Sequential meta-analysis of single-case experimental data

Sofie Kuppens · Mieke Heyvaert ·
Wim Van den Noortgate · Patrick Onghena

Published online: 1 April 2011
© Psychonomic Society, Inc. 2011

Abstract In this article, sequential meta-analysis is presented as a method for determining the sufficiency of cumulative knowledge in single-case research synthesis. Sufficiency addresses the question of whether there is enough cumulative knowledge on a topic to yield convincing statistical evidence. The method combines cumulative meta-analysis of single-case experimental data with formal sequential testing. After describing the underlying statistical techniques, a strategy for conducting a sequential single-case meta-analysis is illustrated using a real meta-analytic database. The sequential methodology may serve as a valuable tool for behavioral researchers to guide them in making optimal use of limited resources.

Keywords Meta-analysis · Sequential testing · Single-case · Cumulative meta-analysis · Group sequential boundaries · Sequential meta-analysis

Single-case experimental designs have a long tradition in behavioral research for testing the effectiveness of behavioral interventions. These kinds of designs are characterized by repeated observations of the *case* (usually a human participant) over different levels of at least one manipulated independent variable (Onghena, 2005). The experimental nature of single-case designs and the possibility of testing causal effects is particularly appealing, but statements about a treatment cannot be generalized on statistical grounds beyond the specific case.

In order to explore the generalizability of results from single-case experiments, meta-analytic techniques have been proposed for aggregating results over multiple entities (e.g., Busk & Serlin, 1992; Van den Noortgate & Onghena, 2003). Although highly informative for obtaining an estimate of an overall treatment effect or for identifying case or study characteristics that moderate a treatment effect, these meta-analytic techniques fail to answer questions pertaining to the sufficiency of cumulative knowledge. *Sufficiency* addresses the question of whether there is enough cumulative knowledge on the same topic to yield convincing statistical evidence or, in other words, whether there are enough pieces to unravel the puzzle. Information on whether or not sufficiency has been attained could make a unique and valuable contribution to a research field for two reasons. First, the decision to initiate a new study should depend on the expected added value of such a study to the existing knowledge base. Second, in order to guide evidence-based practices, it is vital to identify the benefit or failure of a treatment as early as possible.

Cumulative meta-analysis using group sequential boundaries — or sequential meta-analysis (SMA), for short — has been proposed as a valuable tool for gauging sufficiency when large-scale studies are synthesized (Pogue & Yusuf, 1997; Wetterslev, Thorlund, & Gluud, 2008), but its usefulness in single-case research synthesis remains unexplored. In this article, we extend the sequential meta-analytic approach to determine sufficiency when aggregating single-case experimental results. In doing so, we hope to stimulate behavioral researchers to use this method in order to decide whether sufficient cumulative knowledge has already been obtained to render future studies redundant and to guide evidence-based practices.

S. Kuppens (✉) · M. Heyvaert · W. Van den Noortgate ·
P. Onghena
Methodology of Educational Sciences Research Group,
Faculty of Psychology and Educational Sciences,
Katholieke Universiteit Leuven,
Andreas Vesaliusstraat 2, P.O. Box 3762, 3000, Leuven, Belgium
e-mail: Sofie.Kuppens@ped.kuleuven.be

Sequential meta-analysis

Sequential meta-analysis combines the methodology of cumulative meta-analysis with the technique of formal sequential testing.

A cumulative meta-analytic approach entails conducting separate meta-analyses at interim points by successively adding study effect sizes (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2009). When studies are arranged in a chronological sequence, from the earliest to the latest, such repeated pooling reveals the evolution of cumulative knowledge over time. Lau et al. (1992) have proposed using this methodology to identify the benefit or harm of interventions across several randomized clinical trials as early as possible. Unfortunately, because successive analyses are conducted, the problem of multiple testing is inevitably associated with cumulative meta-analysis. These *multiple looks* as evidence accumulates consequently lead to inflated Type I error rates.

The issue of inflated Type I errors due to multiple testing has extensively been addressed in the context of monitoring individual studies in several disciplines—particularly, in industrial process control and clinical trials (Gosh & Sen, 1991; Todd, 2007). Sequential testing was introduced more than 6 decades ago (Wald, 1947) as an economical alternative to traditional statistical decision making. When employing the latter procedure, one decides to accept or reject a hypothesis on the basis of an a priori determined number of measurements. In order to avoid gathering an unnecessary large amount of measurements, Wald suggested sequentially evaluating the available evidence at consecutive interim steps during the data collection. At each interim analysis, an a priori stopping rule is used to decide whether there is enough information to reject or accept a posited hypothesis, while retaining an overall Type I error rate throughout the interim analyses. If the extant information is insufficient, the data collection is continued.

