

Assessing Change in Intervention Research: The Benefits of Composite Outcomes



David Moreau^{ID} and Kristina Wiebels^{ID}

School of Psychology and Centre for Brain Research, The University of Auckland

Advances in Methods and Practices in Psychological Science
January-March 2021, Vol. 4, No. 1,
pp. 1–14

© The Author(s) 2021

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/2515245920931930

www.psychologicalscience.org/AMPPS



Abstract

Intervention research is often time- and resource-intensive, with numerous participants involved over extended periods of time. To maximize the value of intervention studies, multiple outcome measures are often included, either to ensure a diverse set of outcomes is being assessed or to refine assessments of specific outcomes. Here, we advocate for combining assessments, rather than relying on individual measures assessed separately, to better evaluate the effectiveness of interventions. Specifically, we argue that by pooling information from individual measures into a single outcome, composite scores can provide finer estimates of the underlying theoretical construct of interest while retaining important properties more sophisticated methods often forgo, such as transparency and interpretability. We describe different methods to compute, evaluate, and use composites depending on the goals, design, and data. To promote usability, we also provide a preregistration template that includes examples in the context of psychological interventions with supporting R code. Finally, we make a number of recommendations to help ensure that intervention studies are designed in a way that maximizes discoveries. A Shiny app and detailed R code accompany this article and are available at <https://osf.io/u96em/>.

Keywords

training study, clinical trial, measurement error, factor scores, theoretical construct, latent variable, behavioral dynamics, open materials

Received 8/19/19; Revision accepted 5/6/20

Individuals change continuously across the life span. We learn new skills, refine our knowledge about the world, and change habits. Our minds evolve, assemble, structure, categorize, integrate; we learn and improve. Given these dynamic properties of brains and behaviors, an important area of research in psychology focuses on studying change, and in particular how it can be engineered in the form of interventions, to improve health, skills and abilities, or to nudge behavior. The potential implications of intervention research are far-reaching, yet this line of work remains challenging for two primary reasons. First, intervention studies are often costly, either in terms of financial resources or with respect to the time and effort invested. Studies typically involve a number of individuals followed for a period of time, with unavoidable challenges such as protocol adherence (Johnson & Remien, 2003) and large attrition rates (Davis & Addis, 1999). Second, precise measurement in psychological intervention research is crucial yet typically difficult.

Most phenomena in psychology cannot be observed directly (e.g., Borsboom, 2008); for example, psychologists rarely infer well-being, personality, or intelligence solely from observing an individual's behavior. Rather, they typically rely on standardized tests that have been previously validated and that as a result are thought to accurately reflect latent traits or abilities. Because these latent variables are not directly observed, estimating them precisely can be problematic. Measurement is especially challenging in intervention studies given that the focus is not only on accurate estimation at a given point in time but also on assessing change across a number of time points. In this context, the line between subtle change and no change

Corresponding Author:

David Moreau, School of Psychology and Centre for Brain Research,
The University of Auckland

E-mail: d.moreau@auckland.ac.nz



Creative Commons NonCommercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0

License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits noncommercial use, reproduction, and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

at all can often be a very thin one, calling for sophisticated designs and analyses.

These characteristics may be discouraging, especially when the focus of an intervention is on a very specific outcome measure. What if the hypothesized effect cannot be observed, perhaps because a measure is too noisy or imprecise or because the intervention influenced outcomes that were not measured as part of the study? Given the resources invested, the cost of measuring the wrong outcome, or the right outcome in the wrong way, is often higher than for nonintervention research. A number of solutions to these challenges exist, including the use of large sample size and of tightly controlled experimental settings as well as careful dosage manipulations within study arms or conditions. Here, we focus on refining measurement, an aspect that has traditionally been underappreciated but has gained traction in the field of psychology recently (e.g., Flake & Fried, 2019). Specifically, we advocate the use of composite outcomes to enable inferences about change at the construct level (Cronbach & Meehl, 1955) rather than the level of individual measures. In this context, we first discuss the limitations and benefits of multiple outcome measures as well as the opportunity they provide to refine intervention designs.

Limitations and Benefits of Multiple Outcomes in Interventions

In part to mitigate the aforementioned challenges inherent to psychological intervention studies, designs often include multiple outcome measures targeting a variety of domains. For example, interventions designed to enhance executive function may include well-validated tests of executive function but also additional measures such as short-term memory and reaction time (Takacs & Kassai, 2019). Likewise, interventions targeting well-being often include direct or indirect measures of well-being (e.g., life satisfaction, depressive symptoms) together with more restrictive outcome variables such as illness severity or suicidal thoughts (Bolier et al., 2013; Sin & Lyubomirsky, 2009); regimens focused on alleviating symptoms of clinical conditions such as anxiety disorders, phobia, or schizophrenia may use tests of secondary outcomes alongside symptom assessments (Fedoroff & Taylor, 2001; Mayo-Wilson et al., 2014; Pilling et al., 2002). Yet when multiple outcomes are treated separately at the analysis stage with improvement on any measure being interpreted as support for the effectiveness of an intervention, this approach can be problematic. Increasing the number of outcome variables maximizes the chance that an intervention will be found to have *some* effect, even if it is observed on an outcome measure that was not explicitly hypothesized to relate to the intervention (e.g., in a preregistration or a trial registration).

The reasoning behind this approach is similar to that of “hedging” and comes with a number of problems in the context of interventions. Hedging is a good strategy when uncertainty is important, and there is no underlying “truth.” For example, when investing for retirement, one often spreads potential losses to minimize the influence of one particular event or of specific circumstances. Because investment is stretched across a range of products and placements that are at least partly independent, the risk associated with a portfolio is effectively mitigated. If a given strategy makes money, it is a winning strategy—the goal in investing is to improve the total gains regardless of which particular product within a portfolio is profitable. When designing an intervention, it is perhaps tempting to abide by the same rules, with multiple outcomes spreading the uncertainty associated with the effectiveness of an intervention. However, this strategy can lead to increased error rates, either false positives if no correction for multiple testing is applied or false negatives if one corrects for multiple comparisons. Controlling error rates is important when testing interventions because, in contrast to investments, one typically does not only care about winning (i.e., finding a significant effect somewhere) but also—perhaps more importantly—about being right (i.e., the observed effect reflects a genuine pattern in the population of interest).

This concern is not new—problems with multiple outcomes have long been acknowledged in clinical trials (Pocock, 1997; Pocock et al., 1987), translating into well-defined guidelines for systematically defining primary outcomes (see e.g., the Standard Protocol Items: Recommendations for Interventional Trials [SPIRIT]; Chan et al., 2013). They have persisted in the less regulated, perhaps more lenient space of psychological interventions, but recent developments have been in the right direction, with the publication of specific guidelines for reporting randomized trials of social and psychological interventions (Grant et al., 2018; Montgomery et al., 2018). As a formal and principled way to distinguish exploratory from confirmatory research and, within the latter, to clarify which outcomes are primary and which are secondary, preregistration plays a key role in the context of interventions and should be encouraged (Cybulski et al., 2016).

Despite the aforementioned potential issues, multiple outcomes in intervention research also bring about a number of opportunities to refine assessments of effectiveness by allowing multiple sources of information to be pooled together. The notion of pooling together imperfect measures to get more accurate estimates is ubiquitous: Respectable pollsters do not rely on a single poll to predict the outcome of an election—they use (weighted) aggregates of multiple sources of information; finals in the major American sports leagues are often played to the best of seven games; and undergraduate

college papers typically include a final exam but also a midterm exam and additional coursework. Every measurement in each of these fields, be it a single poll, game, or exam, is imperfect and sometimes a coarse representation of the underlying, unknown reality. Which candidate is ahead, which is the best team, or what is the level or ability of a student on a given topic are questions that typically cannot be reliably answered with one measurement. When multiple measurements are combined, however, one gets a finer estimate of the latent, unobservable reality.

