FORSKNINGSDEL

Forskningsartiklene i Spesialpedagogikk er underlagt strengere form- og innholdskrav enn fagartiklene.

Se forfatterveiledningene www.utdanningsnytt.no/spesialpedagogikk

Artiklene blir vurdert av to anonyme fagfeller (blind review) i tillegg til redaktør.

Single case design i evaluering av spesialpedagogiske tiltak. En gjennomgang av aktuelle analysemetoder for vurdering av effekt

ANNE CATHRINE THURMANN-MOE, seniorrådgiver i Statped og Ph.d.-stipendiat på Institutt for spesialpedagogikk, Universitetet i Oslo.

MONICA MELBY-LERVÅG, professor ved Institutt for spesialpedagogikk, Universitetet i Oslo.

ARNE LERVÅG, professor, Institutt for pedagogikk, Universitetet i Oslo.

Sammendrag

Artikkelen drøfter hvordan single case design (SCD) kan brukes i vurdering av spesialpedagogiske tiltak, og diskuterer spørsmål knyttet til validitet og alternative innfallsvinkler for vurdering av effekt. Det presenteres også et konkret eksempel fra en gjennomført studie. Det argumenteres for at SCD er spesielt egnet til vurdering av effekt av spesialpedagogiske tiltak for elever med lavfrekvente og/eller sammensatte vansker. I skolen kan SCD bidra til å øke kvaliteten på spesialundervisningen gjennom mer nøyaktige vurderinger av hva som virker for den enkelte elev, og også bidra til utvikling av forskningsbasert kunnskap for praksis.

Summary

Single Case Design in Evaluation of Special Needs Education: A Methodological Review and an Example. This paper presents how Single Case Designs (SCD) can be used to evaluate interventions in the field of special needs education and discuss validity and alternative approaches for assessing efficacy. An example from a completed study is also presented. We argue that SCD is particularly suitable for assessing the effect of interventions for students with low-frequency and / or compound difficulties. In schools, SCD can help to improve the quality of special education delivery through more accurate assessments of what works.

Nøkkelord

SPESIALUNDERVISNING
SINGEL CASE DESIGN
EFFEKTSTUDIER
EVALUERING

Innledning

Andelen elever i Norge som mottar spesialundervisning etter enkeltvedtak, er i skoleåret 2019/2020 ca. 8 prosent (UDIR). Samtidig konkluderer forskning med uklart og i noen tilfeller også negativt utbytte av spesialundervisning (Kvande, Bjørklund, Lydersen, Belsky & Wichstrøm, 2018; Opsvik & Haug, 2017; Lekhal 2017; Skorpen, 2017).

Offentlige utredninger, sist utført av Barneombudet (2017) og det regjeringsoppnevnte Nordahl-utvalget (2018) samt melding til Stortinget nr. 6 (2019–2020) har også satt spørsmålstegn ved om bruken av ressurser til spesialundervisning fører fram. Rapportene konkluderer med svikt i mange ledd i tiltakskjeden knyttet til planlegging, gjennomføring, medvirkning fra elev og foresatte, samt evaluering og rapportering av igangsatte spesialpedagogiske tiltak.

Både av hensyn til den enkelte elev og med tanke på bruk av offentlig ressurser synes det i lys av dette å være et stort behov for nærmere vurdering av effekten av konkrete spesialpedagogiske tiltak. Vi vil i denne artikkelen drøfte hvordan Single Case Design (SCD) kan bidra til en sikrere evaluering av hvorvidt igangsatte tiltak for enkeltelever etter vedtak om spesialundervisning har effekt. Vi vil også diskutere hvordan SCD kan bidra til utvikling av mer forskningsbaserte tiltak i praksisfeltet, noe som i økende grad etterspørres av utdanningsmyndighetene (se f.eks. Meld. St. nr. 21 (2016–2017)).

De fleste undersøkelsene av effekt av spesialundervisning er gjort på gruppenivå, der grupper av elever som har vedtak om spesialundervisning, sammenlignes med grupper av elever som ikke har vedtak om spesialundervisning. Denne typen design har flere svakheter. For det første blir gruppene systematisk ulike fordi det er en grunn til at noen fikk enkeltvedtak mens andre ikke fikk det. For det andre kan gruppen som mottar spesialundervisning, være svært heterogen (Skorpen, 2017).

Randomiserte kontrollerte studier (RCT) regnes i utgangspunktet for å være den sikreste metoden for å evaluere effekten av et tiltak, fordi flest mulig alternative forklaringer kan utelukkes (Maggin, Lane & Pustejovsky, 2017; Odom mfl., 2005; WWC, 2017). Av etiske årsaker er det imidlertid vanskelig å designe studier der likeverdige elever med vedtak om spesialundervisning randomiseres i grupper der bare halvparten får tiltaket (Lekhal, 2017). I spesialpedagogikk arbeider man også ofte med lavfrekvente grupper hvor det er stor heterogenitet innad i gruppen. Det kan gjøre det krevende å bruke dette designet til å vurdere effekt av tiltak rettet mot lavfrekvente grupper (Lobo mfl., 2017; Onghena, Michiels, Jamshidi, Moeyaert & Van den Noortgate, 2018).

I et mer klinisk perspektiv har gruppedesign også begrensninger når det gjelder å vurdere effekt av tiltak for enkeltelever, fordi en gjennomsnittseffekt basert på en gruppe bare gir en tentativ indikasjon på hvordan tiltaket faktisk virket for hver enkelt deltaker (Onghena, Michiels, Jamshidi, Moeyaert, & Van Den Noortgate, 2018).

I SCD prøves tiltaket direkte på en deltaker eller eventuelt en liten gruppe deltakere og gir direkte mål på effekt for dem som har deltatt i den situasjonen tiltaket er prøvd ut i.

Grunnlaget for konklusjoner fra gruppestudier og SCD bygger dermed på noe ulik logikk (Onghena mfl., 2018). I gruppedesign går man bredt ut og prøver tiltaket på mange deltakere som basis for å konkludere om antatt effekt for enkeltpersoner fra samme populasjon. I SCD blir sammenhengen mellom et gitt tiltak og eventuell endring av atferd hos den enkelte deltaker målt direkte.

SCD er særlig relevant når det gjelder å vurdere effekt av tiltak for enkeltelever eller små grupper innen vanskeområder som er for lavfrekvente eller for heterogene til at det er hensiktsmessig eller praktisk mulig å rekruttere store og randomiserte utvalg. Dette gjør metoden aktuell for store deler av det spesialpedagogiske fagfeltet (De Bruin, 2017; Horner mfl., 2005; Maggin mfl., 2018). SCD er også egnet som utgangspunkt for pilotering før gruppestudier (Krasny-Pacini & Evans, 2018). Ikke minst er SCD velegnet i studier der målet ikke først og fremst er å predikere effekt for framtiden (generalisert effekt), men der man vil kvalitetssikre utprøving av tiltak for en enkelt person eller en definert klinisk gruppe i en praktisk kontekst (Onghena mfl., 2018).

Ulike Single Case Design

Single Case Design (SCD) er en av de eldste kjente metodene for å vurdere effekt av tiltak, og metoden er særlig brukt i klinisk virksomhet på områder innenfor medisin, psykologi og spesialpedagogikk (Shadish, 2014; Shamseer mfl., 2015; Tate mfl., 2016). Single Case-metodologien ble først etablert for å kunne igangsette kontrollerte eksperimenter med kun en deltaker, såkalte N = 1-studier (Kazdin, 2016; Kratochwill & Levin, 2010). Senere har designet blitt utviklet til også å omfatte flere N = 1-deltakere i en kjede (Multiple baseline design) (Gast & Ledford, 2014; Kazdin, 2016). Grunnlaget for alle SCD er først å definere observerbar atferd som kan måles over tid. Utover dette er metoden fleksibel med hensyn til valg av målemetode, og både direkte observasjon, selvrapporteringsskjemaer og repeterte kartlegginger kan brukes. Av disse er direkte observasjon den tradisjonelt mest brukte (Klingbeil, Van Norman & Nelson, 2017).

Den overordnede logikken i SCD er at hver deltaker tjener som sin egen kontroll gjennom at det foretas repeterte målinger i en baseline-fase (A), før man introduserer fasen med tiltak (B). Resultatet fra målingene i baseline-fasen antas å være representativt for atferden eller prestasjonene før intervensjonen starter, og validiteten styrkes ved at det gjennomføres flere målinger,

ideelt sett minst fem (Kratochwill mfl. 2013; Tate mfl., 2016). Målingene fortsetter gjennom intervensjonsfasen, og deretter sammenlignes resultatene fra de to fasene. Slike enkle tofasedesign er utgangspunktet for alle andre SCD (Ledford, 2018) og kan brukes til å vurdere kvalitative endringer hos enkeltelever over tid (Shadish, 2014), men de fyller ellers ikke forskningsmessige krav til eksperimentell kontroll (Ledford, 2018; Lobo, Moeyaerth, Cunha & Babik, 2017; Tate mfl., 2016).

I konsensusrapporten fra The Single Case Reporting Guidline in Behavioural Interventions (SCRIBE) (Tate mfl., 2016) skilles det derfor mellom tofasedesign og eksperimentell design. Skillet angår først og fremst i hvilken grad designet inkorporerer replikasjoner enten ved å prøve tiltaket på flere elever, situasjoner eller fag i en kjede der tiltaket introduseres til ulik tid (multiple baseline design), eller ved å systematisk «skru» tiltaket av og på for samme deltaker gjennom flere enn to faser. For å kunne konkludere med funksjonell sammenheng mellom tiltaket og resultatet har det vært foreslått at effekten bør repliseres tre ganger, det vil si på tre deltakere, i tre situasjoner eller gjennom tre faser av studien (Ledford, 2018; Lobo mfl., 2017; Maggin, Cook & Cook, 2018). En underliggende forutsetning er også at data fyller anbefalte kvalitetskrav knyttet til både design og egenskaper ved innsamlede data (se nedenfor).

Reversible design – ABAB-design

For undersøkelser der man ønsker å vurdere en umiddelbar atferdsendring hos elever, for eksempel om intensivert muntlig ros fra lærer fører til økt elevaktivitet i timen, er det mulig å bruke design som skrur tiltaket av og på fra fase til fase, såkalte ABAB-design. I ABAB-design er vurdering av effekt knyttet til hvorvidt avhengig variabel (for eksempel økt elevaktivitet) systematisk øker hver gang tiltaket i form av positiv forsterkning settes i gang, og så avtar når tiltaket fjernes. ABAB-design kan også utvides med flere faser (f.eks. ABABAABAAB) eller man kan introdusere et tiltak til (C; så designet blir: ABABACAC) (Gast & Backey, 2014).

Irreversible design – multiple baseline design (MBD)

For tiltak som tar sikte på en varig endring av elevens atferd, for eksempel ved å bedre ferdigheter i et akademisk fag, kan ikke tiltaket reverseres. I multiple baseline design (MBD) (Baer, Wolf & Risley, 1968; Gast, Lloyd & Ledford, 2014) introduserer man samme tiltak til et avgrenset antall elever (som oftest tre eller flere) på ulike tidspunkter langs en tidslinje. Målingene i baseline-fasen foregår kontinuerlig for alle deltakerne, men tiltaket settes i verk til ulik tid. Forutsatt at øvrige kvalitetskrav er oppfylt, konkluderes det med positiv effekt av tiltaket dersom det demonstreres en positiv endring av avhengig variabel som for minst tre deltakere kan assosieres med tidspunktet tiltaket ble satt i verk. Forholdstallet synes å ta utgangspunkt i at MBD-design vanligvis inkluderer 3–4 deltakere (Shadish & Sullivan, 2011), og at noen deltakere potensielt må ekskluderes fra vurderingen på grunn av ustabile data, for få datapunkter og lignende.

