect. Values > 1 imply that the effect or the extraneous variable is associated with greater effect size estimates than white noise data. As Table 2 shows, PND is the procedure that differentiates the most between presence and absence of intervention effect. However, it is also the procedure most affected by trend. PAND distinguishes less between data patterns, except for data series with $n_A = 5$ and $n_B = 15$ where its performance is practically equivalent to PEM's.

Series' Length Effect

In order to explore the variation of the performance of the indices as one of the phases (or both) becomes longer, we divided the effect size estimates obtained for the longer designs with the ones obtained for the shortest one ($n_A = n_B = 5$). Ratios equal to 1 suggest that phase length does not influence the performance of the procedures. Values > or < 1 imply higher or lower effect size estimates, respectively, in comparison to 10-measurement data sets. According to Table 3, increasing series' length leads to a better differentiation between the data patterns. As the example in Figure 6 shows the improvement is expressed basically as lower false alarm rates (i.e., lower percentages for the case of absence of treatment effect) and as higher sensitivity to synergic slope and level changes. Those results highlight the importance of having more measurements of the experimental unit in order to obtain a more precise image of the evolution of its behavior. In accordance with the data simulation method followed, in longer series changes in slope yielded higher effect size estimates than changes in level.

The performance of PAND improves for designs with unbalanced phase lengths. As Figure 7 illustrates for such designs the distinction between data patterns is more pronounced, implying lower effect size estimates for white noise and trend. On the contrary, for PND the presence of trend is more problematic for designs with unequal phase lengths (Figure 8). PEM is the procedure less affected by the amount of data points in the series.

Discussion

In the current investigation we pretended to continue the search of the most appropriate procedure for quantifying treatment effectiveness and summarizing results from single-case designs. The performance of the effect size indices

Table 2. Detection of data patterns in comparison to the case of no effect or trend simulated in independent series

Data pattern	Effect size indices	Series' length					
		5 + 5	5 + 10	7 + 8	5 + 15	10 + 10	15 + 15
Slope change	PND	1.45	2.12	2.01	2.82	2.67	4.89
	PAND	1.06	1.28	1.13	1.61	1.14	1.23
	PEM	1.21	1.42	1.35	1.57	1.44	1.60
Level change	PND	1.42	1.43	1.49	1.43	1.57	1.66
	PAND	1.06	1.11	1.06	1.14	1.05	1.04
	PEM	1.21	1.21	1.21	1.21	1.22	1.23
Level and slope change	PND	1.95	2.66	2.69	3.35	3.60	6.23
	PAND	1.14	1.42	1.21	1.78	1.22	1.31
	PEM	1.40	1.58	1.52	1.70	1.60	1.72
Trend	PND	1.43	1.67	1.77	1.94	2.27	3.59
	PAND	1.06	1.17	1.10	1.31	1.11	1.15
	PEM	1.21	1.30	1.31	1.39	1.42	1.60
Trend and slope change	PND	1.95	2.93	3.07	3.75	4.58	8.71
	PAND	1.14	1.48	1.26	1.92	1.30	1.45
	PEM	1.39	1.61	1.58	1.74	1.70	1.84
Trend and level change	PND	1.92	2.21	2.46	2.51	3.17	4.98
	PAND	1.13	1.30	1.18	1.50	1.18	1.23
	PEM	1.40	1.49	1.50	1.56	1.59	1.73
Trend, level, and slope change	PND	2.50	3.47	3.79	4.17	5.56	9.97
	PAND	1.21	1.62	1.35	2.05	1.38	1.52
	PEM	1.56	1.73	1.71	1.82	1.80	1.90

Table 3. Influence of series' length on pattern detection for independent series – comparison to $n_A = n_B = 5$

Series' length	Effect size indices	Data pattern				
		No effect or trend	Slope change	Level change	Level and slope change	Trend
5 + 10	PND	1.00	1.47	1.00	1.37	1.17
	PAND	0.76	0.92	0.79	0.95	0.84
	PEM	1.00	1.18	1.00	1.13	1.08
7 + 8	PND	0.75	1.04	0.78	1.04	0.93
	PAND	0.91	0.97	0.91	0.98	0.94
	PEM	1.00	1.12	1.00	1.09	1.09
5 + 15	PND	1.00	1.95	1.00	1.72	1.36
	PAND	0.64	0.97	0.69	1.01	0.80
	PEM	1.00	1.31	1.00	1.22	1.15
10 + 10	PND	0.55	1.01	0.60	1.01	0.87
	PAND	0.94	1.00	0.92	1.00	0.97
	PEM	1.00	1.20	1.01	1.15	1.18
15 + 15	PND	0.37	1.26	0.43	1.19	0.94
	PAND	0.91	1.05	0.89	1.05	0.99
	PEM	1.00	1.33	1.01	1.23	1.32

was tested by means of data patterns generated to represent the likely features of real data (i.e., few observations per phase, serially dependent measurements). Among the desirable features those indices can be stated: (a) to detect changes in behavior due to the introduction of an intervention – low miss (Type II error) rates; (b) to produce low, ideally null, effect size estimates in absence of treatment effect – low false alarm (Type I error) rates; (c) to be insensitive to extraneous variables such as general trend; and (d) to remain unaffected by autocorrelation.

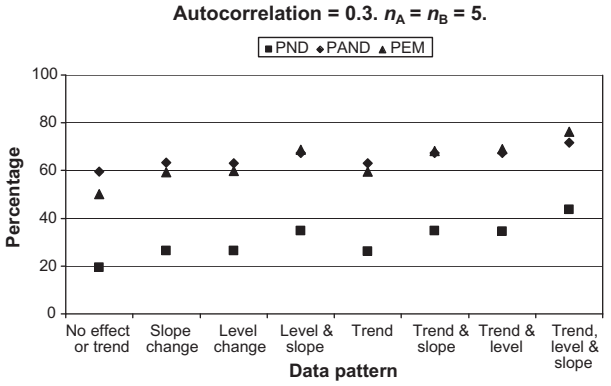


Figure 6. Effect sizes calculated for different data patterns and moderate positive serial dependence in a design with equal phase lengths.

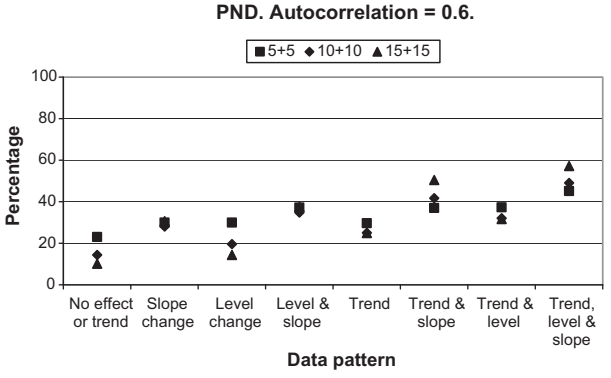


Figure 7. Influence of series' length on PND.

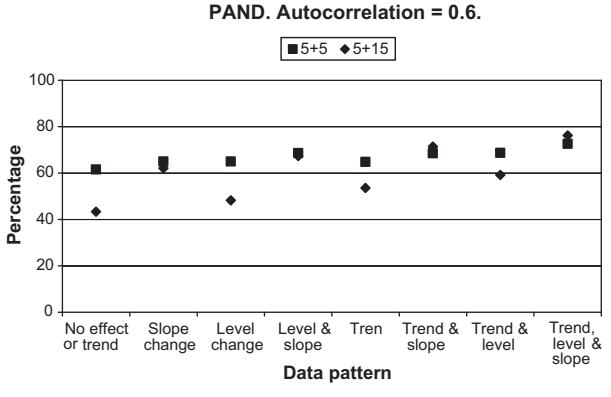


Figure 8. Influence of phase length on PAND.

Taking the first two criteria into consideration simultaneously we can point to PND as the best performer as it produces lowest effect size estimates in presence of solely white noise. Moreover, among the three procedures tested, it presents the highest relative differentiation between effective and ineffective interventions. PEM also shows a good patterns' discrimination, being more sensitive but less specific than PND. PAND is the index that performs less satisfactorily in the cases when baseline and treatment phases have approximately the same number of observations. A positive characteristic of all three indices studied is the discrimination between data patterns even when series consist of only ten data points.

As regards autocorrelation, PEM is the less affected procedure in absence of effect and is conservatively biased by both positive and negative serial dependence in presence of treatment effect. Applied researchers should keep in mind that both overestimation and underestimation of an existing treatment effect are possible when PND and PAND are used, depending on the degree of autocorrelation and on the type of effect (slope change or level change). Out of those two indices PND is the one whose effect size estimates are more distorted by serial dependence.

A shortcoming of the indices is the finding of the distorting impact of trend in data, which makes necessary the visual inspection prior to applying any of the three procedures. PAND was the least affected index, while PND was the most affected one.

In conclusion, what recommendation can be given to applied researchers? To begin with, they ought to keep in mind what each index represents in order to interpret it correctly. In this sense, we consider that the meaning of both PND and PEM is more straightforward than the information given by PAND. In terms of computational accessibility, all three indices can easily be calculated, especially PND. We have to advert that whenever the intervention is supposed to reduce rather than to enhance the behavior measured, the manner of computation of the indices can be adjusted to the needs of the applied researcher. A potential advantage of PAND is the possibility to derive from it a conventional effect size index, like Pearson's *Phi* (Parker et al., 2007). Nonetheless, mathematical-statistical calculations beyond the computation of the percentage itself may make the index less attractive to applied researchers. Applied researchers can be advised to use PND in data sets with no autocorrelation or trend, as it is the procedure that best distinguishes between presence and absence of intervention effect. When there is a high outlier in the baseline phase and the objective of the intervention is to increase the behavior of interest, the use of PND cannot be advised as it would lead to an underestimation of the treatment effect. In cases when the behavioral measurements present general trend or are likely to be sequentially related, PEM ought to be the effect size index chosen. PAND approximates PEM's performance only when the baseline phase is considerably shorter than the treatment phase.

In any case, professionals should not follow the same criteria for labeling the treatment as "effective" when using different procedures (e.g., 70–90% "effective", 50–70% "questionable", in Scruggs and Mastropieri, 1998). This is due to the fact that the effect sizes provided by PEM and PAND are systematically higher than the one provided by PND. Whatever index is utilized, visual inspection should not be replaced as a source of supplementary information (Parker et al., 2006).

As regards meta-analysis of single-case data, applied psychologists ought to be cautious when integrating information

from studies using different number of measurement times, since these may imply different levels of affection by autocorrelation and general trend. That is, the effect size estimates obtained from studies with a specific N may not have the same precision and the same insensitivity to extraneous variables as the estimates obtained for other series and/or phase lengths. This difficulty is, however, not only applicable to effect size procedures based on visual analysis, but also to the ones based on regression or standardized mean difference (Manolov & Solanas, 2008).

A limitation of the present investigation consists in the fact that only two-phase designs were studied. However, as Busse, Kratochwill, and Elliott (1995) claim, the AB designs' results can also be useful for multiple-baseline designs.

Future research may center on calibrating the data generation procedure with the most appropriate values (i.e., β_1, β_2, and β_3) for simulating treatment effects in order to improve real data modeling. In addition, it is necessary to obtain evidence on the performance of the effect size indices in designs consisting of more than two phases.

Acknowledgments

This research was supported by the *Comissionat per a Universitats i Recerca del Departament d'Innovació, Universitats i Empresa* of the *Generalitat de Catalunya*, the European Social Fund, the *Ministerio de Educación y Ciencia* Grant SEJ2005-07310-C02-01/PSIC, and the *Generalitat de Catalunya* Grant 2005SGR-00098.

References

Allison, D. B., Faith, M. S., & Franklin, R. (1995). Antecedent exercise in the treatment of disruptive behavior: A review and meta-analysis. *Clinical Psychology: Science and Practice, 2*, 279–303.

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy, 31*, 621–631.

Bellini, S., Peters, J. K., Benner, L., & Hopf, A. (2007). A meta-analysis of school-based social skills interventions for children with autism spectrum disorders. *Remedial and Special Education, 28*, 153–162.

Borrego, J. Jr., Anhalt, K., Terao, S. Y., Vargas, E. C., & Urquiza, A. J. (2006). Parent-child interaction therapy with a Spanish-speaking family. *Cognitive and Behavioral Practice, 13*, 121–133.

Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment, 10*, 229–242.

Busse, R. T., Kratochwill, T. R., & Elliott, S. N. (1995). Meta-analysis for single-case consultation outcomes: Applications to research and practice. *Journal of School Psychology, 33*, 269–285.

Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification, 28*, 234–246.

Center, B. A., Skiba, R. J., & Casey, A. (1985–1986). A methodology for the quantitative synthesis of intra-subject design research. *The Journal of Special Education, 19*, 387–400.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304–1312.

Cohen, J (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997–1003.

Crosbie, J. (1987). The inability of the binomial test to control Type I error with single-subject data. *Behavioral Assessment, 9*, 141–150.

Faith, M. S., Allison, D. B., & Gorman, D. B. (1997). Meta-analysis of single-case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 245–277). Mahwah, NJ: Erlbaum.

Friedman, H. (1982). Simplified determinations of statistical power, magnitude of effect and research sample sizes. *Educational and Psychological Measurement, 42*, 521–526.

Gorman, B. S., & Allison, D. B. (1997). Statistical alternatives for single-case designs. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 159–214). Mahwah, NJ: Erlbaum.

Gorsuch, R. L. (1983). Three methods for analyzing limited time-series (N of 1) data. *Behavioral Assessment, 5*, 141–154.

Greenwood, K. M., & Matyas, T. A. (1990). Problems with application of interrupted time series analysis for brief single-subject data. *Behavioral Assessment, 12*, 355–370.

Huitema, B. E. (1985). Autocorrelation in behavior analysis: A myth. *Behavioral Assessment, 7*, 107–118.

Huitema, B. E., & McKean, J. W. (1991). Autocorrelation estimation and inference with small samples. *Psychological Bulletin, 110*, 291–304.

Huitema, B. E., & McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement, 60*, 38–58.

Huitema, B. E., & McKean, J. W. (2007). An improved portmanteau test for autocorrelated errors in interrupted time-series regression models. *Behavior Research Methods, 39*, 343–349.

Huitema, B. E., McKean, J. W., & McKnight, S. (1999). Autocorrelation effects on least-squares intervention analysis of short time series. *Educational and Psychological Measurement, 59*, 767–786.

Jones, R. R., Weinrott, M. R., & Vaught, R. S. (1978). Effects of serial dependence on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis, 11*, 277–283.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746–759.

Kromrey, J. D., & Foster-Johnson, L. (1996). Determining the efficacy of intervention: The use of effect sizes for data analysis in single-subject research. *Journal of Experimental Education, 65*, 73–93.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject research: Percentage of data points exceeding the median. *Behavior Modification, 30*, 598–617.

Manolov, R., & Solanas, A. (2008). Comparing $N = 1$ effect size indices in presence of autocorrelation. *Behavior Modification, 32*, 860–875.

Mathur, S. R., Kavale, K. A., Quinn, M. M., Forness, S. R., & Rutherford, R. B. Jr. (1998). Social skills interventions with students with emotional and behavioral problems: A quantitative synthesis of single-subject research. *Behavioral Disorders, 23*, 193–201.

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis for single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23*, 341–351.

Matyas, T. A., & Greenwood, K. M. (1991). Problems in the estimation of autocorrelation in brief time series and some implications for behavioral data. *Behavioral Assessment, 13*, 137–157.

Matyas, T. A., & Greenwood, K. M. (1997). Serial dependency in single-case time series. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 215–243). Mahwah, NJ: Erlbaum.

Parker, R. I. (2006). Increased reliability for single-case research results: Is bootstrap the answer? *Behavior Therapy, 37*, 326–338.

Parker, R. I., & Brossart, D. F. (2003). Evaluating single-case research data: A comparison of seven statistical methods. *Behavior Therapy, 34*, 189–211.

Parker, R. I., Cryer, J., & Byrns, G. (2006). Controlling baseline trend in single-case research. *School Psychology Quarterly, 21*, 418–443.

Parker, R. I., & Hagan-Burke, S. (2007a). Median-based overlap analysis for single case data: A second study. *Behavior Modification, 31*, 919–936.

Parker, R. I., & Hagan-Burke, S. (2007b). Useful effect size interpretations for single case research. *Behavior Therapy, 38*, 95–105.

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data: An alternative to PND. *Journal of Special Education, 40*, 194–204.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276–1284.

Scruggs, T. E., & Mastropieri, M. A. (1994). The utility of the PND statistic: A reply to Allison and Gorman. *Behaviour Research and Therapy, 32*, 879–883.

Scruggs, T. E., & Mastropieri, M. A. (1998). Summarizing single-subject research: Issues and applications. *Behavior Modification, 22*, 221–242.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education, 8*, 24–33.

Scruggs, T. E., Mastropieri, M. A., Forness, S. R., & Kavale, K. A. (1988). Early language intervention: A quantitative synthesis of single-subject research. *The Journal of Special Education, 20*, 259–283.

Sharpley, C. F., & Alavosius, M. P. (1988). Autocorrelation in behavior data: An alternative perspective. *Behavior Assessment, 10*, 243–251.

Sierra, V., Solanas, A., & Quera, V. (2005). Randomization tests for systematic single-case designs are not always appropriate. *The Journal of Experimental Education, 73*, 140–160.

Skiba, R. J., Casey, A., & Center, B. A. (1986). Nonaversive procedures in the classroom behavior problems. *The Journal of Special Education, 19*, 459–481.

Suen, H. K. (1987). On the epistemology of autocorrelation in applied behavior analysis. *Behavioral Assessment, 9*, 113–124.

Suen, H. K., & Ary, D. (1987). Autocorrelation in behavior analysis: Myth or reality? *Behavioral Assessment, 9*, 130–150.

Toothaker, L. E., Banz, M., Noble, C., Camp, J., & Davis, D. (1983). $N = 1$ designs: The failure of ANOVA-based tests. *Journal of Educational Statistics, 4*, 289–309.

White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta-analysis in individual subject research. *Behavioral Assessment, 11*, 281–296.

Wilkinson, L., & The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 694–704.

Rumen Manolov

Departament de Metodologia de les Ciències del Comportament
Universitat de Barcelona
Passeig de la Vall d'Hebron 171
8035 Barcelona
Spain
Tel. +34 933 125 844
Fax +34 934 021 359
E-mail rrumenov13@ub.edu

Estudi 3©

Percentage of nonoverlapping corrected data

Rumen Manolov i Antonio Solanas

Departament de Metodologia de les Ciències del Comportament
Facultat de Psicologia
Universitat de Barcelona

Resum. L'Estudi 3 es centra en proposar una millora al percentatge de no solapament entre les dades (PND), índex que s'ha mostrat com un dels més utilitzats i més adients per quantificar la magnitud de l'efecte en dissenys de cas únic. La proposta consisteix en introduir un pas inicial de correcció de les dades que pretén controlar la possible existència de tendència durant la fase de línia base. Aquesta modificació és necessària atès que una de les principals fonts de distorsió de les estimacions produïdes pel PND és la tendència general. A banda de presentar i exemplificar el càlcul del nou procediment, es dur a terme un estudi de simulació per tal de contrastar si es produeix una millora en el rendiment del PND. La comprovació de la influència de la tendència general es realitza en una àmplia varietat de condicions experimentals, incloent dades generades per processos autoregressius i de mitjanes mòbils, variables aleatòries amb distribució exponencial, normal i uniforme, diferents tipus d'efecte del tractament, diversos nivells de dependència serial. Els resultats mostren que les estimacions de la magnitud de l'efecte produïdes pel procediment proposat no varien en presència de tendència, és a dir, la correcció de les dades assoleix la seva finalitat. A més, la tècnica proposada és menys afectada per l'autocorrelació que el PND.