The importance of combining outcome variables is especially salient in intervention studies. If change is observed in one measure after an intervention, one cannot distinguish between two competing options: that the theoretical construct genuinely has changed or that the intervention modulated variance in that particular measure that is not related to the construct itself (see Shiny app for a comparison between these two options). For example, a computerized intervention designed to alleviate Parkinson's symptoms on cognition may elicit changes in a computerized test battery of cognitive function because cognition has improved or because patients gained computer fluency over the course of the treatment. When multiple measures are used for construct estimation, ideally based on different testing modalities, the psychometric structure linking the different outcome measures (e.g., variance-covariance matrix) can help distinguish between these options (see Fig. 1). For example, the psychometric structure is likely preserved if improvements are observed on none of the measures (Fig. 1, postintervention scenario 1) or on all of the measures (Fig. 1, postintervention scenario 3). In the latter case, given that improvements are not only restricted to a single measure but also result in gains in other measures tapping the same construct, it is likely that the intervention genuinely influenced the theoretical construct. In contrast, the psychometric structure is typically not preserved when improvements are observed on at least one but not all of the measures (Fig. 1, postintervention scenario 2). In this case, the intervention likely modified variance unrelated to the construct (an outcome especially plausible if the improved measure is a poorer estimate of the construct, as is the case with M1 in Fig. 1) given that this task-specific variance is probably not shared across all measures. If only one measure is used to estimate a construct, these three scenarios are impossible to disentangle.

The field of brain training (for a review, see Simons et al., 2016), and in particular its implementation via working memory regimens (for a meta-analysis, see Melby-Lervåg et al., 2016), offers a compelling example. More than a decade ago, an influential article suggested that working memory training could improve fluid intelligence by as many as five points in a brief 4-week intervention (Jaeggi et al., 2008). Although it claimed

enhancement at the construct level (fluid intelligence), improvements were reported on a single task (either the Raven's Progressive Matrices or the Bochum Matrix Test, depending on training length), thus allowing for the possibility that working memory gains translated to better performance on a fluid intelligence *task* but not necessarily in fluid intelligence at the construct level (Moody, 2009; Moreau & Conway, 2014; Shipstead et al., 2012). In line with this idea, subsequent work showed that the findings could not be replicated when modeling latent improvements (Chooi & Thompson, 2012; Colom et al., 2013; Redick et al., 2013; Schmiedek et al., 2010). The field of brain training has since increased its standards, with routine use of multiple assessment tasks per construct.

The Case for Composite Outcomes

Because it helps reduce measurement error, modeling change at the construct level increases the probability that a genuine effect will be detected for a given sample size (Weintraub, 2016) and allows more efficient use of resources (e.g., fewer participants for equivalent statistical power; Ross, 2007). Assessments of change at the construct level can be easily accommodated within the general structural equation modeling (SEM) framework, for example via latent curve models (LCM) and latent change score models (LCSM), while allowing multigroup comparisons (for a tutorial, see Kievit et al., 2018; for a detailed description, see Little, 2013). These techniques are appealing because they allow modeling of change directly at the latent level, in line with typical assumptions in many domains of psychology, according to which constructs can be explained by unobserved common causes (McCrae & Costa, 1987; Spearman, 1904; van Bork et al., 2019) and thus are most meaningful in the latent space (McArdle, 2009). However, LCM and LCSM can be problematic in some respects: (a) They typically require more than two time points for accurate modeling (Duncan & Duncan, 2009), whereas many psychological interventions include only a baseline and a postintervention session; (b) they may not provide much transparency about the underlying computations, especially given the flexibility afforded by the approach (McArdle, 2009; Tomarken & Waller, 2005); and (c) they can prevent, or at least complicate, meaningful comparisons across studies because of the inherent disparities between models (for a review, see Tomarken & Waller, 2005). Other methods have been developed recently that do not make assumptions about latent abilities; for example, psychometric networks have been proposed as an alternative to latent variable models for representing psychological constructs (van Bork et al., 2019), including to model change in longitudinal assessments (Blanken et al., 2019; Greene et al., 2018). These are promising developments; however, evidence for the benefits of network models over those of latent models

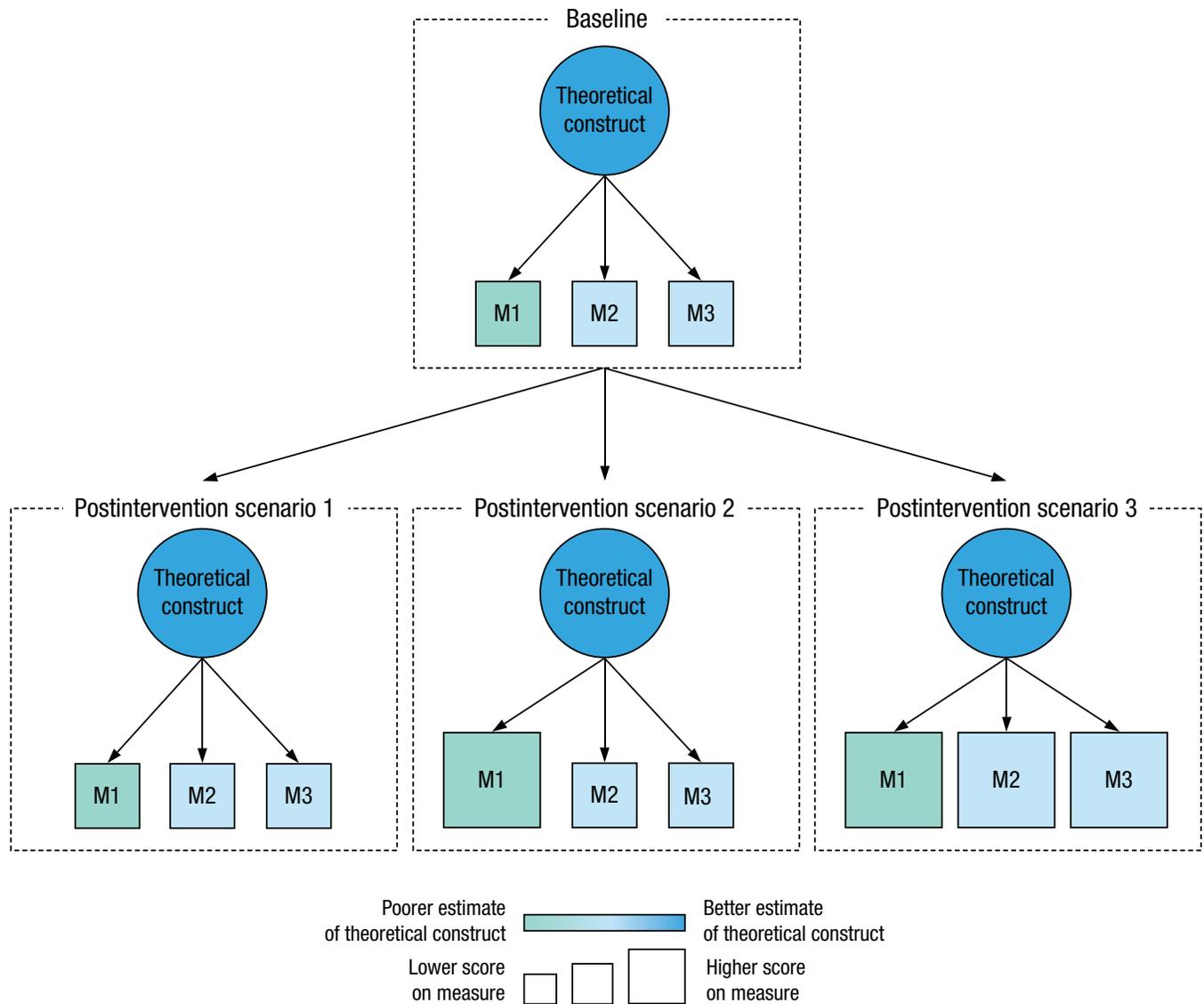


Fig. 1. Psychometric structure linking latent and observed measures in three hypothetical scenarios. In the first scenario (left), the intervention has no effect, and the relationship between the three measures remains unchanged from baseline to postintervention. In the second scenario (middle), the intervention elicits an increase in just one of the measures (M1). As a result, the relationship between measures changes between the two testing phases—a plausible scenario if the intervention modulates task variance that is not related to the theoretical construct but instead is specific to that measure. In the third scenario (right), the intervention elicits improvements on all measures, suggesting that the intervention tapped variance shared by all tasks and is thus part of the construct.

remains mixed, with the two approaches appearing to be complementary to one another rather than redundant (Guyon et al., 2017; van Bork et al., 2019).