En variant av multiple baseline design er multiple probe design (Horner & Baer, 1978) der målingene langs tidslinjen ikke foregår kontinuerlig, men organiseres i planlagte bolker. Denne typen design er spesielt egnet hvis det er usannsynlig med store variasjoner i baseline-fasen. Fordelen med denne varianten er også at den minsker risikoen for at elevene blir enten lei av eller for godt kjent med kartleggingsprosedyrene (Horner & Baer, 1978). MBD/probe design kan også brukes til å undersøke om et tiltak for samme elev virker i flere fag eller situasjoner, og tiltaket introduseres da på ulikt tidspunkt i de ulike settingene (minst tre).

Kausalitet og evidens

I SCD er den indre validiteten knyttet til i hvilken grad man har eksperimentell kontroll med at eventuell endring fra baseline til intervensjonsfase skyldes det igangsatte tiltaket, og ikke andre forhold. Nedenfor vil vi gå gjennom de viktigste truslene mot den indre validiteten i SCD og konkretisere hvordan disse truslene kan møtes (Gast, 2014; Tate mfl., 2016).

Trusler mot indre validitet og begrepsvaliditet

Historie: I den perioden tiltaket foregår, kan det også inntre andre hendelser som påvirker den variabelen man ønsker å undersøke. Det kan være at konkurrerende tiltak som man ikke kontrollerer, settes i verk relativt samtidig. Dette kan dreie seg om større endringer i elevens liv eller mindre forbigående episoder (Gast, 2014). Slike hendelser er ikke lett å kontrollere for, men nøyaktige nedtegnelser, gjerne som loggføring både fra undervisning og kartlegging, kan bidra til å forklare endringer som skyldes eksterne faktorer.

Modning: Hvis tiltaket foregår over lang tid, vil man måtte ta i betraktning at det kan skje en modning av ferdigheter ettersom tiden går, uavhengig av tiltaket. Ifølge Gast (2014) er dette spesielt aktuelt dersom tiltaket gjelder unge barn og strekker seg over fire til seks måneder.

Test-/retest-effekt: Repeterte målinger av samme fenomen er kjernen i ethvert single case-eksperiment, men innebærer samtidig en trussel mot validiteten. Repeterte målinger kan føre til at eleven går lei og presterer svakt som følge av dette. En annen mulighet er at kjennskap til testprosedyren fører til bedrede prestasjoner (Gast, 2014). For å unngå at eleven blir for godt kjent med oppgavene, anbefales det å randomisere rekkefølgen av oppgaver og testledd fra gang til gang (ibid.). Ved visuell inspeksjon (se nedenfor) regnes endring av data som kan assosieres med tidspunkt for oppstart av intervensjonen, og som ikke forekommer eller er mindre synlige i baseline-fasen, som indikasjon på at resultatet i intervensjonsfasen ikke er påvirket av testeffekt (Gast, Lloyd & Ledford, 2014).

Hawthorne-effekt: Repetere målinger innebærer økt oppmerksomhet rundt eleven. Eleven blir også kjent med at han er med på et eksperiment, og det er relevant å diskutere hvorvidt det å være gjenstand for oppmerksomhet i seg selv virker inn på prestasjonene (Shadish, Cook, & Campbell, 2002). Det at prosedyrene i forbindelse med kartleggingen er nye for eleven, kan innvirke både i positiv og negativ retning. For å styrke validiteten anbefales det å la eleven bli kjent med prosedyrene og eventuelle fremmede testledere før man starter datainnsamlingen (Gast, 2014). Såkalt performance-effekt kan også oppstå både hos elev og andre involverte, for eksempel testledere og lærere. Dette innebærer at atferden økes i positiv retning fordi man blir observert. Dette kan bety at den målte effekten ikke kun er knyttet til tiltaket, men også til det å delta i et eksperiment.

Stabilitet: Ustabile data er en trussel mot validiteten og kan forstyrre tolkningen av resultatene. Dette kan være enten som følge av faste sykliske variasjoner, for eksempel hvis data ved daglige målinger alltid er bedre på en fredag enn en mandag, eller data med stor variasjon fra måling til måling internt i fasen (Gast, 2014). Stabilitet i data er også en sentral forutsetning for at en studie skal oppnå kvalitetskravene for at man kan vurdere potensiell effekt av et tiltak (Kratochwill mfl., 2010).

Implementeringsvaliditet: Implementeringsvaliditet er avhengig av at prosedyrer i forbindelse med datainnsamling og selve tiltaket gjennomføres som planlagt. For å sikre dette bør det benyttes opptak enten med lyd/bilde eller bare lyd, som gjør det mulig for en annen fagperson å vurdere og tolke data (Gast, 2014).

Sosial og økologisk validitet: I SCD legges det vekt på sosial validitet, altså hvorvidt tiltaket oppfattes positivt og relevant i det aktuelle miljøet, og økologisk validitet, som dreier seg om hvorvidt tiltaket er prøvd ut på en måte og i en situasjon som er representativ for praksisfeltet (Gast, 2014). Det er derfor viktig at tiltak prøves ut i en naturlig kontekst. Sosial validitet kartlegges ved å benytte intervju/spørreskjema til viktige nærstående personer, for eksempel lærere og foreldre. Deres oppfatninger må tas i betraktning når det gjelder vurdering av videre bruk av tiltaket og anbefales også brukt for å validere resultatene (Tate mfl., 2016).

Ytre validitet og effektbegrepet i SCD

Historisk har SCD vært et verktøy for evaluering av tiltak i en klinisk kontekst. Tradisjonelt er en enkeltstående SCD-studie også først og fremst et eksempel på hvordan et aktuelt tiltak virker i den situasjonen det er prøvd ut. Resultatene fra en enkelt studie kan derfor i utgangspunktet ikke generaliseres. Ved å sikre at antall målepunkter i baseline-fasen er tilstrekkelige, og ved å møte aktuelle trusler mot den indre validiteten, kan imidlertid enkeltstående SCD oppnå god eksperimentell kontroll (Tate mfl., 2016). Når studien også fyller krav til presise casebeskrivelser og nøyaktige beskrivelser av det utprøvde tiltaket, kan resultatene gi et forskningsbasert grunnlag for utprøving på nye elever med lignende status (Maggin mfl., 2018).

Dersom SCD skal bidra til å dokumentere generalisert effekt av et gitt spesialpedagogisk tiltak, kreves det at effekt repliseres gjennom gjentatte SCD-studier og metaanalyser. For at et tiltak vurdert med SCD skal kunne regnes som forskningsbasert, har det vært foreslått at positiv effekt av tiltaket må være dokumentert i minst fem uavhengige studier med til sammen minst 20 N = 1-design, utført av minst tre ulike forskningsmiljøer («3–5–20-kriteriet») (Hitchcook, Kratochwill & Chezan, 2015; Maggin mfl., 2018; WWC, 2017). Inkludert i disse kriteriene er det også kvalitetskrav knyttet både til overordnet design og til egenskaper ved innsamlede data. For nærmere diskusjon om kvalitetskrav og kriterier for vurdering av effekt i SCD se ellers Zimmerman mfl. (2018a) for beskrivelse og sammenligning av ulike evidensstandarder.

Som en konsekvens av etterspørselen etter såkalt forskningsbaserte undervisningstiltak har det i USA, og etter hvert også i Europa, blitt etablert åpne databaser over undervisningstiltak som fyller kriterier for å regnes som effektive (se f.eks. What Works Clearing House (WWC)). Siden SCD ofte foregår i praksisnære kontekster har det vært ønskelig å kunne inkludere resultater fra SCD i vurdering av konkrete undervisningsprogrammer (Horner mfl., 2005; Kratochwill mfl., 2013; Shadish, 2012; 2014). Flere har imidlertid vært kritiske til at databasenes standarder primært har tatt utgangspunkt i kvalitetskrav innen gruppedesignmetodologi, og beklaget konsekvensen dette har fått for metodeutviklingen i SCD i retning av økte krav til generaliserbarhet (de Bruin, 2017; Horner mfl., 2005; Ledford, Wolery & Gast, 2014; Maggin mfl., 2017). Hovedinnvendingene har vært at lavfrekvente vanskeområder risikerer å bli underrepresentert i forskningen, og at fordelene ved småskalastudier blir oversett (Onghena mfl., 2018).

Parallelt har det, primært med tanke på å kunne gjennomføre metaanalyser, blitt lansert statistiske analyseverktøy for vurdering av samlet deltakereffekt i SCD (Shadish, Hedges, & Pustejovsky, 2013). Dette har hatt som formål å kunne presentere synteser fra flere enkeltstående SCD på en slik måte at de kan sammenlignes med resultater fra gruppestudier (se f.eks. Shadish mfl., 2015). I oppdaterte retningslinjer fra WWC (2020) er bruk av design-uavhengige effektmål et krav, og det antydes at resultater fra denne typen mål delvis kan erstatte det tidligere 3–5–20-kriteriet.

Metoder for dataanalyse

Analyse av data i SCD har tradisjonelt basert seg på metoder som ikke forutsetter at man har tilgang på programvare eller kompetanse som ikke er tilgjengelig i en klinisk hverdag. Dataanalysen baseres på visuelle inspeksjoner av grafiske framstillinger av resultatene og/eller ulike kvantitative effektmål. De siste årene har det også blitt utviklet nettbaserte kalkulatorer for utregning av effekt. Nedenfor vil vi presentere noen aktuelle visuelle analysemetoder og et utvalg kvantitative effektmål. Listen er absolutt ikke fullstendig, og i utvelgelsen er det blitt lagt vekt på metoder som ofte er referert i publiserte studier, og som har god tilgjengelighet ved at de utføres enten for hånd, ved hjelp av ordinær programvare på PC eller via gratis nettressurser.

Visuell inspeksjon av data

Visuell inspeksjon har historisk sett vært den mest brukte metoden for å vurdere effekt i SCD (Gast & Ledford, 2014; Lane & Gast 2014; Shadish, 2014). Den

visuelle inspeksjonen av data har primært to formål. Det første er å undersøke hvorvidt egenskaper ved data er tilfredsstillende som grunnlag for å vurdere eventuell effekt. Her vurderes særlig stabilitet og trender i data. I klinisk sammenheng kan dette også inngå som et arbeidsredskap i gjennomføring av studien, for eksempel ved at man avventer oppstart av intervensjonsfasen til data i intervensjonsfasen er stabile (Gast & Spriggs, 2014). Det andre er å vurdere hvorvidt det kan påvises en funksjonell sammenheng mellom oppstart av tiltak og endring i avhengig variabel. Denne vurderingen angår grad av overlapp i datapunkter mellom fasene, eventuell umiddelbar endring i grafens helningsgrad ved oppstart av intervensjon og nivåforskjeller mellom fasene. Visuelle analyser angir ikke grad eller størrelse på sammenhengen, men er dikotom (Gast & Ledford, 2014). Det er utviklet retningslinjer for vurdering av trender, stabilitet og overlapp av data, som bidrar til å øke validiteten av de visuelle analysene (Gast & Spriggs, 2014; Lane & Gast, 2014). Hovedpunkter når det gjelder disse retningslinjene presenteres nedenfor (Kratochwill mfl., 2010; Lane & Gast, 2014).

Stabilitet

Stabilitet i data regnes ut for hver fase ved at man først finner medianverdien i baseline-fasen og så etablerer et intervall rundt medianen, som utgjør +/- 25 prosent av verdien. Dette intervallet brukes også i intervensjonsfasen (Gast & Spriggs, 2014; Lane & Gast, 2014). For at data i fasen skal vurderes som stabile, kreves det at 80 prosent av datapunktene skal være innenfor intervallet. Stabilitet internt i fasene er viktig for å kunne foreta en gyldig sammenligning av det gjennomsnittlige nivået i hver fase. Ideelt sett er det anbefalt å fortsette baseline-målingene og vente med å introdusere tiltaket inntil man har oppnådd et stabilt resultat (ibid.).