Since the introduction of the idea of sequential analysis over 60 years ago, the methodology has been further developed. Pogue and Yusuf (1997) proposed using the technique of group sequential boundaries to indicate whether sufficiency in cumulative meta-analysis of group comparison studies has been obtained. The key feature of group sequential testing (Armitage, 1967) is that the cumulative data are analyzed at intervals or with a group of measurements. At each interim analysis, the test statistic is compared with a boundary point (critical value), which is chosen such that the overall significance level does not exceed the desired α. Group sequential testing was originally designed for an a priori planned number of equally spaced interim analyses. Afterward, Lan and DeMets (1983) extended this methodology using an alpha spending function to construct group sequential boundaries when the number of interim analyses is unknown and/or unequally spaced. Given that there is a fundamental uncertainty in cumulative meta-analysis about the total number of studies that will ever be conducted on a certain topic and given the fact that the amount of information will differ for each study, the flexible alpha spending function provides a way to obtain a stopping boundary in cumulative meta-analysis that controls the Type I error. The boundary points are characterized by the rate at which α is spent and by past decision times but are independent of the number of future decision times. The function itself is monotone nondecreasing and is indexed by the accumulating information. The sequential meta-analytic approach first entails calculating an a priori optimal information size (OIS). This is the amount of information required to have a high probability of detecting an a priori specified effect while minimizing false positive results. Estimating the OIS for a sequential meta-analysis is thus very similar to an a priori sample size calculation for an individual study. The OIS is then used to construct group sequential boundaries b at each interim analysis $q = 1, \ldots, Q$, employing the Lan–DeMets alpha spending function. These boundaries are compared with the interim standardized test statistic that represents the Z-value of the pooled effect size at interim step q of a cumulative meta-analysis, denoted by Z_q. Information accumulation is continued as long as $|Z_q| < b_q$.

The technique of SMA for group comparison studies cannot simply be transferred to determine sufficiency in single-case research synthesis in case raw data are combined instead of effect sizes. Also an additional source of heterogeneity emerges when aggregating across single-case studies. In this article, we propose adjustments with regard to conducting the cumulative meta-analysis and estimating the optimal information size. In the following, a strategy for conducting an SMA of single-case data is outlined and illustrated using a real meta-analytic database.

Sequential meta-analysis of single-case data

In the following, cumulative meta-analysis of single-case experimental data is introduced, and a strategy for adding group sequential boundaries is presented. We label the combination of both methods *sequential single-case meta-analysis* (SSCMA).

Cumulative meta-analysis of single-case experimental data To obtain the interim standardized test statistic, Z_q, a cumulative meta-analysis is typically performed by successively adding effect sizes from group comparison studies at each interim analysis. In contrast to group comparison research, raw data are often available and can be combined in a single-case research synthesis, which hampers the use of standard cumulative meta-analytic techniques to obtain the interim standardized test statistic. To overcome this problem, we recommend using a cumulative approach to multilevel meta-analysis of single-case data. Raw data from several single-case studies, often including multiple cases within each study, have a multilevel or hierarchical structure. The repeated measures (level 1) are nested within the same case (level 2), and cases are nested within studies (level 3). Van den Noortgate and Onghena (2003) have proposed using a three-level linear model, which partitions the variation in outcome variable at three levels, to synthesize single-case studies.

Variation within participants (level 1) when the treatment condition is compared with the baseline condition is described by the following equation:

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}(Treatment)_{ijk} + e_{ijk} \tag{1}$$

The value of the dependent variable at measurement occasion i for case j of study k, represented by Y_{ijk}, is regressed on a dummy variable *treatment*. This first-level predictor equals 1 when the measurement occasion pertains to the treatment condition and 0 otherwise. While β_{0jk} represents the expected score for case j from study k during the baseline condition, β_{1jk} can be interpreted as the treatment effect for case j from study k. Because the value for this measure of effect depends on the scale of the dependent variable, raw scores can be standardized to permit direct comparison of scores across studies. This can easily be done by performing an ordinary regression analysis, with a single predictor that indicates the treatment condition, for each participant and dividing the scores by the estimated root mean square error. The regression coefficient β_{1jk} can now be interpreted as a standardized mean difference, although not directly comparable to the standardized mean difference used in group comparison studies (Van den Noortgate & Onghena, 2008). The first level of the multilevel model (Eq. 1) can easily be extended to model trends in the data by including a time predictor in addition to the treatment predictor and the interaction between both (Center, Skiba, & Casey, 1985; Van den Noortgate & Onghena, 2003).

Variation between cases within studies (level 2) is described as follows:

$$\begin{aligned} \beta_{0jk} &= \theta_{00k} + u_{0jk} \\ \beta_{1jk} &= \theta_{10k} + u_{1jk} \end{aligned} \tag{2}$$

Both equations indicate that the expected score for case j from study k equals a mean score for study k plus a random deviation from this mean.

Expected scores at the study level (level 3) comprise an overall score across scores plus random deviations from this mean:

$$\begin{aligned} \theta_{00k} &= \gamma_{000} + v_{00k} \\ \theta_{10k} &= \gamma_{100} + v_{10k} \end{aligned} \tag{3}$$

The multilevel model is a random-effects model that is equivalent to a three-level extension of commonly used random-effects models for the meta-analysis of effect sizes (DerSimonian & Laird, 1986; Hedges & Olkin, 1985; Van den Noortgate & Onghena, 2008). Apart from estimating and testing the between-study and between-case variance, this flexible procedure also allows one to add study or case characteristics that may moderate the size of the treatment effect. These predictors are included in the equations that describe the variation in the effect at the particular level. For example, to examine gender differences in the treatment effect, a dummy predictor variable is added to Eq. 2:

$$\beta_{1jk} = \theta_{10k} + \theta_{11k}(male)_{jk} + u_{1jk} \tag{4}$$

Now θ_{10k} reflects the treatment effect for females in study j, and θ_{11k} indicates the additional treatment effect for males in study j. This model is equivalent to a mixed-effects model commonly used in meta-analysis (Raudenbush & Bryk, 1985).