Behavior Research Methods
2009, 41 (4), 1262-1271
doi:10.3758/BRM.41.4.1262

Percentage of nonoverlapping corrected data

Rumen Manolov and Antonio Solanas
University of Barcelona, Barcelona, Spain

In the present study, we proposed a modification in one of the most frequently applied effect-size procedures in single-case data analysis: the percentage of nonoverlapping data. In contrast with other techniques, the calculus and interpretation of this procedure are straightforward and can be easily complemented by visual inspection of the graphed data. Although the percentage of nonoverlapping data has been found to perform reasonably well in $N = 1$ data, the magnitude of effect estimates that it yields can be distorted by trend and autocorrelation. Therefore, the data-correction procedure focuses on removing the baseline trend from data prior to estimating the change produced in the behavior as a result of intervention. A simulation study was carried out in order to compare the original and the modified procedures in several experimental conditions. The results suggest that the new proposal is unaffected by trend and autocorrelation and that it can be used in case of unstable baselines and sequentially related measurements.

Single-case designs are useful for obtaining scientific evidence about intervention effectiveness in different behavioral fields of knowledge (Crane, 1985; Gedo, 2000; Tervo, Estrem, Bryson-Brockman, & Symons, 2003). Recent methodological research on single-case data analysis has centered on effect-size measures instead of on statistical techniques yielding *p* values exclusively. The increased interest in quantifying the magnitude of effect might be due to the recommendations for reporting studies' results (Wilkinson & the Task Force on Statistical Inference, 1999) based on the advantages of effect sizes over statistical significance, such as the focus on the strength of relationship between the intervention and behavior of interest, the possibility of establishing different degrees of treatment effectiveness, and the avoidance of sample-size dependence (Cohen, 1990, 1994; Kirk, 1996; Kromrey & Foster-Johnson, 1996; Rosnow & Rosenthal, 1989). The importance of effect-size measurements in single-case designs has been reflected in an increased number of recent publications, answering the need for evidence-based interventions in the behavioral sciences (Jenson, Clark, Kircher, & Kristjansson, 2007; Schlosser & Sigafoos, 2008; Shadish, Rindskopf, & Hedges, 2008).

From the perspective of an applied researcher in clinical, educational, or social settings, a potentially useful effect-size index needs to meet several criteria: (1) it must perform well in a short data series, producing low estimates in absence of a treatment effect and higher ones in its presence; (2) it must be easy to interpret in applied rather than in statistical terms; (3) related to the previous criterion, it is desirable that the procedure be designed specifically for $N = 1$ data in order to avoid interpretations that are based on group designs terminology; (4) it should be simple to compute, not requiring expertise, commercial statistical software packages, or an excessive amount of time; (5) it should be easily complemented by visual inspection, considering its utility (Parker, Cryer, & Byrns, 2006) and its frequent application (Kratochwill & Brody, 1978; Parker & Brossart, 2003).

As regards the first criterion mentioned, several regression-based techniques have been found to have unacceptable statistical properties (Beretvas & Chung, 2008; Manolov & Solanas, 2008; Parker & Brossart, 2003). These procedures also require a greater amount of knowledge and calculus on the part of the researcher in comparison with the proposed indices related to visual analysis. With respect to the latter procedures, Ma's (2006) percentage of data points exceeding the median and Parker, Hagan-Burke, and Vannest's (2007) percentage of all nonoverlapping data (PAND) were designed to improve the performance of the percentage of nonoverlapping data (PND; Scruggs, Mastropieri, & Casto, 1987) procedure, but it has been shown that this does not always work (Manolov, Solanas, & Leiva, in press). Additionally, the magnitude of the effect estimate produced by PAND has a less straightforward interpretation, whereas Pearson's Φ^2, which can be obtained from it, requires several steps to be carried out with different software programs (Schneider, Goldstein, & Parker, 2008).

Given these considerations, the PND procedure that was designed for single-case data can be regarded as a procedure that works well (i.e., better than its most similar alternatives, although not optimally), being simple to interpret and to compute, and closely related to visual inspection. In recent studies, PND has been the most frequently applied procedure for quantifying treatment effectiveness in single-case studies and also in meta-analyses (Schlosser, Lee, & Wendt, 2008). Nevertheless, despite its attractiveness to psychologists, PND is not trouble-free (Allison &

R. Manolov, rrumenov13@ub.edu

Gorman, 1994; Manolov & Solanas, 2008). Therefore, the main objective of the present investigation was to propose a modification of the PND procedure that was intended to overcome some of its limitations. The performance of the modified index is tested in the context of data sets with different characteristics, such as presence or absence of confounding variables (i.e., trend, serial dependence) and of intervention effects. In order to contrast the percentages obtained against known data attributes, Monte Carlo methods were used to construct the data series.

Overcoming the Drawbacks of PND

The present study proposed a data-correction procedure to be implemented prior to applying the PND. The main aim of the procedure was to eliminate from the data a possible preexisting trend that was not related to the introduction of the intervention. Since the proposal is basically a modification of PND—adding an initial data-correction step—we refer to the procedure as the "percentage of nonoverlapping corrected data" (PNCD). Before a treatment is introduced (i.e., in an AB design's initial phase), it can be reasonably assumed that the behavior of the individual (y) or group studied is randomly fluctuating around a certain value; that is, $y_t = \varepsilon_t$. If there is a trend in the behavior, then $y_t = \beta \cdot t + \varepsilon_t$, where β is the trend coefficient (equal to 0 in the absence of trend), and t is the value of the time variable. The original phase A consists of n_A data points, which, when differenced, leads to a new series of $n_A - 1$ values: $\Delta y_{t+1} = y_{t+1} - y_t$. In case there is a trend in data,

$$\begin{aligned}\Delta y_{t+1} &= [\beta \cdot (t+1) + \varepsilon_{t+1}] - [\beta \cdot t + \varepsilon_t] \\ &= \beta \cdot t + \beta + \varepsilon_{t+1} - \beta \cdot t - \varepsilon_i \\ &= \beta + \varepsilon_{t+1} - \varepsilon_t.\end{aligned}$$

ε_{t+1} and ε_t are supposed to be independent and randomly and identically distributed, and their mathematical expectancy is assumed to be 0. Given that $E[\hat{\beta} + \varepsilon_{t+1} - \varepsilon_t] = \hat{\beta}$, an estimate of β can be obtained averaging the differenced data series; that is, $\overline{\Delta y}$ is used as $\hat{\beta}$. After the trend in the baseline phase is estimated, the whole series (both phases A and B) can be corrected by subtracting $\hat{\beta} \cdot t$ (the trend estimate multiplied by the measurement time) from the original data points. This operation is expected to remove any trend from the data and thus to avoid inflation in the percentages obtained by means of PND. Trend is not estimated from the whole data series, since a change in level between the phases may be confounded for trend, and such a correction may remove the intervention effect. The steps necessary for computing both PND and PNCD are illustrated in a following section. Additionally, R codes were developed for computing both indices and are presented in Appendices A and B for interventions aiming to increase and decrease the response rate, respectively.

Concerning autocorrelation, a difference needs to be established between positive serial dependence and negative serial dependence. Higher degrees of positive autocorrelation can be represented by upward or downward trends and, therefore, one can conjecture that a correction focusing on trend may also have influence on it and attenuate its impact on the effect-size index. Negative autocorrelation, however, is related to alternations of dissimilar measurements. In this case, the effect of the correction procedure proposed cannot be foreseen and needs to be explored.

Outliers represent another data feature that can distort the magnitude of effect estimates provided by PND. For instance, a single extremely high value in phase A can mask a behavioral change taking place after the treatment is introduced. Outliers can be detected using statistical calculi and can be controlled by means of elimination, winsorization, and so on. However, it has to be taken into account that in a single-case study, the applied researcher possesses a thorough knowledge of the client and is able to identify which measurement is an extreme and potentially anomalous one and interpret it (e.g., seek for its reason) from a clinical, educational, social, and so on, point of view. Such a theoretical interpretation may be more meaningful than an arbitrary statistical treatment of the unexpected datum.

METHOD

AB Series' Lengths

Short data series ($N = n_A + n_B$) were included in the present study, since those are more feasible in applied settings: (1) $N = 10$, with $n_A = n_B = 5$; (2) $N = 15$, with $n_A = 5$; $n_B = 10$; (3) $N = 15$, with $n_A = 7$; $n_B = 8$; (4) $N = 20$, with $n_A = n_B = 10$; (5) $N = 30$, with $n_A = n_B = 15$; and (6) $N = 40$, with $n_A = n_B = 20$.

Data Generation

For each combination of n_A and n_B, the data were according to the model proposed by Huitema and McKean (2000, 2007b): $y_t = \beta_0 + \beta_1 \cdot T_t + \beta_2 \cdot D_t + \beta_3 \cdot SC_t + \varepsilon_t$, where y_t is the value of the dependent variable at moment t; β_0 is the intercept set to 0; β_1, β_2, and β_3 are the coefficients associated with trend, level change, and slope change, respectively; T_t is the value of the time variable at moment t (taking values from 1 to N); D_t is a dummy variable for level change (equal to 0 for phase A and to 1 for phase B); SC_t is the value of the slope change variable being equal to 0 for phase A, and taking values from 0 to ($n_B - 1$) for phase B; and ε_t is the error term.

The error term (ε_t) was generated following two different models. One of these was the commonly used first-order autoregressive model $\varepsilon_t = \phi_1 \cdot \varepsilon_{t-1} + u_t$, with ϕ_1 ranging from −.9 to .9 in steps of .1. Since there is evidence that other models—especially a first-order moving average—can be used to represent behavioral data (Harrop & Velicer, 1985), the MA(1) model, $\varepsilon_t = u_t - \theta_1 \cdot u_{t-1}$, presented in McCleary and Hay (1980) was studied using 19 values of θ_1: −.9(.1).9. According to the formula $\phi_1 = -\theta_1/(1 + \theta_1^2)$, this meant that the degrees of autocorrelation ranged from −.4972 to .4972.

For both models, the random variable u_t was generated following $N(0,1)$ and an exponential and a uniform distribution with the same mean and standard deviation, since normal distributions are not always appropriate models for behavioral measurements (Bradley, 1977; Micceri, 1989). The aforementioned distributions are relevant, since they differ in terms of skewness and kurtosis from the Gaussian distribution.

The values of β_1, β_2, and β_3 (.06, .3, and .15, respectively) were chosen by trial and error—a procedure that was also followed by Parker and Brossart (2003) and Brossart, Parker, Olson, and Mahadevan (2006), aiming to avoid floor and ceiling effects in the percentages obtained (Manolov & Solanas, 2008). In addition, the values of those coefficients were determined in a way to produce equivalent mean shifts in the case of trend, a change in slope, and a change in level for the $n_A = n_B = 5$ data series. In any case, the specific beta values are not essential, since they only serve to construct data series with and without trend or intervention effect and thus create a common background for comparing PND and PNCD.

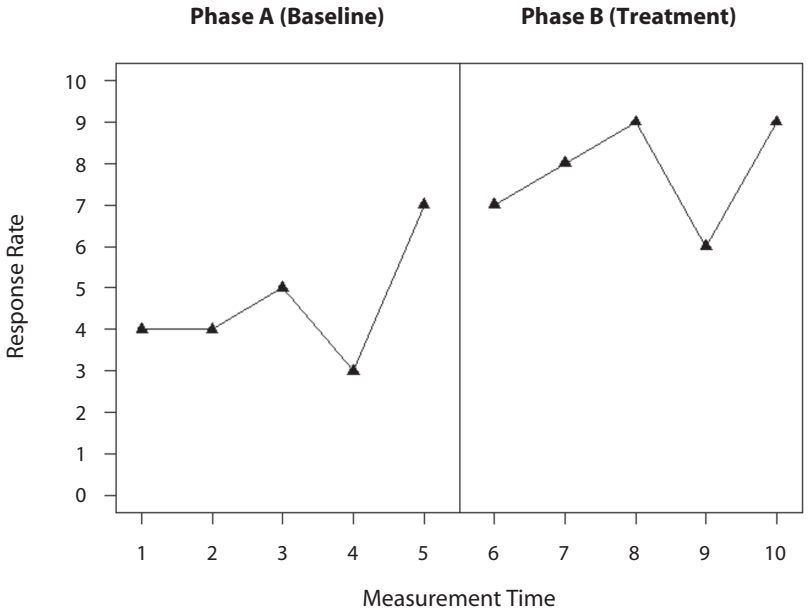


Figure 1. A fictitious example of an AB data series with $n_A = n_B = 5$.

Analysis

Prior to presenting in detail the steps needed to carry out the two effect-size procedures included in the present study, an example of a fictitious data set is presented. Consider a psychological single-case study educating a parent to interact with children who have been diagnosed with autism, counting a child's desirable behavior of interest (e.g., communication) in each session (Symon, 2005). The data gathered using the AB design structure (4, 4, 5, 3, and 7 positive communications during baseline and 7, 8, 9, 7, and 9 during treatment phase) can be represented graphically, as is shown in Figure 1. In the following section, the original and the proposed procedures are applied to the data set presented in order to illustrate their calculus.

PND

(1) Identify the highest measurement in phase A. In the example, it is 7 positive communications corresponding to baseline day 5.

(2) Calculate the number of phase B data points that exceed the value identified in the previous step. The measurements corresponding to days 7, 8, and 10 are greater than 7, so there are 3 values exceeding phase A's highest value.

(3) Divide the value obtained in step 2 by the number of observations in phase B. The number of phase B observations is 5, and the result of the division is 3/5 = .6.

(4) Multiply the value obtained in step 3 by 100 in order to convert the proportion into a percentage. The percentage obtained for the example is .6 · 100 = 60%.

PNCD

(1) Difference the phase A data points and obtain the differenced series with the length $n_A - 1$. In the example, the differenced series has the following 5 − 1 = 4 data points: 0 (4 − 4), 1 (5 − 4), −2 (3 − 5), and 4 (7 − 3).

(2) Compute the mean of the differenced series. The average of 0, 1, −2, and 4 is 0.75.

(3) Compute the trend-correction factor for each data point: the mean of the differenced series, multiplied by T_t. In the example, the values of the correction factor are .75 · 1, .75 · 2, . . . , .75 · 10.

(4) Perform the data correction subtracting the corresponding correction factor from each original data point. After the correction phase, A consists of 3.25 (4 − .75 · 1); 2.5 (4 − .75 · 2); 2.75 (5 − .75 · 3); 0 (3 − .75 · 4); and 3.25 (7 − .75 · 5). Phase B consists of the following data points: 2.5 (7 − .75 · 6); 2.75 (8 − .75 · 7); 3 (9 − .75 · 8); .25 (7 − .75 · 9); and 1.5 (9 − .75 · 10).

(5) Apply PND: None of the phase B data points is greater than the phase A highest value (3.25) and, therefore, PNCD = 0%.

Simulation

The specific steps that were implemented in the Fortran programs (one for each of the six series' length) were the following:

(1) Systematic selection of each of the 19 values of ϕ_1 or θ_1.

(2) Systematic selection of the (β_1, β_2, and β_3) parameters for data generation, leading to eight different data patterns: autoregressive or moving average model with no effect or trend, trend, level change, slope change, trend and level change, trend and slope change, combined level and slope change, and trend and combined level and slope change.

(3) 100,000 iterations of steps 4–15.

(4) Generate the u_t term according to an exponential, normal, or uniform distribution, eliminating the first 50 random numbers using the next N ones.

(5) Establish $\varepsilon_1 = u_1$.

(6) Obtain the error term ε_t out of the random variable u_t using the AR(1) model $\varepsilon_t = \phi_1 \cdot \varepsilon_{t-1} + u_t$ or the MA(1) model $\varepsilon_t = u_t - \theta_1 \cdot u_{t\text{-}1}$.

(7) Obtain the time array $T_t = 1, 2, \ldots, N$.

(8) Obtain the dummy treatment variable array D_t, where $D_t = 0$ for phase A and $D_t = 1$ for phase B.

(9) Obtain the slope change array according to $SC_t = [T_t - (n_A + 1)] \cdot D_t$.

(10) Obtain the y_t array containing measurements (i.e., dependent variable): $y_t = \beta_0 + \beta_1 \cdot T_t + \beta_2 \cdot D_t + \beta_3 \cdot SC_t + \varepsilon_t$.

(11) Calculate PND on the original data (i.e., the y_t array).

(12) Correct the data according to the procedure proposed.

(13) Calculate PNCD on corrected data.

(14) Average the obtained percentages from the 100,000 replications of each experimental condition.

For data generation NAG libraries, *nag_rand_neg_exp*, *nag_rand_normal*, and *nag_rand_uniform* were used. In order to guarantee suitable simulated data, the 50 values previous to each simulated data series were eliminated in order to reduce artificial effects (Greenwood & Matyas, 1990) and to avoid dependence between successive data series (Huitema, McKean, & McKnight, 1999).

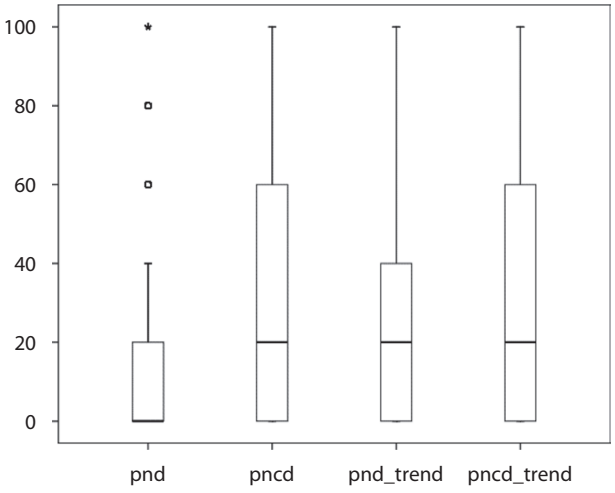


Figure 2. Distribution of the percentages provided by the percentage of nonoverlapping data (PND) and the percentage of nonoverlapping corrected data (PNCD) in the absence (the two boxplots on the left) and presence (the two boxplots on the right) of trend. The percentages were taken from 100,000 samples of independent $n_A = n_B = 5$ data, with no treatment effect simulated and normal error.

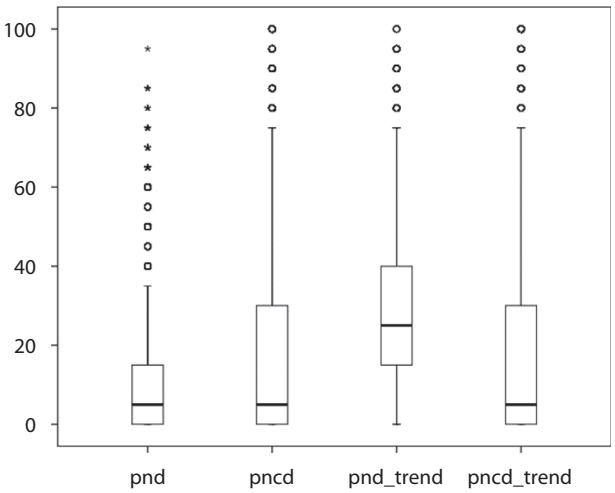


Figure 3. Distribution of the percentages provided by the percentage of nonoverlapping data (PND) and the percentage of nonoverlapping corrected data (PNCD) in the absence (the two boxplots on the left) and presence (the two boxplots on the right) of trend. The percentages were taken from 100,000 samples of $n_A = n_B = 20$ data, with level change simulated and uniform error in moving average processes with autocorrelation of .5.

RESULTS

When the data series represent solely random fluctuation (i.e., there is no trend, autocorrelation, or treatment effect), the percentages provided by PNCD are systematically larger than the ones provided by PND, as is illustrated by Figure 2. This finding implies that PND may be a better filter for ineffective interventions in the absence of trend and serial dependence. In the aforementioned conditions, higher effect-size estimates were also obtained for PNCD than for PND when treatment effects existed. However, if the data represent a trend, the PND estimates increase and may become superior to the PNCD estimates for both independent (Figure 2) and serially related (Figure 3) data series, as the within-figure comparisons show.