Trender

Trender i data vurderes gjennom at man tegner opp trendlinjer for hver fase. Det er relevant å avdekke hvorvidt det skjer en endring i trendens helningsgrad ved oppstart av intervensjonen, og det er også relevant å vurdere hvorvidt det er tydelige og stabile trender i baseline-fasen, fordi dette kan være en begrensing når det gjelder vurdering av effekt.

Ved «split-middle»-metoden danner medianverdien fra hver halvpart av hver fase utgangspunkt for å trekke opp trendlinjer (Gast & Spriggs, 2014). Stabiliteten vurderes ved å trekke opp et intervall rundt trendlinjen og regne ut hvor stor prosentandel av punktene som faller innenfor intervallet (Gast & Spriggs, 2014; Lane & Gast, 2014; Manolov mfl., 2018). Størrelsen på intervallet og utregningsmåten er den samme som beskrevet i avsnittet om stabilitet over. Intervallet angir +/-25 prosent av medianverdien i baseline-fasen og avvik fra dette intervallet for mer enn 20 prosent av datapunktene tolkes som at observert trend er upålitelig. En pålitelig trend skal være visuelt observerbar, men kan også regnes ut ved en enkel regresjonsanalyse. Her vil en lav korrelasjon / forklart variasjon (R2) mellom tidsvariabelen og avhengig variabel angi en upålitelig trend og vice versa (Manolov, mfl., 2018). Bruk av regresjonsanalyse forutsetter imidlertid at trenden er lineær. I tilfeller man er usikker, anbefales den manuelle metoden (Manolov mfl., 2018).

En metode for vurdering av trender er også å predikere trendlinjen basert på data i baseline-fasen og så se i hvilken grad den observerte trenden i intervensjonsfasen avviker fra den predikerte (Kazdin, 2011). Vurdering av effekt av et tiltak basert på ekstrapolering av trendlinjen fra baseline-fasen forutsetter imidlertid at observert trend i baseline-fasen automatisk videreføres over tid. Undersøkelser av publiserte SCD-studier har vist at dette i mange tilfeller ikke er tilfelle (Parker, Vannest, Davis & Sauber, 2011).

En begrensning ved trendanalysene er at det ikke finnes gyldige standarder for hva som regnes som en akseptabel endring av grafenes helningsgrad for vurdering av effekt i ulike design og med ulike målemetoder (Busse, McGill & Kennedy 2015; Manolov & Vannest, 2019). Det har også blitt påpekt at kriteriet om umiddelbar endring av trend i terapeutisk retning ved oppstart av intervensjonen ikke er relevant for studier der målet er

bedrede akademiske ferdigheter, siden dette krever tid å etablere (Klingbeil, Norman & Nelson, 2017).

Nivåendringer
Nivåendringer mellom fasene kan uttrykkes gjennom sammenligning av gjennomsnittet eller medianverdien av resultatene for hver fase.

Overlapp
Graden av overlapp mellom datapunkter i hver fase regnes ut ved at man teller opp antall punkter i intervensjonsfasen som er høyere (eventuelt lavere) enn det høyeste/laveste datapunktet i baseline-fasen (se også PND nedenfor).

Kvantitative effektmål
Effektstørrelser har betydning for å kunne kvantifisere størrelsen eller graden av effekt og skal ideelt fungere slik at det er mulig å sammenligne ulike studier (Hedges, 2008). Effektmålene i SCD er ikke direkte sammenlignbare på tvers av de ulike måleverktøyene, men bruker egne skalaer. Det skilles mellom tre «familier» av kvantitative mål i SCD: overlappingsteknikker, parametriske effektmål og mål for samlet deltakereffekt (Pustejovsky & Ferron, 2017).

Overlappingsteknikker
Det grunnleggende felles prinsippet for overlappingsteknikkene er å vurdere graden av overlapp mellom målinger gjort i baseline-fasen sammenlignet med intervensjonsfasen. Dette prinsippet er basis for en rekke utregningsmetoder og teknikker for vurdering av effekt (se Parker, Vannest & Davis, 2011 for en gjennomgang).

Percent of Nonoverlapping Data (PND) (Scruggs, Mastropieri, & Casto, 1987). Prosentandelen av datapunkter fra intervensjonsfasen som ikke overlapper med datapunkter i baseline-fasen (PND), er den enkleste av overlappingsteknikkene og regnes ut ved at man først finner det høyeste datapunktet i baseline-fasen (eller laveste hvis målet for studien er reduksjon av verdiene). Deretter teller man opp hvor mange datapunkt i intervensjonsfasen som ligger over (eller under) dette. Prosenttallet regnes ut ved at dette antallet divideres på antall målepunkter i intervensjonsfasen. For PND er den foreslåtte standarden at 50 og mindre viser dårlig effekt, skårer mellom 50 og 70 usikker effekt, skårer fra 70 til 90 god effekt og skårer over 90 svært god effekt (Scruggs & Mastropieri, 1998). Kritikk mot PND er knyttet til at et isolert (og mulig lite representativt) målepunkt i baseline-fasen tillegges for stor vekt ved utregning (Maggin mfl., 2011). Som et alternativ er det derfor foreslått å bruke medianverdien i baseline-fasen, PEM (The Percentage of datapoints in phase B exceeding the Median of the Baseline Phase), som sammenligningsgrunnlag (Ma, 2006; Parker mfl., 2011). Andre aktuelle varianter av PND er PAND (Percentage of All Nonoverlapping Data) og NAP (Nonoverlap of All Pairs (Parker mfl., 2011)).

Tau-U (Parker, Vannest, Davis & Sauber, 2011) er en effektstørrelse som kombinerer graden av overlapp mellom fasene med kontroll for trend i baseline-fasen. Det anbefales å kun kontrollere for trend i baseline-fasen dersom trenden er statistisk signifikant (Parker mfl., 2011). Tau-U er ikke-parametrisk og forutsetter dermed ikke normalfordelte data. Utregningen baseres på kombinasjon av to metoder: Mann-Whitney U og Kendalls Tau (Parker mfl., 2011). Begge disse utregningsmåtene baseres på sammenligning av alle teoretisk mulige kombinasjoner av dataverdier fra de to fasene, og Tau U-estimatet er uttrykket for prosentverdien av alle datapunkter som har økt verdi i intervensjonsfasen sammenlignet med baseline-fasen. Det er mulig å regne konfidensintervall og signifikansnivå for effektstørrelsen.

Effektstørrelsen angis som et estimat mellom 0 og 1. For Tau-U regnes estimat under 0.20 som liten effekt, mellom 0.20 og 0.60 som moderat effekt, fra 0.60 til 0.80 som stor effekt og over 0.80 som svært stor effekt (Vannest & Ninci, 2015). Diskusjoner om Tau-U som effektmål har blant annet omhandlet vurderingen av baseline-trender (Manolov mfl., 2018; Tarlow, 2017), og andre utregningsmetoder enn de som er gjort tilgjengelige av opphavspersonene, er foreslått (se f.eks. Tarlow

2016; 2017 og Pustejovsky & Swan, 2018) for alternativer). Et viktig moment i denne diskusjonen er imidlertid at de alternative metodene for trendkontroll forutsetter en lineær baselinetrend, mens den originale Tau-U kontrollerer for trender med både lineær og ikke-lineær variasjon.

Nettbaserte kalkulatorer for utregning av PND og Tau-U er tilgjengelig både via Vannest, Parker, Gonen & Adiguzel, (2016) og fra Pustejovsky & Swan (2018).

Parametriske effektmål

Standardized Mean Difference (SMD) og varianter av dette (Busk & Serlin, 1992; Cohen, 1988; Hegdes, Pustejovsky & Shadish, 2012) angir gjennomsnittlig endring fra baseline-fasen til intervensjonsfasen og regnes ut ved at gjennomsnitt for baselinemålingene for hver elev trekkes fra gjennomsnittsverdien fra intervensjonsfasen og deles på standardavviket i baseline-fasen eller på standardavviket for både baseline- og intervensjonsfasen. Tolkning av effektstørrelsen for SMD basert på gjennomgang av 52 publiserte SCD fra 2010-årgangen av Journal of Applied Behavior Analysis (Harrington & Velicer, 2015) har foreslått følgende tommelfingerregler: 0–1 er liten effekt, 1–2,5 er middels effekt og større enn 2,5 er stor effekt. Kalkulator for utregning av SMD er tilgjengelig online (Pustejovsky & Swan, 2018).

Log Respons Ratio (LRR) (Pustejovsky, 2018) er et effektmål som også sammenligner gjennomsnittet fra intervensjonsfasen med gjennomsnittet fra baseline-fasen. Det er ikke utviklet standard for LRR, men effektestimatet kan konverteres til et mål på prosentvis økning eller reduksjon av målt atferd i intervensjonsfasen sammenlignet med baseline. LRR er kun aktuell dersom måleskalaen som brukes, har et naturlig nullpunkt og er primært prøvd ut i studier basert på direkte atferdsobservasjon (ibid.). I situasjoner med gulv- og takeffekter, for eksempel hvis man måler utvikling av en ny ferdighet og baseline-målene er på null, vil enhver økning i intervensjonsfasen innebære en stor prosentvis endring og derfor være misvisende (ibid.). Også for Log Respons Ratio (LRR) er det kalkulator tilgjengelig (Pustejovsky & Swan, 2018).

Mål for samlet deltakereffekt

Between Case – Standardized Mean Difference (BC-SMD) er et nylig lansert effektmål som bruker samme standard som i gruppestudier (Valentine, Tanner Smith, Pustejovsky & Lau, 2016). Denne utregningsmåten angir ikke individuell effekt, men en samlet effektstørrelse for alle deltakerne. BC-SMD er foreløpig kun tilgjengelig for ABAB-design og for MBD/probe-design mellom deltakere. Effektstørrelsen regnes ut ved flernivåanalyse med mulighet for å la de individuelle deltakerne variere både med hensyn til utgangspunkt, oppnådd effekt og eventuelle trender i data. Utregningen av effektstørrelse baseres på Restricted Maximum Likelihood (REML) (Valentine mfl., 2016). Effektmålet oppgis med konfidensintervall og signifikansnivå og bruker samme skala som for Cohens d. Det vil si at d = 0.20 regnes som liten effekt, d =0.50 som moderat effekt og d = 0.80 som stor effekt (Cohen, 1988). Denne utregningsmåten gjør at resultater med BC-SMD teoretisk kan sammenlignes med resultater fra gruppestudier på samme skala (Shadish, Hedges, Horner & Odom, 2015). Det finnes en nettbasert kalkulator for BC-SMD der man laster inn filer med egne data (se Pustejovsky, 2016).

Også ved bruk av den nettbaserte kalkulatoren for TAU-U (se over) er det mulig å regne ut en samlet deltakereffekt. Denne regnes ut som en vektet gjennomsnittsverdi av de individuelle effektstørrelsene for et gitt utvalg av deltakere i samme studie. Det kombinerte effektestimatet bruker samme standard som Tau-U for individuell effekt og oppgis også med konfidensintervall og signifikansnivå.

Vurdering av de ulike effektmålene

Et stort utvalg av ulike analysemetoder og teknikker for vurdering av effekt i SCD er altså tilgjengelig, og flere studier har sammenlignet de ulike metodenes egnethet både med bruk av simulerte data og ved reanalyse av allerede publiserte studier (Manolov mfl., 2010; Wolfe, Dickenson, Miller & McGrath, 2019; Zimmerman mfl., 2018a).

Manolov og kolleger (2010) undersøkte i hvilken grad effektmålene PND og variantene PEM og PAND ble

påvirket av grad av autokorrelasjon og antall målepunkter (serielengde) i hver fase. De brukte simulerte data der de manipulerte både autokorrelasjon og serielengde, og vurderte så hvordan dette påvirket effektstørrelsen fra de forskjellige målemetodene. Resultatene viste at PND og PAND begge ble påvirket av autokorrelasjon, og at dette førte til både overestimering og underestimering av resultatene. PEM var mer robust både for autokorrelasjon og serielengde.

