Using multiverse analysis to highlight differences in convergent correlation outcomes due to data analytical and study design choices

Sam S. Webb[1] & Nele Demeyere[1]

[1] *University of Oxford, United Kingdom*

Author Note

Sam S. Webb iD https://orcid.org/0000-0002-0029-4665

Nele Demeyere iD https://orcid.org/0000-0003-0416-5147

Correspondence concerning this article should be addressed to Sam S. Webb Department of Experimental Psychology, New Radcliffe House, Radcliffe Observatory Quarter, Oxford, OX2 6AE. Tel: +44 07707047433. Email: sam.webb@psy.ox.ac.uk

Abstract

In neuropsychological research there are a near limitless number of different approaches researchers can choose when designing studies. Here we showcase the multiverse technique to establish the robustness of analytical pathways choices within classic psychometric test validation in an example test of executive function.

We examined the impact of choices regarding sample groups, sample sizes, test metrics, and covariate inclusions on convergent validation correlations between tests of executive function. Data were available for 87 neurologically healthy adults and 117 stroke survivors, and a total of 2,220 different analyses were run in a multiverse analysis.

We found that the type of sample group, sample size, and test metric used for analyses impacted validation outcomes. Covariate inclusion choices did not impact the observed coefficients in our analyses.

The present analysis demonstrates the importance of carefully justifying every aspect of a psychometric test validation study a priori with theoretical and statistical factors in mind. It is essential to thoroughly consider the purpose and use of a new tool when designing validation studies.

Keywords: multiverse analysis, specification curve analysis, executive function, validation, clinical population

Using multiverse analysis to highlight differences in convergent correlation

outcomes due to data analytical and study design choices

Depending on the way in which neuropsychological data are collected, processed, or scored, there are potentially hundreds of final data sets that could be compiled which may lead down different analytical paths to different findings (Silberzahn et al., 2018). Steegen et al. (2016) modeled how to account for choices regarding data processing, cleaning, and analysis, through specification curve multiverse analysis (see Simonsohn et al., 2020 for a general introduction), now referred to as *multiverse analysis*. Multiverse analysis is a statistical technique that allows researchers to view the impact of analysis choices on statistical outcomes, both graphically and via inferential statistics (Simonsohn et al., 2020). If a researcher has an a priori prediction of a particular statistical outcome, running multiple versions of that statistical analysis, adapting for factors such as sample size, covariate inclusion, outlier exclusion and other cleaning techniques, can aid in determining the robustness of their finding (Steegen et al., 2016). For instance, if in most iterations of the analysis the statistical outcomes are non-significant, assuming all the factors in the multiverse analysis were reasonable, then it casts doubt on the replicability or fidelity of the finding.

The multiverse analysis technique is a variant of existing techniques, such as sensitivity analysis (Pichery, 2014) or robustness analysis (Rosenhead, 2013). These techniques each take decisions that are drawn from uncertainty and examine the impact of these on final models or decisions (Hall et al., 2022). There are a variety of mathematical, statistical, and or graphical versions of sensitivity analysis (Frey & Patil, 2002). The

multiverse analysis technique is designed to address variance in hypotheses testing outcomes and includes a statistical test of the multiverse analysis outcome to a null distribution (Simonsohn et al., 2020). In doing this, the multiverse is a superior analysis technique to assess the impact of analytical choices on the outcomes of psychometric assessment as it provides researchers with statistics to infer whether the analysis on the whole shows robust or inconsistent outcomes.

Here, we propose and demonstrate a pathway to apply this technique to convergent validation analysis and provide an example of tests of executive function. In this paper, *validation* refers to the process of establishing evidence for the interpretation of tests used in their intended populations/settings etc, and *validation outcomes* refer to the hypothesised and intended outcomes of a study (e.g., that there will be a strong positive correlation between two scales). In the current study we use an example of establishing convergent validity in tests of executive functioning (see Goldstein et al. 2014 for an overview capturing the complexity of defining the term). Currently, the multiverse technique is not widely used as a newly developed technique (Simonsohn et al., 2020), but it is picking up traction regarding its use in psychological research. Limited research is published with neuropsychological data, except in the case of Moore & Demeyere, 2022, which employs the statistical techniques and code used in the current paper, though applied to a question of robustness of regression analyses. We anticipate future research to use this technique more widely, supported by the openly available code.

The focus of the current paper is on convergent validity as one of the most common methods to investigate the validity of neuropsychological assessment in general

(Pawlowski et al., 2013), and particularly so with regards to the example of Multiple

Errands Tasks we will present here (Webb et al., 2021). The 2014 Standards for

Educational and Psychological Testing (standards 1.12 &1.16) state that if a cognitive

process is being assessed via a test, in this case executive functioning, then evidence must

be generated that the test does assess that cognitive process (American Educational

Research Association et al., 2014), and to generate evidence using well construct-validated

tools. We note that best practice for a full validation requires multiple examining different

aspect of the validity of the test and test interpretation (Kane, 2013), though we have

limited this multiverse to construct validation.

