Constructing Composites to Optimise Cognitive Outcomes

The relative insensitivity of traditional cognitive outcome measures to describe the more subtle and selective cognitive impairment associated with neurologic disorders such as mild cognitive impairment/prodromal Alzheimer's disease, and psychiatric disorders such as depression, schizophrenia and ADHD, as well as track treatment-related changes, has resulted in a recent boon in cognitive composite measures. Composites such as these are typically created in an effort to reduce Type 1 error by reducing the number of outcome measures to a more manageable level, and ultimately to improve signal detection by being more sensitive to disease state and treatment effects while reducing sample size. Composite endpoints characteristically have several other advantages, including being more highly correlated with putative biomarkers such as neuroimaging and CSF measures, and being better at predicting disease progression. Only rarely are composite measures employed to guarantee that appropriate cognitive domains of interest are sampled in a practical and efficient manner, ensuring adequate psychometric properties (such as sufficient reliability and avoiding celling/floor effects); or employed as a method to characterise the cognitive profile of a drug in an a priori fashion that is associated with a disease state longitudinally and/or with treatment intervention. While it is relatively easy for many clinical triallists to acknowledge that specific cognitive domains are more likely to be associated with particular CNS conditions,[1] few appreciate that even widely recognised cognitive enhancers typically affect multiple cognitive domains: preferentially improving some domains while possibly causing impairments in others, even against a backdrop of improved overall cognitive function. One method for ensuring that this variability is adequately captured is through the proficient construction and analysis of cognitive composite measures.

Although many researchers use the terms composite score and summary score interchangeably, composites differ from summary scores in that composite typically represent small sets of data points that are highly related to one another, both conceptually and – importantly – statistically in terms of collinearity. As such, reducing several cognitive outcome measures to a single composite essentially reduces the amount of information representing a single underlying construct. Although there can be room for improvisation, composites are typically based on well-established methodologies and are calculated in a very fixed and consistent manner using standard analytic tools. On the other hand, summary scores often combine many different types of measures into a single unified score, even though these measures may be related to various outcomes and several underlying constructs that are often unrelated. This flexibility permits the combination of data across several theoretical constructs to gain a wider appreciation of variable domains.

Although there are advantages to using both summary and composites scores as noted above, there are also relevant disadvantages including the fact that these derived scores may mask important differences apparent in individual component scores. Even with well-constructed composites, it is highly unlikely that all component measures will be equally reliable, have equal variances, be equally inter-correlated and be equally correlated with the underlying construct which the composite is attempting to measure.[2] In order to remedy this issue, many researchers have chosen to weight individual components of the composite score.

Weighting Composites

Although the weighting process can be very formal or quite arbitrary, the goal of differential weighting of composite scores should always be to improve the reliability of the composite and provide more valid and useful composites than obtained when component measures are simply summed and averaged, with all components having equal weight. The latter type of weighting is often referred to as using *natural weights*, in which raw scores are simply summed or averaged to form a composite measure where differences in the variances of component variables and differences in their inter-correlations determine the weights. On the other hand *a priori weights* can be assigned on the basis of judgments or ratings, or based on more empirical analytic methods. For example, weights may be chosen to maximise certain internal criteria such as the reliability of the composite measure. In this case, more weight is given to components with higher reliability and less weight to those with lower reliability. However, one of the most common methods is to weight components is by maximising the validity of the composite in relation to a pre-specified external criterion using multiple regression techniques, which includes canonical variate analysis, principal component analysis, maximum reliability and canonical factor analysis.