Pustejovsky (2019) undersøkte i hvilken grad ulike effektmål påvirkes av datainnsamlingsprosedyrer. Han sammenlignet fem forskjellige effektmål basert på overlapp (PND, PEM, NAP, RIRD, Tau) med to parametriske effektmål (SMD og LRR). Han gjennomførte analyser med simulerte data og vurderte så de ulike effektmålenes sensitivitet for variasjon i antall målepunkter i hver fase, observasjonslengde og system for datainnsamling. Han fant at de parametriske metodene generelt var mer robuste når det gjaldt påvirkning fra forhold knyttet til prosedyrene enn effektmålene basert på overlapp. Dette gjaldt særlig for LRR, som i denne undersøkelsen viste seg å gi mer presise estimater enn SMD. En annen undersøkelse som sammenlignet de samme effektmålene med utgangspunkt i publiserte SCD-studier som vurderte effekten av sensorisk stimulering i autistiske utvalg (Zimmerman mfl., 2018 b), kom til samme konklusjon. Også her ble LRR vurdert som det effektmålet som samlet sett var det mest valide, mens Tau-U ble vurdert som den mest valide av teknikkene basert på overlapp.

I en annen undersøkelse som brukte simulerte data til å vurdere ulike effektmåls sensitivitet i forhold til autokorrelasjon, kom imidlertid SMD godt ut (Manolov, Solanas & Sierra, 2010). Denne undersøkelsen indikerte også at SMD er egnet til å fange opp effekt også i datasett med relativt få målepunkter.

Andre undersøkelser har konkludert med at det til tross for ulikheter synes å være relativt stort samsvar både mellom ulike kvantitative effektmål og også mellom visuelle analyser og kvantitative effektmål. Wolfe mfl. (2019) undersøkte i hvilken grad konklusjoner fra visuelle analyser samsvarte med resultater fra fire utvalgte statistiske metoder (IRD, Tau-U, Hedges g og BC-SMD). De rekrutterte et utvalg på 52 eksperter, og disse ble bedt om å gjøre selvstendige visuelle analyser av 25 grafer fra publiserte Multiple Baseline Design-studier. Forskerne gjorde selv de statistiske analysene ved å gå inn i det publiserte materialet. De sammenlignet så råskårene fra de fire statistiske analysene for å vurdere i hvilken grad disse kom til samme resultat. De vurderte også i hvilken grad ekspertene var samstemte i konklusjonene og i hvilken grad resultatene fra de visuelle analysene og de statistiske analysene samsvarte. Resultatene viste at for 67 prosent av de publiserte studiene kom de visuelle analysene og samtlige statistiske analyser til samme konklusjon når det gjaldt hvorvidt studien hadde hatt effekt. Det var også relativt høy grad av samsvar mellom de ulike måleverktøyene (Spearmans $r = 0.65–0.89$). For de resterende studiene, der ekspertene var uenige i konklusjonene, viste de statistiske målene både stor grad av samstemmighet og en gjennomgående mindre streng vurdering.

Som beskrevet over kan visuelle analyser fange opp detaljer i studien som ikke kommer fram ved statistisk analyse (Lane & Gast 2014). Samtidig har undersøkelser vist at både hvem som tolker dataene, og hvordan grafene teknisk framstilles, virker inn på vurderingene, særlig i tilfeller med variable data (Busse, McGill & Kennedy, 2015; Maggin & Odom, 2014; Pustejovsky, 2019). Når det gjelder kvantitative effektmål, er den viktigste kritikken mot de parametriske effektmålene at de fleste effektmålene i utgangspunktet forutsetter egenskaper ved data som er vanskelig å realisere i SCD. Dette gjelder særlig utfordringer knyttet til autokorrelasjon, trender og data som ikke er normalfordelte (Lobo mfl., 2017). De siste årene har det blitt lansert flere alternative metoder for å inkludere kontroll for autokorrelasjon og trend i de parametriske statistiske modellene (f.eks. Pustejovsky, Hegdes & Shadish, 2014; Swaminathan mfl., 2014). Et problem er imidlertid at mange av modellene forutsetter lineære trender og gir upresise estimater dersom de brukes på data som ikke er lineære (Manolov, Solanas & Sierra, 2018). Disse metodene forutsetter også

tilgang til spesiell programvare og statistisk kompetanse, noe som kan gjøre dem lite anvendelige for praksisfeltet. Et unntak er den omtalte BS-SMD, der det finnes en effektkalkulator. Denne er imidlertid kun for utregning av samlet effekt og kan ikke brukes for å vurdere effekt hos enkeltdeltakere.

De ikke-parametriske overlappingsteknikkene, som tradisjonelt har vært antatt å være mindre påvirket av autokorrelasjon (Kratochwill mfl., 2013), har på sin side blitt kritisert for å være upresise. Overlappingsteknikker påvirkes også i større grad av forhold knyttet til prosedyrer for datainnsamling, for eksempel antall målepunkter, observasjonsmetode og observasjonsintervall (Manolov, Solanas & Leiva, 2010; Pustejovsky, 2019).

Som vist har både visuelle analyser og samtlige kvantitative effektmål begrensninger. I henhold til standarder for SCD (Kratochwill mfl., 2013; Tate mfl., 2016) er det per i dag heller ikke enighet om hvilke effektmål som er best egnet. Det synes imidlertid å være enighet om at visuelle analyser og kvantitative mål utfyller hverandre og at det derfor er nyttig å bruke begge metoder i vurdering av resultatene (Kratochwill mfl., 2010; Maggin & Odom, 2014; Manolov mfl., Sierra, 2018; Shadish mfl., 2015). Det har også blitt foreslått at visuelle analyser brukes først, og at man dersom man finner at studien har hatt effekt, går videre med statistiske analyser for å kunne kvantifisere effektstørrelsen (Lobo mfl., 2017). Når det gjelder statistiske løsninger, har det blitt foreslått å bruke Tau-U for vurdering av individuell effekt dersom det er tydelige observerte trender i baseline-data (Rakap, 2015). I nylig oppdaterte retningslinjer for evidensstandarder i SCD (WWC, 2020) er det også inkludert krav om vurdering av samlet deltakereffekt.

SCD – et eksempel

I det følgende vil vi presentere data fra en multiplebaseline/probe-studie der vi vurderte effekten av et konkret spesialpedagogisk tiltak. Elevene som presenteres her, kom fra fire forskjellige skoler og ble henvist Statped på grunn av vedvarende lese- og skrivevansker (dysleksi) og lite utvikling til tross for langvarig spesialpedagogisk bistand. Elevenes leseproblemer var særlig knyttet til upresis fonologisk avkoding, og tiltaket ble rettet mot å øke elevenes lesenøyaktighet. Tiltaket ble gjennomført som individuell trening fire timer per uke over åtte uker og tok sikte på å lære elevene å bruke artikulasjon som støtte i ordavkoding. Intervensjonsprogrammet besto av konkrete læringsaktiviteter knyttet til å opparbeide artikulatorisk bevissthet. Studien ble realisert gjennom et samarbeid mellom lokal skole, Pedagogisk-psykologisk tjeneste (PPT) og Statped.

Vi vil her kort redegjøre for den praktiske gjennomføringen av denne studien. Presentasjonen inneholder bare et utvalg data fra den originale undersøkelsen og de aktuelle elevene er valgt ut for å illustrere problemstillinger ved visuelle og statistiske analyser knyttet til ulikheter i individuelle data. Lærervurderinger, nærmere casebeskrivelser og presentasjon av tiltaket er utelatt.

Beskrivelse av design

Mål for studien var å vurdere effekten av et konkret tiltak på et avgrenset område knyttet til elevenes lesestrategier. Vi ønsket også å vurdere hvorvidt den eventuelle effekten holdt seg eller ble borte når tiltaket ble avsluttet. Tiltaket ble gjennomført med utgangspunkt i elevenes ordinære timer til spesialundervisning, men for noen av elevene som deltok, måtte skolene omdisponere ressurser i åtte uker for å realisere individuell trening. For å oppnå et sikkert grunnlag for å vurdere effekt ble det gjennomført fem kartlegginger i baseline-fasen (Tate mfl., 2016). Målingene i denne fasen er uttrykk for elevenes utgangspunkt før tiltaket startet. Kartleggingene ble gjennomført av PPT og tidspunktet ble delvis tilpasset deres rutiner for besøk på skolene. Siden intervensjonen tok sikte på en varig endring av ferdigheter ble det valgt et multiple baseline design, men med elementer fra multiple probe design ved at målingene i baseline og postfasen ikke ble foretatt kontinuerlig, men organisert som blokker.

FIGUR 1 Kartleggingsplan, Multiple baseline/ probe design for fire elever med totalt 18 målinger pr elev.

Elev/ Uke	39	40	41	42	43	44	45	46	47	48	49	50	51	52	1	2	3	4	5	6	7	8	9	1	11	12	13	14	15	16	17	18
2	X	X		X	X	X	X	X	X	X	X	X			X	X	X	X					X	X	X							
3		X		X		X	X	X	X	X	X	X			X	X	X	X	X	X					X	X	X					
7	X		X	X	X	X	X	X	X	X	X	X			X	X	X	X							X	X	X					
10								X	X	X	X	X			X	X	X	X	X	X	X		X	X	X					X	X	X

Merknad: Uke 40 (41), 51, 52 og 8 skolefri på grunn av ferie. Baselinefase markert i lys grått og postfasen markert i mørk grått.

Operasjonalisering av avhengig variabel

Målet for tiltaket var å bedre elevenes fonologiske lesestrategi. Non-ordlesing regnes som et valid mål på fonologisk lesing (Pennington & Bishop, 2009), og det ble valgt å operasjonalisere begrepet «fonologisk lesestrategi» til «antall non ord lest korrekt på ett minutt». Dette er dermed avhengig variabel, mens uavhengig variabel er treningen som ble gitt gjennom de åtte ukene tiltaket varte. Som mål ble det brukt både antall korrekt leste ord per minutt (råskåre) og prosentvis andel av ord lest riktig.

Tiltak for å styrke validiteten

I planlegging av studien ble det gjort tiltak for å styrke validiteten. Det ble særlig lagt vekt på å minske risikoen for såkalt testeffekt av de gjentatte målingene. For å redusere risiko for at elevene skulle gå lei av testingen, ble det som nevnt benyttet en «probe»-variant av multiple baseline design, noe som innebar færre målinger i pre- og postfasen av studien sammenlignet med et klassisk multiple baseline design (Horner & Baer, 1978). For å redusere risiko for at elevene skulle huske oppgavene fra gang til gang, laget vi unike ordlister til hver kartlegging, som besto av likeverdige, men ikke identiske testledd. Siden lesing av ordlister med tidsbegrensning kan være sårbart for forbigående uoppmerksomhet leste elevene to unike lister hver gang de ble kartlagt, og et gjennomsnitt av disse to ble lagt til grunn. For å minske risiko for «performance»-effekt ble kartleggingen foretatt av en annen fagperson enn den som hadde ansvaret for undervisningen.

Vurdering av resultatene

I dette eksemplet vil vi legge vekt på å vise noen grunnleggende metoder som kan være egnet som utgangspunkt for evaluering av spesialpedagogiske tiltak i praksis.

Vi vil først vise hvordan visuelle analyser kan brukes for å vurdere kvaliteten på data som er samlet inn, og gi tentative indikasjoner på virkning av tiltaket. Vi vil deretter vise hvordan statistisk analyse av data kan supplere visuelle analyser både for vurdering av individuell effekt og for samlet deltakereffekt (Lobo mfl., 2017; Pustejovsky mfl., 2014).