An appealing alternative is to group theoretically related variables into a composite score (Ard et al., 2015). By pooling different imperfect measures of an underlying ability, one is less prone to measurement error. Thus, composites can help tap shared variance and reduce task-specific variability and come with many desirable properties—they allow meaningful comparisons across studies and are often straightforward to interpret (Proust-Lima et al., 2019). For example, an intervention study that investigates the effect of a

particular treatment on a composite depression score can be compared against another study using the same composite as well as against population norms measured in nonintervention settings. As long as composite outcomes are computed in the same way, and assuming adequate validity and reliability, comparisons across studies remain meaningful.

In addition, composites allow finer assessments of the outcomes of interest at the construct level—beyond increasing the probability that any observed effect is genuine, assessing change at the construct level can help maximize statistical power (Freemantle & Calvert, 2010; Freemantle et al., 2003). We present a visual illustration

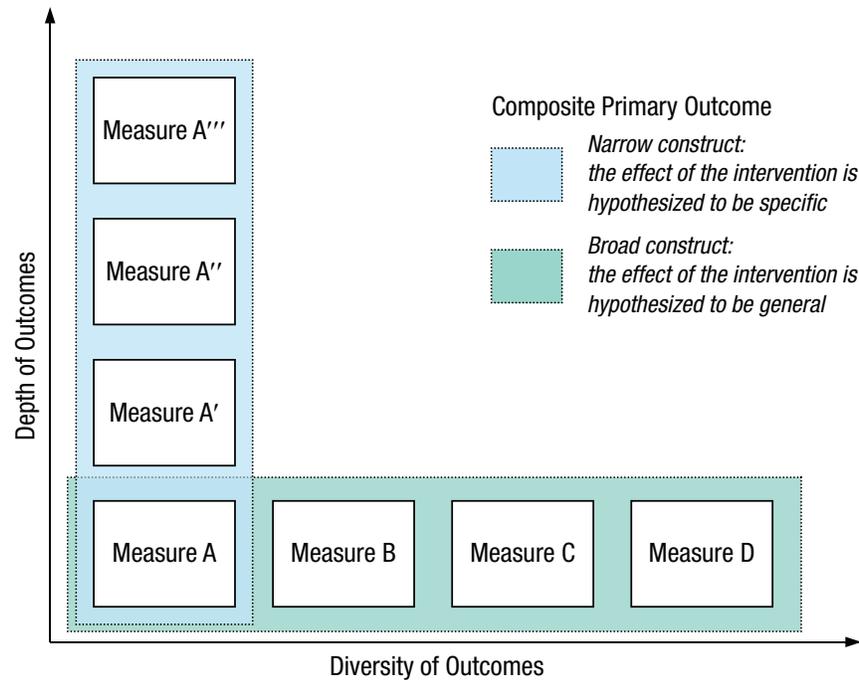


Fig. 2. Two strategies to maximize the value of intervention studies. If the effect of the intervention is hypothesized to be general (green box)—for example, an intervention targeting general cognition—a diverse set of measures can be used to form a composite primary outcome. In this case, the composite makes clear that the effect on each individual measure (A, B, C, D) stems from shared mechanisms. If the effect of the intervention is hypothesized to be specific (blue box)—for example, an intervention targeting working memory—a more restricted set of measures (A, A', A'', A''') can be used to form a composite primary outcome that is narrower but explored in more depth. This strategy helps refine assessments of change by tapping shared variance across outcome measures. Ideally, an intervention that is thought to influence a broad construct would include a diverse set of measures and additional assessments within each narrower construct (i.e., A, A', A'', A''', B, B', . . . , D'''); occupying the top right corner of the figure) but limited resources often preclude intervention researchers from using optimal designs.

of the benefits of composite outcomes over single measures in the Shiny app that accompanies this article. Therefore, and rather than adding variety in intervention outcomes or focusing on single-test assessments, a more sensible approach is often to measure fewer constructs, ideally with a primary outcome explored in depth with multiple assessments (see Fig. 2, blue box; Freemantle & Calvert, 2007a; Ross, 2007). If, however, a variety of measures are hypothesized to benefit from the intervention, it is likely that they are all thought to be influenced by the intervention via a common mechanism. When this is the case, the measures should be grouped together to form a composite spanning a wide construct (Fig. 2, green box). Both of these approaches have merit: Outcomes focused on narrower constructs with multiple measurements allow greater precision, whereas outcomes spread out over a broader construct can increase the impact and generalizability of the study. Irrespective of the particular type of construct one chooses to assess—narrow or broad—the composite outcome should be the primary outcome of the intervention.

Assessing change at the construct level using composite scores involves several steps, all of which need careful consideration by the researcher before the start of the intervention study. These include the computation of a composite following one of several methods available (*estimation*); assessments of validity, reliability, and other psychometric properties (*evaluation*); and the inclusion of the resulting composite in subsequent analyses to assess the effectiveness of the intervention (*application*). In the remainder of the article, we discuss the practicalities of using composite scores, including how to compute, evaluate, and preregister them, as well as their limitations.

Estimating Composites From Theory and Data

After identifying the underlying construct of interest and the outcome measures that will be used for estimation (advice on how to select appropriate measures is available elsewhere; e.g., Borsboom et al., 2003; Flake &

Fried, 2019), one needs to decide how the composite should be calculated. Broadly speaking, composites are created by combining multiple measures into a single score. Assume that we have three different outcome measures aiming to measure the same underlying construct, M_1 , M_2 , and M_3 . We can use these three outcome measures to create a composite score \hat{C} of the following form:

$$\hat{C} = w_1 M_1 + w_2 M_2 + w_3 M_3, \quad (1)$$

in which w_1 , w_2 , and w_3 are the weights, or coefficients, assigned to variables M_1 , M_2 , and M_3 , respectively. The art of creating composite scores is in large part about estimating these weights because there are many different ways to determine which values should be assigned, all of which may be more or less suitable depending on the specific circumstances.

Note that most of us are already familiar with composite scores because they encounter them almost every day. Whenever one makes claims such as that people watch television for 3 hr a day, or when one is calling a specific state in the United States a “blue state” or a “red state,” one does not mean that those statements are valid for all in the entire population of interest. These statements are correct on average when aggregating across individuals. The average (i.e., the arithmetic mean) is a special type of composite score that is unit-weighted, that is, it gives equal weight to all observations. In our example, to get the average, w_1 , w_2 , and w_3 would be set to one third.

