THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Are we measuring the same thing? Psychometric and research considerations when adopting new testing modes in the times of COVID-19

Citation for published version:
Booth, T, Murray, AL & Terrera, GM 2020, 'Are we measuring the same thing? Psychometric and research considerations when adopting new testing modes in the times of COVID-19', *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*. https://doi.org/10.1002/alz.12197

Digital Object Identifier (DOI):
10.1002/alz.12197

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Alzheimer's & Dementia: The Journal of the Alzheimer's Association

Publisher Rights Statement:
This is the peer reviewed version of the following article: Booth, T, Murray, A, Muniz-Terrera, G. Are we measuring the same thing? Psychometric and research considerations when adopting new testing modes in the time of COVID-19. Alzheimer's Dement. 2020; 1– 4. https://doi.org/10.1002/alz.12197, which has been published in final form at https://doi.org/10.1002/alz.12197. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 10. Aug. 2024

Are we measuring the same thing? Psychometric and research considerations when adopting new testing modes in the times of COVID-19

Tom Booth[1], Aja Murray[1], Graciela Muniz-Terrera[2]

[1]Department of Psychology, University of Edinburgh, Scotland, UK

[2]Centre for Dementia Prevention, University of Edinburgh, Scotland, UK

Corresponding author:

Dr. Graciela Muniz Terrera

9A Bioquarter, Little France

Edinburgh,

EH16 4 UX, UK.

Phone number: +44 131 651 7828

Email: g.muniz@ed.ac.uk

Declarations of interest: none

Abstract.

As the world navigates unchartered territories and witnesses the overwhelming impact of COVID-19, investigators face important challenges to ensure continuity of research studies in a scientifically sound manner. Given the susceptibility of the older population to COVID-19, research in the field of ageing and dementia may be more severely impacted than other areas. With in-person testing halted, researchers are considering remote testing to collect data on questionnaires and functioning, including cognitive functioning. This is not without challenges. Here, we discuss psychometric properties of the scales that need to be considered and evaluated when implementing remote testing to ensure the quality of the studies is preserved. We encourage the community to join efforts to improve practice sharing and facilitating access to item level data.

The context

Dementia and cognitive ageing studies are deemed highly sensitive to COVID-19 effects. Most studies have halted in-person data collections and there is uncertainty about when the use of traditional in-person data collection modes will resume. However, as a disruption in the continuation of the studies increases the risk of participants' drop out and of missing critical data about events of particular interest, researchers are increasingly considering alternative data collection modes to continue their studies. Advances in the use of technologies for the remote administration of questionnaires and the evaluation of performance in some cognitive tasks may facilitate the continuation of data collection in key areas such as cognitive ability and cognitive function. However, alternative data collection modes require the consideration of a series of points with respect to the tests administered.

The problem

Whilst changing the mode of assessment may practically resolve a number of imminent and important problems in the area, it is not without its challenges. A primary aim of longitudinal studies in the field of Alzheimer's and dementia, is the identification of changes in cognitive status and test performance of individuals within a cohort, which in turn are central to efforts to try to understand the precursors and course of both diseases.

The identification of change requires scores from at least two time points. However, the observation of a difference or decline in test scores across time could be driven by a number of factors, including:

1. Properties of the test, such as its reliability. If the magnitude of the difference is small enough, it may simply reflect measurement error, not true change.
2. Differential performance of test questions across time, e.g., due to practice effects.

3. True change in the underlying cognitive function being assessed.

To effectively explain and understand change, it is critical to evaluate options (1) and (2), such that it is clear whether or not it is reasonable to consider the change to be meaningful true change. This is true whenever we consider change in a test score across groups, time or both, and is not specific to the COVID-19 pandemic. However, changing the mode of administration of a test, as would be the case in a continuing cohort through COVID-19, introduces an additional source of variation that may lead to differential test performance. A change in mode of assessment, from in person to remote, also brings with it additional changes to the testing environment. For example, familiarity with technologies used may vary, as might the equipment required which will be constrained to that which is available in people's homes such as internet speed and quality of display screens. Home environments may be noisier and contain greater distractions that lab environments. Taken together, researchers simply have far less control over the testing environment. Ultimately, all such factors add additional complication to identifying true change.

The methodologies

The field of psychometrics is vast, but discussion of change across time and mode of assessment highlights one area of particular interest, namely the assessment of the equivalence of measurement (see [1] for a technical introduction).

Measurement invariance (from Classical Test Theory, CTT), or differential item functioning (from Item Response Theory, IRT), are sets of statistical tools that evaluate whether a given psychometric test is performing in the same way across groups, time, testing format etc. – essentially any other variable deemed important. Crucially, it allows us to test whether scores can be validly compared across these groups [2], for example,

whether declines or improvements in scores on a cognitive test over time can be interpreted as genuine changes in the underlying cognitive ability measured by the test.