Multiple regression methods provide a set of weights optimal for minimising the error of prediction for the group on which the weights are derived under certain assumptions of normality and linearity. In the standard multiple regression equation predictors maximise the correlation between the composite score and the actual criterion, and can be used to derive the actual weights. The linear combination of the component scores can also be used to derive composite scores that maximise the correlation between the external criterion and the composite.[3] Although it is relatively easy to calculate these weights, researchers caution that using multiple regression requires considerable thoughtfulness in interpretation as predictor weights are those which maximise the multiple correlation R, within the sample from which they were derived.[4] However, typically these weights are derived in one sample and then applied to another sample in a prospective manner, which has more often than not resulted in poorer performance of that composite measure on the new sample. One method to avoid this problem is simply to utilise the multiple regression methodology on each specific sample under investigation in a blinded manner by using screening or baseline before treatment intervention. This would guarantee that the weights would be fully applicable to the study sample and, with enough patients, these composites would likely generalise to other samples of similar patients from which they were derived.

Current Cognitive Composites

Many cognitive composites are based on observational neuropsychological test data from measures that were originally intended for use in other populations. Alternatively, composites can be based on a theory of neuropsychological dysfunction specific to a disease state based on well accepted principles of neuropsychological function and localisation.[1]

Of course, these principles can change over time with increased understanding, but oftentimes the cognitive domains fall into one of several well-known categories including episodic memory, executive function, motor function and language. The actual cognitive test variables that comprise each of these domains have multiple outcome variables associated with them, and some of these outcome variables are better suited to one cognitive domain versus another. For example, although total delayed recall of a word list may conceptually best fit into a cognitive domain representing episodic memory, the number of intrusions or perseverative errors made on that task, and derived measures of signal detection may be better suited for inclusion into the executive function domain. Decisions regarding appropriate assignment such as this are subjective and based not only on the idiosyncratic group of tests employed but also on the attribution of these tests to a specific cognitive domain.

For instance, the agreement regarding essential cognitive domains required to adequately assess the well accepted construct of Cognitive Impairment Associated with Schizophrenia (CIAS) came about only after a very prolonged process facilitated by the RAND corporation to drive consensus amongst numerous stakeholders including regulatory bodies, the NIH, an assortment of key opinion leaders from academia, and various industry leaders.[5] This herculean effort resulted in the cognitive battery called the Measurement and Treatment Research to Improve Cognition in Schizophrenia (MATRICS) battery which was designed to help facilitate a methodology for developing and registering potential nootropic agents in schizophrenic populations. Neuropsychological test expert vary appreciably regarding the assignment of any given outcome measure into a single cognitive domain, with many choosing instead to weight each outcome variable into several categories as most cognitive tests involve complex attention, psychomotor speed, and language minimally. Even test authors and testing companies have suggested alternate views on what exactly each outcome variable is intended to measure. In short, it is widely agreed that although cognitive tests are often reported to be sensitive to a single cognitive domain, their measurement always includes variance associated with other common cognitive and non-cognitive factors. Even the well vetted and agreed upon MATRICS cognitive battery continues to evolve from data acquired from various trials regarding its psychometric properties and clinical utility.

Unlike the mammoth MATRICS initiative, composite measures related to the cognitive domains associated with early AD (MCI, prodromal AD) and mild AD have been taken on in a less onerous fashion, with several individual pharma companies and public-private partnerships leading the way. Some of these, such as the ADComs, TriAD and ProAD are based on a subset of measures typically given in later stages of illness such as the ADASCog, MMSE and CDR, while other composites are based on a prior group of standardised neuropsychological tests classically administered across a host of different patient populations. In the earlier example, researchers started with standard AD scales and selected items that were most relevant to earlier stage patients, choosing items that exhibited a large decline and low variability via a variety of techniques, such as item analyses, partial least squares regression and multiple correlations.