Analogous to standard multilevel modeling, fixed parameters of the model (the γs) can be tested using the Wald test, which compares the difference in parameter estimate and the hypothesized population value divided by the standard error with a standard normal distribution. By performing a cumulative multilevel meta-analysis of single-case experimental data, the Z-value of the Wald test can be used as the interim standardized test statistic Z_q. It should be noted that early in the meta-analytic database, when the number of studies is limited, estimates of population parameters over studies will be poor, because estimation procedure and statistical tests in multilevel modeling are based on large sample properties. The large sample requirement is most problematic at the highest level (i.e., study level). At lower levels, it might be less problematic, because the number of units is usually larger (by definition the number of units at lower levels is at least as large as the number of units at higher levels). If a small sample is used, fixed parameter estimates will be unbiased, but standard errors and variance estimates may be biased. Van den Noortgate and Onghena (2007) have suggested that combining single-case studies using a multilevel framework should be postponed until at least about 20 entities are available at the highest level. Hence, in the context of SSCMA, we recommended that the initial interim analysis should at least comprise about 20 single-case studies including single or multiple cases.

Group sequential boundaries In order to construct group sequential boundaries, an optimal information size required to have a high probability of detecting a pooled effect (1- β, or power) of a presumed effect size while minimizing false positive results (α) has to be computed a priori. Because the OIS should at least equal that of a well-designed individual study, Pogue and Yusuf (1997) have suggested that standard methods of sample size calculation can be used as a starting point. However, since the aim of meta-analyses is to provide authoritative evidence (Sutton & Higgins, 2008), we recommend using a more stringent strategy to determine the OIS in meta-analysis.

To compute the OIS, reasonable values have to be specified for the treatment effect size as well as the Type I (α) and Type II (1 - β) error rates. Previous meta-analyses or primary studies in the area on a similar topic can serve as a source of information to determine a reasonable effect size, or the smallest effect size deemed to be of practical significance in a particular context can be used. It should be noted that the effect size in previous meta-analyses of group comparison studies cannot directly be compared with effect sizes in single-case meta-analyses. The comparability of effect sizes from both designs has been discussed by Van den Noortgate and Onghena (2008), but as a guideline researchers could keep in mind that the effect sizes obtained in meta-analyses of group comparison studies likely reflect a conservative estimate of the effect in single-case research syntheses. Although a widely accepted cut-point for declaration of statistical significance (α) in meta-analysis seems to be .05, we recommend using at least the more stringent .01 criterion. In a related vein, we suggest adopting a more conservative .90 criterion for power, instead of the commonly used .80 criterion.

Apart from sampling variability (level 1), between-case (level 2) and between-study (level 3) variability is likely to emerge in meta-analyses of single-case data. In contrast to standard sample size calculation, such heterogeneity should be taken into account when specifying the OIS. In SMA of group comparison studies, Wetterslev et al. (2008) have proposed adjusting the OIS according to the amount of between-study variability expressed by the I^2 index, using the formula OIS/(1 - I^2). The I^2 index is similar to an intraclass correlation in multilevel models (Higgins & Thompson, 2002). Given that in single-case meta-analysis, between-study and between-case variability should be accounted for, we propose adjusting the OIS according to the intracase correlation ($\rho_{\mathrm{IntraCase}}$). On the basis of the formula for estimating an intraclass correlation at the second level of a three-level model (Siddiqui, Hedeker, Flay, & Hu, 1996) — namely, $\rho = \left(\sigma^2_{v0} + \sigma^2_{u0}\right) / \left(\sigma^2_{v0} + \sigma^2_{u0} + \sigma^2_e\right)$ — the intracase correlation can be interpreted as the expected correlation between two randomly chosen measurements of the same case, taking into account that the two measurements of the same case also belong to the same study. Since a single-case meta-analysis is a random-slopes model, the intracase correlation also depends on the value of the treatment predictor for each of the measurements, which complicates the expression of the intracase correlation (for details on the formula, see Goldstein, Browne, & Rasbash, 2002). However, because we are interested in the intracase correlation conditional upon the treatment condition, the formula for estimating the expected correlation between two randomly chosen measurements in the treatment condition (treatment = 1) of the same case within the same study reduces to the following:

$$\rho_{IntraCase} = \frac{\sigma^2_{v00} + 2\sigma_{v01} + \sigma^2_{v10} + \sigma^2_{u0} + 2\sigma_{u01} + \sigma^2_{u1}}{\sigma^2_{v00} + 2\sigma_{v01} + \sigma^2_{v10} + \sigma^2_{u0} + 2\sigma_{u01} + \sigma^2_{u1} + \sigma^2_e} \quad (5)$$

In case moderators are added as fixed predictors, the residual intracase correlation can be estimated using the same formula based on the variance components of the model including moderators. The residual intracase correlation reflects the expected correlation between two randomly chosen measurements in the treatment condition of the same case within the same study after controlling for the effect of the moderators.