Visuell inspeksjon av data

Som nevnt har visuell inspeksjon av data to funksjoner. For det første å vurdere om innsamlede data har den nødvendige stabiliteten i baseline-fasen, slik at det kan foretas en nivåmessig sammenligning mellom baseline og intervensjonsfase. For det andre kan visuell inspek-

sjon også indikere hvorvidt det har skjedd en endring som en følge av intervensjonen, og sammen med loggføringer fra kartlegging og undervisning generelt bidra til tolkning av data.

Stabilitet

Det ble lagt vekt på å vurdere hvorvidt data for hver deltaker var stabile i baseline-fasen. Utregning av dette fulgte prosedyrene som beskrevet over (Lane & Gast, 2014). Resultatene vise at tre av elevene hadde stabile data i baseline-fasen, ved at 80 prosent av datapunktene var innenfor 25-prosentintervallet av medianverdien i fasen. For elev 7 var data i fasen ikke tilstrekkelig stabile, da to av datapunktene avvek med mer enn 25 prosent fra medianverdien.

Trender

Siden intervensjonen i dette tilfellet handlet om å lære nye strategier i ordavkoding forventet vi ingen umiddelbar effekt ved oppstart av intervensjonen, men en eventuell gradvis progresjon. Kriteriet om umiddelbar endring i helningsgrad ble derfor vurdert irrelevant (Klingbeil mfl., 2017). I trendanalysen ble det derfor primært undersøkt hvorvidt det var uønskede trender i baseline-fasen, siden dette kan svekke validiteten av vurderingene (Gast & Ledford, 2014). Som støtte i vurderingen ble trendlinjen lagt til ved hjelp av Microsoft Excel (figur 2). Stabilitet rundt trendlinjen ble så vurdert gjennom å regne ut korrelasjonen mellom tidsvariabelen og avhengig variabel ved hjelp av Excel. Denne metoden forutsetter som nevnt at data er lineære, noe som bare delvis er tilfelle i dette eksemplet. Inspeksjonen avdekket en tydelig, men ikke lineær trend i baseline-fasen for elev 7, mens trendene for de øvrige var enten nøytral (elev 2 & 3) eller svakt positiv (elev 10). Stabilitet rundt trendlinjen ble målt med R2 (figur 2). Her viser resultatene en moderat stabilitet for tre av elevene (2, 3 og 7) og en stabil lineær trend for elev 10.

For øvrig sees en viss endring i data assosiert med tidspunkt for oppstart av intervensjonen, og for elev 3, 7 og 10 en svak nedgang i postfasen. For elev 7 sees også en markant nedgang etter seks ukers intervensjon, assosiert med avbrekk i treningen som følge av juleferie (figur 1). For elev 2 sees generelt en større variasjon i data i intervensjons- og postfasen sammenlignet med baseline.

Nivåendringer

Nivåendringer ble vurdert gjennom utregning av gjennomsnittsskår for hver fase både når det gjaldt råskåre, og når det gjaldt målet for i hvilken grad elevene utviklet seg positivt i retning av å tilegne seg en mer nøyaktig lesestrategi. Som vist i Figur 3 er det for tre av elevene en positiv endring i form av at prosentandelen av ord som blir lest riktig, øker i intervensjonsfasen sammenlignet med baseline. Dette gjelder også for råskårene, om enn i mindre tydelig grad. For to av elevene (elev 7 og elev 10) opprettholdes endringen også i postfasen av studien, mens elev 2 viser en svak nedgang og elev 3 en tydelig reduksjon i postfasen. Resultatene for elev 3 må generelt

Merk:

Rød linje: Baselinefase. **Grå linje:** Intervensjonsfase. **Lys rosa linje:** Postfase. Grafikken gir et inntrykk av sammenhengen av datapunkter i de tre fasene uten at opphold mellom målingene er visualisert (Se figur 1 for korrekt oversikt over tidsplan for kartlegging for hver elev). **Stiplet linje:** Lineær representasjon av trendlinjen basert på data i baselinefasen. **Alle effektmål:** Utregningene er basert på samlet utvikling i intervensjonsfasen fra og med uke tre samt postfasen.

SMD = Standardized Mean Difference
LRR = Log Response Ratio – Increased

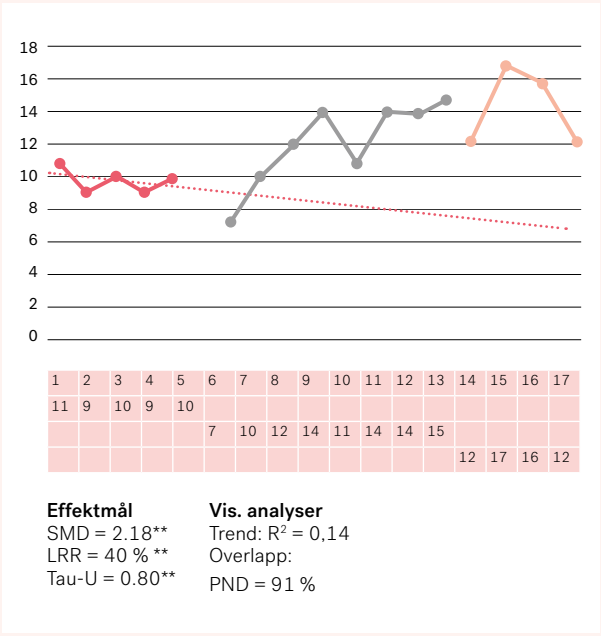
Utregninger av SMD og LRR:
https://jepusto.shinyapps.io/SCD-effect-sizes/

Utregninger av Tau-U:
http://www.singlecaseresearch.org/calculators

Utregning av BC-SMD:
https://jepusto.shinyapps.io/scdhlm

** = p < 0,05, n.s. = p >0,05
(t) = Kontrollert for baseline trend

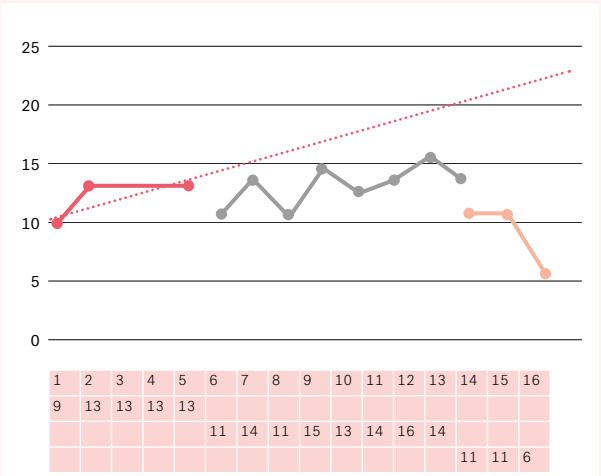
FIGUR 2 Visuell framstilling av resultater i tre faser (ett minutts non-ord lesing) for fire elever.



Effektmål
SMD = 2.18**
LRR = 40 % **
Tau-U = 0.80**

Vis. analyser
Trend: R^2 = 0,14
Overlapp:
PND = 91 %

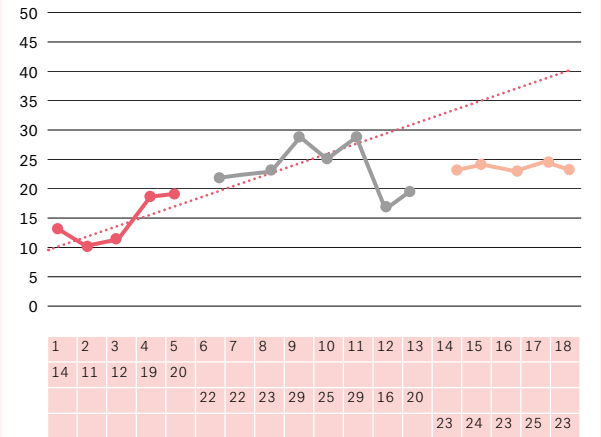
Elev 2



Effektmål
SMD = 0,01 n.s.
LRR = 1 % n.s.
Tau-U = 0,15 n.s.

Vis. analyser
Trend: R^2 = 0,50
Overlapp:
PND = 62%

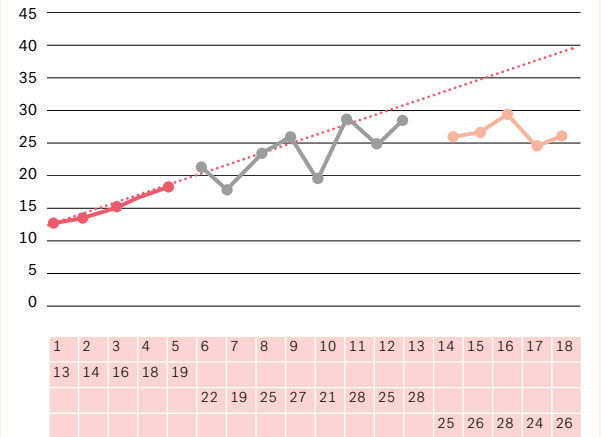
Elev 3



Effektmål
SMD = 2,10**
LRR = 55%**
Tau-U = 0.90**

Vis. analyser
Trend: R^2 = 0,59
Overlapp:
PND = 62%

Elev 7

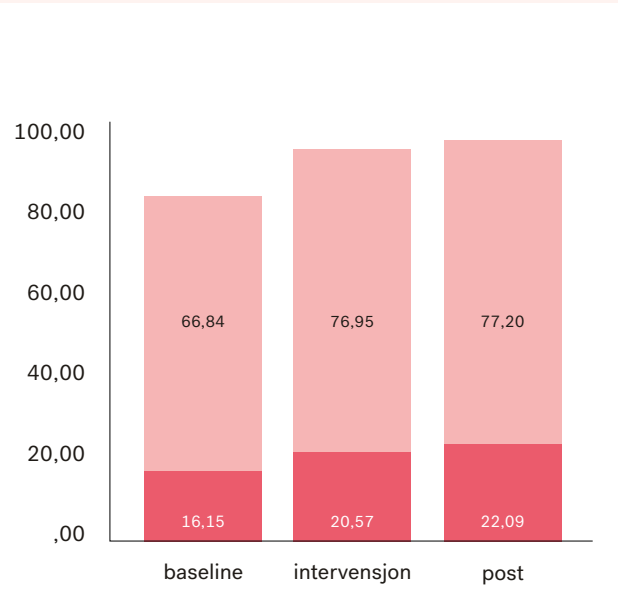


Effektmål
SMD = 4,27**
LRR = 63%**
Tau-U = 0.83* (t)

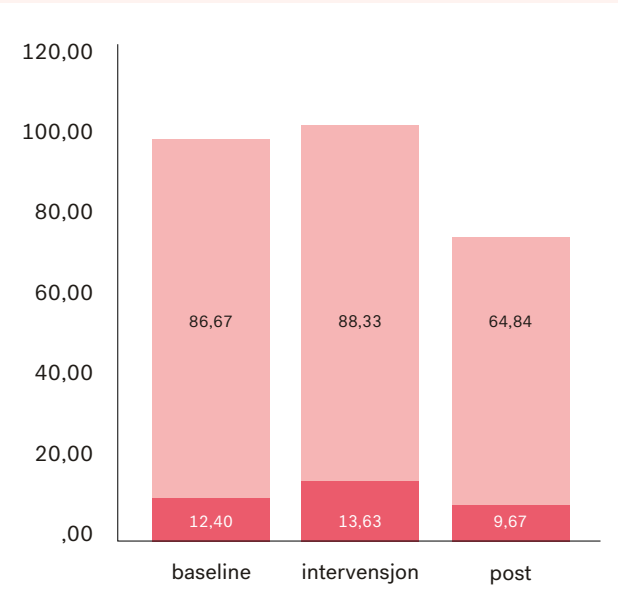
Vis. analyser
Trend: R^2= 0,98
Overlapp:
PND = 87%

Elev 10

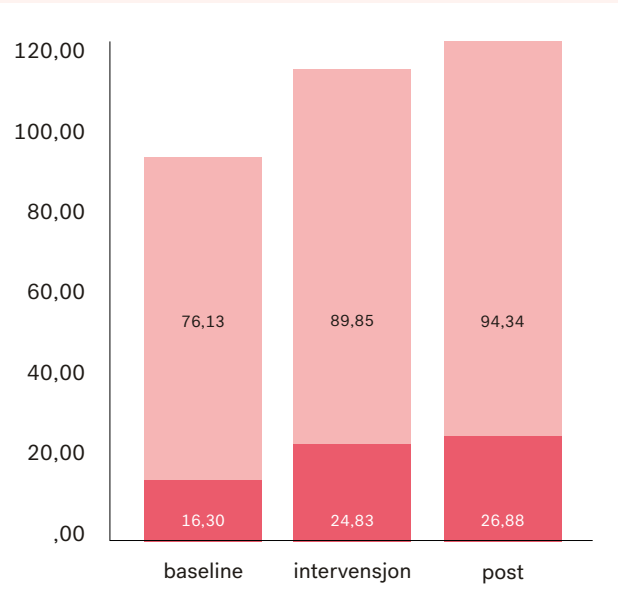
FIGUR 3 Endring i gjennomsnittsskår gjennom tre faser av studien for fire elever – ett minutts lesing av non ord.