Trend Effect

In order to quantify the distortion of effect-size estimates produced by trend, the ratio between percentages with and without trend in data was computed. Therefore, a ratio close to 1 would indicate that trend does not introduce distortion, whereas values greater than 1 would imply an overestimation of the magnitude of effect. In the experimental conditions with no treatment effect simulated (Table 1), ratios > 1 entail an increment in false alarms, which is the case for PND in contrast with PNCD, which maintains approximately the same magnitude estimates in both the presence and absence of trend. This finding is applicable to all series lengths and errors' distributions tested.

When there is a treatment effect (slope change, level change, or both), the presence of trend leads to the overestimation of the effect size obtained through PND, as Table 2 shows. In contrast, the estimates provided by PNCD are not affected by the confounding variable.

The ratios presented in Tables 1 and 2 show that the PND estimates become more distorted by trend when the number of measurements N increases. PNCD seems to deal effectively with trend for both shorter and longer data series.

Autocorrelation Effect

The distortion of effect-size estimates produced by serial dependence was quantified by means of the ratio

Table 1
Distortion As a Result of Trend in an Independent Data Series

Phase Length		Ratio Trend/ Random Fluctuations	
n_A	n_B	PND	PNCD
	Exponential		
5	5	1.336	0.996
5	10	1.576	1.002
7	8	1.570	1.003
10	10	1.807	0.999
15	15	2.431	1.000
20	20	3.293	0.995
	Normal		
5	5	1.429	0.998
5	10	1.674	1.000
7	8	1.772	0.997
10	10	2.279	1.000
15	15	3.601	1.003
20	20	5.511	1.005
	Uniform		
5	5	1.517	1.002
5	10	1.747	0.997
7	8	2.003	0.995
10	10	2.761	1.005
15	15	4.590	0.991
20	20	6.952	0.993

Note—The values represent the ratio between the presence of trend/absence of trend in experimental conditions without treatment effect. PND, percentage of nonoverlapping data; PNCD, percentage of nonoverlapping corrected data.

Table 2
Distortion As a Result of Trend in Independent Data Series

Phase Length		Ratio Trend and Level/ Level Change Only		Ratio Trend and Slope/ Slope Change Only		Ratio Trend and Both Effects/ Both Effects	
n_A	n_B	PND	PNCD	PND	PNCD	PND	PNCD
				Exponential			
5	5	1.338	0.999	1.340	0.996	1.301	1.004
5	10	1.547	1.003	1.380	0.994	1.298	1.002
7	8	1.544	0.991	1.507	1.003	1.433	1.002
10	10	1.803	0.992	1.723	1.005	1.605	1.000
15	15	2.433	1.001	1.962	0.999	1.782	0.998
20	20	3.240	0.998	1.992	0.998	1.808	1.007
				Normal			
5	5	1.353	1.002	1.348	1.002	1.287	0.995
5	10	1.546	1.010	1.385	0.999	1.301	1.002
7	8	1.627	1.005	1.523	1.005	1.413	0.995
10	10	2.016	0.999	1.703	1.003	1.547	0.998
15	15	2.985	1.004	1.779	0.996	1.601	1.003
20	20	4.220	0.998	1.681	1.003	1.525	0.997
				Uniform			
5	5	1.325	1.001	1.339	0.998	1.246	0.998
5	10	1.506	0.996	1.354	0.997	1.276	0.997
7	8	1.596	0.995	1.456	0.996	1.350	1.005
10	10	1.909	1.008	1.562	0.997	1.432	0.998
15	15	2.530	0.996	1.611	1.003	1.473	1.001
20	20	3.119	0.989	1.505	1.000	1.374	0.994

Note—The values represent the ratio between the presence of trend/absence of trend in experimental conditions with single or combined treatment effect. PND, percentage of overlapping data; PNCD, percentage of nonoverlapping corrected data.

Table 3
Distortion As a Result of an AR(1) Process

Phase Length		Ratio $\phi_1 = -.3$/ Random Fluctuations		Ratio $\phi_1 = .3$/ Random Fluctuations		Ratio $\phi_1 = .6$/ Random Fluctuations	
n_A	n_B	PND	PNCD	PND	PNCD	PND	PNCD
				Exponential			
5	5	0.941	0.926	1.135	1.121	1.302	1.250
5	10	0.948	0.943	1.169	1.110	1.422	1.234
7	8	0.955	0.943	1.167	1.147	1.482	1.365
10	10	0.958	0.940	1.164	1.157	1.559	1.455
15	15	0.956	0.950	1.138	1.157	1.591	1.511
20	20	0.981	0.953	1.141	1.152	1.614	1.545
				Normal			
5	5	0.933	0.953	1.158	1.065	1.379	1.121
5	10	0.933	0.968	1.167	1.042	1.441	1.093
7	8	0.946	0.955	1.171	1.067	1.503	1.173
10	10	0.944	0.953	1.174	1.069	1.579	1.207
15	15	0.965	0.954	1.178	1.075	1.637	1.221
20	20	0.983	0.957	1.168	1.065	1.634	1.212
				Uniform			
5	5	0.929	0.961	1.158	1.060	1.378	1.114
5	10	0.929	0.967	1.166	1.032	1.428	1.079
7	8	0.936	0.954	1.185	1.050	1.497	1.130
10	10	0.932	0.946	1.195	1.050	1.561	1.146
15	15	0.949	0.920	1.189	1.006	1.602	1.088
20	20	0.971	0.912	1.188	0.994	1.619	1.035

Note—The values represent the ratio between serially dependent data and an independent series with no trend or intervention effect. PND, percentage of nonoverlapping data; PNCD, percentage of nonoverlapping corrected data.

between percentages computed for autocorrelated and independent data. Once again, ratios of 1 imply no distortion, and values greater than 1 are indicative of elevated false alarm rates in the absence of an intervention effect. In the case of exponential errors, for both AR(1) and MA(1) models, PNCD performs worse than PND when there is negative autocorrelation, and performs only slightly better for positive serial dependence. In contrast, for the normal and uniform errors, PNCD outperforms PND. For these two error distributions and AR(1) processes (Table 3) with $\phi_1 > 0$, the difference between PNCD and PND increases for longer data series, whereas for $\phi_1 < 0$ PNCD performs better only for $N \leq 20$. For the MA(1) processes (Table 4) with negative values of θ_1 (i.e., positive autocorrelation), PNCD shows less distortion than PND, whereas for $\theta_1 > 0$, it outperforms PND only for $N \leq 15$, always referring to normal and uniform errors.

Combined Effect

In addition to the individual effects of each of these data features, their combined effect was studied following the same procedure for quantifying distortion. Table 5 shows that for AR(1) processes with trend, PNCD is much less affected by the confounding variables than is PND, whose effect size estimate is quintupled in certain experimental conditions. For MA(1) processes (Table 6), the findings are similarly favorable for PNCD.

Table 4
Distortion As a Result of an MA(1) Process

Phase Length		Ratio $\theta_1 = -.5$/ Random Fluctuations		Ratio $\theta_1 = .5$/ Random Fluctuations	
n_A	n_B	PND	PNCD	PND	PNCD
			Exponential		
5	5	0.913	0.887	1.232	1.177
5	10	0.903	0.914	1.242	1.151
7	8	0.912	0.886	1.248	1.190
10	10	0.915	0.882	1.257	1.196
15	15	0.925	0.910	1.240	1.206
20	20	0.947	0.920	1.235	1.196
			Normal		
5	5	1.203	1.077	0.887	0.927
5	10	1.217	1.063	0.880	0.947
7	8	1.226	1.080	0.901	0.919
10	10	1.221	1.077	0.910	0.902
15	15	1.207	1.067	0.931	0.905
20	20	1.197	1.058	0.945	0.899
			Uniform		
5	5	1.194	1.066	0.882	0.938
5	10	1.200	1.046	0.868	0.951
7	8	1.207	1.047	0.881	0.925
10	10	1.205	1.038	0.906	0.894
15	15	1.218	0.990	0.929	0.860
20	20	1.179	0.947	0.940	0.842

Note—The values represent the ratio between nonnull and null θ_1 parameters in series with no trend or intervention effect. PND, percentage of nonoverlapping data; PNCD, percentage of nonoverlapping corrected data.

Table 5
Distortion As a Result of Combined Presence of Trend and an AR(1) Process

Phase Length		Ratio Trend and $\phi_1 = -.3$/ Random Fluctuations		Ratio Trend and $\phi_1 = .3$/ Random Fluctuations		Ratio Trend and $\phi_1 = .6$/ Random Fluctuations	
n_A	n_B	PND	PNCD	PND	PNCD	PND	PNCD
				Exponential			
5	5	1.267	0.929	1.518	1.131	1.671	1.239
5	10	1.484	0.942	1.785	1.116	2.007	1.240
7	8	1.489	0.941	1.793	1.140	2.012	1.284
10	10	1.734	0.939	2.074	1.164	2.566	1.459
15	15	2.336	0.951	2.734	1.150	3.368	1.524
20	20	3.211	0.960	3.679	1.141	4.369	1.520
				Normal			
5	5	1.325	0.953	1.581	1.056	1.783	1.126
5	10	1.565	0.969	1.839	1.036	2.043	1.089
7	8	1.663	0.950	1.961	1.068	2.223	1.170
10	10	2.130	0.944	2.457	1.068	2.763	1.210
15	15	3.359	0.944	3.768	1.068	3.961	1.200
20	20	5.100	0.955	5.594	1.075	5.585	1.215
				Uniform			
5	5	1.370	0.959	1.600	1.054	1.765	1.115
5	10	1.587	0.968	1.826	1.032	1.980	1.076
7	8	1.795	0.952	2.056	1.049	2.205	1.133
10	10	2.415	0.945	2.711	1.054	2.758	1.139
15	15	4.027	0.917	4.360	1.015	4.033	1.092
20	20	6.132	0.904	6.590	0.995	5.766	1.037

Note—The values represent the ratio between serially dependent data with trend and independent series with no trend. PND, percentage of nonoverlapping data; PNCD, percentage of nonoverlapping corrected data.

Table 6
Distortion As a Result of Combined Presence of Trend and an MA(1) Process

Phase Length		Ratio Trend and $\theta_1 = -.5$/ Random Fluctuations		Ratio Trend and $\theta_1 = .5$/ Random Fluctuations	
n_A	n_B	PND	PNCD	PND	PNCD
			Exponential		
5	5	1.217	0.888	1.597	1.174
5	10	1.406	0.910	1.837	1.153
7	8	1.418	0.892	1.874	1.196
10	10	1.658	0.877	2.153	1.184
15	15	2.268	0.910	2.839	1.201
20	20	3.099	0.920	3.727	1.182
			Normal		
5	5	1.608	1.075	1.253	.924
5	10	1.850	1.062	1.462	.947
7	8	1.962	1.078	1.562	.919
10	10	2.412	1.072	1.970	.909
15	15	3.619	1.066	3.074	.900
20	20	5.231	1.052	4.601	.895
			Uniform		
5	5	1.602	1.082	1.271	.933
5	10	1.831	1.054	1.485	.949
7	8	1.974	1.056	1.617	.924
10	10	2.539	1.039	2.150	.891
15	15	3.938	0.992	3.503	.857
20	20	5.815	0.949	5.281	.836

Note—The values represent the ratio between moving average data with trend and data series with $\theta_1 = 0$ and no trend. PND, percentage of nonoverlapping data; PNCD, percentage of nonoverlapping corrected data.

Discrimination Between Data Patterns

In general, the desirable characteristics of an effect-size procedure are to be sensitive to intervention effects and are not to be affected, for instance, by trend or serial dependence. Hence, an optimal performance (illustrated by Figure 4) would imply (1) low effect-size estimates in the absence of a treatment effect; (2) low effect-size estimates when there is only a general trend; and (3) higher estimates when there are actual changes in the response rate because of an intervention.

Comparing this ideal discrimination with the estimates obtained by means of PND and PNCD, it can be seen that there is a greater resemblance in the case of the latter procedure. That is, a combined effect (change both in level and in slope) yields a greater effect-size estimate than does an individual effect, and the percentage obtained in the absence of an intervention effect is even lower. Additionally, trend does not shift estimates up, as is the case for PND, which detects trend as an intervention effect. Figure 5 illustrates these findings for the shortest series length studied.

DISCUSSION

The present investigation proposed a data correction step to be introduced prior to applying the PND as a tech-

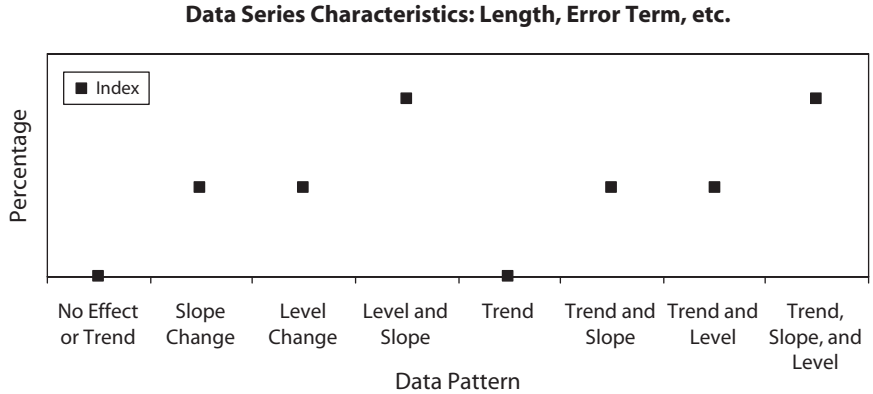


Figure 4. Ideal discrimination between data patterns.

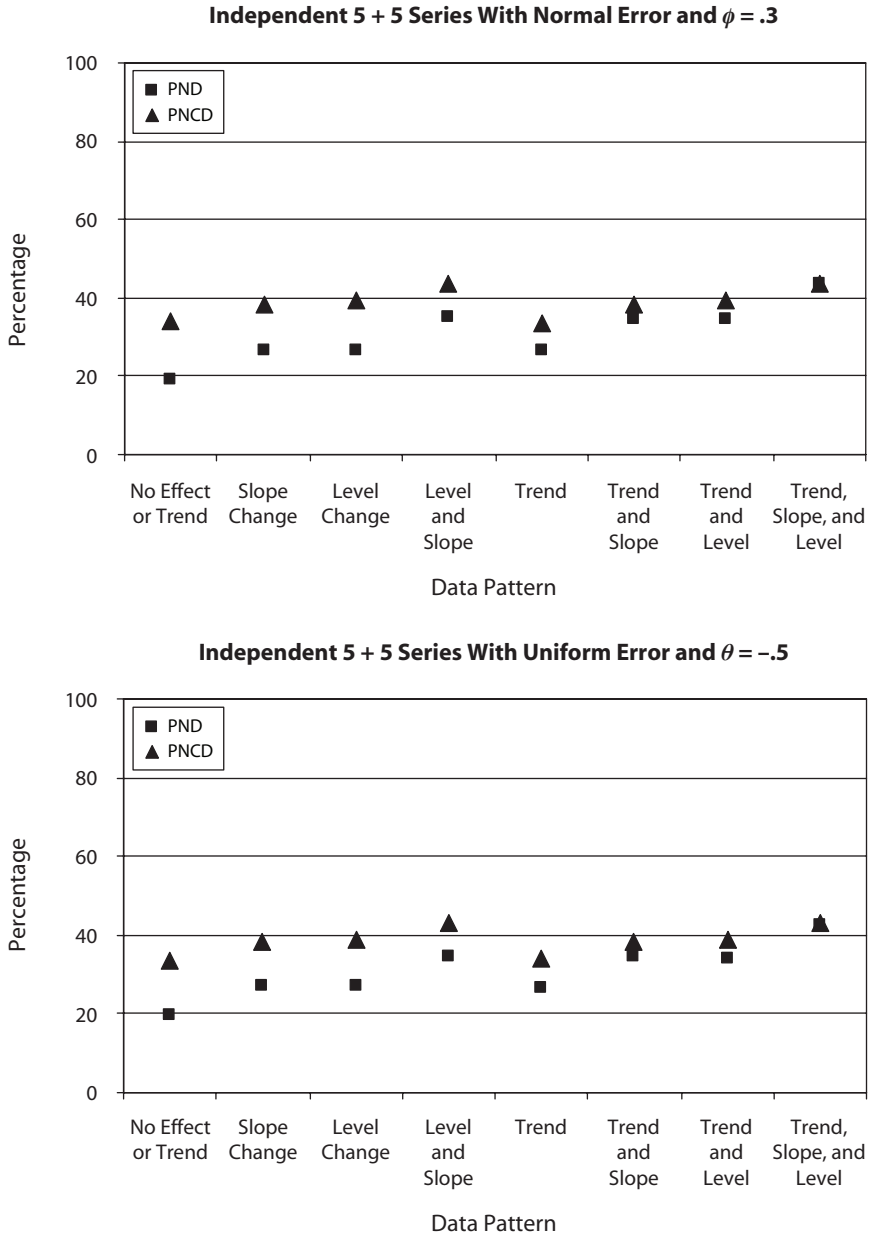


Figure 5. Discrimination between data patterns for both indices in different experimental conditions. Upper panel: $N = 10$ series generated from an AR process with normal error and $\phi_1 = .3$. Lower panel: $N = 10$ series generated from an MA process with uniform error and $\theta_1 = -.5$.

nique for quantifying treatment effectiveness. The modified procedure was compared with the original in the context of data sets generated with known attributes, such as trend, autocorrelation, and treatment effect. For applied researchers, the results that were obtained suggest that PNCD is an effective method to deal with trend, and that it can, therefore, be used in situations when preintervention measurements are not pure random fluctuation. Unstable baselines have been regarded as undesirable, but they can be common in applied settings in which the introduction of the treatment is subjected to factors that cannot always be controlled by the practitioners. Although a professional might be reluctant to initiate the intervention when there is trend in data, treatment administration may be imposed by institutional time schedules, a client's availability, and so on. In such a case, some kind of statistical control is advisable (Kazdin, 1978), and it can be achieved by means of the procedure proposed in the present article. Apart from behavioral data with baseline trends, another potential context for application of PNCD consists of studies in which the data points are not sufficiently spaced in time and can present a sequential relation. PNCD ought to be preferred to PND in these cases, because of the fact that autocorrelation is more problematic for the latter.

Whenever the behavioral measurements are not serially dependent and do not present a trend, PND may be a better option than PNCD, since it produces a lower magnitude of effect estimates. This difference in the estimates implies that in the aforementioned cases, PND is less likely to label an intervention as effective when it is not. It has already been discussed that different effect-size procedures may lead to different conclusions about the degree of treatment effectiveness for the same data set (McGrath & Meyer, 2006; Parker et al., 2005). In the particular case of PND and PNCD, the difference in estimates implies that the interpretation benchmarks proposed by Scruggs and Mastropieri (1998) cannot be applied directly to PNCD. On the other hand, there is evidence that PND is conservative, as compared with other procedures for estimating magnitude of effect (Jenson et al., 2007). Therefore, the effect-size estimates provided by PNCD may resemble more the ones obtained by other models.

From a methodological perspective, PNCD can be regarded as an attempt to improve a procedure that is attractive to applied psychologists and is frequently employed by them. The aim is not only to achieve a better performance, but also to maintain the simplicity of the technique. Therefore, we consider that the modifications balancing statistical properties improvements and low levels of calculus/interpretative complexity have to be encouraged. Furthermore, the present study followed the practice of offering data analysis programs for single-case designs in freeware, such as R (see, e.g., Bulté & Onghena, 2008)—a practice that we feel ought to be promoted.