There are different levels of strictness with which invariance can be tested but for evaluating whether a test is invariant over time in cohort studies; a relatively strict test is required. At the lowest level, we can check each item response or subtest score relates to the same hypothesized construct across time, e.g., whether verbal ability items relate to a verbal ability factor; and spatial ability to a spatial ability factor, etc. This is typically referred to as configural invariance. The next level involves checking whether the strength of that relationship between item/subtest scores and the underlying cognitive ability are the same, e.g., whether the same verbal ability item has the same correlation with a verbal ability factor across all waves of a cohort study. In CTT, this is referred to as metric invariance, and in IRT concerns testing the equivalence of item discrimination parameters. Finally, we can check if an item/subtest has the same level of difficulty across time, e.g., that the same level of verbal ability is needed to pass the same item, irrespective of the measurement wave. In CTT, this is referred to as scalar invariance, and in IRT concerns testing of the equivalence of item difficulty parameters.

Scalar invariance is the critical test to in studying comparability of scores over time. If an item/subtest becomes easier or harder over time then the scores on that test will not provide an accurate indication of whether individuals are declining or improving. It is not necessary for every single item/subtest in a test to show this kind of invariance over time to make valid inferences about change [3]; however, it is important to test invariance to discriminate between invariant and noninvariant items/tests so that so that this can be appropriately modelled and taken account of. It is also important to note that invariance tests are necessary, but not sufficient, to conclude that observed change is true change [4].

That is, much like any statistical test, it is not possible to guarantee the underlying processes of responding have not changed over time.

The research

The literature on teleneuropsychology has reported positive results regarding the use of videoconference technology for remote cognitive assessments [5] in healthy and participants with mild cognitive impairment. A recent systematic review by Marra and colleagues [6] supported the test level validity across a variety of relevant types of measure (e.g. screening measures, cognitive tests) using teleneuropsychology methodology. Some over the phone assessments have also been deployed and their psychometric properties tested [7]. Recent work has also demonstrated the equivalence of in person versus web-based administration of a number of items assessing a variety of political and sociological constructs [8].

However, very few recent studies have considered the equivalence of performance on cognitive tests between web-based and in-person modes of delivery. Gooch [9] applied item response theory to a cognitive test battery and found that modest-to-difficult questions were easier when delivered by web-based methods than face-to-face, but that the rank ordering of item difficulties remained the same. Others [10, 11] similarly found that respondent performance was generally better via web-based modes of assessment than face-to-face. Research on why these differences may exist in cognitive data is very limited. Across different types of constructs, researchers [e.g. 8] have explored a variety of individual differences and situational factors (e.g. presence of an experimenter in face-to-face) as potential explanations. To date, there are no consistent findings as to the source of mode differences in cognitive data.

Thus though there has been limited research on this question, there is accumulating evidence that tests may perform differently across modes of delivery. Consistently identifying such differences is a crucial first step in accounting for them in studies of change. Once it is possible to draw conclusions on the extent of any potential issue, it then becomes possible to extend the exploration of the reasons differential performance, and make decisions as to how to modify tests or account for the known issues.

However, whilst these studies have begun to tackle the issue of the comparability of assessments across modes of delivery, they have not presented comprehensive assessments of the measurement properties required to fully investigate change across time and mode of assessment. Many previous studies provide examples of testing longitudinal invariance to ensure that valid inferences about change over time in cohort studies can be made [12] and the impact this may have on the study of change [13, 14]; however, there remain few examples in relation to cognitive tests.

A proposal

To protect the integrity of the research findings from longitudinal cohorts, it is important we now make best efforts to rigorously assess the measures being used across studies. A major complication in the current situation is that longitudinal changes will be potentially confounded with modality changes and the associated changes in environmental factors. As it is currently not possible to conduct in-person testing, the ideal psychometric evaluation whereby the same test is administered at multiple time points in different modes, is no longer possible. However, we can as a research community collaborate on best approximations.

First, it is crucial that investigations into the psychometric properties of tests across time and mode of delivery are conducted. As some in-person testing resumes under strict

safety conditions, it may be possible to collect information of this kind in samples of participants for whom in-person testing is likely to present minimal risk. More generally, we may be able to conduct studies in new samples that will be informative to historically collected data during COVID. While findings in such samples cannot be assumed to generalise to older adults, this is a reasonable practical first step. This information would be of value for the continued use of historical data on specific tests where only subscale or total scores are available.

A critical component of psychometric evaluations discussed here, is that the responses to individual questions are required. For many cohort studies, this data may not be electronically available. Accordingly, and given the difficulties surrounding new data collection, it is imperative as a field that we engage in the large-scale data collation and collaborative research initiatives, such that the item level data that is available is accessible.

There is existing precedent of successful psychometric development of measures for the study of health in initiatives such as the Patient-Reported Outcomes Measurement Information System (PROMIS). PROMIS has made extensive use of item level CTT and IRT analyses in the production of a number of adaptive tests with well-characterised psychometric properties. Such adaptive tests have been highlighted as being advantageous to older populations, though it must be acknowledged that development has greater upfront costs [15]. Thus, PROMIS and adaptive testing in general, can provide a model for continued efforts in the field of Alzheimer's and Dementia.