In line with Wetterslev et al. (2008), a heterogeneity-adjusted OIS (HOIS) could be computed using the formula OIS/(1-$\rho_{\mathrm{IntraCase}}$), which results in a larger optimal information size as between-study and between-case variability increases. The intracase correlation obtained in the actual meta-analytic database or in an existing meta-analysis of single-case studies on a similar topic can be used to determine a realistic value. When the actual meta-analytic database contains a limited number of studies and previous meta-analyses are lacking, we recommend using a conservative value for the intracase correlation that reflects a high amount of heterogeneity, since an underestimation of the heterogeneity will fail to render the authoritative evidence expected from a meta-analysis. Using the classification proposed for the I^2 index (Higgins & Thompson, 2002), which is similar to an intraclass correlation in multilevel models, a value of .75 can be used to reflect high heterogeneity. In a related vein, we recommend using a medium (.50) amount of heterogeneity as a conservative value for the residual intracase correlation, because moderators rarely explain all between-case and between-study variability.

The (H)OIS is then used to construct group sequential boundaries *b* at each interim analysis *q* = 1, …, *q* using the Lan–DeMets alpha spending function. The alpha spending function, denoted α*, allocates the allowable Type I error through a function based on the information fraction *t*. In the context of meta-analysis, the information fraction *t* (0 ≤

$t \leq 1$) represents the proportion of the optimal information size that has been accumulated at a particular interim analysis, $t_q = i_q/\mathrm{OIS}$, where i_q is the information available at the qth interim analysis. The spending function $\alpha^*(t)$ equals 0 for $t = 0$ and α for $t = 1$ and is monotone nondecreasing in between. This flexible procedure guarantees a fixed significance level and power when the optimal information size is achieved. Several functions can be fit into the Lan–DeMets alpha spending function, including the O'Brien and Fleming (1979) rules:

$$\alpha^*(t) = \begin{cases} 0 & \text{when } t = 0 \\ 2\left[1 - \Phi\left(Z_{\alpha/2}/\sqrt{t}\right)\right] & \text{when } 0 < t \leq 1 \end{cases} \tag{6}$$

where Φ is the standard normal cumulative distribution function. The Type I error that is allocated to each interim analysis through the alpha spending function in turn determines the boundary point b_q as the critical Z-value that corresponds to the allocated α at interim analysis q.

The interim standardized test statistic at the qth interim step, denoted by Z_q, is the Z-value obtained by performing a Wald test of the fixed parameters at interim analysis q of the cumulative meta-analysis. At each interim analysis, the sufficiency of the cumulative knowledge is evaluated by comparing Z_q with the boundary point b_q. As long as $|Z_q| < b_q$, sufficiency is not yet attained, and further studies are needed to establish convincing statistical evidence. When $|Z_q| \geq b_q$, the boundary is crossed at interim analysis q, which indicates that sufficient evidence favoring at least the specified effect is established. In case the OIS is reached at the final analysis Q and boundaries have not been crossed, sufficiency of the cumulative knowledge is achieved, since the SSCMA is unable to detect the anticipated effect despite the specified level of power.

Illustration

The method is applied to a single-case meta-analytic dataset of 271 participants (116 male and 150 female) from 138 studies on the effect of contingency management (e.g., reward, praise, and attention for positive behaviors) of challenging behavior among persons with intellectual disabilities. Raw scores were extracted from graphs or retrieved electronically and were standardized as outlined above. Apart from the overall treatment effect, the moderating effect of participants' sex and the moderator effect of an additional contextual intervention (e.g., informing, educating, and training the environment or adapting the environment to the person's needs) were also estimated. The results of a multilevel meta-analysis of the single-case studies are presented in Table 1. Analysis revealed evidence for a significant reduction in challenging behavior after contingency management, $Z = -3.47$, $p < .0001$. There was also evidence of systematic differences in the treatment effect between cases, $\chi^2(2) = 6{,}774.8$, $p < .0001$, and between studies, $\chi^2(2) = 7.5$, $p = .0235$. Furthermore, the moderator effect of participants' sex was nonsignificant, $Z = -0.30$, $p = .764$. The use of a contextual intervention also did not significantly moderate the treatment effect of contingency management, $Z = -1.10$, $p = .271$. We used this meta-analytic dataset to illustrate the sequential single-case meta-analysis employing a four-step approach. The goal was to determine the sufficiency of cumulative knowledge with regard to the overall treatment effect and the moderator effects.

Step 1: Compute OIS The OIS required in order to have a high probability of detecting a prespecified effect while minimizing false positive results was first computed. We estimated the OIS assuming that even a small treatment

Table 1 Results of the multilevel meta-analysis of all single-case studies ($k = 138$)