One of the most notable novel composites is based on data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and applies psychometric methods to various cognitive tests utilised across approximately 800 subjects in a series of studies. Outcome measures related to many cognitive tests including the ADAS-Cog, the Rey Auditory Verbal Learning Test (RAVLT), Category Fluency, Trail Making, Clock Drawing, Digit Span and Logical Memory were available for analysis. Of note, ADNI authors reviewed the entire baseline ADNI neuropsychological battery to identify items which could be considered indicators of either executive function (EF) or memory (MEM), both known to be important in early AD, then refined item selection using an iterative process in which they constructed a model using confirmatory factor analysis, reviewed findings as a group, and then constructed a revised model.[6,7] Specifically, a confirmatory factor analysis played an important role in composite development. In the case of the EF composite, the single factor model was not a good representation of the data and a bi-factor structure that included a secondary domain for correlations between category fluency items, and that included a methods factor for the clock drawing items, produced a better measure of model fit. The researchers utilised several statistical techniques in order to assess the fit of the model, including the confirmatory fit index (CFI), the Tucker Lewis Index (TLI) and the root mean squared error of approximation (RMSEA).

Importantly, the team then compared ADNI-EF with individual component measures in 390 subjects with mild cognitive impairment (MCI) with respect to the composite's ability to detect change over time; to predict conversion to dementia; to be correlated with MRI-derived measures of structures involved in frontal systems; and with cerebrospinal fluid (CSF) levels of amyloid β1–42, total tau, and phosphorylated tau. The ADNI-EF composite showed the greatest changes over time, followed closely by the component category fluency measure but notably the ADNI-EF composite required a 40% smaller sample size to detect change. The ADNI-EF composite was also the strongest predictor of conversion to AD and was the only measure significantly associated with all of the frontal regions on MRI. However, other measures were more strongly associated in a few other regions and with CSF measures. Thus ADNI-EF appears to be a useful composite measure of EF in MCI, as good as or better than any of its component parts.[6]

Not surprisingly, given the pattern of early cognitive deficits reported in the literature favouring executive dysfunction over memory, the ADNI-MEM composite did not fare as well, performing only slightly better at detecting change than total RAVLT recall scores. However, ADNI-MEM did do as well as or better than its component scores at predicting conversion from MCI to AD, and was associated with all selected imaging parameters. As noted by several researchers, MCI subjects who exhibited the characteristic AD CSF signature of high tau and low beta amyloid in this study also exhibited a more rapid decline than did those without such a CSF signature, and although all of the component cognitive measures suggested faster rates of decline among subjects with the CSF signature, the difference was largest for the ADNI-MEM composite.[7]

Constructing Novel Cognitive Composites

As many existing clinical trials of MCI, prodromal AD and mild AD utilise a large and varied number of neuropsychological tests with innumerable associated variables, it is virtually impossible for drug developers to choose a single test or test item most likely to show changes associated with treatment. Of course, selecting multiple measures welcomes criticism related to multiplicity and inflating Type 1 error, or rejecting a true null (also known as a false positive error). However, one crucial advantage of utilising composites is that it helps to control for Type I error, given the large number of neuropsychological tests and associated outcome measures.

An exploratory factor analysis can be utilised to uncover the underlying structure of a relatively large set of variables and allow the researcher to select which test variables should enter into a composite in a totally objective fashion. This differs from a confirmatory factor analysis which is used to verify the de facto structure of a set of observed variables created by more subjective means such as expert opinion. Either of these will help to reduce data efficiently, but only exploratory efforts will fully allow the data to be unconfined. Importantly, before conducting any factor analyses it is imperative that all of the raw data be transformed to a common metric based on normative data. This does not require the use of normal healthy cohorts to be part of the study but does demand access to normative data. Unfortunately, for some novel cognitive measures this data may be lacking.

Transforming data into a common metric based on standard scores such as Z scores helps to ensure that the psychometric properties of the various components that comprise the composite have similar psychometric properties, keeping in mind that the goal is to include measures that are highly related to one another, based on the test properties influencing reliability and on true score variance. Chapman and Chapman[8] suggest that true score variance is also influenced by other test parameters, such as the item difficulty and the number of items which can vary greatly across component tests. For example, important differences in difficulty levels have long been noted between versions of RAVLT, with one version being systematically more difficult than others. Failing to account for differences in things like difficulty level (probably caused in this specific case by differences in word frequency, imagery, number of syllables and serial position of certain words in the list) could result in a misinterpretation of results when standard scores are used without knowledge of the test's discriminatory power.[2] Even though component tests may be similar in content and length, alternate versions of these tests are only considered equivalent if they have the similar means, variances and discriminatory power. Additionally, alternate forms of cognitive tests are required in longitudinal settings in order to help reduce practice effects. Even if the alternate forms are psychometrically equivalent, practice effects can still be attributable to the general testing factors that arise from repeated exposure to the same type of task and not just the specific content of the tests. There are several methods for equating alternate forms of tests that do not have similar psychometric properties. One method referred to as equipercentile equating is accomplished by identifying the subject's scores on two measures with the same percentile rank and transforming the score on a new test to the corresponding score on the reference with the same percentile rank.[9]