The present investigation focused only on AB designs, although the results are potentially applicable to multiple-baseline designs (Busse, Kratochwill, & Elliott, 1995). The data sets used in the present study were constructed using permanent linear trend, constant variance, and constant autocorrelation throughout the whole series. These data assumptions are common to simulation studies on $N = 1$ designs (see, e.g., Brossart et al., 2006; Huitema & McKean, 2007a, 2007b; Matyas & Greenwood, 1990; Parker & Brossart, 2003). Thus, future studies may explore the performance of PNCD for ABAB designs with curvilinear trends, computing the percentage for each change in the condition, as was suggested by Kromrey and Foster-Johnson (1996). Additionally, comparative studies such as the present one, which center on finding the technique that performs better, need to be complemented by precision studies in order to identify techniques that perform well—that is, yield accurate estimates of the effect sizes simulated.

AUTHOR NOTE

The present research was supported by the Comissionat per a Universitats i Recerca del Departament d'Innovació, Universitats i Empresa of the Generalitat de Catalunya, and the European Social Fund. The authors thank the anonymous reviewers for their useful comments and suggestions, which contributed to improving the manuscript. Correspondence concerning this article should be addressed to R. Manolov, Departament de Metodologia de les Ciències del Comportament, Facultat de Psicologia, Universitat de Barcelona, Passeig de la Vall d'Hebron, 171, 08035-Barcelona, Spain (e-mail: rrumenov13@ub.edu).

REFERENCES

ALLISON, D. B., & GORMAN, B. S. (1994). "Make things as simple as possible, but no simpler": A rejoinder to Scruggs and Mastropieri. *Behaviour Research & Therapy*, **32**, 885-890.

BERETVAS, S. N., & CHUNG, H. (2008). An evaluation of modified R^2-change effect size indices for single-subject experimental designs. *Evidence-Based Communication Assessment & Intervention*, **2**, 120-128.

BRADLEY, J. V. (1977). A common situation conducive to bizarre distribution shapes. *American Statistician*, **31**, 147-150.

BROSSART, D. F., PARKER, R. I., OLSON, E. A., & MAHADEVAN, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification*, **30**, 531-563.

BULTÉ, I., & ONGHENA, P. (2008). An R package for single-case randomization tests. *Behavior Research Methods*, **40**, 467-478.

BUSSE, R. T., KRATOCHWILL, T. R., & ELLIOTT, S. N. (1995). Meta-analysis for single-case consultation outcomes: Applications to research and practice. *Journal of School Psychology*, **33**, 269-285.

COHEN, J. (1990). Things I have learned (so far). *American Psychologist*, **45**, 1304-1312.

COHEN, J. (1994). The earth is round ($p < .05$). *American Psychologist*, **49**, 997-1003.

CRANE, D. R. (1985). Single-case experimental designs in family therapy research: Limitations and considerations. *Family Process*, **24**, 69-77.

GEDO, P. M. (2000). Single case studies in psychotherapy research. *Psychoanalytic Psychology*, **16**, 274-280.

GREENWOOD, K. M., & MATYAS, T. A. (1990). Problems with application of interrupted time series analysis for brief single-subject data. *Behavioral Assessment*, **12**, 355-370.

HARROP, J. W., & VELICER, W. F. (1985). A comparison of alternative approaches to the analysis of interrupted time-series. *Multivariate Behavioral Research*, **20**, 27-44.

HUITEMA, B. E., & MCKEAN, J. W. (2000). Design specification issues in time-series intervention models. *Educational & Psychological Measurement*, **60**, 38-58.

HUITEMA, B. E., & MCKEAN, J. W. (2007a). Identifying autocorrelation generated by various error processes in interrupted time-series progression designs: A comparison of AR1 and portmanteau tests. *Educational & Psychological Measurement*, **67**, 447-459.

HUITEMA, B. E., & MCKEAN, J. W. (2007b). An improved portmanteau test for autocorrelated errors in interrupted time-series regression models. *Behavior Research Methods*, **39**, 343-349.

HUITEMA, B. E., MCKEAN, J. W., & MCKNIGHT, S. (1999). Autocorrelation effects on least-squares intervention analysis of short time series. *Educational & Psychological Measurement*, **59**, 767-786.

JENSON, W. R., CLARK, E., KIRCHER, J. C., & KRISTJANSSON, S. D. (2007). Statistical reform: Evidence-based practice, meta-analyses, and single subject designs. *Psychology in the Schools*, **44**, 483-493.

KAZDIN, A. (1978). Methodological and interpretive problems of single-case experimental designs. *Journal of Consulting & Clinical Psychology*, **46**, 629-642.

KIRK, R. E. (1996). Practical significance: A concept whose time has come. *Educational & Psychological Measurement*, **56**, 746-759.

KRATOCHWILL, T. R., & BRODY, G. H. (1978). Single subject designs: A perspective on the controversy over employing statistical inference and implications for research and training in behavior modification. *Behavior Modification*, **2**, 291-307.

KROMREY, J. D., & FOSTER-JOHNSON, L. (1996). Determining the efficacy of intervention: The use of effect sizes for data analysis in single-subject research. *Journal of Experimental Education*, **65**, 73-93.

MA, H. H. (2006). An alternative method for quantitative synthesis of single-subject research: Percentage of data points exceeding the median. *Behavior Modification*, **30**, 598-617.

MANOLOV, R., & SOLANAS, A. (2008). Comparing $N = 1$ effect size indices in presence of autocorrelation. *Behavior Modification*, **32**, 860-875.

MANOLOV, R., SOLANAS, A., & LEIVA, D. (in press). Comparing "visual" effect size indices for single-case designs. *Methodology—European Journal of Research Methods for the Behavioral & Social Sciences*.

MATYAS, T. A., & GREENWOOD, K. M. (1990). Visual analysis for single-case time series: Effects of variability, serial dependence, and mag-

nitude of intervention effects. *Journal of Applied Behavior Analysis*, **23**, 341-351.
McCleary, R., & Hay, R. A., Jr. (1980). *Applied time series analysis for the social sciences.* Beverly Hills, CA: Sage.
McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of *r* and *d*. *Psychological Methods*, **11**, 386-401.
Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, **105**, 156-166.
Parker, R. I., & Brossart, D. F. (2003). Evaluating single-case research data: A comparison of seven statistical methods. *Behavior Therapy*, **34**, 189-211.
Parker, R. I., Brossart, D. F., Vannest, K. J., Long, J. R., Garcia De-Alba, R., Baugh, F. G., & Sullivan, J. R. (2005). Effect sizes in single case research: How large is large? *School Psychology Review*, **34**, 116-132.
Parker, R. I., Cryer, J., & Byrns, G. (2006). Controlling baseline trend in single-case research. *School Psychology Quarterly*, **21**, 418-443.
Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data: An alternative to PND. *Journal of Special Education*, **40**, 194-204.
Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, **44**, 1276-1284.
Schlosser, R. W., Lee, D. L., & Wendt, O. (2008). Application of the percentage of non-overlapping data (PND) in systematic reviews and meta-analyses: A systematic review of reporting characteristics. *Evidence-Based Communication Assessment & Intervention*, **2**, 163-187.
Schlosser, R. W., & Sigafoos, J. (2008). Meta-analysis of single-subject designs: Why now? *Evidence-Based Communication Assessment & Intervention*, **2**, 117-119.
Schneider, N., Goldstein, H., & Parker, R. (2008). Social skills interventions for children with autism: A meta-analytic application of percentage of all non-overlapping data (PAND). *Evidence-Based Communication Assessment & Intervention*, **2**, 152-162.
Scruggs, T. E., & Mastropieri, M. A. (1998). Summarizing single-subject research: Issues and applications. *Behavior Modification*, **22**, 221-242.
Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial & Special Education*, **8**, 24-33.
Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment & Intervention*, **2**, 188-196.
Symon, J. B. (2005). Expanding interventions for children with autism: Parents as trainers. *Journal of Positive Behavior Interventions*, **7**, 159-173.
Tervo, R. C., Estrem, T. L., Bryson-Brockman, W., & Symons, F. J. (2003). Single-case experimental designs: Application in developmental–behavioral pediatrics. *Developmental & Behavioral Pediatrics*, **24**, 438-448.
Wilkinson, L., & the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, **54**, 694-704.

APPENDIX A

An example of R code computing PND and PNCD as output. The input required from the user is (1) the data for phase A in the expression `phaseA <- c(1:10)`, replacing "`1:10`" with the measurements obtained separated by commas; and (2) the data for phase B placed instead of "`11:20`" in the expression `phaseB <- c(11:20)`. After introducing the behavioral measurements, the text is copied and pasted into the R console, and the estimates are printed out.

```
# Data input
phaseA <- c(1:10)
phaseB <- c(11:20)
n_a <- length(phaseA)
n_b <- length(phaseB)

# Data correction: phase A
phaseAdiff <- c(1:(n_a-1))
for (iter1 in 1:(n_a-1))
phaseAdiff[iter1] <- phaseA[iter1+1] - phaseA[iter1]
phaseAcorr <- c(1:n_a)
for (iter2 in 1:n_a)
phaseAcorr[iter2] <- phaseA[iter2] - mean(phaseAdiff)*iter2

# Data correction: phase B
phaseBcorr <- c(1:n_b)
for (iter3 in 1:n_b)
phaseBcorr[iter3] <- phaseB[iter3] - mean(phaseAdiff)*(iter3+n_a)

# PND on corrected data
countcorr <- 0
for (iter4 in 1:n_b)
if (phaseBcorr[iter4] > max(phaseAcorr)) countcorr <- countcorr+1
pndcorr <- (countcorr/n_b)*100
print ("The percentage of nonoverlapping corrected data is"); print(pndcorr)

# PND on original data
count <- 0
for (iter5 in 1:n_b)
if (phaseB[iter5] > max(phaseA)) count <- count+1
pnd <- (count/n_b)*100
print ("The percent of nonoverlapping data is"); print(pnd)
```

APPENDIX B

R code computing PND and PNCD as output used, as is described in Appendix A. This is useful when the objective of the behavior of interest is an undesirable one and the treatment pretends to eliminate or reduce it.

```
# Data input
phaseA <- c(1:10)
phaseB <- c(11:20)
n_a <- length(phaseA)
n_b <- length(phaseB)

# Data correction: phase A
phaseAdiff <- c(1:(n_a-1))
for (iter1 in 1:(n_a-1))
phaseAdiff[iter1] <- phaseA[iter1+1] - phaseA[iter1]
phaseAcorr <- c(1:n_a)
for (iter2 in 1:n_a)
phaseAcorr[iter2] <- phaseA[iter2] - mean(phaseAdiff)*iter2

# Data correction: phase B
phaseBcorr <- c(1:n_b)
for (iter3 in 1:n_b)
phaseBcorr[iter3] <- phaseB[iter3] - mean(phaseAdiff)*(iter3+n_a)

# PND on corrected data
countcorr <- 0
for (iter4 in 1:n_b)
if (phaseBcorr[iter4] < min(phaseAcorr)) countcorr <- countcorr+1
pndcorr <- (countcorr/n_b)*100
print ("The percentage of nonoverlapping corrected data is"); print(pndcorr)

# PND on original data
count <- 0
for (iter5 in 1:n_b)
if (phaseB[iter5] < min (phaseA)) count <- count+1
pnd <- (count/n_b)*100
print ("The percent of nonoverlapping data is"); print(pnd)
```

(Manuscript received February 26, 2009;
revision accepted for publication July 14, 2009.)

Estudi 4©

Estimating slope and level change in N=1 designs

Antonio Solanas[1], Rumen Manolov[1] i Patrick Onghena[2]

[1] Departament de Metodologia de les Ciències del Comportament
Facultat de Psicologia
Universitat de Barcelona

[2] Department of Educational Sciences
Faculty of Psychology and Educational Sciences
Katholieke Universiteit Leuven

Estimating slope and level change in single-case designs

Resum. L'estudi 4 es centra en proposar un nou procediment (SLC) per estimar el canvi de nivell i el canvi de pendent de forma separada. Aquest procediment inclou diversos passos, el primer dels quals pretén eliminar l'impacte de la tendència general en les dades que no es relaciona amb la introducció del tractament. Seguidament s'estima el canvi de pendent i, finalment, s'estima el canvi de nivell, havent controlat la tendència i el canvi de pendent. L'aplicació de l'SLC s'exemplifica mitjançant dades fictícies, mostrant la possibilitat de dur a terme els càlculs a mà. El biaix i la variància dels dos estimadors s'avalua en un estudi de simulació Monte Carlo i es compara amb un model de regressió múltiple que es correspon perfectament amb el model de generació de les dades i que, per tant, serveix com a *gold standard*. Les condicions experimentals inclouen dos processos de generació de dades (autoregressiu i de mitjanes mòbils), tres distribucions de la variable aleatòria (exponencial, normal i uniforme), diferents graus de dependència serial i diferents tipus d'efecte de la intervenció. Els resultats indiquen que cap de les dues tècniques estudiades presenta biaix. Quant a l'eficiència, l'SLC té un rendiment menys variable en casos d'autocorrelació positiva. La nova tècnica és especialment recomanable quan hi ha un canvi progressiu en la conducta, cosa que sembla lògica en els lents processos psicològics, ja que el estimador del canvi de pendent és més eficient que el estimador del canvi de nivell.

Behavior Modification
34(3) 195–218

DOI: 10.1177/0145445510363306
http://bmo.sagepub.com
SAGE

Estimating Slope and Level Change in $N = 1$ Designs

Antonio Solanas,[1] Rumen Manolov,[1] and Patrick Onghena[2]

Abstract

The current study proposes a new procedure for separately estimating slope change and level change between two adjacent phases in single-case designs. The procedure eliminates baseline trend from the whole data series before assessing treatment effectiveness. The steps necessary to obtain the estimates are presented in detail, explained, and illustrated. A simulation study is carried out to explore the bias and precision of the estimators and compare them to an analytical procedure matching the data simulation model. The experimental conditions include 2 data generation models, several degrees of serial dependence, trend, and level and/or slope change. The results suggest that the level and slope change estimates provided by the procedure are unbiased for all levels of serial dependence tested and trend is effectively controlled for. The efficiency of the slope change estimator is acceptable, whereas the variance of the level change estimator may be problematic for highly negatively autocorrelated data series.

Keywords

single-case designs, level change, slope change, autocorrelation, trend

[1]Universitat de Barcelona, Spain
[2]Katholieke Universiteit Leuven, Belgium

Corresponding Author:
Rumen Manolov, Departament de Metodologia de les Ciencies del Comportament, Facultat de Psicologia, Universitat de Barcelona, Passeig de la Vall d'Hebron 171 08035 Barcelona, Spain
Email: rrumenov13@ub.edu

In recent years it has become widely accepted that psychological studies based on single-case designs need to provide scientific evidence on the interventions applied, which would allow subsequent $N = 1$ studies to have a solid foundation (Jenson, Clark, Kircher, & Kristjansson, 2007; Olive & Smith, 2005; Parker & Brossart, 2006; Parker & Hagan-Burke, 2007; Schlosser & Sigafoos, 2008; Shadish, Rindskopf, & Hedges, 2008). The evidence on the effectiveness of intervention techniques can be accumulated integrating individual studies in a quantitative way, for instance, considering different levels of analysis and incorporating study characteristics as effect moderators (Van den Noortgate & Onghena, 2003, 2008).

The availability of useful summary measures is essential for quantifying the results of an individual study and also for conducting meta-analyses (Busk & Serlin, 1992; Cohen, 1990, 1994; Kirk, 1996; Kromrey & Foster-Johnson, 1996; Rosnow & Rosenthal, 1989; Wilkinson & Task Force on Statistical Inference, 1999). A "useful" effect size index needs to meet the following criteria: (a) to represent correctly the true data characteristics, (b) to offer valuable and easily interpretable information, and (c) to be easily applicable by researchers with scarce statistical expertise. Regression-based techniques (e.g., Allison & Gorman, 1993; Center, Skiba, & Casey, 1985-1986; Gorsuch, 1983; White, Rusch, Kazdin, & Hartmann, 1989) have been found to perform unsatisfactorily with respect to criterion 1 (Beretvas & Chung, 2008b; Brossart, Parker, Olson, & Mahadevan, 2006; Manolov & Solanas, 2008; Parker & Brossart, 2003), while also being deficient regarding criterion 2, due to the large effect size estimates yielded (Campbell, 2004; Scruggs & Mastropieri, 1998). Interpretability is also problematic for ITSE (Gottman, 1981) and ITSACORR (Crosbie, 1993), since the meaningfulness of their results is compromised by conceptual and computational issues (Huitema, 2004; Huitema, McKean, & Laraway, 2007). As regards criterion 3, the procedures based on visual analysis (i.e., Ma, 2006; Parker, Hagan-Burke, & Vannest, 2007; Scruggs, Mastropieri, & Casto, 1987) can be easily applied even by hand calculation. Although the percentages yielded by these techniques are also readily interpretable (criterion 2), trend and autocorrelation have been shown to jeopardize the completion of criterion 1 (Manolov, Solanas, & Leiva, in press).

In the following section a new procedure for assessing treatment effectiveness is proposed, taking into consideration the need for separate quantification of slope change and level change to represent more fully the information contained in single-case studies (Beretvas & Chung, 2008a). The procedure is designed to yield meaningful results (criterion 2) expressed in terms of the behavioral measurement units used in each individual study. Subsequently, the interpretation of the estimates produced by the procedure is discussed and

illustrated with an example. To meet criterion 3 different software solutions are proposed to researchers and practitioners. Monte Carlo simulation is used to evaluate how well the estimates represent known data parameters, that is, to obtain information about the degree of achievement of criterion 1.

A Procedure for Estimating Slope and Level Change (SLC)

Rationale

The objective of the proposed procedure is to estimate slope and level change (being present either of them or both) eliminating baseline data linear trend whenever it is present. After a potential phase A trend is removed, it can be logically assumed that the level of behavior in that phase presents zero slope. It is conjectured that the procedure may also deal with positive serial dependence, since the presence of large positive autocorrelation in data can be represented by an upward or a downward trend. In contrast, when measurements are negatively autocorrelated, data present greater variability rather than trend. The slope and level change estimates obtained for the detrended data express the shifts in terms of the measurement units. For instance, if the frequency of behavior is measured, a slope change of 3 would mean that at each measurement time during phase B the experimental unit produces an average of three behaviors more than in the previous observation point. A level change of 3 would imply that with the introduction of the intervention (i.e., with the change in phase) the experimental unit produces an average of three behaviors more in the treatment phase than in the baseline phase. It has to be remarked that this average change in level is computed after estimating and controlling slope change.

Steps Required for Computing SLC

In the present paragraph the computation of the SLC estimates is explained verbally, whereas the mathematical expressions and an example can be found in Appendix A. Since SLC is designed to control general trend before assessing intervention effectiveness, an initial data correction step involves eliminating phase A trend from data. Trend is estimated only for the baseline phase, as it allows avoiding confusion between trend and potential intervention effects taking place in phase B (Allison & Gorman, 1993). Trend estimation is not carried out by means of ordinary least squares, since unstable estimates are expected on theoretical basis when few data points are available

(Weisberg, 1980). Additionally, there is evidence on the inappropriate performance of Gorsuch's (1983) trend analysis which uses this kind of estimation (Manolov & Solanas, 2008; Parker & Brossart, 2003). Instead, the phase A trend is estimated as the mean of the differenced phase A measurements. Afterwards trend is removed from both phase A and phase B, using a method that has been shown to be useful for dealing not only with trend but also with autocorrelation (Manolov & Solanas, 2009).

The second step involves estimating slope change as the trend present in the phase B data, from which baseline trend has already been removed. The average value is assumed to represent an estimation of slope change, considering that the phase A trend has been previously removed and, thus, the baseline is supposed to have a zero slope.

The third step consists in the estimation of level change. Firstly, the already estimated change in slope is removed from the treatment phase data, without removing the intercept. That is, the phase B slope is eliminated from the detrended phase B data, while maintaining potential shifts taking place at the first measurement time of the treatment phase. Level change is estimated by subtracting the detrended baseline data mean from the detrended and slope-change-controlled treatment data mean.