	Notation	Parameter Estimate (*SE*)	
		Model 1	Model 2
Fixed effects			
Mean effect treatment	γ_{100}		−2.95 (0.62)
Moderator effect sex	γ_{110}	−3.47 (0.42)	−0.25 (0.85)
Moderator effect contextual intervention	γ_{120}		−0.93 (0.85)
Variance between studies			
Intercept	$\sigma_{\nu 00}{}^2$	3.25 (1.79)	2.87 (1.85)
Effect treatment	$\sigma_{\nu 10}{}^2$	3.44 (1.85)	3.06 (1.93)
Covariance	$\sigma_{\nu 01}$	−3.31 (1.81)	−2.92 (1.88)
Variance between participants			
Intercept	$\sigma_{\mu 0}{}^2$	43.41 (4.01)	44.49 (4.17)
Effect treatment	$\sigma_{\mu 1}{}^2$	37.01 (3.51)	37.93 (3.66)
Covariance	$\sigma_{\mu 01}$	−39.84 (3.73)	−40.82 (3.89)
Residual variance	$\sigma_e{}^2$	1.00 (0.01)	1.00 (0.01)

effect (d = .20) would be worthwhile. Using the freeware program G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), the OIS required in order to have 90% probability of accurately detecting a difference between two dependent means of .20 while minimizing false positive results to 1% (two-sided) equaled 376 participants. Using the aforementioned formula, the intracase correlation for the model without moderators (model 1) based on all included studies (Table 1) equaled:

$$\rho_{IntraCase} = \frac{3.25 - 6.62 + 3.44 + 43.41 - 79.66 + 37.01}{3.25 - 6.62 + 3.44 + 43.41 - 79.66 + 37.01 + 1.00} = .46$$

Given that our dataset comprised a substantial number of studies, we accounted for about a medium (46%) amount of heterogeneity. As a result, the HOIS equaled $376/(1 - .46) = 696$ participants.

For illustrative purposes, the OIS for the moderator effect of participants' sex was computed assuming a difference in treatment effect between independent groups of .80, since a smaller effect of sex may have limited practical significance. This resulted in OIS = 98 participants. For the moderator effect of a contextual intervention, a small difference (d = .20) was used because even a small surplus effect could be of practical importance. This yielded an OIS of 1,482 participants. The OIS for the moderator effects was adjusted according to the residual intracase correlation computed using the variance components of model 2 (Table 1). Hence, the HOIS for the moderator effect of participants' sex was 182 and for the moderator effect of a contextual intervention was 2,744 participants.

Step 2: Perform a cumulative single-case meta-analysis Studies were arranged in a chronological sequence according to year of publication, and a meta-analysis was conducted after each year. The restricted maximum likelihood procedure implemented in the MIXED procedure from SAS was used to fit two models at each intermediate point, resulting in a cumulative single-case meta-analysis with 11 interim analyses (Table 1). First, a three-level model (level 1, within case; level 2, between cases/within study; level 3, between study) with treatment as the first-level predictor was fitted to the meta-analytic database. Second, the two predictor variables were added to the previous model to examine whether these case characteristics would moderate the treatment effect. A Wald test was computed for the overall and moderator effects at each interim analysis, which represent the cumulative Z-values in the sequential meta-analyses (Figs. 1, 2 and 3).

Step 3: Construct group sequential boundaries The Lan–DeMets alpha spending function was used to construct two-sided group sequential boundaries for α = .01 and power = .90, using the open-source software package created by Reboussin, DeMets, Kim, and Lan (2003). The program truncates boundaries at b_q = 8, because it is highly unlikely to obtain a value greater than or equal to that point due to chance, but one could compute the exact boundary using the aforementioned formula. As is presented in Table 2, the information fraction t at each interim analysis was calculated as the proportion of the HOIS. Combinations of group sequential boundaries with cumulative Z-values were used to determine the sufficiency of cumulative knowledge for the overall treatment effect (Fig. 1), the moderator effect for sex (Fig. 2), and the moderator effect for a contextual intervention (Fig. 3).

Step 4: Determine Sufficiency As is presented in Fig. 1, the boundary for a two-sided α-value of .01, assuming a treatment effect of .20 and 90% power, was crossed at the 8th interim analysis with the inclusion of 209 cases (99 studies). The absolute cumulative Z-value larger than the boundary indicates that at that point of the cumulative meta-analysis, sufficiency was reached and convincing evidence for at least a small effect contingency management for challenging behavior among persons with intellectual disabilities was obtained. With regard to the

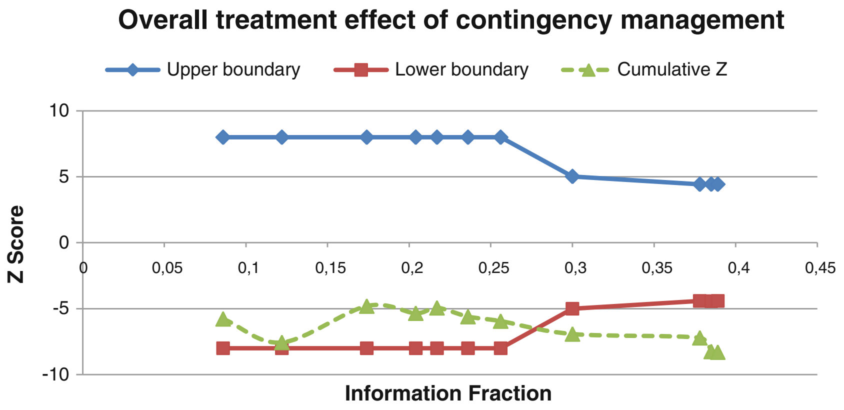

Fig. 1 Sequential single-case meta-analysis of the overall effect of contingency management