Simple averages often represent a stark improvement over single measures (Bakal et al., 2015; see also the Shiny app that accompanies this article). Averaging helps decrease measurement error (for an example in the context of psychological interventions, see Moreau et al., 2016), thus providing better estimates of theoretical constructs. In some circumstances, however, one might want to allow more flexibility than merely adding or averaging outcome measures. On the basis of theoretical or empirical knowledge, one might know that some measures are excellent assessments of a construct, whereas others might be more noisy (but—and this is key—still useful for the estimation of a construct). This can be done by assigning different weights to each outcome variable forming the composite. Weights can be informed theoretically (e.g., using established population parameters or known psychometric structures of the tests) or determined via statistical methods, such as exploratory factor analysis (EFA) or principal component analysis (PCA).¹

Theoretical information is typically available when using a standardized test battery to measure a well-defined construct. For example, the Wechsler Adult Intelligence Scale (Wechsler, 2008) and the Wechsler Intelligence Scale for Children (Wechsler, 1949) are test

batteries of intelligence that consist of several subtests and either include the subtests’ validity and reliability scores or allow estimating them as well as their psychometric structure. In addition, a formula for how the subtests ought to be combined to form a composite score based on these properties is typically available. If no prespecified formula is available as part of the test battery or if a researcher decides to select measures that are not part of a standardized battery to estimate a construct, some theoretical information might still be available. For example, if the measures have documented evidence about their validity or reliability, this information can be used to inform the weights. More valid, reliable measures would be given more weight than less valid or reliable ones.

Although psychometric batteries designed to assess theoretical constructs often use simple, unit-weighted averages to aggregate subtest scores, other, more sophisticated procedures involve a scaling factor. When multiple outcome measures are combined together to form a unit-weighted average, the scale of the resulting composite is different from that of the original measures. This is because increased precision around estimates results in smaller overall standard deviations, a phenomenon that becomes especially exacerbated as the number of subtests increases (Moreau et al., 2016). To correct for this inconsistency between the original standard deviations and the standard deviation of the average, composites are sometimes scaled, typically by expressing the average in terms of units of the square root of the sum of the correlation matrix of all subtests, such that

$$\hat{C} = \frac{M_1 + M_2 + M_3 - \kappa\mu}{\sqrt{\theta}} + \mu, \quad (2)$$

where κ is the number of measures included in the composite \hat{C} and μ is the population mean. In the case of standardized measures, θ might also be known, rather than estimated, from population-level estimates. For a three-variable composite, the sum of the correlation matrix θ is simply:

$$\theta = \kappa + 2cor(M_1, M_2) + 2cor(M_1, M_3) + 2cor(M_2, M_3). \quad (3)$$

Differences between simple and scaled averages are exacerbated when intermeasure correlations get weaker or with increases in the overall number of measures included in the composite (see Fig. SM1 in the Supplemental Material available online and the Shiny app for a visual illustration of this phenomenon). Thus, the advantages of scaled composites over simple averages are exacerbated as designs become more complex and measurements get more noisy. Scaling composite outcomes helps provide finer estimates of an underlying effect, especially in the dynamic context of interventions,

while reducing error inflation and preserving statistical power.

When reliable theoretical information is not readily available—for example, when one combines measures to form an ad hoc set of measures that is not part of a standardized battery or to combine subtests of a standardized battery in a way that was not anticipated by the original developers—another possibility is to determine the weights empirically using information either from previous studies that have administered the same outcome measures in the same way or from the intervention data themselves. One of the most common ways to do this is via EFA (for a primer, see Yong et al., 2013). In an EFA, a key step involves determining factor loadings, that is, indices of the strength or association between a particular measure and each latent factor. In basic implementations of the method, these factor loadings can be used as weights in the calculation of composite scores, following the general case of Equation 1. Specifically, the factor loadings for a given factor are simply multiplied by the standardized observed scores before summing (for a tutorial, see DiStefano et al., 2009). As a result, the contribution of each measure to the composite score is a function of how well it relates to the latent factor, weighted according to the data at hand. In the EFA framework, those composites are referred to as *coarse* factor scores.

Although coarse methods to estimate composite score have the advantage of being relatively robust and stable (Grice & Harris, 1998), simply weighting scores as a function of factor loadings comes with a number of limitations. First, the loading is a regression coefficient in the prediction of a measure from a factor, not in the prediction of a factor from a measure. The two are not necessarily interchangeable, similarly to how $Y = \beta_0 + \beta_1 X + \varepsilon$ is not equivalent to $X = \beta_0 + \beta_1 Y + \varepsilon$. Second, coarse factor scores might not be accurate representations of a theoretical construct given that they are heavily influenced by the specific extraction and rotation methods used. For example, varimax rotations assume that factors are orthogonal (i.e., uncorrelated), whereas oblimin and promax allow for nonorthogonal factors. Which method is appropriate depends on the psychometric structure of the latent factors, but even when assumptions of orthogonality are correct, orthogonal solutions can nevertheless produce correlated factor scores (Glass & Maguire, 1966). Therefore, the relationship between estimated factors might not be reflected in the factor scores themselves. Moreover, it is important to note that differences in within-measure variability across the measures included in an EFA can lead to erroneous estimations of the resulting factor scores—measures for which variability is high (i.e., large standard deviations) will tend to be overrepresented in the calculation of factor

loadings relative to other less variable measures (Gorsuch, 2014; Grice, 2001; Grice & Harris, 1998). This results in higher factor loadings but is purely an artifact of the variability in the data and does not reflect a true relationship with the latent factor. Oftentimes, measures are standardized before EFA to prevent variability from unduly influencing factor loadings (DiStefano et al., 2009).²

More advanced methods, known as *refined* methods, allow circumventing the aforementioned limitations. This class of methods allows maximizing the validity of composite scores by ensuring that they correlate highly with the latent factors. In general, weights are calculated according to the following form (Thurstone, 1935):

$$W = R^{-1}F, \quad (4)$$

in which R is the correlation matrix of the measures included in the composite and F is the factor-loading matrix. As with coarse factor scores, these weights are then multiplied by the standardized observed scores before summing. Scores calculated using these weights are usually referred to as *regression scores* (Thurstone, 1935) and take into account the correlation between factors and measures but also the correlation between measures. This allows maximal validity but comes with several disadvantages, including the possibility of low correlation between observed measures and factors and of nonorthogonality between factor scores (for details, see Grice, 2001). Alternative methods exist to account for these limitations, such as those described by Bartlett (1937) or Anderson and Rubin (1956), although the associated benefits of these methods come at the cost of validity (for an overview, see DiStefano et al., 2009). The R package *psych* (Revelle, 2018) includes several functions that provide an easy way to extract composite scores from EFA, including all major extraction methods described herein.

Which procedure is optimal to compute composites differs according to the outcome measures included, their psychometric properties, and the goal of the study. Simple averages are straightforward to understand and interpret; scaled averages come with interesting properties for comparisons with other measures. Theoretically informed composites have important advantages over composites determined empirically, most notably a high robustness because they are less dependent on specific characteristics of the intervention data. Within data-driven approaches, coarse factor scores are often easier to compute and relatively robust, whereas refined factor scores require additional analytic decisions but provide an additional scaling component based on the correlation structure, similar to that of scaled averages. Irrespective of the method one decides to use, it is important to

note that the weights should be consistent between baseline and postintervention phases. If weights are altered between time points, the estimated constructs will differ in nature, which makes it difficult to assess whether change is genuine at the latent level or just an artifact of differences in composite estimation. Note that the increase in precision from single measures to a composite estimated from multiple measures is often substantial, whereas differences in performance between the various types of composites matter only in specific instances. As a result, using composites is key in our opinion, whereas the specific estimation method is less important as long as it can be justified.

Evaluating and Using Composites

When combining measures in intervention studies, it is important to make sure the composites are valid and reliable estimates of the underlying theoretical construct. Questions of validity can typically be addressed by examining the psychometric structure between the outcome measures that are being combined (i.e., the correlation between measures) at baseline. Construct estimation is valid only if the outcome variables that are aggregated are related, in the sense that they are the observable manifestation of a common factor. If the composite is based on robust theoretical information, its validity can be determined in advance; in cases in which the composite is based on the observed data, validity can be determined according to the specific aims of the intervention study and the norms of the field.