A 2013 systematic review of validation procedures for neuropsychological batteries

aimed at clinical populations, examined 147 studies that were published between 2005 and

2012 (Pawlowski et al., 2013). The clinical populations in the reviewed studies

predominantly included dementia and Alzheimer's disease, mild cognitive impairment, and

acquired brain injury, though other clinical groups were also present. In terms of sample

group used, the systematic review found that half of studies included both healthy controls

and a clinical group, followed by 35.40% that included just clinical groups, and 13.50% that

included just healthy controls (Pawlowski et al., 2013). Sample sizes ranged from 13 to 705

people for clinical groups and three to 13,000 healthy controls, with the majority of papers

having fewer than 100 participants per group (Pawlowski et al., 2013).

Sample group and size should be pre-defined depending on the intended target

group for the test (e.g., if the test is developed for diagnosing dementia, it is appropriate to

validate the test on those with dementia), and whether it is used to determine impairment

compared to normative cut offs (e.g., generating validity evidence from healthy control participants, and using the normative cut offs to diagnose impairment in a clinical group) (American Educational Research Association et al., 2014). Sample size is often arbitrary in clinical research with no clear a priori justification (Anthoine et al., 2014), and sample size justifications differ dependent on the type of validity evidence being collected (e.g., factor analysis to investigate internal factor structure of a scale versus regressions for predicting outcomes) (Hobart et al., 2012). Sample size impacts validation outcomes because the power to detect effects will be drastically reduced the smaller the sample is (Cohen, 1992a; Jones et al., 2003), and additionally small samples could reveal a large correlation given a less reliable estimate of the correlation (Schönbrodt & Perugini, 2013).

When examining whether or not a test measures a construct, or cognitive process, it is important to acknowledge a priori which test metric is used and for what purpose (American Educational Research Association et al., 2014). When assessing tests of executive functioning, researchers should specify which component of executive function is to be tested or whether multiple processes are assessed (Miyake et al., 2000) Specific metrics (i.e., overall score, sub score etc) should be carefully selected for purpose and use and validity evidence collected separately.

The Trail Making Test (TMT) (Reitan, 1971), as adapted for the shape-based analogue OCS-Plus version (Demeyere et al., 2021), is a popular test of executive function (Lin et al., 2021) and provides a good example where one test has multiple outcome metrics. The original TMT (Reitan, 1971) required connecting the letters (part A) or connecting numbers and letters in ascending order (part B) as fast as possible without

lifting the pen and without making any errors (Specka et al., 2021), and used time and errors made. Further scores can also be derived from the test, including a ratio of B/A errors or time (Arbuthnott & Frank, 2000; Reitan, 1971), and a subtraction of B-A time or errors (Heaton et al., 1985), or even B-A/A (Sánchez-Cubillo et al., 2009) among others. The ratio of B/A is thought of as a purer measure of executive functioning, removing variance which is attributable to visual scanning or motor speed measured in part A (Specka et al., 2021). The OCS-Plus version uses a ratio score of mixed over baseline trails, and also a processing measure accounting for both accuracy and time taken on the baseline trails.

In practice, authors do not often justify the choice of a metric explicitly, by discussing either the psychometrics of the score or the intended use of the metric (Kopp et al., 2015). Typically, researchers refer to previous studies or clinical practice, although research may mention that there are alternative metrics (Arbuthnott & Frank, 2000). Metric justification is particularly important where different metrics differentially relate to assessments of the same construct that the TMT measures (Sánchez-Cubillo et al., 2009). For example, Arbuthnott & Frank (2000) found that subtraction scores from Trail B and Trail A did not relate to reaction time on a switching paradigm, but ratio scores did. The use of a different metric in this case affected whether or not the task would be considered convergent or not with a switching construct, illustrating the importance of choosing measures carefully a priori with justification from theoretical expectations and psychometric/intended use.

Norms for tests of executive function are often stratified by age and education (Tombaugh, 2004) due to the known influence of these variables on cognitive performance.

Similarly, the inclusion of known covariates is important and common in neuropsychological assessments (Kabuba et al., 2018; cf Lowe & Reynolds, 1999). Importantly, not including age or education in neuropsychological test validation analyses can incorrectly inflate the correlation coefficients observed as they are both related to performance on the tests, in other words, the inclusion of known covariates will appropriately attenuate correlations, impacting validation outcomes.

In the present multiverse demonstration, we therefore chose to focus on 1) sample group used, 2) sample size, 3) metrics used for outcome measures, and 4) covariate inclusion whilst acknowledging there are many potential other aspects that could impact validation outcomes. We use data from a recently developed multiple errands based test, the Oxford Digital Multiple Errands Test (OxMET; Webb et al., 2021) which involved comparing the OxMET to two executive functioning tests (i.e., trail making test, and rule finding test), taken from the recently validated Oxford Cognitive Screen – Plus (OCS-Plus; see Demeyere et al., 2021 for healthy control validation, and Webb et al., 2022 for subacute and chronic stroke validation).

Methods

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

Ethics

This project represents a retrospective secondary analysis of existing data collected as part of two overarching research projects on neurologically healthy ageing control participants (Oxford University ethics reference 'MSD-IDREC-C1-2013-209') and data collected within

the Oxford Cognitive Screening (OCS) programme (Demeyere et al., 2015) which recruited a consecutive sample of stroke survivors during acute hospitalisation and conducted follow up neuropsychological assessments at six-months. The OCS programme received NHS ethics approval (OCS-Tablet and OCS-Recovery studies, NHS RECs 14/LO/0648 and 18/SC/0044) and all patients provided written or witnessed informed consent.