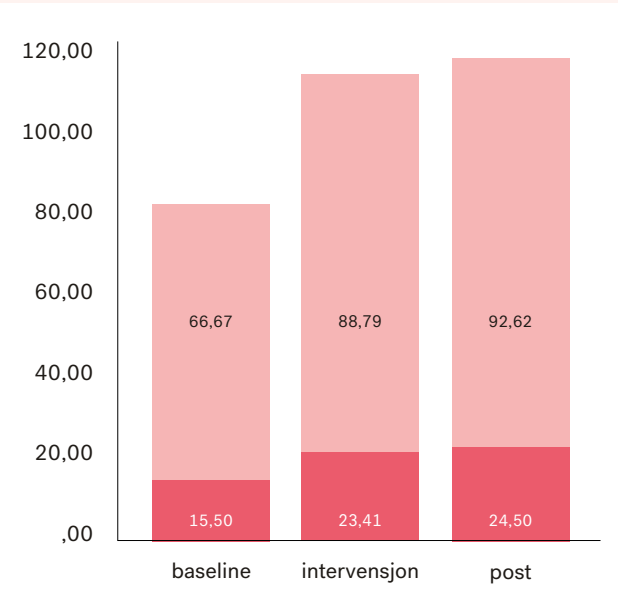
Elev 2



Elev 3



Elev 7



Elev 10

Merk: Rød kolonne = Råskåre. Lys rød kolonne = Prosentvis antall riktige pr antall leste ord.

sees i lys av at denne eleven opplevde store endringer i livssituasjon i løpet av intervensjonsperioden, noe som i henhold til informasjon fra lærer gjorde eleven mindre mottakelig for treningen.

Overlapp

For hver elev ble det telt opp antall datapunkter som økte i henhold til den høyeste verdien i baseline-fasen (figur 2). Prosentandelen av datapunkter i intervensjonsfasen, som ikke overlapper (PND), er vist i figur 2.

Statistiske analyser

For å vurdere hvorvidt den observerte endringen i intervensjonsfasen var signifikant og for å vurdere størrelsen av eventuell effekt for hver elev brukte vi to parametriske mål, SMD og LRR, og et ikke-parametrisk mål, Tau-U. For vurdering av samlet deltakereffekt brukte vi vektet Tau-U og BC-SMD. Alle utregninger av effekt ble basert på råskårer. Siden effekt av tiltaket innebar å lære å bruke et ukjent treningsmateriale, beregnet vi, ut fra tidligere erfaring med treningsmaterialet, en treningsfase på to uker ved oppstart og inkluderte derfor kun data fra og med uke tre av intervensjonen. LRR og SMD ble regnet ut ved hjelp av den nettbaserte kalkulatoren (Pustejovsky & Swan 2018).

Vi valgte å konvertere effektmålet fra LRR til prosentvis økning, og i utregningen av SMD ble det brukt samlet (pooled) standardavvik. Resultater er vist i figur 2. Tau U ble regnet ut med den nettbaserte kalkulatoren (Vannest, Parker, Gonen & Adiguzel, 2016). Vi vurderte først hvorvidt det var signifikante trender i baseline-fasen. Til tross for at visuell inspeksjon viste at elev 7 hadde en tydelig oppadgående trend, var denne trenden ikke signifikant på 95-prosentnivå (p = 0,14). Trenden for elev 10 var derimot signifikant, og det ble derfor kontrollert for baseline-trend i utregning av effektestimatet (Parker mfl., 2011) (figur 2).

Samlet effekt for deltakerne basert på vektet gjennomsnittlig effektstørrelse med Tau-U gir et effektestimat på 0,72 (Konfidensintervall (95 %) = 0,40–1,0). Dette regnes i henhold til den foreslåtte standarden (se over) som stor effekt.

Samlet effekt ble også målt med BC-SMD. Siden RMLE-estimatoren er regnet som mest fleksibel i vurdering av data i små og heterogene utvalg, valgte vi denne (Valentine mfl., 2016). Utregningsmetoden i kalkulatoren krever at det spesifiseres både gjennomsnittsnivå (fixed level) og individuell variasjon (random level) i baseline-fasen, samt gjennomsnittlig effekt (fixed level) i intervensjonsfasen. For videre spesifikasjoner fulgte vi kriteriene foreslått av Wolfe mfl. (2019). I henhold til disse skal det inkluderes spesifikasjon av individuell variasjon i intervensjonsfasen (random level) dersom den gjennomsnittlige intervensjonseffekten mellom deltakerne varierer med mer enn 10 prosent av verdien på y-aksen. I dette tilfellet avviker elev 3 med mer enn 10 prosent fra de andre. For å vurdere om det også skulle spesifiseres trender brukte vi visuell inspeksjon av data som utgangspunkt (Wolfe mfl., ibid.). Siden en av elevene (elev 10) viste en tydelig trend i baseline-fasen, spesifiserte vi random trend i denne fasen. Trendene i intervensjonsfasen var uklare og delvis inkonsistente, og det ble derfor valgt å ikke spesifisere trender i denne fasen. Samlet effekt med BC-SMD viser d = 0,64 med konfidensintervall -8.81–10.11. Dette resultatet er ikke signifikant.

Oppsummering av resultater

Kort oppsummert kan man si at de visuelle analysene angir hvorvidt individuelle data er egnet til å vurdere effekt, og også om det gir et overblikk over fluktuasjoner i data som sammen med logg fra lærer kan avdekke om variasjonen er tilfeldig eller systematisk. De individuelle effektestimatene kan brukes til å vurdere effektstørrelsen hos en gitt elev, mens det samlede effektmålet indikerer eventuell videre utprøving på nye elever. Dersom effektvurdering av tiltaket skulle gjøres på bakgrunn av visuelle analyser alene, ville trolig elev 7 bli tatt ut siden data i baseline-fasen er ustabile. Til en viss grad kan imidlertid de statistiske analysene kompensere for

ustabile data og gir generelt en mindre restriktiv vurdering enn visuelle analyser (Wolfe mfl., 2019).

De gjennomgående positive resultatene for elev 7 på tvers av de ulike effektmålene tydeliggjør dermed at statistiske analyser kan fange opp effekt som ikke framkommer visuelt (ibid.).

Effektmålene synes å samsvare i henhold til vurdert effekt hos tre av elevene (elev 2, 7 og 10), mens for elev 3 er endringen marginal og ikke signifikant. Det samlede målet (BC SMD) viser også et ikke signifikant resultat med stort konfidensintervall. Det samlede resultatet med Tau-U viser imidlertid et signifikant resultat, noe som eksemplifiserer at vurderingen basert på overlapp i dette tilfellet gir en mindre konservativ vurdering enn effektmålet basert på flernivåanalyse.

Utregning av samlet effekt med BC-SMD kan, ifølge veilederen (Valentine mfl., 2016), i prinsippet gjøres med multiple baseline design med så få som 3–4 deltakere. Tidligere simuleringsstudier med flernivåanalyser i SCD har imidlertid indikert at det oppnås usikre resultater med så få deltakere, og at det er nødvendig med 6–8 deltakere for å minske feilmarginene (Ferron mfl., 2009). Siden vi i dette eksemplet har valgt ut enkeltdeltakere fra den originale studien, primært for å vise kontraster i oppnådd effekt, gir utregning av samlet deltakereffekt med disse fire elevene et upresist mål med stort konfidensintervall og dermed ikke signifikante resultater.

Vi har likevel valgt å eksemplifisere bruk av BC-SMD her fordi vi antar at dette effektmålet kan ha positive kvaliteter når det gjelder mulighet til å vurdere effekt av konkrete tiltak på tvers av deltakere (se nedenfor). I den originale studien (Thurmann-Moe, Melby-Lervåg & Lervåg) eksemplifiseres bruk av BC-SMD med 11 deltakere basert på de samme kriteriene for utregning som vist her. Siden BC-SMD er et relativt nylig utviklet verktøy for SCD, er både kriterier for bruk og retningslinjer når det gjelder spesifisering av modeller, under utvikling. Vi har her valgt å bruke kriterier som har vist seg å ha stor grad av samsvar med visuelle analyser og individuelle effektmål (Wolfe mfl., 2019), men også andre kriterier er tilgjengelige (se f.eks. WWC, (2020) for alternative innfallsvinkler).

Diskusjon

Vi har i denne artikkelen drøftet og eksemplifisert hvordan SCD kan brukes som verktøy i evaluering av spesialpedagogiske tiltak. Vi har også diskutert utfordringer knyttet til validitet og presentert forskjellige innfallsvinkler for vurdering av effekt. Eksemplet viser hvordan det er mulig å designe en SCD-studie innenfor rammen av spesialpedagogiske tiltak i skolen. I eksemplet brukes repeterte tester som utgangspunkt for vurdering av effekt. Også andre målemetoder er mulige for å vurdere utbytte av intervensjoner rettet mot å øke akademiske ferdigheter. I en nylig publisert studie som evaluerte effekten av et tiltak med sikte på å øke elevenes bruk av relevante planleggingsstrategier i tekstproduksjon («The STOP and LIST strategy») (Grünke, Nobel & Bracht, 2020), ble det brukt en mer kvalitativ innfallsvinkel ved at utvalg av elevtekster ble analysert og kodet etter bestemte kriterier både i baseline-fasen og intervensjonsfasen. Resultatene ble så sammenlignet og effekt vurdert ut fra differansen mellom tekstene i de to fasene ut fra de målte kriteriene.

Dersom tiltak rettes mot atferdsendring, er design som bruker direkte observasjon en egnet innfallsvinkel. Da kan både reversible design, der tiltaket «skrus» av og på gjennom gjentatte faser, og multiple baseline design brukes. Ved multiple baseline design med bruk av atferdsobservasjon som målemetode kan for eksempel den uavhengige variabelen introduseres som tiltak først i et skolefag og deretter i de neste fagene suksessivt, og effekt vurderes ut fra hvorvidt det forekommer en systematisk økning eller reduksjon av observert atferd assosiert med tiltaket (Gast mfl., 2014).

Som vist i eksemplet kan SCD med relativt enkle grep tilpasses den situasjonen man arbeider i. For å undersøke om et tiltak virker for en elev i en konkret undervisningssituasjon, vil det være tilstrekkelig å bruke et enkelt tofasedesign (AB) der man gjør et visst antall målinger (minst tre, men ideelt sett fem) før man starter tiltaket, og så fortsetter målingene etter at tiltaket har startet. Resultatet vil da kunne si noe om tiltaket virker i akkurat denne situasjonen, men vil ikke ha gyldighet for

andre fag, i andre klasserom, med andre lærere eller for andre elever. Likevel innebærer en slik metode en systematisk måling av atferd som gir et grunnlag for vurdering av om og hvordan tiltaket har virket på eleven, og på om tiltaket bør fortsette eller erstattes med et annet tiltak (Gast & Ledford, 2014).