Problems often arise when composites include a variety of measures with no clear relation to one another, sometimes even defined a posteriori (Cordoba et al., 2010). If outcomes tap different underlying constructs than hypothesized (e.g., if an intelligence test taps substantially into a motivation construct; Duckworth et al., 2011), the interpretation of intervention outcomes is compromised. In this context, selecting adequate outcome measures at the onset is crucial (Coster, 2013), and guidelines exist to help investigators define composites that are sound and well defined (Chan et al., 2013; Moher et al., 2010). In addition to epistemological justifications, a number of statistical procedures can help decide on whether to combine measures into a construct (Raykov, 1997). As a rule of thumb, different measures are probably not tapping a common construct if they do not correlate relatively well with one another. Ideally, this should be determined according to known literature *and* the correlation matrix of the data. Theoretical and empirical approaches can typically inform each other: A firm grounding in theory is a necessary requirement when estimating a construct, yet confirming assumptions with empirical data, such as internal consistency measures

(Peterson & Kim, 2013), is often critical (Harwood et al., 2017; Raykov, 1997).

Another important component in the evaluation of composites relates to their reliability. One way to assess reliability is by comparing the psychometric structure between the outcome measures at baseline with the psychometric structure after the intervention. If the composite is a reliable estimate of the construct and the intervention does not solely modulate task-specific variance, the psychometric structure is likely to be very similar between the two time points. As illustrated in Figure 1, the extent to which this structure changes offers additional insight into the effects of the intervention on the construct of interest. For example, changes in the underlying construct are likely to affect all individual measures within a relatively preserved psychometric structure. In contrast, if the psychometric structure has changed drastically between the two time points, it is likely that the intervention has modulated task variance that is not related to the underlying construct. Generally, poor test-retest reliability within measures between baseline and postintervention sessions may invite caution in interpreting potential improvements. Ideally, this information should be provided in the form of a variance-covariance or correlation matrix.

After validity and reliability have been evaluated, composites can be substituted for individual measures in statistical analyses. Many intervention studies use *t* tests or repeated measures analyses of variance to compare change between baseline and postintervention time points across groups or conditions (e.g., treatment vs. control) or analyses of covariance to model differences in postintervention scores with baseline scores as a covariate (van Breukelen, 2013; Wright, 2006). Given that these models postulate that the effect of the intervention is the same for every participant—an almost definitely untenable assumption—the use of mixed models is generally preferable (Hilbert et al., 2019; McElreath, 2020) because they allow accounting for differences in change both at the group level and at the individual level. This property is especially important in the context of interventions because participants are typically exposed to a number of unsystematic variables that are not experimentally manipulated but can show substantial individual variability and thus greatly influence observed change in measured outcomes. Thus, models that account for this inherent variation can facilitate robust inferences with enhanced statistical power and fewer false positives (Baayen et al., 2008). Mixed models can also seamlessly accommodate unbalanced designs, designs with various time lengths between time points, missing data, nonlinear relationships, continuous covariates, or complex correlational structures among individual observations (Baayen et al., 2008; McElreath,

Table 1. Questions to Consider When Planning Interventions Involving Construct Estimation via Composite Scores

Step	Question	Section in preregistration template
Estimation	<ul style="list-style-type: none"> • Which construct will be estimated? • How many and which tests/tasks/measurements will be used to estimate the construct? • How will the weights be determined (e.g., using theoretical information, empirical information, or both)? 	Variables, Section 17.1 Variables, Section 17.1; Appendix Variables, Section 17.1; Appendix
Evaluation	<ul style="list-style-type: none"> • How will validity and reliability be determined? • How will the psychometric structure at baseline be evaluated? • What will be done to determine whether the psychometric structure was altered substantially by the intervention? 	Variables, Section 17.2 Variables, Section 17.2; Appendix Variables, Section 17.2
Application	<ul style="list-style-type: none"> • Which analyses will be conducted using the composite score? • How will psychometric properties affect interpretation of the results? • What heuristics (if any) will be used to interpret the results? (e.g., thresholds) 	Analysis Plan, Section 18.1; Appendix Analysis Plan, Section 18.3 Analysis Plan, Section 20.1

2020). Because increased model complexity comes with additional flexibility and degrees of freedom in the analyses, data sharing should be encouraged to allow additional analyses with varying parameters, thus ensuring greater impact within the scientific community.

Preregistering Composite Outcomes

The whole process of estimating, evaluating, and using composite outcomes should preferably be preregistered to prevent bias and undisclosed flexibility (Mellor & Nosek, 2018; Nosek et al., 2018). Many of the more general aspects of preregistration apply to psychological interventions using composite outcomes; however, there are also unique specificities that should be considered. To facilitate the preregistration of composites, we provide a template with examples in the context of psychological interventions (see online material at <https://osf.io/u96em/>). Table 1 summarizes typical questions at each of the three steps discussed in this article and the corresponding sections in the preregistration template.

Each step includes a number of questions that need to be addressed before the intervention starts. Some of these are rather straightforward, although very important (e.g., “Which construct will be estimated and how?” “Which analyses will be conducted?”); others require more detail. For example, the question of determining the weights for the composite should be carefully described. If the weights are determined on the basis of recommendations or reliable theoretical information, researchers simply need to disclose what the weights will be and on what information these are based, ideally with supporting references from the available literature. If, however, weights are determined empirically, great care should be taken to explain and justify decisions. In the case of EFA or PCA, preregistration should specify the number of factors or components that will be extracted³ (or the threshold that will be used to determine this

number, e.g., eigenvalue > X), the method that will be used for rotation (e.g., varimax, promax, oblimin, equamax, quartimax), the method that will be used to calculate factor scores (e.g., Thurstone, Bartlett, Anderson-Rubin), how missing values will be handled, and how EFA/PCA assumptions will be checked (e.g., outlier detection, multicollinearity). More generally, aspects such as validity and reliability should be discussed as well as their influence on subsequent statistical analyses. This can be done by inspecting variance-covariance or correlation matrices between measures, between measures across time points, and within measures across time points. Ideally, preregistration should be accompanied by supporting programming code detailing each of these steps, allowing transparent and reproducible results.

Limitations of Composite Outcomes

Although they typically represent improvements over single outcome measures, composites are no panacea. In clinical trials, in which composites have become very popular to increase design efficiency, a body of literature has also pointed out potential limitations. For example, composites can make it more difficult to identify sources of change or improvement (Freemantle & Calvert, 2007b; Ross, 2007) and can make variables difficult to interpret (Freemantle et al., 2003; Ross, 2007). By and large, these are not limitations of composites per se, however, but of their misuse in specific instances (Cannon, 1997; McCoy, 2018; Ross, 2007).

Other considerations perhaps warrant further consideration, however. In cases in which the intervention involves tests or tasks of a latent ability as part of the training regimen and measures of that same latent ability as an estimated construct at baseline and at postintervention—for example, when training working memory with the goal of improving working memory capacity—one should be

aware that gains are likely to reflect distortions in the psychometric structure of the tests (e.g., measurement-construct relationship) rather than actual improvements at the latent level (Shipstead et al., 2012). As a result, it is likely that the psychometric structure that has been reported in the relevant literature and observed at baseline is not representative of the structure at postintervention. These potential discrepancies should be directly investigated by examining variance-covariance or correlation matrices and be factored into the interpretation of findings—for example, important departures from known psychometric properties for a specific field should invite caution. Advanced modeling techniques, especially those from the SEM framework such as LCM and LCSM, can help in situations in which this is a concern (Moreau et al., 2016), although note that these methods are best suited to longitudinal designs with more than two time points (Duncan & Duncan, 2009) and have a number of other limitations (e.g., lack of transparency and interpretability), as mentioned previously.