Participants

The neurologically healthy adult participants were recruited through existing research volunteer databases, drawing on relatives of stroke survivors, volunteers recruited through our website and center open days, or those who self-referred to the group. Participants were included if they had no self-reported neurological diagnoses or symptoms. For the stroke survivors, no selection criteria regarding specific behaviour or lesion characteristics were used with the only two exclusion criteria an inability to stay alert for the duration of testing and incapacity to provide informed consent.

Materials

We used data from the OxMET psychometric investigation (Webb et al., 2021). The OxMET, described in detail in Webb et al., (2021), is a simplified shopping scenario computer-tablet based errands task. A single score on overall accuracy is the primary outcome, though full performance metrics including detailed error related data are also calculated automatically and stored in the task. These include the number of rules broken, the frequency of rule breaks, perseverative errors, commissions, omissions, and total errors (Webb et al., 2021).

The tests used to establish convergent validity in the original validation (Webb et al., 2021) were the shape-analogue Trail Making Test and the Rule Finding Test from the Oxford Cognitive Screen - Plus (OCS-Plus) (Demeyere et al., 2015). To keep participant numbers consistent in this exemplar, we only use the OxMET, and OCS-Plus tasks.

The OCS-Plus Trails uses a ratio score of the baseline Trail A test trials (both a circle and square condition, connecting small to large circles and large to small squares respectively) and Trail B (alternating between large squares and small circles), where an executive score is Trail B accuracy / (Trail A circles + Trail A squares) as a percentage. The OCS-Plus further includes 'thinking time' which is the time elapsed from the start of a trail to the first pen stroke. The rule finding task from the OCS-Plus involves participants predicting next placement of a red dot which moves around the shapes and changes consecutively without notice between five placement rules. The task uses accuracy and number of rules learned, as well as time to completion as outcome metrics. The test is based on the Brixton Spatial Anticipation test (Burgess & Shallice, 1997) which uses incorrect guesses/errors as the main metric.

Identification of Specifications

For each test, different metrics were identified from the literature and test manuals which were conceivable for inclusion in a convergent validity analysis, in addition to theoretically plausible metrics. Each of the plausible metrics in the current project is detailed in Table 1, split by time, accuracy, error, and composite/ratio parameters. As discussed, metrics chosen can depend on the purpose of the test, however given the many

potential choices, it is likely that a variation in choices would occur by different researchers doing the same study (Silberzahn et al., 2018).

Plausibility of metric choice or analysis depends on many factors (Del Giudice & Gangestad, 2021), but can be summarized as whether the metric would actually be used in an analysis testing a specific hypothesis in a single data set. The aim of running the multiverse is to check the robustness of results answering the same question, so completely arbitrary metric choices in analysis would deflate or inflate the observed effects as it would not be a valid way to answer the same research question. For illustration, when looking at convergent validity of executive function tests a researcher would not plausibly include a metric of the number of cats owned by each participant as a covariate in analysis, as this is not theoretically or statistically justified. In the same way, if investigating the speed of reaction times on executive function tests, only speed should be used and not another metric of pure accuracy or errors etc if speed is specific to the hypothesis. Speed could be defined in multiple ways and taken from multiple parts of the same test (e.g., from all blocks of an experiment or just the last block etc.). Note, the definition of theoretically justified or valid choices is debatable (see Del Giudice & Gangestad, 2021, for an excellent discussion and possible solution).

[INSERT TABLE 1]

Planned analysis

When comparing the OxMET (i.e., totaling 10 different outcome variables) to the Trails (i.e., 30 variables) and Rule finding (i.e., 7) metrics, across inclusion or exclusion of covariates (i.e., 2 conditions of with or without covariates), and across healthy adults,

stroke survivors, or both combined, there are a total of 2220 possible analyses that could be conducted on the data.

We ran Spearman's Rho correlation and partial correlations for covariates, collecting outcomes of uncorrected *p*-values, 95% confidence intervals, and Spearman's Rho effect size per 2220 analyses. Note, correlations are justified here to establish the relationship and associations between tests for validity evidence for convergence of constructs (Krabbe, 2017).

Following Simonsohn et al. (2020), the *p*-value for significance of a difference between medians from the observed and null effects multiverse was extracted by calculating the proportion of the 500 permutated null analyses which had a test statistic that was the exact same median value or greater, as in the observed data, and then dividing that by two (significant if below .<05 or if no null values are equal or greater we report p <.002). For this median comparison we rounded median coefficients to two decimal places to maximize the chance of matching medians rather than inflating the risk of non-matching medians if they are >.001 decimal place out.

We also examined the difference in significance levels across the observed effects and null effects models by calculating the proportion of significant results in the null and observed effects models. In this way, we were able to statistically compare the curves of real and null effects by looking at the median coefficient and also the rate of significant findings in both multiverses. After illustrating the output of the multiverse analysis, we compared multiverse outputs first by which sample was used (e.g., healthy adults only, stroke survivors only, or combined sample), then by the size of the samples in each

analysis, then by outcome categories of specifications (e.g., accuracy, time, error, composite/ratio scores), and finally by inclusion or exclusion of covariates.