The procedure described is not restricted to AB designs and can be applied to any combination of a baseline and treatment phase which is included in more complex design structures (e.g., multiple-baselines designs, ABAB designs).

Software Availability

The example in Appendix A shows that SLC can be applied using hand calculations. However, for longer data series this can become tedious. Therefore, the procedure has been implemented in an MS-DOS executable file and in two well-known and widely used statistical packages: R (version 2.9.2) and SAS (version 9.1).

The *slc.exe* file (available from the authors) asks for series length, baseline phase length, and the name of the *.dat* file containing the data points separated by spaces. The estimates are printed in an output file, whose name needs to be specified by the user. However, this program does not permit graphing data and, thus, the R and SAS/IML codes were developed.

R is a freeware language which has already been used to make automatic the application of other techniques for analyzing single-case data (Bulté & Onghena, 2008; Manolov & Solanas, 2009). For SLC, a package and a plug-in (available upon request and from the *http://cran.r-project.org* Web site) containing the R function presented in Appendix B were developed. Both the

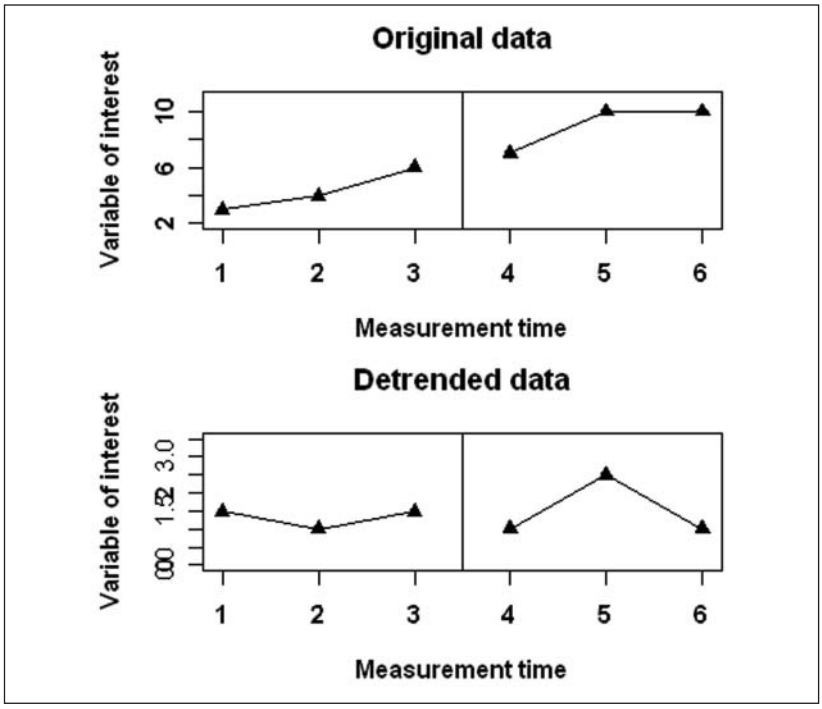


Figure 1. Example of the graphical output of the R function and package "SLC"
Note: SLC = slope and level change

function and the packages compute the slope and level change estimates and represent graphically the original and detrended data (e.g., Figure 1). One of the versions of the package requires using three commands to obtain the results, whereas the other one is based on menus. Appendix B explains the use of the two packages

The SAS/IML code presented in Appendix C also permits obtaining the slope change and level change estimates, by simply inputting the measurements and baseline phase length. The graph of the detrended data which is drawn using the code allows complementing numerical analysis with visual inspection.

The diversity of possibilities mentioned earlier makes the application of SLC straightforward. In fact, the use of any piece of software requires fewer steps than computing, for instance, the percentage of all nonoverlapping data effect size estimate, as described in Schneider, Goldstein, and Parker (2008).

Method

AB Series Lengths

Short data series were included in the present study, since those are more feasible in applied settings: (a) $N = 10$ with $n_A = n_B = 5$; (b) $N = 15$ with $n_A = 5$, $n_B = 10$; (c) $N = 15$ with $n_A = 7$, $n_B = 8$; (d) $N = 20$ with $n_A = n_B = 10$; (e) $N = 30$ with $n_A = n_B = 15$; and (f) $N = 40$ with $n_A = n_B = 20$, where N denotes whole series length and n_A and n_B represent the number of measurements in phase A and B, respectively.

Data Generation Models

For each combination of n_A and n_B data were generated according to the model proposed by Huitema and McKean (2000): $y_t = \beta_0 + \beta_1 \cdot T_t + \beta_2 \cdot LC_t + \beta_3 \cdot SC_t + \varepsilon_t$, where y_t is the value of the dependent variable at moment t, β_0 is the intercept (set to zero), β_1, β_2, and β_3 are the coefficients associated with trend, level change, and slope change respectively, T_t is the value of the time variable at moment t (taking values from 1 to N), LC_t is a dummy variable for level change (equal to 0 for phase A and to 1 for phase B), SC_t is a dummy variable for slope change being equal to 0 for phase A, and taking values from 0 to ($n_B - 1$) for phase B, and ε_t is the error term. This data generation and variables specification method has been previously used in studies related to simulating single-case behavioral data (Beretvas & Chung, 2008b; Huitema & McKean, 2007; McKnight, McKean, & Huitema, 2000).

The beta parameters related to trend, level and slope change were set to 1 and 10 to represent a small and a large effect, respectively. The aim of these parameter values was to explore the impact of effect size on the bias and variance of the estimators. It has to be adverted that a permanent level change of one behavior implies a smaller effect than a progressive slope change of one behavior per measurement time.

The error term (ε_t) was generated following two different models which are assumed to represent adequately the greater part of behavioral data (Harrop & Velicer, 1985): (a) the first-order autoregressive model $\varepsilon_t = \varphi_1 \cdot \varepsilon_{t-1} + u_t$, with φ_1 ranging from –.9 to .9 in steps of .1 and (b) the first-order moving average model $\varepsilon_t = u_t - \theta_1 \cdot u_{t-1}$ (as presented in McCleary & Hay, 1980) with 19 values of θ_1: −.9 (.1) .9, leading to autocorrelation ranging from .4972 to −.4972.

For both models the random variable u_t was generated following three different distributions (exponential, normal, and uniform) with mean equal to zero and standard deviation equal to 1, 2, and 3 respectively, constituting a total of nine different u_t distributions. These conditions permit studying the

effect of skewness, kurtosis, and data variability on the performance of the procedure proposed.

Data Analysis

For each experimental condition defined by the combination of ε_t model, u_t distribution, and β_1, β_2, and β_3 values the mean and variance of the two estimators were computed on the basis of 100,000 samples. The SLC level and slope change estimates were compared to simultaneous multiple regression (SMR; Huitema & McKean, 2000) coefficient estimates in terms of bias and variance. The bias of the estimators was obtained as the difference between the simulation parameters and the estimates for slope and level change. The variance of the estimators was computed as an indicator of the efficiency and a comparison was performed between SLC and SMR. The SMR procedure can be considered a gold standard, as it matches perfectly the data generation model used.

Simulation

The specific steps that were implemented in the FORTRAN programs (one for each of the six series' length) were the following ones:

1) Systematic selection of each of the 19 values of φ_1 or θ_1 autoregressive or moving average models, respectively.
2) Systematic selection of the (β_1, β_2, and β_3) parameters for data generation, leading to eight different data patterns—absence of effect or trend, presence of trend, level change, slope change, trend and level change, trend and slope change, combined level and slope change, and trend and combined level and slope change.
3) 100,000 iterations of steps 4 through 11.
4) Generate the u_t term according to an exponential, a normal, or a uniform distribution with different values of the standard deviation parameter.
5) Establish $\varepsilon_1 = u_1$.
6) Obtain the error term ε_t out of the random variable u_t using the AR(1) model $\varepsilon_t = \varphi_1 \cdot \varepsilon_{t-1} + u_t$ or the MA(1) model $\varepsilon_t = u_t - \theta_1 u_{t-1}$.
7) Obtain the time array $T_t = 1, 2, \ldots, N$.
8) Obtain the dummy treatment variable array LC_t, where $LC_t = 0$ for phase A and $LC_t = 1$ for phase B.
9) Obtain the slope change array according to: $SC_t = (T_t - [n_A + 1])\ LC_t$.

10) Obtain the y_t array containing measurements (i.e., dependent variable): $y_t = \beta_0 + \beta_1 \cdot T_t + \beta_2 \cdot LC_t + \beta_3 \cdot SC_t + \varepsilon_t$.
11) Apply SLC and SMR.
12) Obtain the mean and variance of the slope change and level change estimates the 100,000 replications of each experimental condition.
13) Compute the bias of the estimates.

For data generation NAG libraries *nag_rand_neg_exp*, *nag_rand_normal*, and *nag_rand_uniform* were used. To guarantee suitable simulated data, the 50 values previous to each simulated data series were eliminated to reduce artificial effects (Greenwood & Matyas, 1990) and to avoid dependence between successive data series (Huitema, McKean, & McKnight, 1999).

Results

Detection of Treatment Effects

Both SLC and SMR proved to be unbiased, that is, null estimates were obtained when the simulation parameter beta was zero. Complementarily, when $\beta_2 = 1$ (or 10) or $\beta_3 = 1$ (or 10), the estimates were equal to 1 or 10, respectively. In fact, the estimators' bias and variance for small and large effect differed only at the third decimal level and in the results presented in following sections are applicable to both cases. Thus, SLC detects both level and slope changes, whenever either of them or both are present. Moreover, when the intervention is ineffective, the estimates are equal to zero, indicating that the change in phase is not associated with an alteration in the behavior of interest. The size (0, 1, or 10) of either type of treatment effect does not alter the variance of the estimates of SLC and SMR. This implies that the importance of the variability decreases for greater effects as the relative efficiency of the procedure increases for stronger interventions. For instance, for first-order autoregressive series with normal unitary error, 15 measurements in each phase, level change of 1, and $\varphi_1 = .3$, the coefficient of variation of the SLC level change estimate is 1.241/1 = 124.1%. For the same case and a level change of 10, the coefficient of variation is 1.245/10 = 12.45%, according to the results obtained.

Distortion Due to Trend

According to the results obtained, it can be stated that SLC controls effectively for general trend, as it does not distort the estimates obtained. Whenever the

Table 1. Variance of the SLC Estimators for Data Series Generated Using an AR(1) Model With Exponential Error and With Null Beta Parameters

	$n_A = n_B = 5$		$n_A = n_B = 10$		$n_A = n_B = 15$	
Autocorrelation	LC	SC	LC	SC	LC	SC
–.9	4.71	.30	10.44	.09	6.30	.04
–.6	3.15	.22	2.63	.05	2.32	.02
–.3	2.38	.23	1.73	.05	1.54	.02
.0	2.04	.25	1.45	.05	1.30	.02
.3	1.92	.31	1.41	.06	1.24	.03
.6	1.95	.41	1.69	.09	1.51	.04
.9	1.79	.52	2.36	.20	2.89	.10

Note: LC = level change estimators; SC = slope change estimators; SLC = slope and level change.

intervention is not effective, the presence of trend in the desired direction does not lead to erroneously inferring treatment effectiveness. Additionally, when treatment is effective, its effect is not overestimated because of the presence of trend. The level and slope change estimates of SMR are also not affected by trend, whose magnitude is estimated precisely, without bias. The presence or absence of trend does not alter the variance of the estimates of SLC and SMR.

Distortion Due to Autocorrelation

The presence of autocorrelation in data is not associated with biased estimates. In contrast, serial dependence is relevant for the variance of the estimates obtained. The variance of the estimators is greater for higher degrees of negative and lesser extent for positive autocorrelation (see Tables 1 and 2 for SLC and Tables 3 and 4 for SMR). Another finding common to SLC and SMR and both data generation processes is the greater variability of the level change estimator in comparison to the slope change estimator.

Regarding variability, SLC and SMR differ to a greater degree in the case of the level change estimate where the sign of the serial dependence becomes especially relevant. Negative autocorrelation distorts more the SLC estimator and positive one has greater effect on the SMR estimator for both autoregressive (Figure 2) and moving average (Figure 3) processes. Apart from the greater distortion in relative terms (comparing sequentially related to independent data), SMR also shows greater absolute estimator variability for $\varphi_1 \geq .3$ in AR(1) processes and for $\varphi_1 \geq 0.275$ for $n_A = n_B = 5$ data generated using a MA(1) process.

Table 2. Variance of the SLC Estimators for Data Series Generated Using an MA(1) Model With Normal Error and With Null Beta Parameters.

	$n_A = n_B = 5$		$n_A = n_B = 10$		$n_A = n_B = 15$	
Autocorrelation	LC	SC	LC	SC	LC	SC
−.500	4.40	.31	3.14	.06	2.80	.03
−.400	2.80	.23	2.01	.05	1.79	.02
−.275	2.39	.23	1.72	.05	1.52	.02
−.099	2.10	.24	1.51	.05	1.35	.02
.000	2.03	.25	1.45	.05	1.29	.02
.099	1.95	.26	1.40	.05	1.24	.02
.275	1.98	.30	1.41	.06	1.24	.03
.400	2.11	.36	1.50	.07	1.29	.03
.500	3.03	.56	2.14	.11	1.86	.05

Note: LC = level change estimators; SC = slope change estimators; SLC = slope and level change.

Table 3. Variance of the SMR Estimators for Data Series Generated Using an AR(1) Model With Exponential Error and With Null Beta Parameters.

	$n_A = n_B = 5$		$n_A = n_B = 10$		$n_A = n_B = 15$	
Autocorrelation	LC	SC	LC	SC	LC	SC
−.9	2.38	.13	1.15	.01	0.49	.00
−.6	1.65	.12	0.57	.01	0.32	.00
−.3	1.56	.15	0.62	.02	0.38	.01
.0	1.70	.20	0.81	.02	0.54	.01
.3	1.93	.29	1.23	.04	0.89	.01
.6	2.07	.41	2.02	.09	1.80	.03
.9	1.92	.53	2.96	.22	3.92	.12

Note: LC = level change estimators; SC = slope change estimators; SMR = simultaneous multiple regression.

Influence of Series Length

Although both estimators are unbiased for all series lengths, it is important to have as much measurements as possible. That is so, because of the fact that the efficiency of the estimators of both SMR and SLC improves for longer series, as Tables 1 to 4 show. The increase of N is associated with especially important reduction of variance for the slope change estimator.

Table 4. Variance of the SMR Estimators for Data Series Generated Using an MA(1) Model With Normal Error and With Null Beta Parameters

	$n_A = n_B = 5$		$n_A = n_B = 10$		$n_A = n_B = 15$	
Autocorrelation	LC	SC	LC	SC	LC	SC
−.500	2.52	.12	0.65	.01	0.29	.00
−.400	1.72	.12	0.53	.01	0.28	.00
−.275	1.63	.14	0.60	.01	0.35	.00
−.099	1.64	.18	0.73	.02	0.47	.01
.000	1.70	.20	0.82	.02	0.54	.01
.099	1.79	.23	0.91	.03	0.62	.01
.275	2.05	.29	1.17	.04	0.81	.01
.400	2.45	.36	1.48	.05	1.05	.02
.500	3.98	.60	2.51	.09	1.82	.03

Note: LC = level change estimators; SC = slope change estimators; SMR = simultaneous multiple regression.

Influence of Error Model and Random Variable Distribution

The difference between AR(1) and MA(1) data is mainly in the variance of the estimators, which is somewhat greater in the case of the latter. Generating the random variable u_t following an exponential, a normal, or a uniform distribution does not seem to affect the performance of SLC or SMR in terms of bias or variance.

Discussion

The present investigation proposes a new procedure for estimating slope change and level change, in that order, after controlling for linear baseline phase trend. The estimates obtained are expressed in terms of the measurement units used to quantify the dependent variable, rather than in terms of standardized mean difference (d) or association (R^2). The performance of the procedure proposed is assessed in the context of two-phase data series representing a wide set of experimental conditions including autocorrelation, trend, and two different types of treatment effect. Considering the potential usefulness of SLC it was implemented in both free and commercial software requiring few inputs to produce the magnitude of change estimates.

The results suggest that the procedure proposed is practically unbiased both for first-order autoregressive and moving average processes and regardless of

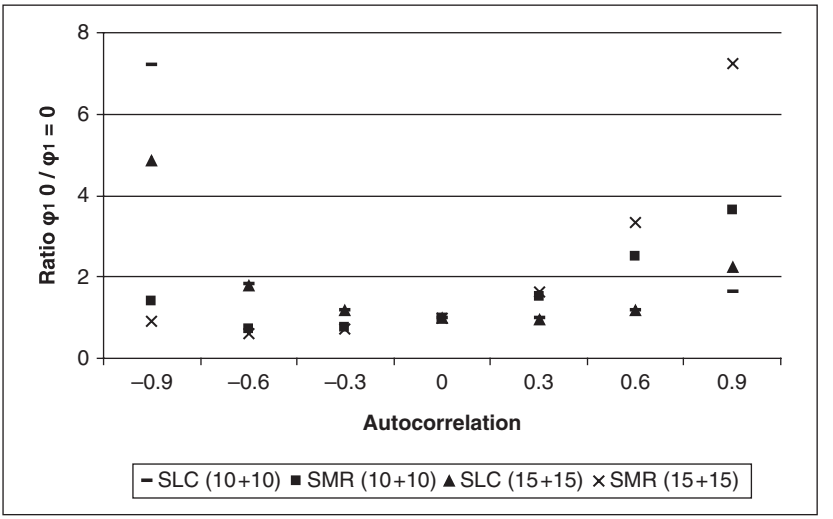


Figure 2. $\varphi_1 \neq 0$ / $\varphi_1 = 0$ ratios of the variability of the level change estimators of SMR and SLC for two series lengths and data generated using an AR(1) process, exponential error, and null beta parameters.
Note: SLC = slope and level change; SMR = simultaneous multiple regression.

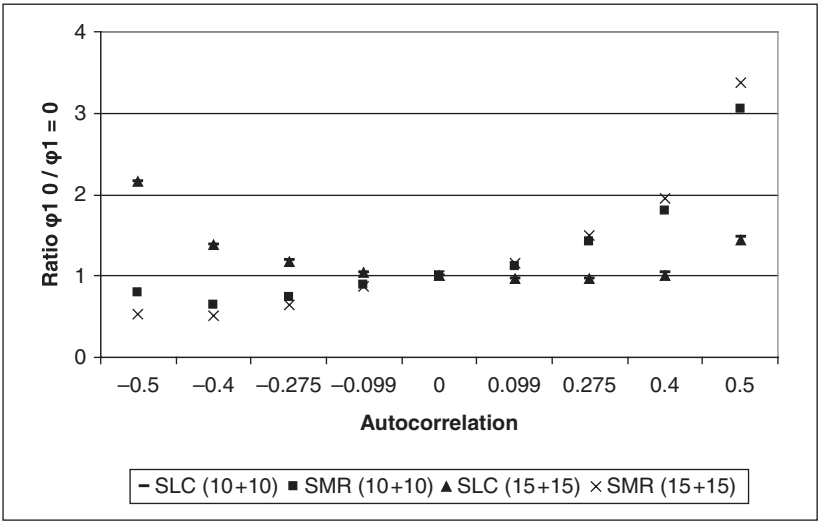


Figure 3. $\varphi_1 \neq 0$ / $\varphi_1 = 0$ ratios of the variability of the level change estimators of SMR and SLC for two series lengths and data generated using an MA(1) process, normal error, and null beta parameters.
Note: SLC = slope and level change; SMR = simultaneous multiple regression.

the distributional shape of the random variable; its initial data correction step controls effectively for linear trend. In the case of unbiased estimators, variance is an important indicator of their efficiency and the SLC estimators are generally more efficient for positively autocorrelated data and less efficient for $\varphi_1 \leq .0$. When comparing both procedures it has to be kept in mind that SMR implies a perfectly specified regression model corresponding exactly to the data parameters. It has to be highlighted that the least favorable conditions for SLC are defined by (a) high negative serial dependence and (b) immediate and permanent changes in the response rate with the introduction of the intervention. Regarding point "a," evidence suggests that high negative autocorrelation is not frequent in behavioral data (Matyas & Greenwood, 1997; Parker, 2006). On a substantive basis, it is also more logical to expect an individual or a group to show consistent behavior over time (i.e., positive autocorrelation). As far as point "b" is concerned, in psychological studies an abrupt and sustained (level) change in the behavior is less likely to occur than a progressive change representing a gradual improvement of the individual or group treated. Another point to be remarked is that the relevance of the estimators' variability is subjected to the magnitude of the intervention effect. For large effects, which seem to be typical for $N = 1$ studies (Campbell, 2004; Matyas & Greenwood, 1990; Parker et al., 2005), a variability of two behavioral units, such as the one observed for the level change estimator for most levels of serial dependence, may not be crucial for determining treatment effectiveness. Taking into account these considerations, it can be stated that SLC will perform well for the majority of applied studies, where positive autocorrelation, slope changes, and/or sizable level changes are likely to be present.