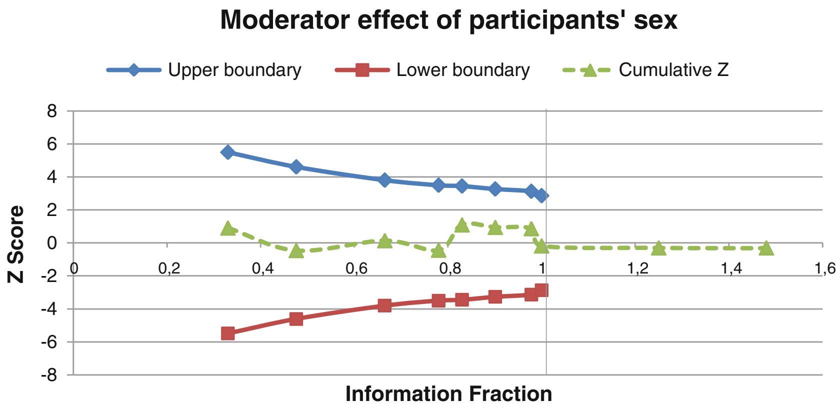

Fig. 2 Sequential single-case meta-analysis of the moderator effect for participants' sex

moderator effect of sex (Fig. 2), the SSCMA indicates that sufficiency was attained at the 8th interim analysis. Given that the HOIS was reached at that point and no boundaries were crossed, there is sufficient evidence to refute at least a large effect of participants' sex with 90% power. For the moderator effect of a contextual intervention, sufficiency is not attained, because no boundaries were crossed at the final interim analysis. Given that the HOIS has not been reached, the SSCMA may still be underpowered (power < 90%) to detect or reject a small moderating effect. Hence, additional single-case studies investigating the additional effect of a contextual intervention are needed to attain sufficiency.

Discussion

Meta-analysis of single-case experimental studies has great potential to combine results from individual studies. Unfortunately, formal guidelines on the interpretation of meta-analytic findings are lacking, which leaves an important question unanswered: Is there enough cumulative knowledge available to draw firm statistical conclusions? We proposed an adaptation of formal sequential boundaries to determine the sufficiency of cumulative knowledge in single-case research synthesis, which we labeled SSCMA.

The first step in conducting an SSCMA is to determine the amount of information that would be required to draw convincing statistical conclusions. In line with the high standards imposed on meta-analysis, we recommended using conservative, but realistic, criteria to determine this optimal information size (OIS). After establishing this OIS, alpha spending functions can be tied with a cumulative approach to a multilevel meta-analysis of single-case experimental data. The alpha spending is a flexible sequential method based on the proportion of the optimal information size available at a particular interim analysis, and thus the boundaries are independent of future interim analyses.

As illustrated in our example on the effect of contingency management for challenging behavior among persons with intellectual disabilities, an SSCMA allows one to determine the sufficiency of the overall treatment effect and potential moderator effects. So far, applications of sequential meta-analysis in group comparison studies have considered only the overall treatment effect (e.g., Devereaux et al., 2005; Keus, Wetterslev, Gluud, Gooszen, & van Laarhoven, 2010). Given that researchers and practitioners are often interested not only in knowing whether, but also under which conditions a treatment is effective, it is just as important to gauge the sufficiency of moderator effects. As holds for our example,

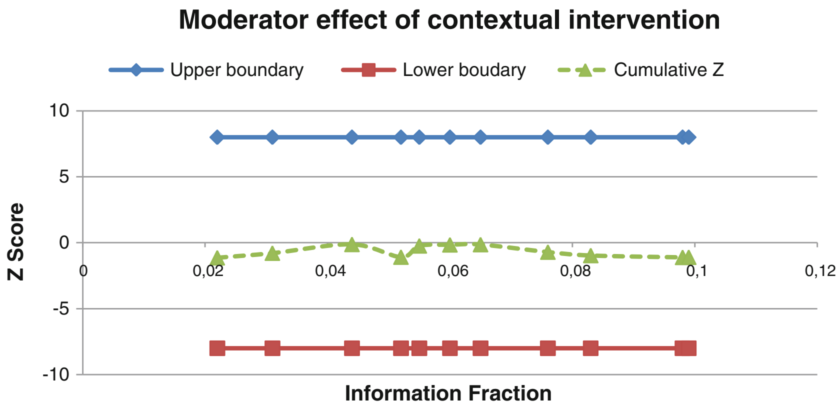

Fig. 3 Sequential single-case meta-analysis of the moderator effect for contextual intervention

Table 2 Overview of the sequential single-case meta-analysis

Interim analysis	Publication Year	k	Cumulative N	Overall Effect Contingency Management		Moderator Effect Sex		Moderator Effect Contextual Intervention	
				t	$Z\gamma_{100}$	t	$Z\gamma_{110}$	t	$Z\gamma_{120}$
1	2000	25	60	.086	−5.77	.329	0.91	.022	−1.14
2	2001	40	85	.122	−7.56	.467	−0.46	.031	−0.79
3	2002	52	121	.174	−4.80	.665	0.14	.044	−1.28
4	2003	64	142	.204	−5.34	.780	−0.44	.052	−1.09
5	2004	70	151	.217	−4.93	.829	1.10	.055	−0.24
6	2005	77	154	.236	−5.60	.901	0.95	.060	−0.16
7	2006	86	178	.256	−5.94	.978	0.87	.065	−0.14
8	2007	99	209	.300	−6.92	>1.00	−0.18	.076	−0.71
9	2008	113	228	.378	−7.21		−0.30	.083	−0.97
10	2009	135	268	.385	−8.25		−0.32	.098	−1.11
11	2010	138	271	.389	−8.30		−0.30	.099	−1.10

k=number of studies; t=information fraction; γ_{100}=overall treatment effect; γ_{110}=moderator effect of sex; γ_{120}=moderator effect of contextual intervention

obtaining sufficient evidence for questions pertaining to moderator effects requires a larger amount of cumulative knowledge than does establishing firm evidence on the overall treatment effect. If sufficiency has not been attained, the optimal information size provides researchers with an indication of the number of additional case studies needed to establish convincing statistical evidence.