More generally, composite scores can be influenced by the method used for estimation. This is especially true when estimation is purely data-driven, as in the case of EFA and PCA. Note that the estimation of composite scores via EFA is known to be indeterminate, in the sense that multiple solutions can fit the observed pattern of data equally well (the “indeterminacy problem”; see e.g., Grice, 2001). This aspect should be taken into account, and researchers should recognize its influence on the computation of composites. In addition, extracting a single factor or component from an EFA or PCA to represent a composite can be a questionable analytic choice in cases in which additional factors or components include a lot of information (i.e., variance) about the underlying measures (Greco et al., 2019).

Finally, assessing a construct with multiple outcomes can exacerbate common challenges in intervention studies, such as carryover effects (i.e., effects that persist from one experimental condition to another) or fatigue effects (i.e., decline in performance on an experimental task because of tiredness or boredom; see e.g., Pan et al., 1994). The former is often controlled with counterbalancing or with fixed testing order at all time points, whereas the latter is typically managed by limiting the total number of tests administered. The multiplicity of tests necessary to create composites can also make it difficult to blind participants to the main hypothesis: If a testing session includes a variety of tasks all measuring the same construct, participants are more likely to become aware of the main hypothesis being tested, or at least of the general trend expected (Boot et al., 2013). If left uncontrolled, for example when studies fail to match expectations across groups, improvements may reflect expectancy effects or participants’ awareness of

the hypothesis rather than intervention effects (e.g., Foroughi et al., 2016). These factors should be considered when designing intervention studies so that the use of multiple measures for construct estimation does not interfere with overall validity or reliability.

Practical Recommendations

Given the aforementioned advantages and limitations of multiple outcome measures in interventions, we would like to make five simple recommendations to help strengthen intervention designs in the field of psychology.

Recommendation 1: preregister primary outcomes and confirmatory hypotheses

Preregistration plays a key role in the context of interventions; it separates exploratory from confirmatory research and, within the latter, makes clear which outcomes are primary and which are secondary. Before starting the intervention, studies should preregister a primary outcome and should state confirmatory hypotheses explicitly. The primary outcome can be broad (e.g., general cognitive assessment) or rather narrow (e.g., working memory assessment). Other measures for which no clear prediction is made should be labeled as exploratory. Exploratory measures that appear to change as a result of the intervention should be preregistered as primary outcomes in subsequent confirmatory research to allow conclusive evidence.

Recommendation 2: assess change at the construct level

Postintervention change does not necessarily imply that the intervention has affected the hypothesized theoretical construct. In some instances, interventions can elicit improvements in test scores or increases in task performance by modulating variance specific to individual measures despite being designed to influence latent abilities, traits, or characteristics. To ensure claims of effectiveness are founded, postintervention changes should thus be assessed at the construct level, not at the level of individual outcome measures.

Recommendation 3: combine multiple outcome measures to create composites

Whenever possible, one should strive to combine multiple measures to test the primary hypothesis of an intervention. This approach is more likely to help detect genuine changes, especially when they are subtle and

embedded within complex, noisy systems. There are a number of ways to compute a composite score—scaling helps preserve distribution properties, allowing meaningful comparisons with the original variable units, whereas weighting helps ensure that valid and reliable variables are contributing to a greater extent to the composite than poorer measures. Different designs call for different procedures, which can produce different composites and affect subsequent analyses. Selection of the construct of interest and of its assessment should be based on theoretical understanding, ideally combined with empirical data, and should be set before the intervention begins.

Recommendation 4: assess validity and reliability of the composite

Because they often do not uniformly affect all outcome measures, interventions can alter the psychometric structure relating outcomes and constructs. This can affect the validity and reliability of the composite. To help gauge the extent to which an intervention perturbed known psychometric structures, authors should report psychometric properties for all testing sessions in the form of variance-covariance or correlation matrices for all outcome variables. This information gives readers an indication of the validity of the outcome measures in the specific context of the intervention and allows putting claims of improvement in context. In addition, test-retest reliability should also be reported—this information helps determine whether change is fairly uniform across individuals or whether it has affected some individuals more than others. In the latter case, attention should be paid to ensure validity has been preserved. Together, indices of validity and reliability help determine whether improvements are likely to be task-specific or reflect broader change at the construct level.

Recommendation 5: follow general guidelines for intervention studies

Finally, a number of suggestions have been made elsewhere about the importance of implementing interventions with adequate experimental designs (Boot et al., 2013; Moreau & Conway, 2014; Shipstead et al., 2012; Simons et al., 2016) or with respect to the use of appropriate statistical techniques for the analysis of intervention outcomes (McArdle, 2009; Moreau et al., 2016). More generally, it is essential to follow best practices in the design of interventions such as those outlined in the SPIRIT statement (Chan et al., 2013), which provides a number of suggestions for the type of information that needs to be included in clinical trial protocols, as well as guidelines for reporting clinical trials and psychological

interventions such as the Consolidated Standards of Reporting Trials statement (Schulz et al., 2011) and the extension for social and psychological interventions (Grant et al., 2018; Montgomery et al., 2018). These documents summarize recommendations to alleviate common problems in the reporting of randomized controlled trials and allow standardized communication for greater impact and outreach. The Journal Article Reporting Standards for Quantitative Research in Psychology (<https://apastyle.apa.org/jars/quantitative>) can also provide general methodological guidelines of relevance for interventions.

Concluding Remarks

In this article, we have shown that the multiplicity of outcomes measures in intervention research comes with a number of advantages and limitations that are important to consider. When assessed separately, using multiple measures can lead to inflated error rate or loss of statistical power, but it can also enable construct estimation via composite outcomes pooling individual measures together. In our view, this approach has undeniable strengths that should be harnessed in intervention research. We have discussed a number of methods to compute, evaluate, and use composites and provided practical recommendations tailored to the field of psychology and a visual, interactive tool to help build an intuition for the benefits of this approach. We hope this contribution can facilitate the design of valid and informative interventions in psychological research.

Transparency

Action Editor: Alexa Tullett

Editor: Daniel J. Simons

Author Contributions

D. Moreau developed the idea for the manuscript and wrote the initial draft. D. Moreau and K. Wiebels programmed the simulations and examples and created the Shiny app. Both authors collaboratively revised the manuscript and approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

D. Moreau is supported by a Marsden grant from the Royal Society of New Zealand, funding from the Neurological Foundation of New Zealand, and a University of Auckland Early Career Research Excellence Award. K. Wiebels is supported by a Kate Edger Educational Charitable Trust Dame Dorothy Winstone Doctoral Completion Award.

Open Practices

Open Data: not applicable

Open Materials: <https://osf.io/u96em/>

Preregistration: not applicable

All materials have been made publicly available via OSF and can be accessed at <https://osf.io/u96em/>. The Shiny app is available at https://kwiebels.shinyapps.io/Multiple_outcomes_in_interventions/. This article has received the badge for Open Materials. More information about the Open Practices badges can be found at <http://www.psychologicalscience.org/publications/badges>.



ORCID iDs

David Moreau  <https://orcid.org/0000-0002-1957-1941>
 Kristina Wiebels  <https://orcid.org/0000-0002-5360-5965>

Notes

- Note that it is also possible to use confirmatory factor analysis (CFA) to incorporate the hypothesized psychometric structure among variables into the empirical analysis. Thus, it can be viewed as an intermediate approach, between fully theoretical composites and those determined statistically. CFA is not discussed further in the present article because it does not differ fundamentally from EFA for the purpose of creating composites and is not commonly used in this way. When reliable theoretical information is available to inform the weights of the composite, approaches such as standardized battery indices or scaled averages should typically be preferred.
- Alternatively, it is also possible to factor the correlation matrix rather than the covariance matrix, which achieves the same goal of standardization.
- In certain cases, a researcher might intend to extract more than one factor, for example when the intervention is thought to influence more than one underlying construct. Ideally, this should be determined in advance, based on theoretical knowledge, and confirmed with a CFA.