Flere av det nettbaserte kalkulatorene som er tilgjengelige, baserer også utregning av effekt på tofasedesign, slik at det selv med et enkelt AB-design er mulig å regne ut effektstørrelse for en elev, slik det er vist i eksemplet over. Selv om AB-designet ikke er et fullverdig forskningsdesign, kan likevel kvantifisering av effekt og bruk av veiledende normer for vurdering av effektstørrelse være nyttig i en klinisk sammenheng og supplere læreres subjektive vurderinger av hva som virker for eleven. Ikke minst vil bruk av kvantitative effektmål være egnet til å sammenligne effekt av ulike tiltak på samme elev.

Effektmålinger i SCD har også den fordelen at de relateres til individuell utvikling ut fra elevens eget utgangspunkt på et definert utviklingsfelt. Denne fleksibiliteten er særlig relevant i spesialpedagogisk virksomhet der man ofte arbeider med konkrete delferdigheter innenfor et fagområde. Fagbaserte måleverktøy som er standardisert på klassetrinnet, for eksempel kartleggingsprøver for lesing og matematikk, vil på sin side sjelden være detaljerte nok til å fange opp endringer hos en elev som følge av et avgrenset spesialpedagogisk tiltak. Individuelle effektmål av denne typen er også egnet som grunnlag for årlige rapporteringer og også for samarbeid med elev og foresatte, samt eventuelle andre aktuelle faginstanser rundt eleven.

For PPT vil det å motivere for bruk av SCD i det spesialpedagogiske arbeidet på skolene eller i barnehagene også danne utgangspunkt for å kunne forske i egen praksis. Ved å sikre at kravene til eksperimentell kontroll følges, for eksempel ved å følge retningslinjene presentert i konsensusrapporten fra SCRIBE (Tate et.al 2016), og ved å replisere resultater på flere elever, vil PPT også kunne bidra til utvikling av forskningsbasert kunnskap. Dette vil både være relevant som grunnlag for videre rådgivning på feltet og samtidig kunne bidra til å knytte praksisfeltet tettere til forskningsmiljøene. Statistiske analyseverktøy som gjør det mulig å vurdere samlet effekt basert på replikasjoner, vil her trolig kunne være nyttige verktøy (se f.eks. Pustejovsky, 2016).

Erfaring viser også at småskalaforskning av denne typen kan ha positive ringvirkninger i miljøet der de settes i gang, og føre til økt faglig engasjement i virksomheten (Dexter & Seden, 2012). I en norsk kontekst vil bruk av SCD i evaluering av spesialpedagogiske tiltak trolig kunne bidra både til økt engasjement i planlegging og gjennomføring av enkeltvedtak og indirekte også fungere som et styringsverktøy for økt kvalitet i alle ledd i tiltakskjeden.

Begrensninger

Vi har i denne artikkelen beskrevet grunnleggende trekk ved SCD-metodologien og diskutert validitet knyttet til design, effektmål og analysemetoder. En viktig begrensning ved denne gjennomgangen er selvsagt at den kun omhandler et utvalg metoder og problemstillinger. Utvalget gjengir heller ikke den historiske utviklingen på området. Metodeutviklingen innen SCD er i stadig endring, og gamle metoder nyanseres og videreutvikles parallelt med at tradisjonelle metoder relanseres og metodiske nyvinninger testes ut og diskuteres (Shadish, 2014). Når det gjelder analysemetodenes validitet, er det noen verktøy som i litteraturen er hyppigere diskutert enn andre, og det preger også denne framstillingen. Det kan gi et skjevt bilde av at metoder som er oftest omtalt og dermed oftest kritisert, også er de som har svakest validitet og kvalitet. Dette er ikke nødvendigvis tilfelle. □

REFERANSER

BAER, D.M., WOLF, M.M. & RISLEY, T.R. (1968). Some current dimensions of applied behavior analysis 1. *Journal of Applied Behavior Analysis, 1*(1), s. 91–97. doi: *10.1901/jaba.1968.1-91*

BARNEOMBUDET (2017). *Uten mål og mening.* Hentet fra: *http://barneombudet.no/wp-content/uploads/2017/03/Bo_rapport_enkeltsider.pdf*

BUSK, P.L. & SERLIN, R.C. (1992). Meta-analysis for single-case research. I: T.R. Kratochwill & J.R. Levin (Red.), *Single-case research design and analysis: New directions for psychology and education* (s. 187–212). Lawrence Erlbaum Associates, Inc.

BUSSE, R., MCGILL, R. & KENNEDY, K. (2015). Methods for Assessing Single-Case School Based Intervention Outcomes. *The Official Journal of the California Association of School Psychologists, 19*(3), s. 136–144. doi: *10.1007/s40688-014-0025-7*

COHEN, J. (1988). *Statistical power analysis for the behavioral sciences* (2. utg.). Hillsdale, N. J: Laurence Erlbaum.

DE BRUIN, C.L. (2017). Conceptualizing effectiveness in disability research. *International Journal of Research & Method in Education, 40*(2), s. *113–136. doi:10.1080/1743727X.2015.1033391*

DEXTER, B. & SEDEN, R. (2012). 'It's really making a difference': how small-scale research projects can enhance teaching and learning. *Innovations in Education and Teaching International, 49*(1), s. 83–93. doi: *10.1080/14703297.2012.647786*

FERRON, J.M. BELL, B.A, HESS, M.R. RENDINA-GOBIOFF, G. & HIBBARD, S.T. (2009). Making. treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods, 41*, s. 372–384. Hentet fra: https://doi.org/10.3758/BRM.41.2.372

GAST, D. & BAEKEY, D.H. (2014). Withdrawal and Reversal Designs. I: D.L. Gast & J. R. Ledford, *Single case research methodology: applications in special education and behavioral sciences* (s. 211–251) (2. utg.). New York: Routledge.

GAST, D. & SPRIGGS, A.M. (2014). Visual Analysis of Graphic Data. I: D.L. Gast, & J. R. Ledford, *Single case research methodology: applications in special education and behavioral sciences* (s. 176–211) (2. utg.). New York: Routledge.

GAST, D.L. & LEDFORD, J.R. (2014). *Single case research methodology: applications in special education and behavioral sciences* (s. 176–211) (2. utg.). New York: Routledge.

GAST, D. (2014). General Factors in Measurement and Evaluation. I: D.L. Gast & J. R. Ledford, *Single case research methodology: applications in special education and behavioral sciences* (2. utg.). New York: Routledge, s. 85–105.

GAST, D., LLOYD, B.P. & LEDFORD, J. (2014). Multiple Baseline and Multiple Probe Designs. I: D.L. Gast & J. R., *Single case research methodology: applications in special education and behavioral sciences* (s. 251–297) (2. utg.). New York: Routledge.

GRÜNKE, M. & NOBEL, K. & BRACHT, J. (2019). Effects of the STOP and LIST Strategy on the Writing Performance of Struggling Fourth Graders. *Insights into Learning Disabilities 17*(1), s. 73–85. Hentet fra: *https://files.eric.ed.gov/fulltext/EJ1258308.pdf*

HARRINGTON, M. & VELICER, W.F. (2015). Comparing Visual and Statistical Analysis in Single-Case Studies Using Published Studies. *Multivariate Behavioral Research, 50*(2), s. 162–183. doi: *10.1080/00273171.2014.973989*

HAUG, P. (2019). *Spesialundervisning: innhald og funksjon.* Samlaget, Oslo.

HEDGES, L.V. (2008). What Are Effect Sizes and Why Do We Need Them? *Child Development Perspectives, 2*(3), s. 167–171. doi: *10.1111/j.1750-8606.2008.00060.x*

HEDGES, L.V., PUSTEJOVSKY, J.E. & SHADISH, W.R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods, 3*(3), s. 224–239. doi: *10.1002/jrsm.1052*

HORNER, R.D. & BAER, D.M. (1978). Multiple probe technique: a variation Of the multiple baseline 1. *Journal of Applied Behavior Analysis, 11*(1), s. 189–196. doi: *10.1901/jaba.1978.11-189*

HORNER, R.H., CARR, E.G., HALLE, J., MCGEE, G., ODOM, S. & WOLERY, M. (2005). The Use of Single-Subject Research to Identify Evidence-Based Practice in Special Education. *Exceptional Children, 71*(2), s. 165–179. doi: *10.1177/001440290507100203*

KAZDIN, A.E. (2016). *Methodological issues & strategies in clinical research* (4. utg.). Washington, DC: American Psychological Association.

KAZDIN, A.E. (2011). Evidence-based treatment research: Advances, limitations, and next steps. *The American psychologist, 66*(8), s. 685–698. doi: *10.1037/a0024975*

KLINGBEIL, D., NORMAN, E. & NELSON, P. (2017). Precision of Curriculum-Based Measurement Reading Data: Considerations for Multiple-Baseline Designs. *Journal of Behavioral Education, 26*(4), s. 433–451. doi: *10.1007/s10864-017-9282-7*

KRATOCHWILL, T.R. & LEVIN, J.R. (2010). Enhancing the Scientific Credibility of Single-Case Intervention Research: Randomization to the Rescue. *Psychological Methods, 15*(2), s. 124–144. doi: *10.1037/a0017736*

KRASNY-PACINI, A. & EVANS, J. (2018). Single-case experimental designs to assess intervention effectiveness in rehabilitation: A practical guide. *Annals of Physical and Rehabilitation Medicine, 61*(3), s. 164–179. Hentet fra: *http://www.sciencedirect.com/science/article/pii/S1877065717304542*. doi: *https://doi.org/10.1016/j.rehab.2017.12.002*

KRATOCHWILL, T.R., HITCHCOCK, J.H., HORNER, R.H., LEVIN, J.R., ODOM, S.L., RINDSKOPF, D. M. & SHADISH, W.R. (2013). Single-Case Intervention Research Design Standards. *Remedial and Special Education, 34*(1), s. 26–38. doi: *10.1177/0741932512452794*

KRATOCHWILL, T.R., HITCHCOCK, J., HORNER, R.H., LEVIN, J.R., ODOM, S.L., RINDSKOPF, D.M. & SHADISH, W.R. (2010). Single-case designs technical documentation. Hentet fra: What Works Clearinghouse website: *http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.*

KVANDE, M.N., BJØRKLUND, O., LYDERSEN, S., BELSKY, J., & WICHSTRØM, L. (2019). Effects of special education on academic achievement and task motivation: a propensity-score and fixed-effects approach. *European Journal of Special Needs Education, 34*(4), s. 409–423. doi: *10.1080/08856257.2018.1533095*

LANE, J.D. & GAST, D.L. (2014). Visual analysis in single case experimental design studies: Brief review and guidelines. *Neuropsychological Rehabilitation*, 24 (3–4), s. 445–463. doi: *10.1080/09602011.2013.815636*

LEDFORD, J.R. (2018). No Randomization? No Problem: Experimental Control and Random Assignment in Single Case Research. *American Journal of Evaluation, 39*(1), s. 71–90. doi: *10.1177/1098214017723110*

LEDFORD, J.R., WOLERY, M. & GAST, D.I. (2014). Controversial and Critical Issues in Single Case Research. I: D.L Gast & J.R. Ledford, *Single case research methodology: applications in special education and behavioral sciences* (2. utg.). New York: Routledge, s. 377–397.

LEKHAL, R. (2017). Elever med vedtak om spesialundervisning: hva vet vi, hvordan har de det, og trives de på skolen? I: P. Haug, (2019). *Spesialundervisning: innhald og funksjon.* s. 368–385. Oslo: Samlaget.