The results of estimators' greater variance for φ_1 values diverging from 0 can be explained considering the fact that the error term (and consequently the data series) variability increases when the degree of (negative and positive) autocorrelation is higher. These relationship can be explained through the following steps for the AR(1) processes. The data generation model for the error term is $\varepsilon_t = \varphi_1 \cdot \varepsilon_{t-1} + u_t$, which implies that $\mathrm{Var}(\varepsilon_t) = \varphi_1^2 \cdot \mathrm{Var}(\varepsilon_{t-1}) + \mathrm{Var}(u_t)$. Since the variability of the series at each measurement time is constant $\mathrm{Var}(\varepsilon_t) = \varphi_1^2 \cdot \mathrm{Var}(\varepsilon_t) + \mathrm{Var}(u_t)$ and thus $\mathrm{Var}(\varepsilon_t) - \varphi_1^2 \cdot \mathrm{Var}(\varepsilon_t) = \mathrm{Var}(u_t)$. Therefore, $\mathrm{Var}(\varepsilon_t) \cdot (1 - \varphi_1^2) = \mathrm{Var}(u_t)$ and $\mathrm{Var}(\varepsilon_t) = \mathrm{Var}(u_t) / (1 - \varphi_1^2)$. Thus, the greater the absolute value of the autocorrelation parameter, the greater the variance of the error term. An additional implication of this expression is that a greater variability of the u_t term would have resulted in more variable series. In both cases this increased data variability entails lower reliability in the estimation of the behavioral change.

Apart from the greater variance observed for $\varphi_1 \neq 0$, the results suggest that the level change estimates vary to a considerably greater degree than the slope change estimates in both SLC and SMR. The fact that this finding is common

to both procedures implies that it is not a consequence of the stepwise nature of SLC in which the level change is estimated after controlling for a potential slope change.

Although testing the performance of SMR was not the main aim of the present study, the results obtained suggested that it can be useful for estimating behavioral change. However, our preliminary results on the statistical significance of the regression coefficients suggest that conventional 5% alpha does not permit using *p* values as a reliable criterion for assessing the existence of level changes as small as one behavior. Additionally, autocorrelation was also shown to be problematic, as negative one makes more difficult the rejection of the null hypothesis and positive one makes SMR too liberal, doubling the probability of committing a Type I error. Until more evidence is available on statistical decision making with SMR, the assessment of treatment effectiveness can be done using solely the regression coefficients.

In summary, SLC can readily be applied to single-case data presenting trend and/or positive autocorrelation, since it is not affected by these data features providing unbiased estimates. The procedure can be complemented by the visual inspection of the graphed original and detrended data to enhance the assessment of treatment effectiveness. In fact, the technique only quantifies the amount of slope and level change in measurement units and, thus, allows the decision on the relevance of the changes to be made according to practitioner's substantive criteria, considering the behavior of interest, the particular client and setting, and so on.

A specific drawback of the procedure proposed is that it is conceived for correcting linear trend in data and, therefore, its performance might not be optimal when behavioral data present curvilinear trends. Further research is needed to explore whether the data correction present in SLC can attenuate the effect of nonlinear trends. Future efforts may also focus on estimating the sampling distribution of the slope and level change estimators, due to its utility for obtaining statistical significance and confidence intervals. The transformation of the SLC estimates into common effect size metrics is another issue to be tackled in the subsequent investigations.

Appendix A

Formulae for computing SLC

Equation (1): Differencing

$$\Delta A_t = A_{t+1} - A_t$$

where A_t represents the original phase A measurement at time t and A_t represents the differenced phase A data.

(continued)

Appendix A (continued)

Equation (2): Obtaining the phase A trend estimate $\hat{\beta}_A$ as the mean of the differenced phase A measurements

$$\hat{\beta}_A = \frac{\sum_{t=1}^{n_{A-1}} \Delta A_t}{(n_A - 1)}$$

where n_A is the number of observations in the baseline phase.

Equation (3): Obtaining the detrended phase A data ($\tilde{A}_t$)

$$(\tilde{A}_t) = A_t - \hat{\beta}_A \cdot t$$

Equation (4): Obtaining the detrended phase B data ($\tilde{B}_t$)

$$(\tilde{B}_t) = B_t - \hat{\beta}_A \cdot t$$

Equation (5): Differencing the detrended phase B data

$$\Delta\tilde{B}_t = \tilde{B}_{t+1} - \tilde{B}_t$$

Equation (6): obtaining the phase B slope estimate $\widehat{sc}$ as the mean of the differenced phase B measurements

$$\widehat{SC} = \frac{\sum_{t=n_A+1}^{n_A+n_B-1} \Delta\tilde{\tilde{B}}_t}{(n_B - 1)}$$

where n_B is the number of observations in the treatment phase.

Equation (7): Eliminating phase B slope from the detrended phase B data, maintaining potential level changes at time $t = n_A+1$.

$$\tilde{\tilde{B}}_t = \left\{\hat{B}_t - \widehat{SC} \cdot [t-1]\right\}$$

Equation (8): Estimating level change subtracting the detrended baseline data mean from the detrended and slope-change-controlled treatment data mean

$$\widehat{LC} = \bar{\tilde{\tilde{B}}} - \bar{\tilde{A}} = \frac{\sum_{t=n_A+1}^{n_A+n_B} \tilde{B}_t}{n_B - \left(\sum_{t=1}^{n_A+n_B} \tilde{A}_t \Big/ n_A\right)}$$

(continued)

Appendix A (continued)

An illustrative example

In order to illustrate the application of SLC, consider the fictitious data set consisting of the following measurements: 1, 2, 3, 4, and 5 for phase A and 7, 9, 11, 13, and 15 for phase B.

Data correction. First, the phase A data are differenced: $\Delta A_1 = A_2 - A_1 = 2-1 = 1$, $\Delta A_2 = A_3 - A_2 = 3-2 = 1$, $\Delta A_3 = A_4 - A_3 = 4-3 = 1$, $\Delta A_4 = A_5 - A_4 = 5-4 = 1$. Then, the average of the differenced phase A data is used to represent phase A trend $\hat{\beta}_A = (1+1+1+1)/(5-1) = 1$. The phase A trend is removed from the whole data series, obtaining the detrended baseline data as $\tilde{A}_1 = A1-1\cdot1= 1-1 = 0$, $\tilde{A}_2 = A_2-1\cdot2= 2-2 = 0$, $\tilde{A}_3 = A_3-1\cdot3= 3-3 = 0$, $\tilde{A}_4 = A_4-1\cdot4= 4-4 = 0$, and $\tilde{A}_5 = A_5-1\cdot5= 5-5 = 0$. The detrended treatment data are computed through $\tilde{B}_1 = B_1-1\cdot6 = 7-6 = 1$, $\tilde{B}_2 = B_2-1\cdot7= 9-7 = 2$, $\tilde{B}_3 = B_3-1\cdot8= 11-8 = 3$, $\tilde{B}_4 = B_4-1\cdot9 = 13-9 = 4$, and $\tilde{B}_5 = B_5-1\cdot10 = 15-10 = 5$.

Slope change estimation. First, the detrended phase B data are differenced: $\Delta\tilde{B}_1 = \tilde{B}_2 - \tilde{B}_1 = 2-1 = 1$, $\Delta\tilde{B}_2 = \tilde{B}_3 - \tilde{B}_2 = 3\text{-}2 = 1$, $\Delta\tilde{B}_3 = \tilde{B}_4 - \tilde{B}_3 = 4\text{-}3 = 1$, and $\Delta\tilde{B}_4 = \tilde{B}_5 - \tilde{B}_4 = 5\text{-}4 = 1$. Then, the average of the differenced and previously detrended phase B data is computed and used to represent the slope change estimate: $\widehat{SC} = (1+1+1+1)/(5-1) = 1$. Hence, in the treatment phase the experimental unit produces an average of one behavior more in each successive measurement time, considering that baseline trend has been removed.

Level change estimation. Phase B slope is removed from the detrended phase B data, without removing the intercept. The detrended slope-change-controlled treatment data are obtained through: $\tilde{\tilde{B}}_1 = \tilde{B}_1 - \hat{\beta}_{\tilde{B}} \cdot (1\text{-}1) = 1-1\cdot0 = 1$, $\tilde{\tilde{B}}_2 = \tilde{B}_2 - \hat{\beta}_{\tilde{B}} \cdot (2-1) = 2-1\cdot1 = 1$, $\tilde{\tilde{B}}_3 = \tilde{B}_3 - \hat{\beta}_{\tilde{B}} \cdot (3-1) = 3-1\cdot2 = 1$, $\tilde{\tilde{B}}_4 = \tilde{B}_4 - \hat{\beta}_{\tilde{B}} \cdot (4-1) = 4-1\cdot3 = 1$, and $\tilde{\tilde{B}}_5 = \tilde{B}_5 - \hat{\beta}_{\tilde{B}} \cdot (5-1) = 5-1\cdot4 = 1$. The level change estimate is computed subtracting the A-detrended phase A mean from the detrended slope-change-controlled phase B data mean: $\widehat{LC} = \bar{\tilde{B}} - \bar{\tilde{A}} = (1+1+1+1+1) / 5 - (0+0+0+0+0) / 5 = 1-0 = 1$. Therefore, after controlling for potential baseline trend and slope change between the two phases, the experimental unit produces an average of one behavior more during phase B than during phase A (level change).

The data were constructed to present no random fluctuations, only general trend in data—on each measurement time the experimental unit increases its

(continued)

Appendix A (continued)

response rate by one behavior. When the treatment phase starts, there is an immediate and permanent change in level of one behavior: the data point at time 6 is 7 instead of 6, as the mere continuation of the phase A trend would imply. Additionally, during phase B the response rate increases by 2 behaviors at a time and, therefore, there is a slope change of one behavior. In summary, the level change and slope change parameters are known and are both equal to 1 behavior. As it can be seen, both slope change and level change were precisely estimated.

Appendix B

The R code for SLC requires copy-pasting the function it in the R console. Afterwards, data can be input reading a file "filename.det" containing the measurements separated by spaces in *info <- array(scan("filename.dat"))* or writing the values directly in *info <- array(c(value1,value2,value3)).* Then, the baseline phase length is specified in *n_a <- length.* Finally, the expression *results <- slcestimates(info,n_a)* needs to be written in order to obtain the output. The same specifications need to be done after installing and loading the *SLC_0.1.tar.gz* package based on commands.

```
# R function for estimating slope and level change
slcestimates <- function(info,n_a)  {
slength <- length(info)
n_b <- slength-n_a
phaseA <- info[1:n_a]
phaseB <- info[(n_a+1):slength]
# Estimate phase A trend
phaseAdiff <- c(1:(n_a-1))
for (iter in 1:(n_a-1))
phaseAdiff[iter] <- phaseA[iter+1] - phaseA[iter]
trendA <- mean(phaseAdiff)
# Remove phase A trend from the whole data series
phaseAdet <- c(1:n_a)
for (timeA in 1:n_a)
phaseAdet[timeA] <- phaseA[timeA] - trendA * timeA
phaseBdet <- c(1:n_b)
for (timeB in 1:n_b)
phaseBdet[timeB] <- phaseB[timeB] - trendA * (timeB+n_a)
# Compute the slope change estimate
phaseBdiff <- c(1:(n_b-1))
```

(continued)

Appendix B (continued)

```
for (iter in 1:(n_b-1))
phaseBdiff[iter] <- phaseBdet[iter+1] - phaseBdet[iter]
trendB <- mean(phaseBdiff)
print ("Slope change estimate = "); print(trendB)
# Compute the level change estimate
phaseBddet <- c(1:n_b)
for (timeB in 1:n_b)
phaseBddet[timeB] <- phaseBdet[timeB] - trendB * (timeB-1)
level <- mean(phaseBddet) - mean(phaseAdet)
print ("Level change estimate = "); print(level)
# Represent graphically
time <- c(1:slength)
par(mfrow=c(2,1))
plot(time,info, xlim=c(1,slength), ylim=c((min(info)-1),(max(info)+1)),
   xlab="Measurement time", ylab="Variable of interest", font.lab=2)
abline(v=(n_a+0.5))
lines(time[1:n_a],info[1:n_a])
lines(time[(n_a+1):slength],info[(n_a+1):slength])
axis(side=1, at=seq(0,slength,1),labels=TRUE, font=2)
axis(side=2, at=seq((min(info)-1),(max(info)+1),2),labels=TRUE, font=2)
points(time, info, pch=24, bg="black")
title (main="Original data")
transf <- c(phaseAdet,phaseBdet)
plot(time,transf, xlim=c(1,slength), ylim=c((min(transf)-1),(max(transf)+1)),
   xlab="Measurement time", ylab="Variable of interest", font.lab=2)
abline(v=(n_a+0.5))
lines(time[1:n_a],transf[1:n_a])
lines(time[(n_a+1):slength],transf[(n_a+1):slength])
axis(side=1, at=seq(0,slength,1),labels=TRUE, font=2)
axis(side=2, at=seq((min(transf)-1),(max(transf)+1),2),labels=TRUE, font=2)
points(time, transf, pch=24, bg="black")
title (main="Detrended data")
list(trendB,level) }
# Input data
info <- array(scan("info.dat"))
n_a <- 3
# Obtain estimates
results <- slcestimates(info,n_a)
```

The plug-in needs both R and R Commander to be installed. The use of the plug-in also requires installing and loading *SLC_0.1.zip* package. The *RcmdrPlugin.SLC_0.1.tar.gz* plug-in needs also to be installed and loaded.

(continued)

Appendix B (continued)

Afterwards, in the R console the expression *library(Rcmdr)* needs to be written in order to open R Commander. The plug-in is loaded in R Commander by choosing *Tools* ► *Load R Cmdr plug-in(s)* and the SLC tab appears in the main menu. Using the menus, the input data file is selected and the length of the baseline phase (obligatorily longer than 1 measurement) is specified prior to executing.

Appendix C

The SAS/IML code for SLC requires copy-pasting the module in the SAS console. Afterwards, data should be input in the statement *measurements = {value1 value2 value3};* between the curly brackets, separating the values by spaces. Then the baseline phase length is specified in *n_a = length*; the slope change and level change estimates are obtained pasting the last three lines and executing the whole code. Furthermore, the detrended data is graphed.

```
proc iml;
*Module SLC;
start slc(series, n_a);
*Obtain phase B length;
n_b = ncol(series)-n_a;
*Difference phase A;
adiff = j(n_a-1,1,1);
do i = 1 to (n_a-1);
adiff[i]=series[i+1]-series[i];
end;
*Estimate trend;
aslope=sum(adiff)/(n_a-1);
*Remove trend from phase A;
adet = j(n_a,1,1);
do i = 1 to n_a;
adet[i]=series[i]-aslope#i;
end;
*Remove trend from phase B;
bdet = j(n_b,1,1);
do i = 1 to n_b;
bdet[i]=series[i+n_a]-aslope*(i+n_a);
end;
*Difference phase B;
```

(continued)

Appendix C (continued)

```
bdiff = j(n_b-1,1,1);
do i = 1 to (n_b-1);
bdiff[i]=bdet[i+1]-bdet[i];
end;
*Graph the detrended data;
time_a=j(n_a,1,1);
do i = 1 to n_a;
time_a[i] = i;
end;
time_b=j(n_b,1,1);
do i = 1 to n_b;
time_b[i] = i+n_a;
end;
dims=j(4,1,1);
dims[1] = 0;
dims[2] = min(min(adet),min(bdet))-1;
dims[3] = n_a + n_b + 1;
dims[4] = max(max(adet),max(bdet))+1;
start_pt = j(2,1,1);
start_pt[1] = 0;
start_pt[2] = min(min(adet),min(bdet))-1;
num_y = max(max(adet),max(bdet)) - min(min(adet),min(bdet)) + 2;
num_x = nrow(time_a) + nrow(time_b) + 1;
call gstart;
call gopen;
call gwindow(dims);
call gport({15 15, 85 85});
call gyaxis(start_pt,num_y,num_y, ,2.,1.5);
call gxaxis(start_pt,num_x,num_x, ,2.,1.5);
call gpoint(time_a,adet,"circle","red");
call gdraw(time_a,adet,1,"red");
call gpoint(time_b,bdet,"square","green");
call gdraw(time_b,bdet,1,"green");
call gshow;
*Estimate slope change;
bslope=sum(bdiff)/(n_b-1);
*Control slope change;
bclear = j(n_b,1,1);
do i = 1 to n_b;
bclear[i]=bdet[i]-bslope*(i-1);
end;
*Estimate level change;
alevel=sum(adet)/n_a;
```

(continued)

Appendix C (continued)

```
blevel = sum(bclear)/n_b;
levelchange=blevel-alevel;
*Save estimates;
estimates = j(2,1,1);
estimates[1] = bslope;
estimates[2] = levelchange;
return(estimates);
finish slc;
*Obtain estimates;
measurements = {1 2 5 6 9};
n_a = 3;
results = slc(measurements,n_a);
print results;
quit;
```

Acknowledgment

The authors would like to thank Dr. David Leiva for improving the SLC software.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

Financial Disclosure/Funding

This research was supported by the Comissionat per a Universitats i Recerca del Departament d'Innovació, Universitats i Empresa of the Generalitat de Catalunya, and the European Social Fund.

References

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy, 31*, 621-631.

Beretvas, S. N., & Chung, H. (2008a). A review of meta-analyses of single-subject experimental designs: Methodological issues and practice. *Evidence-Based Communication Assessment and Intervention, 2*, 129-141.

Beretvas, S. N., & Chung, H. (2008b). An evaluation of modified R^2-change effect size indices for single-subject experimental designs. *Evidence-Based Communication Assessment and Intervention, 2*, 120-128.

Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification, 30*, 531-563.

Bulté, I., & Onghena, P. (2008). An R package for single-case randomization tests. *Behavior Research Methods, 40*, 467-478.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research designs and analysis*: *New directions for psychology and education* (pp. 187-212). Hillsdale, NJ: Lawrence Erlbaum.

Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification, 28*, 234-246.

Center, B. A., Skiba, R. J., & Casey, A. (1985-1986). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education, 19*, 387-400.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology, 61*, 966-974.

Gorsuch, R. L. (1983). Three methods for analyzing limited time-series (N of 1) data. *Behavioral Assessment, 5*, 141-154.

Gottman, J. M. (1981). Time-series analysis: A comprehensive introduction for social scientists. New York: Cambridge University Press.

Greenwood, K. M., & Matyas, T. A. (1990). Problems with application of interrupted time series analysis for brief single-subject data. *Behavioral Assessment, 12*, 355-370.

Harrop, J. W., & Velicer, W. F. (1985). A comparison of alternative approaches to the analysis of interrupted time-series. *Multivariate Behavioral Research, 20*, 27-44.