All final analyses were conducted in R studio version 4.1.0, (R Core Team, 2019) using packages *knitr* (Xie, 2014, 2015, 2019) *dplyr* (Wickham, François, et al., 2019), *ggplot2* (Wickham, 2016), *cowplot* (Wilke, 2020), *esc* (Lüdecke, 2019), *remotes* (Hester et al., 2020), *rbbt* (Dunnington & Wiernik, 2021), *Hmisc* (Harrell Jr, 2021), and *rstatix* (Kassambara, 2021). Generation of the multiverse data and graphs used the following additional libraries: *ggm* (Marchetti et al., 2020), *ppcor* (Kim, 2015), *DescTools* (Signorell et al., 2021), *tidyverse* (Wickham, Averick, et al., 2019). The data for all analyses presented here are openly available at DOI 10.17605/OSF.IO/QHKJ2. We present steps to run multiverse analysis following steps described in Simonsohn et al., (2020) in supplementary materials.

Results

Participants

In total 205 participants were included in the analyses (healthy adults = 88, stroke survivors = 117). All stroke survivors had a clinically confirmed diagnosis of stroke. Laterality of stroke; 6.38% bilateral 45.74% left, and 46.81% were right hemisphere. The average time in years between stroke and assessment date was 1.66 (*SD*=1.49, range= 0.01-7.21). Table 2 details the demographic information for the sample.

[INSERT TABLE 2]

Figure 1 illustrates the multiverse for the combined samples group, based on the different specification combinations from Table 1 which generated the different

coefficients. The graph illustrates the correlation coefficients found, ordered numerically, and colour coded according to significance (below .05) and the number of individuals included in analysis. A full multiverse analysis curve with graphical presentation of which variable across sample, covariate, and metric type was included, can be found in Figure S1 in supplementary materials.

[INSERT FIGURE 1]

Sample Group Multiverse Outcomes

We compared the coefficients obtained from the combined and healthy adults and stroke only samples to their respective 500 permutated null effects models to establish whether the found effects were statistically different to no effect. All sample groups were statistically different to the null effects models at the $p<.002$ level (denoted as '***' in Figure 2). Overall, the sample groups were more robust to deviations in analysis technique than would be expected if the null were true, where the null states that there is no relationship between the variables in the validation correlations. Next, we compared the observed specification curves of the healthy adult, stroke survivor, and combined sample groups to see if there was a significant difference between samples used in the hypothetical validation studies. We compared the found correlation coefficients of the healthy adult and stroke sample groups using Welch Two Sample t-tests and found that the groups were not statistically different in correlation coefficients.

[INSERT FIGURE 2]

Sample Size Multiverse Outcomes

Next, we examined the impact of sample size on validation outcome, by pooling all analyses together irrespective of sample group. See Figure 3 for an illustration of coefficients by sample size collapsed across all three sample groups. Evidently, we know that the larger the sample the smaller the p-value (Thiese et al., 2016), however, the extent of sample size impact on observed correlation coefficient is not known when only using one sample group in a single study. We found that sample group in this illustration did not visible alter the observed correlations, with similar correlation strengths seen at all sample sizes.

[INSERT FIGURE 3]

Validation Measure Choice Multiverse Outcomes

Using only the full combined sample, we examined the impact of comparing metrics which are within and between categories, that is, whether the x variable in a correlation is an accuracy, time, error, or composite/ratio score compared to whether the y variable is a accuracy, time, error, or composite/ratio score. Intuitively, one would want to compare within categories, that is, accuracy to accuracy scores and not accuracy to time scores. However, this is not always what is completed in practice. For instance, in a recent validation of the OCS-Plus, some accuracy/error tests were compared to time measures without a clear examination of whether this would impact validation outcomes (Demeyere et al., 2021).

When using a Wilcoxon rank sum test with continuity correction for the difference between the within- and between-category correlations in the multiverse results, we found that the correlation coefficients were statistically different (W= 23163, p<.001, r=.25 95% CIs for r [.16 - .33]. We examined the specific categories of OxMET variable (accuracy, time, error, or composite/ratio) and the effect of between- or within-group analyses types on coefficient outcomes, where 'median between' refers to the specific OxMET variable compared to non-same category variables, and where 'median within' refers to specific OxMET variables compared within same category (e.g., comparing accuracy measures to accuracy measures etc.). See Table 3 for the results of the category comparisons.

[INSERT TABLE 3]

Covariate inclusion

Next, we assessed the inclusion of age and education covariates in each analysis through use of partial correlations with both age and education factored out to see the impact they had on validation outcomes. We compared the combined sample median coefficient without covariates (-.05) and with covariates (-.05) using a Wilcoxon rank sum with continuity and found no significant difference (W= 68860, p= .89, r=.01, 95% CIs for r [0 – .08]).

Discussion

The aim of this paper was to showcase the effects of sample group, sample size, outcome measure metric choice, and covariate inclusion ~~choices which could impact study~~ on validation study outcomes. We used the convergent validation of the Oxford Digital Multiple Errands Test (Webb et al., 2021) to illustrate. A multiverse analysis totaling 2220

possible analyses was conducted on our data. This was not an exhaustive combination,

lending the possibility for many more scoring specifications than those tested in this

example (such as analysis technique used, outlier inclusion/exclusion, sample

inclusion/exclusion criteria etc). The results of the multiverse analysis were in compliment

to the original authors findings which were largely significant in their anticipated direction

in their paper (Webb et al., 2021), and highlights the wider selection of analysis pipelines

the authors could have followed and found very different results.