References

- Anderson, T. W., & Rubin, H. (1956). Statistical inference in factor analysis. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 5, pp. 111–150). California University Press.
- Ard, M. C., Raghavan, N., & Edland, S. D. (2015). Optimal composite scores for longitudinal clinical trials under the linear mixed effects model. *Pharmaceutical Statistics, 14*, 418–426.
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.
- Bakal, J. A., Westerhout, C. M., & Armstrong, P. W. (2015). Impact of weighted composite compared to traditional composite endpoints for the design of randomized controlled trials. *Statistical Methods in Medical Research, 24*, 980–988.
- Bartlett, M. S. (1937). The statistical conception of mental factors. *British Journal of Psychology, 28*(1), 97–104.
- Blanken, T. F., Van Der Zweerde, T., Van Straten, A., Van Someren, E. J. W., Borsboom, D., & Lancee, J. (2019). Introducing network intervention analysis to investigate sequential, symptom-specific treatment effects: A demonstration in co-occurring insomnia and depression. *Psychotherapy and Psychosomatics, 88*, 52–54.
- Bolier, L., Haverman, M., Westerhof, G. J., Riper, H., Smit, F., & Bohlmeijer, E. (2013). Positive psychology interventions: A meta-analysis of randomized controlled studies. *BMC Public Health, 13*, Article 119. <https://doi.org/10.1186/1471-2458-13-119>
- Boot, W. R., Simons, D. J., Stothart, C., & Stutts, C. (2013). The pervasive problem with placebos in psychology: Why active control groups are not sufficient to rule out placebo effects. *Perspectives on Psychological Science, 8*, 445–454.
- Borsboom, D. (2008). Latent variable theory. *Measurement: Interdisciplinary Research and Perspectives, 6*(1–2), 25–53.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*, 203–219.
- Cannon, C. P. (1997). Clinical perspectives on the use of composite endpoints. *Controlled Clinical Trials, 18*, 517–529; discussion 546–549.
- Chan, A.-W., Tetzlaff, J. M., Gøtzsche, P. C., Altman, D. G., Mann, H., Berlin, J. A., Dickersin, K., Hróbjartsson, A., Schulz, K. F., Parulekar, W. R., Krleža-Jerić, K., Laupacis, A., & Moher, D. (2013). SPIRIT 2013 explanation and elaboration: Guidance for protocols of clinical trials. *BMJ, 346*, Article e7586. <https://doi.org/10.1136/bmj.e7586>
- Chooi, W.-T., & Thompson, L. A. (2012). Working memory training does not improve intelligence in healthy young adults. *Intelligence, 40*, 531–542.
- Colom, R., Román, F. J., Abad, F. J., Shih, P. C., Privado, J., Froufe, M., Escorial, S., Martínez, K., Burgaleta, M., Quiroga, M. A., Karama, S., Haier, R. J., Thompson, P. M., & Jaeggi, S. M. (2013). Adaptive n-back training does not improve fluid intelligence at the construct level: Gains on individual tests suggest that training may enhance visuospatial processing. *Intelligence, 41*(5), 712–727. <https://doi.org/10.1016/j.intell.2013.09.002>
- Cordoba, G., Schwartz, L., Woloshin, S., Bae, H., & Gøtzsche, P. C. (2010). Definition, reporting, and interpretation of composite outcomes in clinical trials: Systematic review. *BMJ, 341*, Article c3920. <https://doi.org/10.1136/bmj.c3920>
- Coster, W. J. (2013). Making the best match: Selecting outcome measures for clinical trials and outcome studies. *The American Journal of Occupational Therapy, 67*, 162–170.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302.
- Cybulski, L., Mayo-Wilson, E., & Grant, S. (2016). Improving transparency and reproducibility through registration: The status of intervention trials published in clinical psychology journals. *Journal of Consulting and Clinical Psychology, 84*, 753–767.
- Davis, M. J., & Addis, M. E. (1999). Predictors of attrition from behavioral medicine treatments. *Annals of Behavioral Medicine, 21*, 339–349.
- DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research & Evaluation, 14*(20), 1–11.
- Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in

- intelligence testing. *Proceedings of the National Academy of Sciences, USA*, 108, 7716–7720.
- Duncan, T. E., & Duncan, S. C. (2009). The ABC's of LGM: An introductory guide to latent variable growth curve modeling. *Social and Personality Psychology Compass*, 3, 979–991.
- Fedoroff, I. C., & Taylor, S. (2001). Psychological and pharmacological treatments of social phobia: A meta-analysis. *Journal of Clinical Psychopharmacology*, 21, 311–324.
- Flake, J. K., & Fried, E. I. (2019). *Measurement schmeasurement: Questionable measurement practices and how to avoid them*. PsyArXiv. <https://doi.org/10.31234/osf.io/hs7wm>
- Foroughi, C. K., Monfort, S. S., Paczynski, M., McKnight, P. E., & Greenwood, P. M. (2016). Placebo effects in cognitive training. *Proceedings of the National Academy of Sciences, USA*, 113, 7470–7474.
- Freemantle, N., & Calvert, M. (2007a). Composite and surrogate outcomes in randomised controlled trials [Review of composite and surrogate outcomes in randomised controlled trials]. *BMJ*, 334, 756–757.
- Freemantle, N., & Calvert, M. (2007b). Weighing the pros and cons for composite outcomes in clinical trials. *Journal of Clinical Epidemiology*, 60, 658–659.
- Freemantle, N., & Calvert, M. J. (2010). Interpreting composite outcomes in trials. *BMJ*, 341, Article c3529. <https://doi.org/10.1136/bmj.c3529>
- Freemantle, N., Calvert, M., Wood, J., Eastaugh, J., & Griffin, C. (2003). Composite outcomes in randomized trials: Greater precision but with greater uncertainty? *JAMA*, 289, 2554–2559.
- Glass, G. V., & Maguire, T. O. (1966). Abuses of factor scores. *American Educational Research Journal*, 3, 297–304.
- Gorsuch, R. L. (2014). *Factor analysis: Classic edition*. Routledge.
- Grant, S., Mayo-Wilson, E., Montgomery, P., Macdonald, G., Michie, S., Hopewell, S., & Moher, D., on behalf of the CONSORT-SPI Group. (2018). CONSORT-SPI 2018 explanation and elaboration: Guidance for reporting social and psychological intervention trials. *Trials*, 19, 406.
- Greco, S., Ishizaka, A., Tasiou, M., & Torrisi, G. (2019). On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Social Indicators Research*, 141(1), 61–94.
- Greene, T., Gelkopf, M., Epskamp, S., & Fried, E. (2018). Dynamic networks of PTSD symptoms during conflict. *Psychological Medicine*, 48, 2409–2417.
- Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, 6, 430–450.
- Grice, J. W., & Harris, R. J. (1998). A comparison of regression and loading weights for the computation of factor scores. *Multivariate Behavioral Research*, 33, 221–247.
- Guyon, H., Falissard, B., & Kop, J.-L. (2017). Modeling psychological attributes in psychology—An epistemological discussion: Network analysis vs. latent variables. *Frontiers in Psychology*, 8, Article 798. <https://doi.org/10.3389/fpsyg.2017.00798>
- Harwood, J. M., Weiss, R. E., & Comulada, W. S. (2017). Beyond the primary endpoint paradigm: A test of intervention effect in HIV behavioral intervention trials with numerous correlated outcomes. *Prevention Science*, 18, 526–533.
- Hilbert, S., Stadler, M., Lindl, A., Naumann, F., & Bühner, M. (2019). Analyzing longitudinal intervention studies with linear mixed models. *TPM: Testing, Psychometrics, Methodology in Applied Psychology*, 26(1). <https://www.tpm.org/product/analyzing-longitudinal-intervention-studies-with-linear-mixed-models/>
- Jaeggi, S. M., Buschkuhl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences, USA*, 105, 6829–6833.
- Johnson, M. O., & Remien, R. H. (2003). Adherence to research protocols in a clinical context: Challenges and recommendations from behavioral intervention trials. *American Journal of Psychotherapy*, 57, 348–360.
- Kievit, R. A., Brandmaier, A. M., Ziegler, G., van Harmelen, A.-L., de Mooij, S. M. M., Moutoussis, M., Goodyer, I. M., Bullmore, E., Jones, P. B., Fonagy, P., NSPN Consortium Lindenberger, U., & Dolan, R. J. (2018). Developmental cognitive neuroscience using latent change score models: A tutorial and applications. *Developmental Cognitive Neuroscience*, 33, 99–117. <https://doi.org/10.1016/j.dcn.2017.11.007>
- Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford.
- Mayo-Wilson, E., Dias, S., Mavranzouli, I., Kew, K., Clark, D. M., Ades, A. E., & Pilling, S. (2014). Psychological and pharmacological interventions for social anxiety disorder in adults: A systematic review and network meta-analysis. *The Lancet Psychiatry*, 1, 368–376.
- McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology*, 60, 577–605.
- McCoy, C. E. (2018). Understanding the use of composite endpoints in clinical trials. *The Western Journal of Emergency Medicine*, 19, 631–634.
- McCrae, R. R., & Costa, P. T., Jr. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52(1), 81–90.
- McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and STAN*. CRC Press.
- Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of “far transfer”: Evidence from a meta-analytic review. *Perspectives on Psychological Science*, 11, 512–534.
- Mellor, D. T., & Nosek, B. A. (2018). Easy preregistration will benefit any research. *Nature Human Behaviour*, 2(2), Article 98. <https://doi.org/10.1038/s41562-018-0294-7>
- Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M., & Altman, D. G.; Consolidated Standards of Reporting Trials Group. (2010). CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *Journal of Clinical Epidemiology*, 63(8), e1–e37. <https://doi.org/10.1016/j.jclinepi.2010.03.004>
- Montgomery, P., Grant, S., Mayo-Wilson, E., Macdonald, G., Michie, S., Hopewell, S., & Moher, D. (2018). Reporting randomised trials of social and psychological