In order to calculate these Z transformed standard scores and put outcomes in a common metric, the normative mean (which can be taken from prior observations or from published data) is simply subtracted from each subject's component test score and this difference is divided by the standard deviation for the appropriate normative sample which sometimes varies by age and education. These Z scores can then be entered into principal components or exploratory factor analysis, followed by varimax rotation, in order to yield orthogonal or independent factors with an eigenvalue greater than one. The eigenvalue for a given factor measures the variance in all the variables which is accounted for by that factor. These factors can then be employed as unit weighted cognitive composites labelled to reflect various cognitive domains such as memory, executive function or motor speed, which typically comprise individual test items with factor weights above at least 0.63 (*very good*) or 0.71 (*excellent*) being included based on sample size. Each factor would be considered a cognitive composite.

Assessing the Reliability of Cognitive Composites

As a general rule, a factor/composite would be seen as reliable if it has four or more loadings of at least 0.6, regardless of sample size. The top three to five factors/composites are typically presented in the order in which they were generated from the factor analysis, with those with the highest factor loadings presented first. Each factor/composite can be labelled based on its constituent parts by the researcher, provided a single label can capture the loadings appropriately. Of note, there will likely be numerous individual neuropsychological measures that are included in the original evaluation but are not used for calculating factor/composites as they did not load on any of the factors with a high enough factor loading. This does not preclude them from independent analyses.

It is also important to determine how well the composite performs and specifically how closely related the individual test items in a cognitive composite are to each other, or how well these reflect a unitary construct. This is akin to a reverse engineering of the factor analysis. To do this, a simple coefficient alpha (Cronbach's alpha[10]) can be used to provide a measure of the internal consistency assessing the reliability of the composite, with "higher" values implying good internal consistency but not unidimensionality. Obviously other tools that indicate the fit of the model include the confirmatory fit index (CFI), the Tucker Lewis Index (TLI) and the root mean squared error of approximation (RMSEA) as noted earlier.[6,7] Once the composite is found to be reliable and valid, future studies can be conducted using fewer measures on similar patient samples, greatly decreasing subject and site burden without sacrificing power. More importantly, a grand initiative such as MATRICS, requiring years and hundreds of patients to construct these composites is not necessary, as reliable composite construction is well within the scope of a Phase II study in terms of time, number of patients and cost.

Analysing Treatment Effects on Cognitive Composites

For most researchers, simply constructing the cognitive composite is not enough and there is a need to compare these composites in a rigorous manner in order to draw conclusions not just about how well composites characterise the cognitive performance of patients but how these are differentially affected by drug treatment. In order to assess this, the Z scores corresponding to each individual cognitive measure that were included in the formation of the factor/composite can simply be summated, averaged and compared statistically across factor/composites. An average of all of the domains would reflect a global summary score.

This Z transformed data for each cognitive composite can be analysed in a manner similar to that for non-transformed data through the use of multivariate statistics. Importantly, this type of analysis permits a shape or profile analysis that can help determine if treatment affects one cognitive composite (e.g., executive function) to a greater degree than any or all of the others. If there is no difference across cognitive composites, there would be no difference from zero (corresponding to the mean of the normative data) and this would be represented by a flat line across composite measures. However, if a non-flat line is apparent, a significant within-subject profile shape can be tested for via standard MANOVA or MMRM techniques that detect significant differences between drug and placebo groups for all cognitive composites at baseline and over time. A significant profile shape by treatment group interaction could then be decomposed using univariate ANOVA techniques and Bonferroni corrected t-tests. If desired measure of premorbid intellectual functioning can be used as a covariate to control for general intelligence, which has been shown to correlate highly with performance on almost all cognitive measures.