Using SSCMA will avoid spurious claims of treatment benefit based on limited cumulative knowledge or poor statistical grounds. As is presented in Figs. 1, 2 and 3, the alpha spending function results in very stringent boundaries at the early stages of the meta-analytics database. As such, the high nominal p-values are not considered to reveal sufficiency, since they may reflect systematic error, such as low methodological quality bias, publication bias, or small sample size bias or random error due to repeated testing. Sequential single-case meta-analysis not only provides researchers with a statistical framework to more consistently interpret cumulative knowledge, it also suggests a limit to conducting research on a particular topic.

An SSCMA can be used to determine the sufficiency of cumulative knowledge in retrospect for an initial or extant meta-analytic database. By ordering the studies in a chronological sequence, such applications can be used to decide whether firm evidence has already been obtained or whether additional case studies should be initiated. By extending the approach to determine the sufficiency of moderator effects, the method can also serve as a tool for setting up targeted research in order to gather specific information with regard to moderators for which sufficiency is not yet achieved. In a prospective application, the goal is to pinpoint the earliest time at which sufficiency is achieved by updating the dataset as new studies become available. Some researchers (Chalmers, 2005; Lau et al., 1992) have proposed systematically putting new research in the context of the existing cumulative knowledge by continuously updating a meta-analysis after completing an individual study. A continuous prospective approach could be particularly beneficial when additional research is costly or when obtaining solid evidence is pressing. Others (Pogue & Yusuf, 1998) have, however, recommended a more parsimonious and more practical updating after a substantial amount of additional information has been obtained (e.g., 20% of the OIS). Regardless of the retrospective or prospective nature, sequential single-case meta-analysis would make the accumulation of cumulative knowledge less haphazard.

Three issues regarding the multilevel approach to aggregating single-case experimental data should be noted. First, given that some authors (Bryk & Raudenbush, 1992; Snijders & Bosker, 1999) have argued that, for the fixed effects, comparing the parameter estimates divided by the standard error to a t-distribution yields better results, the alpha spending function could be adapted to yield critical t-values for the boundary points at each interim analysis. Second, when the single-case studies in a meta-analytic database all include one case, the outlined three-level model reduces to a two-level model with measurements nested in cases. Given that each study represents one single case, the second level can also be interpreted as the study level. Third, as holds for the standard multilevel approach to aggregating single-case studies, this method requires at least 20 studies to perform the analysis comfortably, due to the large sample size properties of the multilevel strategy.

To attain sufficiency, even more studies comprising hundreds of cases may be needed. Nevertheless, in the present illustration and recently published syntheses of single-case studies (e.g., Didden, Korzilius, van Oorsouw, & Sturmey, 2007; Herzinger & Campbell, 2007; Morgan & Sideridis, 2006), the number of studies would be large enough to initialize an SSCMA. Although challenging, we therefore feel that the sample size requirements of SSCMA are not unrealistic for the field.

Although sequential meta-analysis of single-case experimental studies has great potential to contribute to research, some issues remain unresolved. First, the outlined alpha spending function is designed to stop early only if an effect emerges, but it will not stop early for futility. This means that in case no effect becomes apparent, as for the moderator effects in our illustration, the accumulation of knowledge will be continued until the optimal information size is reached. Only then will sufficient evidence to refute a treatment effect of a certain magnitude, despite a prespecified power, be obtained. The usefulness of other sequential designs that also stop early for futility, referred to as the (double) *triangular tests* (Whitehead, 2002), should be investigated in the context of single-case research synthesis. Second, although this method could inform researchers about the sufficiency of cumulative knowledge, yields guidelines with regard to the optimal information size, and allows for a more systematic accumulation of knowledge, it cannot provide information with regard to the optimal design of future single-case studies. For example, should multiple studies with one case be conducted or a few studies with multiple cases? Or should more information be gathered on males or females? Third, SSCMA provides a statistical ground for determining the sufficiency of cumulative knowledge. It, however, does not provide information about the generalizability of the cases included in the meta-analytic database. Given that participants in single-case studies are often not randomly sampled, researchers should carefully evaluate the characteristics of the cases involved when determining sufficiency at a certain point of the meta-analytic dataset. In our example, participants were purposefully selected, since cases were included in the meta-analysis only if they referred to persons with intellectual disabilities that displayed challenging behavior for which a contingency management intervention was initialized. As such, it is clear that any conclusions with regard to the sufficiency of cumulative knowledge are not to be generalized to cases that differ from those included in the meta-analysis. Fourth, adding a new piece of information may lead to more heterogeneity, which increases the uncertainty about the treatment effect estimate, as compared with the previous interim analysis. Several approaches have been suggested in the statistical literature for overcoming this problem, such as altering the boundaries, but none has been thoroughly investigated yet (Whitehead, 2002).