Using a combined sample of healthy controls and stroke survivors, we found that

the choice of the specific metric for a variety of outcome measure in the validation, affects

the associations of convergent validation outcomes, with some metrics leading to small and

non-significant correlations, and others leading to larger significant correlations. However,

the largest correlations found in the current study were only just above the benchmark

strong correlation in psychology of >.40 (Funder & Ozer, 2019) and the benchmark for

convergent validity correlation coefficients >.30 (Rotenberg et al., 2020). Using different

sample groups in our analyses did not affect the pattern of observed correlations between

outcome measures, meaning that there was little change in convergent validation outcomes

when using just healthy controls or just a clinical group.

Further, combining groups led to a greater number of reported significant

correlations, likely due to the larger sample size. However, when investigating the different

samples per correlation in the multiverse analysis, we found there was little variance in the

observed correlation coefficients, which is counter intuitive as we would expect greater

variance at smaller samples and more stability with larger samples (Schönbrodt &

Perugini, 2013; Thiese et al., 2016). We interpret this to mean the associations between outcome metrics were robust. This is further evidenced through our covariate analyses which found no differences when adjusting correlations for age and education or not.

When there are many possible metrics to consider for a single test, turning to the literature is not always the simplest way to pick an outcome measure. Literature on the same task still often operationalize the metrics differently and there is often no consensus on the best method for deciding the comparison metrics (e.g., the myriad ways of determining trail making task performance). Authors rarely justify their outcome measure or scoring technique of said outcome measure, making it more difficult for other researchers to decide on their own technique. When a researcher does choose a metric, they may fail to validate their task simply due to the vastly different ways the same test could be scored, given the mixed results of sensitivity of scores (Lange et al., 2005).

All metrics, samples, and comparisons should be justified a priori. The justification is ultimately determined by two points: 1) the purpose of the measure being validated, and 2) the form of validation being investigated. We recommend that researchers pre-register or submit their study as a registered report, when undertaking a psychometric investigation, and follow standards of psychometric validation (American Educational Research Association et al., 2014). To determine the best metric to use from a measure, a researcher must first establish exactly what it is they are hoping to validate. For example, if it is a new screen for mild cognitive impairment, they must only select metrics from comparator measures that have been shown to detect subtle cognitive impairments, rather than metrics from the comparator measure which are sensitive to gross-level cognitive

decline. In this way, a researcher is able to establish which metric(s) to use in their investigation, based on literature and the theoretically relevant constructs measured by the metrics/measures.

To determine the sample group, again the purpose of the measure must be considered, for instance, including those with dementia and neurologically healthy controls when developing dementia screen with normative data, rather than those with depression alone etc. There are multiple ways to determine sample size; researchers must consider how many participants are needed to have sufficient power for analyses chosen (Lakens, 2022), but also how representative the sample are needed to be for the purpose of the measure. For example, if a dementia screen is created, the normative sample should reflect the characteristics of the population of interest. We recommend determining how many participants are needed regarding power, and then stratifying the sample to best reflect the population of interest.

Finally, we recommend that when choosing the analyses, again the focus needs to be on the purpose of the measure, but also the question being answered in the research. If a new screening tool is created in the format of a bi-dimensional questionnaire, it is critical to establish the factor structure of the questionnaire, as such your analysis technique is pre-determined by the research question. In fitting with sample chosen and sample size, when doing convergent validity analysis, there are several guides for estimating sample size for the common validation techniques, such as factor analysis (Gagne & Hancock, 2006; Wolf et al., 2013), intraclass correlations (Bujang, 2017), and correlations or regressions (Cohen, 1992b). In addition there are important statistical and theoretical considerations

for inclusion of covariates in convergent validity analysis, see Raab et al. (2000) for a

discussion of the impacts and calculations of covariate inclusion.

While we present the multiverse technique for use in assessing degrees of freedom

in convergent validity analysis, the technique can be used for any type of analysis, including

factor analysis / latent modelling, or sensitivity/specificity analyses which are all common

convergent analysis techniques (Pawlowski et al., 2013). This multiverse approach has

been previously shown to be informative in other neuropsychological research, which used

adapted code from the current study (Moore & Demeyere, 2022). Importantly, in addition

to a multitude of potential analysis techniques and analytical decisions, core theoretical or

measurement models will drive the interpretation of validation research. It remains

imperative these frameworks are clearly reported and justified.

Study Limitations

We note that not all manipulations may be executed by a researcher based on the

field norms and standards, and therefore the analyses present in the current study extend

beyond the realistic choices. However, each analysis outcome measure or analysis

manipulation in this paper was chosen based on plausibility. In addition, the sample sizes

used in the exemplar is not large enough to detect all potential effects, as correlations tend

to stabilize at 250 participants (Pawlowski et al., 2013). We therefore encourage caution in

interpreting the results of the exemplar by noting that this is an example of a real

validation study and not reflective of best practice. Further, while the most common

method of assessing convergent validity in Multiple Errands Test literature, correlations

are a very basic statistic and perhaps alternative methods such as factor analytical or structural equation modeling may be better suited to assess these relationships.

Foremost, this paper aimed to highlight issues in validation studies and to inspire researchers to consider their analytical pathway choices in future research.

Conclusions

Here we used a multiverse technique to demonstrate the impact of selecting different plausible outcome variables on convergent validity analysis outcomes in healthy adult and clinical samples. We aimed not to instruct researchers to use the multiverse to bypass non-significant findings in their research, but to encourage better practice and full justification of sample size, group, analysis technique and decisions etc as these may all have an impact on the outcomes of validation studies.