- interventions: The CONSORT-SPI 2018 extension. *Trials*, 19(1), 1–14.
- Moody, D. E. (2009). Can intelligence be increased by training on a task of working memory? *Intelligence*, 37, 327–328.
- Moreau, D., & Conway, A. R. A. (2014). The case for an ecological approach to cognitive training. *Trends in Cognitive Sciences*, 18, 334–336.
- Moreau, D., Kirk, I. J., & Waldie, K. E. (2016). Seven pervasive statistical flaws in cognitive training interventions. *Frontiers in Human Neuroscience*, 10, Article 153. <https://doi.org/10.3389/fnhum.2016.00153>
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, USA*, 115, 2600–2606.
- Pan, C. S., Shell, R. L., & Schleifer, L. M. (1994). Performance variability as an indicator of fatigue and boredom effects in a VDT data-entry task. *International Journal of Human-Computer Interaction*, 6(1), 37–45.
- Peterson, R. A., & Kim, Y. (2013). On the relationship between coefficient alpha and composite reliability. *The Journal of Applied Psychology*, 98, 194–198.
- Pilling, S., Bebbington, P., Kuipers, E., Garety, P., Geddes, J., Orbach, G., & Morgan, C. (2002). Psychological treatments in schizophrenia: I. Meta-analysis of family intervention and cognitive behaviour therapy. *Psychological Medicine*, 32, 763–782.
- Pocock, S. J. (1997). Clinical trials with multiple outcomes: A statistical perspective on their design, analysis, and interpretation. *Controlled Clinical Trials*, 18, 530–545; discussion 546–549.
- Pocock, S. J., Geller, N. L., & Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics*, 43(3), 487–498. <https://doi.org/10.2307/2531989>
- Proust-Lima, C., Philipps, V., Dartigues, J.-F., Bennett, D. A., Glymour, M. M., Jacqmin-Gadda, H., & Samieri, C. (2019). Are latent variable models preferable to composite score approaches when assessing risk factors of change? Evaluation of type-I error and statistical power in longitudinal cognitive studies. *Statistical Methods in Medical Research*, 28, 1942–1957.
- Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, 21, 173–184.
- Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., Kane, M. J., & Engle, R. W. (2013). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. *Journal of Experimental Psychology. General*, 142(2), 359–379.
- Revelle, W. (2018). *Psych: Procedures for personality and psychological research*. Northwestern University.
- Ross, S. (2007). Composite outcomes in randomized clinical trials: Arguments for and against. *American Journal of Obstetrics and Gynecology*, 196(2), 119.e1–e6.
- Schmiedek, F., Lövdén, M., & Lindenberger, U. (2010). Hundred days of cognitive training enhance broad cognitive abilities in adulthood: Findings from the COGITO study. *Frontiers in Aging Neuroscience*, 2, Article 27. <https://doi.org/10.3389/fnagi.2010.00027>
- Schulz, K. F., Altman, D. G., & Moher, D., & CONSORT Group. (2011). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *International Journal of Surgery*, 9, 672–677.
- Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin*, 138, 628–654.
- Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. L. (2016). Do “brain-training” programs work? *Psychological Science in the Public Interest*, 17, 103–186.
- Sin, N. L., & Lyubomirsky, S. (2009). Enhancing well-being and alleviating depressive symptoms with positive psychology interventions: A practice-friendly meta-analysis. *Journal of Clinical Psychology*, 65, 467–487.
- Spearman, C. (1904). “General intelligence,” objectively determined and measured. *The American Journal of Psychology*, 15, 201–292.
- Takacs, Z. K., & Kassai, R. (2019). The efficacy of different interventions to foster children’s executive function skills: A series of meta-analyses. *Psychological Bulletin*, 145, 653–697.
- Thurstone, L. L. (1935). *The vectors of mind: Multiple-factor analysis for the isolation of primary traits*. Ulan Press.
- Tomarken, A. J., & Waller, N. G. (2005). Structural equation modeling: Strengths, limitations, and misconceptions. *Annual Review of Clinical Psychology*, 1, 31–65.
- van Bork, R., Rhemtulla, M., Waldorp, L. J., Kruis, J., Rezvanifar, S., & Borsboom, D. (2019). Latent variable models and networks: Statistical equivalence and testability. *Multivariate Behavioral Research*. Advance online publication. <https://doi.org/10.1080/00273171.2019.1672515>
- van Breukelen, G. J. P. (2013). ANCOVA versus CHANGE from baseline in nonrandomized studies: The difference. *Multivariate Behavioral Research*, 48, 895–922.
- Wechsler, D. (1949). *Wechsler Intelligence Scale for Children*. Pearson.
- Wechsler, D. (2008). *Wechsler Adult Intelligence Scale—Fourth edition (WAIS-IV)* (Vol. 22). NCS Pearson.
- Weintraub, W. S. (2016). Statistical approaches to composite endpoints. *JACC: Cardiovascular Interventions*, 9, 2289–2291.
- Wright, D. B. (2006). Comparing groups in a before-after design: When t test and ANCOVA produce different results. *The British Journal of Educational Psychology*, 76(Pt. 3), 663–675.
- Yong, A. G., & Pearce, S. (2013). A beginner’s guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, 9(2), 79–94.