LOBO, A.M., MOEYAERT, A.M., BARALDI CUNHA, A.A. & BABIK, A.I. (2017). Single-Case Design, Analysis, and Quality Assessment for Intervention Research. *Journal of Neurologic Physical Therapy, 41*(3), s. 187–197. doi: *10.1097/NPT.0000000000000187*

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

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

MAGGIN, D.M., LANE, K.L. & PUSTEJOVSKY, J.E. (2017). Introduction to the Special Issue on Single-Case Systematic Reviews and Meta-Analyses. *Remedial and Special Education, 38*(6), s. 323–330. doi: *10.1177/0741932517717043*

MAGGIN, D.M., SWAMINATHAN, H., ROGERS, H.J., AMP, APOS, KEEFFE, B.V., . . . HORNER, R.H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology, 49*(3), s. 301–321. doi: *10.1016/j.jsp.2011.03.004*

MAGGIN, D.M., SWAMINATHAN, H., ROGERS, H.J., O'KEEFFE, B.V., SUGAI, G. & HORNER, R.H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology, 49*(3), s. 301–321. Hentet fra: *http://www.sciencedirect.com/science/article/pii/S0022440511000203.*

MANOLOV, R. & VANNEST, K.J.A. (2019). A Visual Aid and Objective Rule Encompassing the Data Features of Visual Anlysis. *Behavior Modification, 0*(0). Hentet fra: *https://journals.sagepub.com/doi/abs/10.1177/0145445519854323.* doi: *10.1177/0145445519854323*

MANOLOV, R., SOLANAS, A. & LEIVA, D. (2010). Comparing "Visual" Effect Size Indices for Single-Case Designs. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 6*(2), s. 49–58. doi: *10.1027/1614-2241/a000006*

MANOLOV, R., SOLANAS, A., & SIERRA, V. (2018). Extrapolating baseline trend in single-case data: Problems and tentative solutions. Behavior Research Methods, doi: *10.3758/s13428-018-1165-x*

MELD. ST. (2016–2017). Lærelyst – tidlig innsats og kvalitet i skolen. Hentet fra: *https://www.regjeringen.no/no/dokumenter/meld.-st.-2120162017/id2544344/*

MELD. ST. (2019–2020). Tett på – tidlig innsats og inkluderende fellesskap i barnehage, skole og SFO. Hentet fra: *https://www.regjeringen.no/no/dokumenter/meld.-st.-6-20192020/id2677025/*

NORDAHL, T. MFL. (2018). *Inkluderende fellesskap for barn og unge.* Fagbokforlaget. Hentet fra: *https://nettsteder.regjeringen.no/inkludering-barn-unge/files/2018/04/INKLUDERENDE-FELLESSKAP-FOR-BARN-OG-UNGE-til-publisering-04.04.18.pdf*

NOU (2019: 3). Nye sjanser – bedre læring – Kjønnsforskjeller i skoleprestasjoner og utdanningsløp. Regjeringen/Kunnskapsdepartementet. Hentet fra: *https://www.regjeringen.no/no/dokumenter/nou-2019-3/id2627718/sec1*

ODOM, S.L., BRANTLINGER, E., GERSTEN, R., HORNER, R.H., THOMPSON, B. & HARRIS, K.R. (2005). Research in Special Education: Scientific Methods and Evidence-Based Practices. *Exceptional Children, 71*(2), s. 137–148. doi: *10.1177/001440290507100201*

OLIVE, M.L. & SMITH, B.W. (2005). Effect Size Calculations and Single Subject Designs. *Educational Psychology, 25*, s. 313–324.

ONGHENA, P., MICHIELS, B., JAMSHIDI, L., MOEYAERT, M. & VAN DEN NOORTGATE, W. (2018). One by One: Accumulating Evidence by using Meta-Analytical Procedures for Single-Case Experiments. *Brain Impairment, 19*(1), s. 33–58. doi: *10.1017/BrImp.2017.25*

OPSVIK, F. & HAUG, P. (2017). Læringsutbyttet i matematikk. I: P. Haug, (2019). *Spesialundervisning: innhald og funksjon.* (s. 324–349). Oslo: Samlaget.

PARKER, R.I., VANNEST, K.J. & BROWN, L. (2009). The Improvement Rate Difference for Single-Case Research. *Exceptional Children, 75*(2), s. 135–150. doi: *10.1177/001440290907500201*

PARKER, R.I., VANNEST, K.J. & DAVIS, J.L. (2011). Effect Size in Single-Case Research: A Review of Nine Nonoverlap Techniques. *Behavior Modification, 35*(4), s. 303–322.

PARKER, R.I., VANNEST, K.J., DAVIS, J.L. & SAUBER, S.B. (2011). Combining Nonoverlap and Trend for Single-Case Research: Tau-U. *Behavior Therapy, 42*(2), s. 284–299. doi: *10.1016/j.beth.2010.08.006*

PENNINGTON, B.F. & BISHOP, D.V.M. (2009). Relations among speech, language, and reading disorders. *Annual Review of Psychology, 60*(1) s. 283–306.

PUSTEJOVSKY, J.E. (2016). scdhlm: A web-based calculator for between-case standardized mean differences (Version 0.3.1) [Web application]. Hentet fra: *https://jepusto.shinyapps.io/scdhlm*

PUSTEJOVSKY, J. (2018). Using response ratios for meta-analyzing single-case designs with behavioral outcomes. *Journal of School Psychology, 68*, s. 99–112. doi: *10.1016/j.jsp.2018.02.003*

PUSTEJOVSKY, J.E. (2019). Procedural sensitivities of effect sizes for single-case designs with directly observed behavioral outcome measures. *Psychological Methods, 24*(2), s. 217–235. *https://doi.org/10.1037/met0000179*

PUSTEJOVSKY, J.E. & FERRON, J.M. (2017). Research synthesis and meta-analysis of single-case designs. I: J.M. Kaufmann, D.P. Hallahan, & P.C. Pullen (Red.). Handbook of Special Education (2.utg.) New York, NY: Routledge.

PUSTEJOVSKY, J.E., HEDGES, L.V. & SHADISH, W.R. (2014). Design-Comparable effect sizes in multiple baseline designs: A general modeling framework. *Journal of Educational and Behavioral Statistics, 39*(5), s. 368–393. doi: *10.3102/1076998614547577*

PUSTEJOVSKY, J.E. & SWAN, D.M. (2018). Single-case effect size calculator (Version 0.5) Web application. Hentet fra: *https://jepusto.shinyapps.io/SCD-effect-sizes/*

PUSTEJOVSKY, J.E., SWAN, D.M., ENGLISH, K.W. & PUSTEJOVSKY, J.E. (2019). An Examination of Measurement Procedures and Characteristics of Baseline Outcome Data in Single-Case Research. *Behavior Modification*, doi: *10.1177/0145445519864264*

RAKAP, S. (2015), Effect sizes as result interpretation aids in single-subject experimental research: description and application of four non-overlap methods. *British Journal of Special Education, 42*, s. 11–33. doi: *10.1111/1467-8578.12091*

SCRUGGS, T.E. & MASTROPIERI, M.A. (1998). Summarizing Single-Subject Research: Issues and Applications. *Behavior Modification, 22*(3), s. 221–242. Hentet fra: *https://doi.org/10.1177%2F01454455980223001*

SCRUGGS, T.E., MASTROPIERI, M.A. & CASTO, G. (1987). The Quantitative Synthesis of Single-Subject Research: Methodology and Validation. *Remedial and Special Education, 8*(2), s. 24–33. doi: *10.1177/074193258700800206*

SHADISH, W.R. (2014). Statistical Analyses of Single-Case Designs: The Shape of Things to Come. *Current Directions in Psychological Science, 23*(2), s. 139–146. doi: *10.1177/0963721414524773*

SHADISH, W.R., COOK, T.D. & CAMPBELL, D.T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin.

SHADISH, W.R., HEDGES, L.V., & PUSTEJOVSKY, J.E. (2013). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of school psychology, 52*(2), s. 123–147. doi: *10.1016/j.jsp.2013.11.005*

SHADISH, W.R., HEDGES, L.V., HORNER, R.H. & ODOM, S.L. (2015). *The Role of Between-Case Effect Size in Conducting, Interpreting, and Summarizing Single-Case Research.* Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

SHADISH, W. & SULLIVAN, K. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43*(4), s. 971–980. doi: *10.3758/s13428-011-0111-y*

SHAMSEER, L., SAMPSON, M., BUKUTU, C., SCHMID, C.H., NIKLES, J., TATE, R., . . . VOHRA, S. (2015). CONSORT extension for reporting N-of-1 trials (CENT): Explanation and elaboration. British Medical Journal Publishing Group. doi: *10.1136/bmj.h1793*.

SKORPEN, L.B. (2017). Elevar med matematikkvanskar og deira utvikling i løpet av eit år. I: P. Haug (2019). *Spesialundervisning: innhald og funksjon*, s. 296–324. Oslo: Samlaget.

SWAMINATHAN, H., ROGERS, H.J., HORNER, R.H., SUGAI, G. & SMOLKOWSKI, K. (2014). Regression models and effect size measures for single case designs. *Neuropsychological Rehabilitation, 24* (3–4), s. 1–18. doi: *10.1080/09602011.2014.887586*

TARLOW, K.R. (2016). Baseline Corrected Tau-U Calculator. Hentet fra: *http://www.ktarlow.com/stats/tau/*

TARLOW, K.R. (2017). An improved rank correlation effect size statistic for single-case designs: Baseline Corrected Tau. Behavior Modification, 41(4), s. 427–467. Hentet fra: *http://dx.doi.org/10.1177/0145445516676750*

TATE, R.L., PERDICES, M., ROSENKOETTER, U., SHADISH, W., VOHRA, S., BARLOW, D.H., ... WILSON, B. (2016). The Single-Case Reporting guideline in behavioural interventions (SCRIBE). Journal of School Psychology, 56, s. 133–142. doi: *10.1016/j.jsp.2016.04.001*

THURMANN-MOE, A.C., MELBY-LERVÅG, M & LERVÅG, A. The Impact of Articulatory Consciousness Training on Reading and Spelling Literacy in Students with Severe Dyslexia: An Experimental Single Case Study. *Upublisert artikkel under fagfellevurdering*

VALENTINE, J.C., TANNER-SMITH, E.E., PUSTEJOVSKY, J.E., & LAU, T.S. (2016). Between-case standardized mean difference effect sizes for single-case designs: a primer and tutorial using the scdhlm web application. *Campbell Systematic Reviews, 12*(1), s. 1–31. doi: *10.4073/cmdp.2016.1*

VANNEST, K.J. & NINCI, J. (2015). Evaluating Intervention Effects in Single-Case Research Designs. *Journal of Counseling & Development, 93*(4), s. 403–411. doi: *10.1002/jcad.12038*

VANNEST, K.J., PARKER, R.I., GONEN, O. & ADIGUZEL, T. (2016). Single Case Research: Web based calculators for SCR analysis. (Version 2.0) [Web-based application]. Texas: A & M University, College Station.

WWC (What Works Clearinghouse) (2017). *Procedures Handbook Version 4*. Hentet fra: *https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_handbook_v4.pdf*

WWC (What Works Clearinghouse) (2020) *Current standards version 4.1. – summary of changes.* Hentet fra: *https://ies.ed.gov/ncee/wwc/Handbooks*

ZIMMERMAN, K.N., LEDFORD, J R., SEVERINI, K.E., PUSTEJOVSKY, J.E., BARTON, E.E. & LLOYD, B.P. (2018a). Single-case synthesis tools I: Comparing tools to evaluate SCD quality and rigor. *Research in Developmental Disabilities, 79*, s. 19–32.

ZIMMERMAN, K.N., PUSTEJOVSKY, J.E., LEDFORD, J.R., BARTON, E.E., SEVERINI, K.E. & LLOYD, B.P. (2018b). Single-case synthesis tools II: Comparing quantitative outcome measures. *Research in Developmental Disabilities, 79*, s. 65–76. doi: *https://doi.org/10.1016/j.ridd.2018.02.001*