We provide a clear example method and instructions for multiverse analysis, inclusive of all code and data to be adapted by future researchers. Avoiding potentially mis-diagnosing the nature of any cognitive impairments based on new neuropsychological assessments is paramount and researchers must ensure tests are robust and generalizable across experimental manipulations and clinical samples.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Anthoine, E., Moret, L., Regnault, A., Sébille, V., & Hardouin, J.-B. (2014). Sample size used to

validate a scale: A review of publications on newly-developed patient reported

outcomes measures. *Health and Quality of Life Outcomes*, *12*(1), 2.

https://doi.org/10.1186/s12955-014-0176-2

Arbuthnott, K., & Frank, J. (2000). Trail making test, part B as a measure of executive

control: Validation using a set-switching paradigm. *Journal of Clinical and

Experimental Neuropsychology*, *22*(4), 518–528. https://doi.org/10.1076/1380-

3395(200008)22:4;1-0;FT518

Bujang, M. A. (2017). A simplified guide to determination of sample size requirements for

estimating the value of intraclass correlation coefficient: A review. *Archives of

Orofacial Sciences*, *12*, 1–11.

Burgess, P. W., & Shallice, T. (1997). *The hayling and brixton tests*. Thames Valley Test

Company.

Cohen, J. (1992a). Statistical power analysis. *Current Directions in Psychological Science*,

1(3), 98–101. https://doi.org/10.1111/1467-8721.ep10768783

Cohen, J. (1992b). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

https://doi.org/10.1037//0033-2909.112.1.155

Del Giudice, M., & Gangestad, S. W. (2021). A Traveler's Guide to the Multiverse: Promises,

Pitfalls, and a Framework for the Evaluation of Analytic Decisions. *Advances in

Methods and Practices in Psychological Science*, *4*(1), 2515245920954925.

https://doi.org/10.1177/2515245920954925

Demeyere, N., Haupt, M., Webb, S. S., Strobel, L., Milosevich, E., Moore, M. J., Wright, H.,

Finke, K., & Duta, M. (2021). Introducing the tablet-based Oxford Cognitive Screen-

Plus (OCS-Plus) as an assessment tool for subtle cognitive impairments. *Scientific Reports*, *11*(8000). https://doi.org/10.1038/s41598-021-87287-8

Demeyere, N., Riddoch, M. J., Slavkova, E. D., Bickerton, W.-L., & Humphreys, G. W. (2015). The Oxford Cognitive Screen (OCS): Validation of a stroke-specific short cognitive screening tool. *Psychological Assessment*, *27*(3), 883. http://dx.doi.org/10.1037/pas0000082

Dunnington, D., & Wiernik, B. M. (2021). *rbbt: R Interface to the Better BiBTex Zotero Connector* (R Package Version 0.0.0.9000). https://github.com/paleolimbot/rbbt

Frey, H. C., & Patil, S. R. (2002). Identification and Review of Sensitivity Analysis Methods. *Risk Analysis*, *22*(3), 553–578. https://doi.org/10.1111/0272-4332.00039

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245920979282

Gagne, P., & Hancock, G. R. (2006). Measurement Model Quality, Sample Size, and Solution Propriety in Confirmatory Factor Models. *Multivariate Behavioral Research*, *41*(1), 65–83. https://doi.org/10.1207/s15327906mbr4101_5

Goldstein, S., Naglieri, J. A., Princiotta, D., & Otero, T. M. (2014). Introduction: A history of executive functioning as a theoretical and clinical construct. In *Handbook of executive functioning* (pp. 3–12). Springer.

Hall, B. D., Liu, Y., Jansen, Y., Dragicevic, P., Chevalier, F., & Kay, M. (2022). A Survey of Tasks and Visualizations in Multiverse Analysis Reports. *Computer Graphics Forum*, *41*(1), 402–426. https://doi.org/10.1111/cgf.14443

Harrell Jr, F. E. (2021). *Hmisc: Harrell Miscellaneous* (R package version 4.5.0).

https://CRAN.R-project.org/package=Hmisc

Heaton, R. K., Nelson, L. M., Thompson, D. S., Burks, J. S., & Franklin, G. M. (1985).

Neuropsychological findings in relapsing-remitting and chronic-progressive

multiple sclerosis. *Journal of Consulting and Clinical Psychology*, *53*(1), 103.

Hester, J., Csárdi, G., Wickham, H., Chang, W., Morgan, M., & Tenenbaum, D. (2020). remotes:

R Package Installation from Remote Repositories, Including 'GitHub'. *R Package

Version*, *2*(1).

Hobart, J. C., Cano, S. J., Warner, T. T., & Thompson, A. J. (2012). What sample sizes for

reliability and validity studies in neurology? *Journal of Neurology*, *259*(12), 2681–

2694. https://doi.org/10.1007/s00415-012-6570-y

Jones, S., Carley, S., & Harrison, M. (2003). An introduction to power and sample size

estimation. *Emergency Medicine Journal : EMJ*, *20*(5), 453–458.

https://doi.org/10.1136/emj.20.5.453

Kabuba, N., Menon, J. A., Franklin, D. R., Jr, Lydersen, S., Heaton, R. K., & Hestad, K. A. (2018).