Huitema, B. E. (2004). Analysis of interrupted time-series experiments using ITSE: A critique. *Understanding Statistics, 3*, 27-46.

Huitema, B. E., & McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement, 60*, 38-58.

Huitema, B. E., & McKean, J. W. (2007). An improved portmanteau test for autocorrelated errors in interrupted time-series regression models. *Behavior Research Methods, 39*, 343-349.

Huitema, B. E., McKean, J. W., & Laraway, S. (2007). Time series intervention analysis using ITSACORR: Fatal flaws. *Journal of Modern Applied Statistical Methods, 6*, 367-379.

Huitema, B. E., McKean, J. W., & McKnight, S. (1999). Autocorrelation effects on least-squares intervention analysis of short time series. *Educational and Psychological Measurement, 59*, 767-786.

Jenson, W. R., Clark, E., Kircher, J. C., & Kristjansson, S. D. (2007). Statistical reform: Evidence-based practice, meta-analyses, and single subject designs. *Psychology in the Schools, 44*, 483-493.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746-759.

Kromrey, J. D., & Foster-Johnson, L. (1996). Determining the efficacy of intervention: The use of effect sizes for data analysis in single-subject research. *Journal of Experimental Education, 65*, 73-93.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject research: Percentage of data points exceeding the median. *Behavior Modification, 30*, 598-617.

Manolov, R., & Solanas, A. (2008). Comparing N = 1 effect size indices in presence of autocorrelation. *Behavior Modification, 32*, 860-875.

Manolov, R., & Solanas, A. (2009). Percentage of nonoverlapping corrected data. *Behavior Research Methods, 41*, 1262-1271.

Manolov, R., Solanas, A., & Leiva, D. (in press). Comparing "visual" effect size indices for single-case designs. *Methodology—European Journal of Research Methods for the Behavioral and Social Sciences*.

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis for single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23*, 341-351.

Matyas, T. A., & Greenwood, K. M. (1997). Serial dependency in single-case time series. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 215-243). Mahwah, NJ: Lawrence Erlbaum.

McCleary, R., & Hay, R. A., Jr. (1980). *Applied time series analysis for the social sciences.* Beverly Hills: Sage.

McKnight, S. D., McKean, J. W., & Huitema, B. E. (2000). A double bootstrap method to analyze linear models with autoregressive error terms. *Psychological Methods, 3*, 87-101.

Olive, M. L., & Smith, B. W. (2005). Effect size calculations and single subject designs. *Educational Psychology, 25*, 313-324.

Parker, R. I. (2006). Increased reliability for single-case research results: Is bootstrap the answer? *Behavior Therapy, 37*, 326-338.

Parker, R. I., & Brossart, D. F. (2003). Evaluating single-case research data: A comparison of seven statistical methods. *Behavior Therapy, 34*, 189-211.

Parker, R. I., & Brossart, D. F. (2006). Phase contrasts for multiphase single case intervention designs. *School Psychology Quarterly, 21*, 46-61.

Parker, R. I., Brossart, D. F., Vannest, K. J., Long, J. R., Garcia De-Alba, R., & Baugh, F. G. (2005). Effect sizes in single case research: How large is large? *School Psychology Review, 34*, 116-132.

Parker, R. I., & Hagan-Burke, S. (2007). Useful effect size interpretations for single case research. *Behavior Therapy, 38*, 95-105.

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data: An alternative to PND. *Journal of Special Education, 40*, 194-204.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.

Schlosser, R. W., & Sigafoos, J. (2008). Meta-analysis of single-subject designs: Why now? *Evidence-Based Communication Assessment and Intervention, 2*, 117-119.

Schneider, N., Goldstein, H., & Parker, R. (2008). Social skills interventions for children with autism: A meta-analytic application of percentage of all non-overlapping data (PAND). *Evidence-Based Communication Assessment and Intervention, 2*, 152-162.

Scruggs, T. E., & Mastropieri, M. A. (1998). Summarizing single-subject research: Issues and applications. *Behavior Modification, 22*, 221-242.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education, 8*, 24-33.

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention, 2*, 188-196.

Van den Noortgate, W., & Onghena, P. (2003). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, and Computers, 35*, 1-10.

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention, 2*, 142-151.

Weisberg, S. (1980). *Applied linear regression.* New York: John Wiley.

White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta-analysis in individual subject research. *Behavioral Assessment, 11*, 281-296.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 694-704.

Bios

Antonio Solanas, PhD, is full professor at the Faculty of Psychology at the University of Barcelona, Spain. His main research interests include single-case designs analysis, social reciprocity measurement, and multivariate data analysis methods.

Rumen Manolov is an assistant lecturer and PhD student at the Faculty of Psychology at the University of Barcelona, Spain. His investigation is focused on single-case designs data analysis.

Patrick Onghena, PhD, is full professor at the Faculty of Psychology and Educational Sciences at the Katholieke Universiteit Leuven, Belgium. His research centres on meta-analysis, single-case designs, multilevel models, and nonparametric inference.

4. DISCUSSIÓ

La recerca inclosa en la present tesi doctoral es va centrar en l'exploració de diferents índexs per quantificar la magnitud de l'efecte en dissenys de cas únic amb l'objectiu d'identificar tècniques que podrien ser útils per als investigadors aplicats a l'hora de prendre decisions sobre l'eficàcia de les intervencions. Les principals conclusions dels diferents estudis es ressalten a continuació, així com també es presenta la imatge global que es deriva de la tesi.

4.1. Conclusions específiques de cada estudi

En el primer estudi es van comparar tècniques amb fonaments tan diferents com la regressió, la diferència de mitjanes estandarditzades o el solapament entre les dades pertanyents a fases contigües utilitzat com a criteri en la inspecció visual. Les troballes concorden amb les obtingudes per altres autors senyalant el rendiment insatisfactori de les tècniques basades en la regressió quan les dades no són independents (Beretvas i Chung, 2008b). Es va comprovar que aquestes tècniques diferencien menys entre dades amb i sense efecte de la intervenció. En concret, el *Trend analysis* de Gorsuch (1983), va mostrar una mínima variabilitat en els valors d'R^2, tots ells excessivament baixos. Contràriament, els models d'Allison i Gorman (1993) i White et al. (1989) varen produir unes estimacions més altes en absència d'efecte i, per tant, poden comportar que un tractament que no incideix sobre la conducta s'etiqueti com a "efectiu" reforçant la seva aplicació. El PND i les diferències de mitjanes estandarditzades es mostren més conservadors i semblen ser un filtre més apropiat per detectar tractaments potents. Addicionalment, el PND va ser l'índex menys afectat per l'autocorrelació. En canvi, la tendència general sí que introdueix més distorsió en les estimacions produïdes pel PND. En general, tot i les seves limitacions, l'índex de càlcul més simple, PND, es va mostrar com a més adient, degut a que diferenciava millor entre presència i absència d'efecte del tractament.

El segon estudi va comparar aquest índex amb dues propostes recents que pretenen millorar el seu funcionament. Cap de les tres tècniques es va mostrar inequívocament com a superior a la resta. Es va observar que el PEM es veu menys afectat per la dependència serial quan el tractament no és efectiu i per tant aquesta no contribueix a augmentar la probabilitat de falses alarmes. Les estimacions calculades mitjançant el PAND, en canvi, es veuen menys distorsionades per la tendència general en les dades. No obstant, tant el PEM com el PAND produeixen sistemàticament unes estimacions més altes que el PND. Per la seva banda, el PND va destacar per ser l'índex que distingia en major grau entre els tractaments efectius i els inefectius.

El tercer estudi va sorgir del fet que per tots tres índexs "visuals" la tendència és un aspecte problemàtic i va focalitzar el PND, atès que aquesta és la tècnica més freqüentment emprada per mesurar la magnitud de l'efecte en contexts d'*N=1* i per dur a terme integracions quantitatives d'estudis individuals. L'efecte de la tendència es va intentar controlar introduint un pas previ de diferenciació de les dades. Degut a la proximitat entre la tendència i l'autocorrelació positiva es conjecturava que la correcció podria atenuar la influència de la dependència serial. Es va mostrar, per a una àmplia gama de condicions experimentals, que la inclusió de la correcció de les dades abans d'aplicar el PND fa que l'impacte de la tendència general sigui pràcticament zero. Es va observar també una reducció important de la repercussió de l'autocorrelació sobre les estimacions de la magnitud de l'efecte. Per tant, el PNCD va tenir un rendiment millor que el PND segons el criteri de no afectació per la relació seqüencial entre les dades. Quant a la discriminació entre presència i absència d'efecte del tractament, el PND mostrava una diferenciació relativa (en termes de raons) més elevada només en absència de tendència.

El quart estudi va néixer de la necessitat de cercar una tècnica que estimés de manera precisa la quantitat de canvi entre les condicions de línia base i de tractament. El procediment SLC està estretament relacionat amb el PNCD, perquè utilitza la mateixa manera de controlar la tendència, que a més es fa servir per estimar el canvi de pendent, incorporant també l'estimació del canvi de nivell. Les simulacions varen demostrar la manca de biaix en la quantificació dels canvis, el mateix resultat que es va observar per al procediment d'anàlisi que reflectia perfectament el model de generació de les dades. La comparació en termes d'error estàndard indica que el procediment proposat només és inferior quan es tracta de nivells elevats d'autocorrelació negativa; característica poc comuna en dades comportamentals obtingudes del mateix individu o grup d'individus. En canvi, quan les dades presenten dependència serial positiva, els estimadors de l'SLC es mostren més eficients (en termes relatius i en molts casos en termes absoluts) que els estimadors del model de regressió múltiple.

Respecte a l'objectiu d'establir uns criteris de comparació adients entre tècniques que quantifiquen la magnitud de l'efecte en termes diferents (e.g., percentatges, R^2), des de la segona investigació s'utilitzen raons adimensionals. Per exemple, la raó entre la grandària de l'efecte en presència de canvi de nivell i la grandària de l'efecte per a dades sense canvi, controlant la resta de paràmetres, és un indicador relatiu de la capacitat de detecció de l'índex en qüestió. Per tant, una raó més gran implicaria un augment més considerable en l'índex, és a dir, una sensibilitat més gran. Les raons també són útils per comparar el grau de distorsió degut a l'autocorrelació. La raó entre el valor de l'índex per a una dependència serial positiva moderada i el valor en sèries independents, mantenint constant les altres característiques de les condicions experimentals, indicaria el grau de sobreestimació (si fos més gran que 1)

o de subestimació (si fos més petit que 1) de la magnitud de l'efecte relacionada amb l'autocorrelació.

Quant al software desenvolupat, cadascuna de les investigacions comporta la creació de programes per generar dades, aplicar les tècniques estadístiques que es posen a prova, així com calcular les mitjanes i les variàncies dels índexs per a cada condició experimental. La generació d'errors exponencials i uniformes es va fer assegurant de forma analítica que les condicions fossin comparables amb la normal unitària, partint de les expressions per als moments de primer i segon ordre d'aquelles. Per a la distribució exponencial negativa, es van generar dades utilitzant una distribució amb paràmetre de localització $\theta=0$ i paràmetre d'escala $\sigma=1$ i, posteriorment, es va restar 1 a tota la sèrie per centrar la distribució en zero. En el cas de la distribució uniforme, aconseguir una mitjana de zero i una desviació estàndard d'u va requerir establir el mínim $a=-1.732050808$ i el màxim $b=1.732050808$. Aquest mètode de generació de dades es va haver de crear durant la recerca, atès que no s'havien trobat indicacions explícites en la literatura científica. A més dels programes d'ús intern, es van desenvolupar codis en R (una plataforma gratuïta) que permeten aplicar els nous procediments d'una manera ràpida i senzilla. Aquests codis són un complement lògic a les propostes realitzades en el marc de la tesi i potencien la seva aplicabilitat, ja que estan disponibles com a annexes dels articles, sense necessitat ni tan sols de contactar amb els seus autors.

4.2. Conclusions generals

Considerant els resultats en la seva globalitat, es pot afirmar que els procediments basats en l'anàlisi de la regressió inclosos en la present tesi semblen menys apropiats per quantificar la magnitud quan es disposa de relativament poques dades del mateix participant o grup de participants. La principal raó és que aquests procediments no ajuden a diferenciar les dades amb canvi conductual de les dades sense, atès que mostren grandàries de l'efecte molt petites o molt grans. Aquest segon cas és més greu, ja que l'acumulació de falses alarmes en diversos estudis implicaria que les persones que vulguin assolir un canvi rebrien una intervenció amb una base científica inapropiada. De fet, diverses revisions bibliogràfiques, inclosa la feta en el marc de la tesi doctoral, indiquen que les tècniques de regressió s'apliquen en menor mesura que els índexs "visuals" per quantificar el canvi en estudis individuals o en meta-anàlisis de dades recollides mitjançant dissenys de cas únic. Degut al seu origen aquestes últimes tècniques requereixen uns càlculs més simples que es poden dur a terme fins i tot en absència de mitjans informàtics.

A l'hora de triar entre les diferents tècniques relacionades amb el grau de solapament entre les dades, el PND no es mostra inferior al PEM i al PAND que pretenen ser-ne una millora, sinó presenta una distància més gran entre les

estimacions de la magnitud de l'efecte quan aquest existeix i quan no. Addicionalment, els percentatges més baixos que dóna suggereixen que el PND és més restrictiu que el PEM i el PAND i, per tant, seria més apropiat quan es vulguin detectar només tractaments potents. A més, un dels punts forts del PAND – la possibilitat d'obtenir *Phi* o Phi^2 de Pearson – complica el càlcul fins a requerir l'ús de més d'un tipus de software per dur a terme la computació (Schneider, Goldstein i Parker, 2008).

No obstant els avantatges del PND, s'ha demostrat que aquest pot tenir un rendiment optimitzat si s'introdueix un pas inicial de correcció de les dades. D'aquesta manera, aplicant el PNCD, s'elimina o s'atenua l'amenaça que representen factors com la tendència general o l'autocorrelació, sense comprometre la capacitat per detectar els efectes de la intervenció. La quantificació de la magnitud del canvi es pot complementar amb la informació que dóna l'SLC – un procediment que, en el cas de mesurar freqüències d'un comportament d'interès, indica quantes conductes en promig hi ha de diferència entre dues fases i quantes conductes de més o de menys es produeixen per cada moment d'observació en comparació amb la fase anterior. Així doncs, s'ha procurat potenciar la interpretació de l'SLC en termes intel·ligibles a nivell aplicat, ja que les estimacions del canvi de nivell i de pendent s'expressen en la unitat de mesura de les dades enregistrades.

De la mateixa manera que el PND és fàcilment combinable amb la inspecció visual, també ho és el PNCD que, a més, es pot acompanyar amb el gràfic de les dades amb tendència eliminada. Finalment, l'SLC es pot aplicar conjuntament amb el PNCD per potenciar encara més la presa de decisions sobre l'eficàcia d'un tractament. No s'ha d'oblidar que l'ús de qualsevulla de les tècniques no exclou ni tampoc no reemplaça els criteris socials, educatius, clínics, etc. dels professionals sobre quan un canvi comportamental és important i quan no ho és pas, sinó que les quantificacions han de servir com a eina complementària.

4.3. Limitacions de la present recerca

Els estudis inclosos en la present tesi doctoral comparen només una part de la totalitat de procediments creats per quantificar l'efectivitat del tractament en dissenys de cas únic. La recerca focalitza les tècniques que apareixen amb més freqüència en la literatura científica en l'àmbit de la psicologia i dels estudis d'*N=1*. A més a més, la investigació s'ha centrat progressivament en cert tipus de tècniques degut a la seva utilitat potencial. Així doncs, no és possible afirmar amb total certesa quin és el procediment òptim no havent estudiat d'altres propostes realitzades en relació amb el tema, però no disponibles en les bases de dades més comunes. Addicionalment, s'ha d'esmentar que en el transcurs del desenvolupament de la tesi han sorgit propostes noves (e.g., Parker i Hagan-Burke, 2009) que no s'han pogut avaluar, atès que varen coincidir en el temps amb l'elaboració dels articles presentats en la tesi.

Quant al mètode de simulació, els estudis inclosos en aquesta tesi doctoral incorporen també variables aleatòries amb distribucions no normals, seguint les troballes de Bradley (1977) i Micceri (1989), i utilitzen els dos models de generació de dades que potencialment representen millor les mesures conductuals reals (Harrop i Velicer, 1985). No obstant, el mètode de generació de dades té les seves limitacions. D'una banda, només s'ha simulat tendència lineal i seria important comprovar el rendiment de les diferents tècniques (les noves propostes incloses) en el context de dades que presenten tendència curvilínia. D'altra banda, els efectes de tractament simulats són immediats i permanents. Per tant, és necessari modelar altres tipus d'efecte, com per exemple canvis retardats o transitoris en la conducta d'interès.

4.4. Línies de recerca futura

Una part dels estudis futurs es pot centrar en comparar procediments recentment creats amb els que són més àmpliament acceptats en l'àmbit de les ciències del comportament. Es tractaria de comprovar quines tècniques presenten un rendiment més apropiat en una varietat creixent de condicions experimentals. Per tant, s'utilitzarien mètodes Monte Carlo per generar dades que representin les observacions reals i es compararien els paràmetres de la simulació (i.e., la veritat coneguda) amb les estimacions obtingudes pels diferents índexs. Aquest tipus de simulació hauria de mantenir la continuïtat amb les recerques que es duen a terme sobre el tema, però també superar els seus punts febles. S'haurien d'ampliar les condicions experimentals simulades inclouen, per exemple, dissenys de més de dues fases (e.g., ABAB) i una major varietat d'efectes del tractament. A banda dels estudis de simulació, la discussió dels nous procediments també es podria fer des d'una perspectiva teòrica. Les opinions d'altres autors experts en la temàtica d'*N=1*, tant de la vessant metodològica com de l'aplicada, serien molt valuoses per millorar les tècniques proposades dins del marc de la tesi doctoral.

Altre tipus d'estudis es pot dedicar a introduir millores en tècniques ja existents. Objecte d'interès serien tant els procediments proposats dintre del grup de recerca com els creats per altres autors. La finalitat d'aquestes recerques seria controlar diferents factors de confusió que podrien distorsionar les estimacions obtingudes mitjançant els diferents índexs.

Finalment, no s'ha d'excloure la possibilitat de proposar nous procediments quan les modificacions de tècniques existents comporten una complexitat excessiva o quan la idea que hi serveix de fonament és diferent o innovadora. Aquestes recerques respondrien als avenços teòrics i a la necessitat d'obtenir informació inaccessible amb les tècniques existents. En els darrers dos casos seria important introduir les modificacions o propostes realitzades en software d'aplicació fàcil i gratuïta, per tal d'incrementar l'atractivitat i utilitat dels nous desenvolupaments en la quantificació de la magnitud de l'efecte de la intervenció.

5. REFERÈNCIES☯

Algina, J. i Swaminathan, H. (1979). Alternatives to Simonton's analysis of the interrupted and multiple-group time-series designs. *Psychological Bulletin, 86*, 919-926.

Allison, D. B. i Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy, 31*, 621-631.

Arnau, J. i Bono, R. (1998). Short time series analysis: C statistic vs Edgington model. *Quality & Quantity, 32*, 63-75.

Arnau, J. i Bono, R. (2003). Autocorrelation problems in short time series. *Psychological Reports, 92*, 355-364.

Arnau, J. i Bono, R. (2004). Evaluating effects in short time series: Alternative models of analysis. *Perceptual and Motor Skills, 98*, 419-432.

Ato, M. i Vallejo, G. (2007). *Diseños experimentales en psicología*. Madrid: Pirámide.

Barlow, D. H. i Hersen, M. (1973). Single-case experimental designs: Uses in applied clinical research. *Archives of General Psychology, 29*, 319-325.

Barlow, D. H. i Hersen, M. (2008). *Single case experimental designs. Strategies for studying behavior change* (3rd ed.). Boston: Allyn & Bacon.