The sequential methodology for gauging sufficiency in single-case research synthesis may serve as a valuable tool for behavioral researchers, to guide them in making the best use of limited resources. It provides a statistical framework for enhancing the interpretation of treatment effects in research synthesis. This kind of information could serve as a starting point for setting up new single-case studies if the cumulative knowledge is insufficient, whereas it may provide scientific underpinning of clinical practices or a redirection of research goals if sufficiency is attained.

Author Note Sofie Kuppens is a postdoctoral fellow of the Research Foundation Flanders (Belgium). Mieke Heyvaert is a predoctoral fellow of the Research Foundation Flanders (Belgium).

References

Armitage, P. (1967). Sequential analysis in therapeutic trials. *Annual Review of Medicine, 20*, 425–430.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: Wiley.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models. Application and data analysis methods*. Newbury Park: Sage.

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

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

Chalmers, I. (2005). The scandalous failure of science to cumulate evidence scientifically. *Clinical Trials, 2*, 229–231.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7*, 177–188.

Devereaux, P. J., Beattie, W. S., Choi, P. T., Badner, N. H., Guyatt, G. H., Villar, J. C., et al. (2005). How strong is the evidence for the use of perioperative blockers in non-cardiac surgery? Systematic review and meta-analysis of randomized controlled trials. *British Medical Journal, 331*, 313–321.

Didden, R., Korzilius, H., van Oorsouw, W., & Sturmey, P. (2007). Behavioral treatment of challenging behaviors in individuals with mild mental retardation: Meta-analysis of single-subject research. *American Journal of Mental Retardation, 111*, 290–298.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Goldstein, H., Browne, W. J., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics, 1*, 223–232.

Gosh, B. K., & Sen, P. K. (1991). *Handbook of sequential analysis*. New York: Marcel Dekker.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando: Academic Press.

Herzinger, C. V., & Campbell, J. M. (2007). Comparing functional assessment methodologies: A quantitative synthesis. *Journal of Autism and Developmental Disorders, 31*, 1430–1445.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a metaanalysis. *Statistics in Medicine, 21*, 1539–1558.

Keus, F., Wetterslev, J., Gluud, C., Gooszen, H. G., & van Laarhoven, C. J. H. M. (2010). Trial sequential analyses of meta-analyses of complications in laparoscopic vs. small-incision cholecystectomy: More randomized patients are needed. *Journal of Clinical Epidemiology, 63*, 246–256.

Lan, K. K. G., & DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika, 70*, 659–663.

Lau, J., Antman, E. M., Jimenez-Silva, J., Kupelnick, B. A., Mosteller, F., & Chalmers, T. C. (1992). Cumulative meta-analysis of therapeutic trials for myocardial infarction. *The New England Journal of Medicine, 327*, 248–254.

Morgan, P., & Sideridis, G. D. (2006). Contrasting the effectiveness of fluency interventions for students with or at risk for learning disabilities: A multilevel random coefficient modeling meta-analysis. *Learning Disabilities Research & Practice, 21*, 191–210.

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics, 35*, 549–556.

Onghena, P. (2005). Single-case designs. In B. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 4, pp. 1850–1854). New York: Wiley.

Pogue, J. M., & Yusuf, S. (1997). Cumulating evidence from randomized trials: Sequential monitoring boundaries for cumulative meta-analysis. *Controlled Clinical Trials, 18*, 580–593.

Pogue, J. M., & Yusuf, S. (1998). Overcoming the limitations of current meta-analysis of randomized controlled trials. *Lancet, 351*, 47–52.

Raudenbush, S. W., & Bryk, A. S. (1985). Empirical Bayes meta-analysis. *Journal of Educational Statistics, 10*, 75–98.

Reboussin, D. M., DeMets, D. L., Kim, K. M., & Lan, K. K. G. (2003). *WinLD version 2.1* [Computer software and manual]. Retrieved from http://www.biostat.wisc.edu/landemets/.

Siddiqui, O., Hedeker, D., Flay, B. R., & Hu, F. B. (1996). Intraclass correlation estimates in a school-based smoking prevention study: Outcome and mediating variables, by gender and ethnicity. *American Journal of Epidemiology, 144*, 425–433.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advancedmultilevel modeling*. London: Sage.

Sutton, A. J., & Higgins, J. P. T. (2008). Recent developments in meta-analysis. *Statistics in Medicine, 27*, 625–650.

Todd, S. (2007). A 25-year review of sequential methodology in clinical studies. *Statistics in Medicine, 26*, 237–252.

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

Van den Noortgate, W., & Onghena, P. (2007). The aggregation of single-case results using hierarchical models. *Behavior Analyst Today, 8*, 196–208.

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.

Wald, A. (1947). *Sequential analysis*. New York: Wiley.

Wetterslev, J., Thorlund, K., & Gluud, C. (2008). Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. *Journal of Clinical Epidemiology, 61*, 64–75.

Whitehead, A. (2002). *Meta-analysis of controlled clinical trials*. Chichester: Wiley.