Effect of Age and Level of Education on Neurocognitive Impairment in HIV Positive

Zambian Adults. *Neuropsychology*, *32*(5), 519.

https://doi.org/10.1037/neu0000438

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of

Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kassambara, A. (2021). *rstatix: Pipe-Friendly Framework for Basic Statistical Tests* (R

package version 0.7.0). https://cran.r-project.org/web/packages/rstatix/index.html

Kim, S. (2015). *ppcor: Partial and Semi-Partial (Part) Correlation* (R package version 1.1).

  https://CRAN.R-project.org/package=ppcor

Kopp, B., Rösser, N., Tabeling, S., Stürenburg, H. J., de Haan, B., Karnath, H.-O., & Wessel, K.

  (2015). Errors on the Trail Making Test Are Associated with Right Hemispheric

  Frontal Lobe Damage in Stroke Patients. *Behavioural Neurology*, *2015*, e309235.

  https://doi.org/10.1155/2015/309235

Lakens, D. (2022). Sample Size Justification. *Collabra: Psychology*, *8*(1).

  https://doi.org/10.1525/collabra.33267

Lange, R. T., Iverson, G. L., Zakrzewski, M. J., Ethel-King, P. E., & Franzen, M. D. (2005).

  Interpreting the trail making test following traumatic brain injury: Comparison of

  traditional time scores and derived indices. *Journal of Clinical and Experimental

  Neuropsychology*, *27*(7), 897–906. https://doi.org/10.1080/1380339049091290

Lin, Z., Tam, F., Churchill, N. W., Lin, F.-H., MacIntosh, B. J., Schweizer, T. A., & Graham, S. J.

  (2021). Trail Making Test Performance Using a Touch-Sensitive Tablet: Behavioral

  Kinematics and Electroencephalography. *Frontiers in Human Neuroscience*, *15*,

  663463. https://doi.org/10.3389/fnhum.2021.663463

Lowe, P. A., & Reynolds, C. R. (1999). Age, Gender, and Education May Have Little Influence

  on Error Patterns in the Assessment of Set-shifting and Rule Induction Among

  Normal Elderly. *Archives of Clinical Neuropsychology*, *14*(3), 303–315.

  https://doi.org/10.1016/S0887-6177(98)00029-8

Lüdecke, D. (2019). *Esc: Effect Size Computation For Meta Analysis: Vol. 0.5.1*.

  https://doi.org/10.5281/zenodo.1249217

Marchetti, G. M., Drton, M., & Sadeghi, K. (2020). *ggm: Graphical Markov Models with Mixed Graphs* (R package version 2.5). https://CRAN.R-project.org/package=ggm

Miyake, A., Emerson, M. J., & Friedman, N. P. (2000). Assessment of executive functions in clinical settings: Problems and recommendations. *Seminars in Speech and Language*, *21*, 169. https://doi.org/10.1055/s-2000-7563

Moore, M. J., & Demeyere, N. (2022). Multiverse to inform neurological research: An example using recovery outcome of neglect. *Journal of Neurology*, *269*, 233–242. https://doi.org/10.1007/s00415-021-10612-8

Pawlowski, J., Segabinazi, J. D., Wagner, F., & Bandeira, D. R. (2013). A systematic review of validity procedures used in neuropsychological batteries. *Psychology & Neuroscience*, *6*(3), 311. https://doi.org/10.3922/j.psns.2013.3.09

Pichery, C. (2014). Sensitivity Analysis. In P. Wexler (Ed.), *Encyclopedia of Toxicology (Third Edition)* (pp. 236–237). Academic Press. https://doi.org/10.1016/B978-0-12-386454-3.00431-0

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/

Raab, G. M., Day, S., & Sales, J. (2000). How to Select Covariates to Include in the Analysis of a Clinical Trial. *Controlled Clinical Trials*, *21*(4), 330–342. https://doi.org/10.1016/S0197-2456(00)00061-1

Reitan, R. M. (1971). Trail making test results for normal and brain-damaged children. *Perceptual and Motor Skills*, *33*(2), 575–581. https://doi.org/10.2466/pms.1971.33.2.575

Rosenhead, J. (2013). Robustness Analysis. In S. I. Gass & M. C. Fu (Eds.), *Encyclopedia of*

    Operations Research and Management Science (pp. 1346–1347). Springer US.

    https://doi.org/10.1007/978-1-4419-1153-7_200722

Rotenberg, S., Ruthralingam, M., Hnatiw, B., Neufeld, K., Yuzwa, K. E., Arbel, I., & Dawson, D.

    R. (2020). Measurement Properties of the Multiple Errands Test: A Systematic

    Review. *Archives of Physical Medicine and Rehabilitation*, *101*(9), 1628–1842.

    https://doi.org/10.1016/j.apmr.2020.01.019

Sánchez-Cubillo, I., Periáñez, J. A., Adrover-Roig, D., Rodríguez-Sánchez, J. M., Ríos-Lago, M.,

    Tirapu, J., & Barceló, F. (2009). Construct validity of the Trail Making Test: Role of

    task-switching, working memory, inhibition/interference control, and visuomotor

    abilities. *Journal of the International Neuropsychological Society*, *15*(3), 438–450.

    https://doi.org/10.1017/S1355617709090626

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize?