Revealing a treatment by profile shape interaction would represent an achievement which could not be realised when examining a sole composite. The power of the shape analysis stems from the fact that the chosen cognitive composite will benefit differentially from treatment (whether improving, showing a more gradual decline over time or even a worsening for drugs associated with cognitive dysfunction). Ensuring that the analysis contains *control* composites reflecting premorbid intellectual function, language or motor function which should not be associated with practice or treatment-related changes lends validity to the composite of interest. In order to assess the potential contribution of practice effects on change scores (from baseline to follow-up), a series of univariate analyses can be conducted for each cognitive composite, keeping in mind that in early Alzheimer's patient samples, some drug treatments may facilitate a practice effect that may be neutral in the placebo group which would not benefit from practice.

REFERENCES

1. Millan MJ, Agid Y, Brüne M, Bullmore ET et al. Cognitive dysfunction in psychiatric disorders: characteristics, causes and the quest for improved therapy. Nat Rev Drug Discov. 2012 Feb 1;11(2):141-68.
2. Kang SS, MacDonald AW. Limitations of true score variance to measure discriminating power: psychometric simulation study. Abnorm Psychol. 2010 May;119(2):300-6.
3. He Q. Estimating the Reliability of Composite Scores. The Office of Qualifications and Examinations Regulation in 2010. Qualifications and Curriculum Authority 2010. pp 1-39.
4. Wang, M., Stanley J. Differential Weighting: A Review of Methods and Empirical Studies. Review of Educational Studies, 1970: Vol 40 (5) pp 663-705.
5. Breier, A. Developing Drugs for Cognitive Impairment in Schizophrenia. Schizophrenia Bulletin 2005; vol. 31 no. 4 pp. 816–822.
6. Gibbons LE, Carle AC, Mackin RS, Harvey D, Mukherjee S, Insel P, Curtis SM, Mungas D, Crane PK. Alzheimer's Disease Neuroimaging Initiative. A composite score for executive functioning, validated in Alzheimer's Disease Neuroimaging Initiative (ADNI) participants with baseline mild cognitive impairment. Brain Imaging Behav. 2012 Dec;6(4):517-27.
7. Crane PK, Carle A, Gibbons LE, Insel P, Mackin RS, Gross A, Jones RN, Mukherjee S, Curtis SM, Harvey D, Weiner M, Mungas D. Alzheimer's Disease Neuroimaging Initiative. Development and assessment of a composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). Brain Imaging Behav. 2012 Dec;6(4):502-16.
8. Chapman LJ, Chapman JP. The measurement of differential deficit. J Psychiatr Res 1978;14(1-4):303– 311.
9. Gross AL, Inouye SK, Rebok GW, Brandt J, Crane PK, Parisi JM, Tommet D, Bandeen-Roche K, Carlson MC, Jones RN. Parallel but not equivalent: challenges and solutions for repeated assessment of cognition over time. J Clin Exp Neuropsychol. 2012;34(7):758-72.
10. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297–334.

Henry J. Riordan, Ph.D.

Executive Vice President of Medical and Scientific Affairs and Global Lead for Neuroscience at Worldwide Clinical Trials.

Dr Riordan has been involved in the assessment, treatment and investigation of various CNS drugs and disorders in both industry and academia for the past 20 years. Dr Riordan specialises in clinical trials methodology and has advanced training in biostatistics, experimental design, neurophysiology, neuroimaging and clinical neuropsychology. He has over 100 publications, including co-authoring two books focusing on innovative CNS clinical trials methodology.

Email: henry.riordan@worldwide.com