Beretvas, S. N. i Chung, H. (2008a). A review of meta-analyses of single-subject experimental designs: Methodological issues and practice. *Evidence-Based Communication Assessment and Intervention, 2*, 129-141.

Beretvas, S. N. i Chung, H. (2008b). An evaluation of modified R^2-change effect size indices for single-subject experimental designs. *Evidence-Based Communication Assessment and Intervention, 2*, 120-128.

Blampied, N. M. (2000). Single-case research designs: A neglected alternative. *American Psychologist, 55*, 960.

Blumberg, C. J. (1984). Comments on "A simplified time-series analysis for evaluating treatment interventions". *Journal of Applied Behavior Analysis, 17*, 539-542.

Bradley, J. V. (1977). A common situation conducive to bizarre distribution shapes. *American Statistician, 31*, 147-150.

Brossart, D. F., Parker, R. I., Olson, E. A. i Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification, 30*, 531-563.

Busk, P. L. i Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment, 10*, 229-242.

Busk, P. L. i Marascuilo, L. A. (1992). Statistical analysis in single-case research: Issues, procedures, and recommendations, with applications to multiple behaviors. En T. R. Kratochwill i J. R. Levin (Eds.), *Single-case research designs and analysis: New directions for psychology and education* (pp. 159-185). Hillsdale, NJ: Lawrence Erlbaum.

☯ Aquest apartat no inclou les referències dels estudis, atès que cadascú d'ells conté el seu propi apartat anomenat "References".

Busk, P. L. i Serlin, R. C. (1992). Meta-analysis for single-case research. En T. R. Kratochwill i J. R. Levin (Eds.), *Single-case research designs and analysis: New directions for psychology and education* (pp. 187-212). Hillsdale, NJ: Lawrence Erlbaum.

Callahan, C. D. i Barisa, M. T. (2005). Statistical process control and rehabilitation outcome: The single-subject design reconsidered. *Rehabilitation Psychology, 50*, 24-33.

Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification, 28*, 234-246.

Center, B. A., Skiba, R. J. i Casey, A. (1985-1986). A methodology for the quantitative synthesis of intra-subject design research. *The Journal of Special Education, 19*, 387-400.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Crosbie, J. (1987). The inability of the binomial test to control Type I error with single-subject data. *Behavioral Assessment, 9*, 141-150.

Crosbie, J. (1989). The inappropriateness of the C statistic for assessing stability or treatment effects with single-subject data. *Behavioral Assessment, 11*, 315-325.

Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology, 61*, 966-974.

DeCarlo, L. T. i Tryon, W. W. (1993). Estimating and testing correlation with small samples: A comparison of the C-statistic to modified estimator. *Behaviour Research and Therapy, 31*, 781-788.

DeProspero, A. i Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis, 12*, 573-579.

Edgington, E. S. (1980a). Random assignment and statistical tests for one-subject experiments. *Behavioral Assessment, 2*, 19-28.

Edgington, E. S. (1980b). Validity of randomization tests for one-subject experiments. *Journal of Educational Statistics, 5*, 235-251.

Edgington, E. S. i Onghena, P. (2007). *Randomization tests* (4a. ed.). London: Chapman & Hall/CRC.

Evans, J., Emslie, H. i Wilson, B. A. (1998). External cueing systems in the rehabilitation of executive impairments of action. *Journal of the International Neuropsychological Society, 4*, 399-408.

Faith, M. S., Allison, D. B. i Gorman, D. B. (1997). Meta-analysis of single-case research. En R. D. Franklin, D. B. Allison i B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 245-277). Mahwah, NJ: Lawrence Erlbaum Associates.

Ferron, J. (2002). Reconsidering the use of the general linear model with single-case data. *Behavior Research Methods, Instruments & Computers, 34*, 324-331.

Ferron, J., Foster-Johnson, L. i Kromrey, J. D. (2003). The functioning of single-case randomization tests with and without random assignment. *The Journal of Experimental Education, 71*, 267-288.

Ferron, J. i Jones, P. K. (2006). Tests for the visual analysis of response-guided multiple-baseline data. *The Journal of Experimental Education, 75*, 66-81.

Ferron, J. i Onghena, P. (1996). The power of randomization tests for single-case phase designs. *The Journal of Experimental Education, 64*, 231-239.

Ferron, J. i Sentovich, C. (2002). Statistical power of randomization tests used with multiple-baseline designs. *The Journal of Experimental Education, 70*, 165-178.

Ferron, J. i Ware, W. (1995). Analyzing single-case data: The power of randomization tests. *The Journal of Experimental Education, 63*, 167-178.

Fisch, G. S. (2001). Evaluating data from behavioral analysis: Visual inspection or statistical models? *Behavioural Processes, 54*, 137-154.

Friedman, H. (1982). Simplified determinations of statistical power, magnitude of effect and research sample sizes. *Educational and Psychological Measurement, 42*, 521-526.

Gentile, J. R., Roden, A. H. i Klein, P. D. (1972). An analysis of variance model for the intrasubject replication test. *Journal of Applied Behavior Analysis, 5*, 193-198.

Glass, G. V., Wilson, V. L. i Gottman, J. M. (1975). Design and analysis of time-series experiments. Boulder, CO: Colorado Associated Press.

Good, P. (1994). *Permutation tests. A practical guide to resampling methods for testing hypotheses*. New York: Springer-Verlag.

Gorman, B. S. i Allison, D. B. (1997). Statistical alternatives for single-case designs. En R. D. Franklin, D. B. Allison i B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 159-214). Mahwah, NJ: Lawrence Erlbaum.

Gorsuch, R. L. (1983). Three methods for analyzing limited time-series (N of 1) data. *Behavioral Assessment, 5*, 141-154.

Gottman, J. M. (1973). N-of-one and N-of-two research in psychotherapy. *Psychological Bulletin, 80*, 93-105.

Gottman, J. M. (1981). Time-series analysis: A comprehensive introduction for social scientists. New York: Cambridge University Press.

Greenwald, A. G. (1976). Within-subject designs: To use or not to use? *Psychological Bulletin, 8*, 314-320.

Greenwood, K. M. i Matyas, T. A. (1990). Problems with application of interrupted time series analysis for brief single-subject data. *Behavioral Assessment, 12*, 355-370.

Hantula, D. A. (1995). Disciplined decision making in an interdisciplinary environment: Some implications for clinical applications of statistical process control. *Journal of Applied Behavior Analysis, 28*, 371-377.

Harrop, J. W. i Velicer, W. F. (1985). A comparison of alternative approaches to the analysis of interrupted time-series. *Multivariate Behavioral Research, 20*, 27-44.

Hartmann, D. P. (1974). Forcing square pegs into round holes: Some comments on "An analysis-of-variance model for the intrasubject replication design". *Journal of Applied Behavior Analysis, 7*, 635-638.

Hayes, A. F. (1996). Permutation test is not distribution-free: Testing H_0: $\rho = 0$. *Psychological Methods, 1*, 184-198.

Hersen, M. i Barlow, D. H. (1976). *Single case experimental designs: Strategies for studying behavior change.* New York: Pergamon Press.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S. L. i Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*, 165–179.

Hugdahl, K. i Öst, L.-G. (1981). On the difference between statistical and clinical significance. *Behavioral Assessment, 3*, 289-295.

Huitema, B. E. (1985). Autocorrelation in behavior analysis: A myth. *Behavioral Assessment, 7*, 107-118.

Huitema, B. E. (1988). Autocorrelation: 10 years of confusion. *Behavioral Assessment, 10*, 253-294.

Huitema, B. E. (2004). Analysis of interrupted time-series experiments using ITSE: A critique. *Understanding statistics, 3*, 27-46.

Huitema, B. E. i McKean, J. W. (1991). Autocorrelation estimation and inference with small samples. *Psychological Bulletin, 110*, 291-304.

Huitema, B. E., McKean, J. W. i Laraway, S. (2007). Time series intervention analysis using ITSACORR: Fatal flaws. *Journal of Modern Applied Statistical Methods, 6*, 367-379.

Huitema, B. E., McKean, J. W. i McKnight, S. (1999). Autocorrelation effects on least-squares intervention analysis of short time series. *Educational and Psychological Measurement, 59*, 767-786.

Jenson, W. R., Clark, E., Kircher, J. C. i Kristjansson, S. D. (2007). Statistical reform: Evidence-based practice, meta-analyses, and single subject designs. *Psychology in the Schools, 44*, 483-493.

Johnston, J. M. i Pennypacker, H. S. (2008). *Strategies and tactics of behavioral research* (3rd ed.). New York, NJ: Routledge.

Jones, R. R., Vaught, R. S. i Weinrott, M. R. (1977). Time-series analysis in operant research. *Journal of Applied Behavior Analysis, 10,* 151-166.

Jones, R. R., Weinrott, M. R. i Vaught, R. S. (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis, 11,* 277-283.

Kazdin, A. E. (1978a). Methodological and interpretive problems of single-case experimental designs. *Journal of Consulting and Clinical Psychology, 46*, 629-642.

Kazdin, A. E. (1978b). Statistical analyses for single-case experimental designs. En M. Hersen i D. H. Barlow (Eds.), *Single case experimental designs: Strategies for studying behavior change.* New York: Pergamon Press.

Kazdin, A. E. (1980). Obstacles in using randomization tests in single-case experimentation. *Journal of Educational Statistics, 5*, 253-260.

Kazdin, A. E. (1982). *Single-case research design: Methods for clinical and applied settings.* New York: Oxford University Press.

Kazdin, A. E. (2001). *Behavior modification in applied settings* (6th ed.). Belmont, CA: Wadsworth.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746-759.

Knapp. T. J. (1983). Behavioral analysts' visual appraisal of behavior change in graphic display. *Behavioral Assessment, 5*, 155-164.

Kratochwill, T. R. i Brody, G. H. (1978). Single subject designs: A perspective on the controversy over employing statistical inference and implications for research and training in behavior modification. *Behavior Modification, 2*, 291-307.

Kratochwill, T. R. i Levin, J. R. (1980). On the applicability of various data analysis procedures to the simultaneous and alternating treatment designs in behavior therapy research. *Behavioral Assessment, 2*, 353-360.

Kromrey, J. D. i Foster-Johnson, L. (1996). Determining the efficacy of intervention: The use of effect sizes for data analysis in single-subject research. *The Journal of Experimental Education, 65*, 73-93.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject research: Percentage of data points exceeding the median. *Behavior modification, 30*, 598-617.

Mace, F. C. i Kratochwill, T. R. (1986). The individual subject in behavior analysis research. En J. Valsiner (Ed.), *The individual subject and scientific psychology* (pp. 153-180). London: Plenum Press.

Manolov, R. i Solanas, A. (2009). Problems of the randomization test for AB designs. *Psicológica, 30*, 137-154.

Manolov, R., Solanas, A., Bulté, I. i Onghena, P. (2010). Data-division-specific robustness and power for ABAB designs. *The Journal of Experimental Education, 78*, 191-214.

Marascuilo, L. A. i Busk, P. L. (1988). Combining statistics for multiple-baseline AB and replicated ABAB designs across subjects. *Behavioral Assessment, 10*, 1-28.

Mastropieri, M. A. i Scruggs, T. E. (1985). Early intervention for socially withdrawn children. *Journal of Special Education, 19*, 429-441.

Matyas, T. A. i Greenwood, K. M. (1990). Visual analysis for single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23*, 341-351.

Matyas, T. A. i Greenwood, K. M. (1991). Problems in the estimation of autocorrelation in brief time series and some implications for behavioral data. *Behavioral Assessment, 13*, 137-157.

Matyas, T. A. i Greenwood, K. M. (1997). Serial dependency in single-case time series. En R. D. Franklin, D. B. Allison i B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 215-243). Mahwah, NJ: Lawrence Erlbaum.

McGrath, R. E. i Meyer, G. J. (2006). When effect size disagree: The case of *r* and *d*. *Psychological Methods, 11*, 386-401.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166.

Michael, J. (1974). Statistical inference for individual organism research: Mixed blessing or curse? *Journal of Applied Behavior Analysis, 7*, 647-653.

Miller, M. J. (1985). Analyzing client change graphically. *Journal of Counseling and Development, 63*, 491-494.

Noreen, E. R. (1989). *Computer-intensive methods for testing hypotheses: An introduction.* New York: John Wiley & Sons.

Normand, M. P. i Bailey, J. S. (2006). The effects of celeration lines on visual data analysis. *Behavior Modification, 30*, 295-314.

Olive, M. L. i Smith, B. W. (2005). Effect size calculations and single subject designs. *Educational Psychology, 25*, 313-324.

Onghena, P. i Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *Clinical Journal of Pain, 21*, 56-68.

Ottenbacher, K. J. (1990). When is a picture worth a thousand p values? A comparison of visual and quantitative methods to analyze single subject data. *Journal of Special Education, 23*, 436-449.

Parker, R. I. (2006). Increased reliability for single-case research results: Is bootstrap the answer? *Behavior Therapy, 37*, 326-338.

Parker, R. I. i Brossart, D. F. (2003). Evaluating single-case research data: A comparison of seven statistical methods. *Behavior Therapy, 34*, 189-211.

Parker, R. I. i Brossart, D. F. (2006). Phase contrasts for multiphase single case intervention designs. *School Psychology Quarterly, 21*, 46-61.

Parker, R. I., Brossart, D. F., Vannest, K. J., Long, J. R., Garcia De-Alba, R., Baugh, F. G. i Sullivan, J. R. (2005). Effect sizes in single case research: How large is large? *School Psychology Review, 34*, 116-132.

Parker, R. I., Cryer, J. i Byrns, G. (2006). Controlling baseline trend in single-case research. *School Psychology Quarterly, 21*, 418-443.

Parker, R. I. i Hagan-Burke, S. (2007a). Median-based overlap analysis for single case data: A second study. *Behavior Modification, 31*, 919-936.

Parker, R. I. i Hagan-Burke, S. (2007b). Single case research results as clinical outcomes. *Journal of School Psychology, 45*, 637-653.

Parker, R. I. i Hagan-Burke, S. (2007c). Useful effect size interpretations for single case research. *Behavior Therapy, 38*, 95-105.

Parker, R. I. i Vannest, K. J. (2009). An improved effect size for single case research: Non-overlap of all pairs (NAP). *Behavior Therapy, 40*, 357-367.

Parker, R. I., Hagan-Burke, S. i Vannest, K. (2007). Percentage of all non-overlapping data: An alternative to PND. *Journal of Special Education, 40*, 194-204.

Parsonson, B. S. i Baer, D. M. (1986). The graphic analysis of data. En A. Poling i R. W. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and advances* (pp. 157-186). New York: Plenum Press.

Pfadt, A., Cohen, I., Sudhalter, V., Romanczyk, R. i Wheeler, D. (1992). Applying statistical process control to clinical data: An illustration. *Journal of Applied Behavior Analysis, 25*, 551-560.

Pfadt, A. i Wheeler, D. J. (1995). Using statistical process control to make data-based clinical decisions. *Journal of Applied Behavior Analysis, 28*, 349-370.

Rabin, C. (1981). The single-case design in family therapy evaluation research. *Family Process, 20*, 351-366.

Reynhout, G. i Carter, M. (2006). Social stories for children with disabilities. *Journal of Autism and Developmental Disorders, 36*, 445–469.

Richards, S. B., Taylor, R. L. i Ramasamy, R. (1997). Effects of subject and rater characteristics on the accuracy of visual analysis of single subject data. *Psychology in the Schools, 34*, 355-362.

Robey, R. R. i Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology, 45*, 283-288.

Rosnow, R. L. i Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.

Scheffé, H. (1959). *The analysis of variance.* New York: Wiley.

Schlosser, R. W., Lee, D. L. i Wendt, O. (2008). Application of the percentage of non-overlapping data (PND) in systematic reviews and meta-analyses: A systematic review of reporting characteristics. *Evidence-Based Communication Assessment and Intervention, 2*, 163-187.

Schlosser, R. W. i Sigafoos, J. (2008). Meta-analysis of single-subject designs: Why now? *Evidence-Based Communication Assessment and Intervention, 2*, 117-119.

Schneider, N., Goldstein, H. i Parker, R. (2008). Social skills interventions for children with autism: A meta-analytic application of percentage of all non-overlapping data (PAND). *Evidence-Based Communication Assessment and Intervention, 2*, 152-162.

Scruggs, T. E. i Mastropieri, M. A. (1998). Summarizing single-subject research: Issues and applications. *Behavior Modification, 22*, 221-242.

Scruggs, T. E., Mastropieri, M. A. i Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education*, 8, 24-33.

Serlin, R. C. (2000). Testing for robustness in Monte Carlo studies. *Psychological Methods, 5*, 230-240.

Shadish, W. R., Rindskopf, D. M. i Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention, 2*, 188-196.

Sharpley, C. F. i Alavosius, M. P. (1988). Autocorrelation in behavior data: An alternative perspective. *Behavior Assessment, 10*, 243-251.

Shine, L. C. i Bower, S. M. (1971). A one-way analysis of variance for single-subject designs. *Educational and Psychological Measurement, 31*, 105-113.

Sidman, M. (1960). *Tactics of scientific research*. New York, NJ: Basic Books.

Sierra, V., Solanas, A. i Quera, V. (2005). Randomization tests for systematic single-case designs are not always appropriate. *The Journal of Experimental Education, 73*, 140-160.

Simonton, D. K. (1977). Cross-sectional time-series experiments: Some suggested statistical analyses. *Psychological Bulletin, 84*, 489-502.

Solanas, A., Manolov, R. i Sierra, V. (2010). Lag-one autocorrelation in short series: Estimation and hypothesis testing. *Psicológica, 31,* 357-381.

Suen, H. K. (1987). On the epistemology of autocorrelation in applied behavior analysis. *Behavioral Assessment, 9,* 113-124.

Suen, H. K. i Ary, D. (1987). Autocorrelation in behavior analysis: Myth or reality? *Behavioral Assessment, 9*, 150-130.

Toothaker, L. E., Banz, M., Noble, C., Camp, J. i Davis, D. (1983). N = 1 designs: The failure of ANOVA-based tests. *Journal of Educational Statistics, 4*, 289-309.

Tryon, W. W. (1982). A simplified time-series analysis for evaluating treatment interventions. *Journal of Applied Behavior Analysis, 15*, 423-429.

Tryon, W. W. (1984). "A simplified time-series analysis for evaluating treatment interventions": A rejoinder to Blumberg. *Journal of Applied Behavior Analysis, 17*, 543-544.

Vallejo, G. (1994). Evaluación de los efectos de la intervenció en diseños de series temporales en presencia de tendencia. *Psicotema, 6*, 503-524.

Van den Noortgate, W. i Onghena, P. (2003a). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly, 18*, 325-346.

Van den Noortgate, W. i Onghena, P. (2003b). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers, 35*, 1-10.

Van den Noortgate, W. i Onghena, P. (2003c). Multilevel meta-analysis: A comparison with traditional meta-analytical procedures. *Educational and Psychological Measurement, 63*, 765-790.

Van den Noortgate, W. i Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention, 2*, 142-151.

Velicer, W. F. i McDonald, R. P. (1984). Time-series analysis without model identification. Multivariate Behavioral Research, 19, 33-47.

Wampold, B. E. i Furlong, M. J. (1981a). Randomization tests in single-subject designs: Illustrative examples. *Journal of Behavioral Assessment, 3*, 329-341.

Wampold, B. E. i Furlong, M. J. (1981b). The heuristics of visual inference. *Behavioral Assessment, 3*, 79-82.

White, D. M., Rusch, F. R., Kazdin, A. E. i Hartmann, D. P. (1989). Applications of meta-analysis in individual subject research. *Behavioral Assessment, 11*, 281-296.

Wilkinson, L. i Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 694-704.

Ximenes, V. M., Manolov, R., Solanas, A. i Quera, V. (2009). Factors affecting visual inference in single-case designs. *The Spanish Journal of Psychology, 12*, 823-832.