    Journal of Research in Personality, *47*(5), 609–612.

    https://doi.org/10.1016/j.jrp.2013.05.009

Signorell, A., Aho, K., Alfons, A., Anderegg, N., Aragon, T., & Arppe, A. (2021). DescTools:

    Tools for descriptive statistics. *R Package Version 0.99.40*, *18*.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai,

    F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M.

    A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2018).

    Many Analysts, One Data Set: Making Transparent How Variations in Analytic

    Choices Affect Results. *Advances in Methods and Practices in Psychological Science*,

    1(3), 337–356. https://doi.org/10.1177/2515245917747646

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). p-curve and effect size: Correcting for

    publication bias using only significant results. *Perspectives on Psychological Science*,

    9(6), 666–681.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve: Descriptive and

    inferential statistics on all reasonable specifications. *Nature Human Behaviour*

    Volume, *4*, pages1208-1214. https://doi.org/10.1038/s41562-020-0912-z

Specka, M., Weimar, C., Stang, A., Jöckel, K.-H., Scherbaum, N., Hoffmann, S. S., Kowall, B., &

    Jokisch, M. (2021). Trail Making Test Normative Data for the German Older

    Population. *Archives of Clinical Neuropsychology*, *acab027*.

    https://doi.org/10.1093/arclin/acab027

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency

    through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712.

    https://doi.org/10.1177/1745691616658637

Thiese, M. S., Ronna, B., & Ott, U. (2016). P value interpretations and considerations. *Journal

    of Thoracic Disease*, *8*(9), E928. https://doi.org/10.21037/jtd.2016.08.16

Tombaugh, T. N. (2004). Trail Making Test A and B: normative data stratified by age and

    education. *Archives of Clinical Neuropsychology*, *19*(2), 203–214.

    https://doi.org/10.1016/S0887-6177(03)00039-8

Webb, S. S., Anders, J., Chiu, E. G., Payne, F., Basting, R., Duta, M. D., & Demeyere, N. (2021).

    The Oxford Digital Multiple Errands Test (OxMET): Validation of a Simplified

    Computer Tablet Based Multiple Errands Test. *Neuropsychological Rehabilitation*.

    https://doi.org/10.1080/09602011.2020.1862679

Webb, S. S., Hobden, G., Roberts, R., Chiu, E. G., King, S., & Demeyere, N. (2022). Validation of

the UK English Oxford cognitive screen-plus in sub-acute and chronic stroke

survivors. *European Stroke Journal*, 23969873221119940.

https://doi.org/10.1177/23969873221119940

Wickham, H. (2016). *Ggplot2* (2nd ed. 2016.). Springer International Publishing.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G.,

Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M.,

Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019).

Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686.

https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., & Müller, K. (2019). *Dplyr: A Grammar of Data

Manipulation. R package version 0.8. 0.1*.

Wilke, C. O. (2020). *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'* (R

package version 1.1.1). https://CRAN.R-project.org/package=cowplot

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample Size Requirements

for Structural Equation Models: An Evaluation of Power, Bias, and Solution

Propriety. *Educational and Psychological Measurement*, *73*(6), 913–934.

https://doi.org/10.1177/0013164413495237

Xie, Y. (2014). *knitr: A Comprehensive Tool for Reproducible Research in R* (1st ed., pp. 3–31).

CRC Press. https://doi.org/10.1201/9781315373461-1

Xie, Y. (2015). *Dynamic documents with R and knitr* (2. ed.). Chapman and Hall/CRC Press.

http://bvbr.bib-

bvb.de:8991/F?func=service&doc_library=BVB01&local_base=BVB01&doc_number

=028305648&sequence=000002&line_number=0001&func_code=DB_RECORDS&se

rvice_type=MEDIA

Xie, Y. (2019). *knitr: A General-Purpose Package for Dynamic Report Generation in R: Vol.*

1.26.

F**igure 1.**

Multiverse analysis curve for all correlations between the 10 OxMET metrics, and 30 OCS-Plus

Trails and seven Rule Finding tasks in a combined group sample of 88 healthy adults and 117

stroke survivors.

Note. The degree of error is presented in light grey at 95% confidence intervals per

analysis. In red are non-significant correlations and in blue are significant correlations at

the .05 alpha level. The horizontal dotted line represents the median overall correlation

coefficient. Figure available at https://osf.io/rv7k8 under a CC-BY4.0 license.

Figure 2.

The distribution of correlation coefficients across three sample group

Note. N refers to the number of correlations computed per sample group, and sig. refers to the percentage of correlations per group which were statistically significant. '***' refers to the significance level ($p<.002$) when comparing the percentage of significant results to the percentage of significant results in 500 iterations of permutated data to create a null effects model. Horizontal bars between groups reflect Welch Two Sample t-tests p values. Figure available at https://osf.io/mt6xf under a CC-BY4.0 license.

Figure 3.

Illustrates the multiverse correlation coefficients across all sample groups where coefficients are ordered by number of observations on the x-axis.

Note. The x-axis reflects actual sample sizes per correlation. In red are non-significant correlations and in blue are significant correlations at the .05 alpha level. The horizontal dotted line represents the median overall correlation coefficient. To aid in data visualisation the data points are jittered to see overlapping values. There is a gap between 110 and 170 because few analyses had sample sizes between these. Only the specification correlation coefficients are shown rather than which variables go into the curve, as these can be obtained from Figure 1. Figure available at https://osf.io/szwhg under a CC-BY4.0 license.