As a further step, if as a community we are serious about assessing the integrity of our measures, it is important that we improve data storage practice and record electronically all individual item responses. The current context, and a move to online testing, may actually prove to be a positive turning point in this respect, as it becomes a

trivial matter to store individual responses when tests are delivered electronically, without the additional issues of human error in data entry. Given the variety of possible sources of differential performance outlined above, it may also be advantageous to record as much contextual information as possible about the environmental circumstances in which individual completed testing, in order to investigate potential sources of differential performance.

A larger challenge would be a collaborative data archiving project which sought to go back into previous waves of major studies and, where possible, electronically store the item level information.

Finally, where data allows, future research should explicitly include testing measurement properties when studying change. By not testing measurement properties, researchers are making an assumption they hold. As a field, we are used to testing the assumptions of our statistical analyses, we are perhaps less used to testing the assumptions of our measurements.

Summary

The current context presents challenges to the continuation of longitudinal studies. Technology can help us practically overcome loss of data, but will escalate the complexity of the assessment of change. Our focus here has been on the integrity of measurement and the impact on change estimates. We present some possible avenues for maximizing information based on existing data, and to structure future work based on the assessment of measurement equivalence.

Importantly, some may question the ethics of continued testing in the face of possible threats to the validity of data. If the data cannot be trusted, and is not useful for the study of change, should we be testing at all? This is a very reasonable concern. However,

we would argue that there is currently insufficient information on the scale of the potential problem faced. The approaches set out here allow us to begin to gather this information, and thus make informed judgements in the future about how and when we test.

More broadly, COVID-19 may force the research and measurement community to consider carefully our means of testing such that any future threat to our ability to collect data is minimized. Our suggestions here would correspond to a more rigorous approach as compared to the status quo. Gathering in depth data at the finest level of measurement – the item response – to statistically account for lack of equivalence in applied studies, and to provide information that would be potentially useful for revising tests to use across multiple delivery modes. It may also prove valuable to consider again alternative forms of assessment of functioning, perhaps incorporating ecological momentary assessment [16] into typical cohort data collection practices.

There are further challenges to address. For example, the potential sampling bias that may be created in applying technological solutions in populations of older adults and attempting to understand the varied and differing contextual factors in modes of test administration. Beyond measurement, there are additional considerations in examining change over time through COVID-19 which are not specific to measurement; for example, the potential differential impact of COVID-19 at different stages of progression of Alzheimer's and dementia. However, with an organised collective effort, much progress can be made.

References

[1] Van de Schoot, R., Lugtig, P., & Hox, J. A checklist for testing measurement invariance. European Journal of Developmental Psychology, 2012, 9:486-492. https://doi.org/10.1080/17405629.2012.686740

[2] Borsboom D. When does measurement invariance matter? Medical Care, 2006, 44:S176-81. www.jstor.org/stable/41219517

[3] Pokropek, A., Davidov, E., Schmidt, P. A monte carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. Structural Equation Modeling: A Multidisciplinary Journal, 2019, 26:724–744. https://doi.org/10.1080/10705511.2018.1561293

[4] Widaman, K. F., Little, T. D., Geary, D. C., Cormier, P. Individual differences in the development of skill in mental addition: Internal and external validation of chronometric models. Learning and Individual Differences, 1992, 4:167–213. https://doi.org/10.1016/1041-6080(92)90002-V

[5] Munro Cullum, C., Hynan, L. S., Grosch, M., Parikh, M., Weiner, M. F. Teleneuropsychology: evidence for video teleconference-based neuropsychological assessment. Journal of the International Neuropsychological Society : JINS 2014;20: 1028–1033. https://doi.org/10.1017/S1355617714000873

[6] Marra, D.E., Hamlet, K.M., Bauer, R.M., Bowers, D. Validity of teleneuropsychology for older adults in response to COVID-19: A systematic and critical review. The Clinical Neuropsychologist, 2020, 9:1-42. https://doi.org/10.1080/13854046.2020.1769192

[7] Zietemann, V., Kopczak, A., Müller, C., Wollenweber, F.A., Dichgans M. Validation of the Telephone Interview of Cognitive Status and Telephone Montreal Cognitive Assessment against detailed cognitive testing and clinical diagnosis of mild cognitive impairment after stroke. Stroke, 2017;48:2952–2957. https://doi.org/10.1161/STROKEAHA.117.017519

[8] Cernat, A., Revilla, M. Moving from Face-to-Face to a Web Panel: Impacts on Measurement Quality. Journal of Survey Statistics and Methodology, 2020, smaa007. https://doi.org/10.1093/jssam/smaa007

[9] Gooch, A. Measurements of cognitive skill by survey mode: Marginal differences and scaling similarities. Research & Politics 2015;2:1-11. https://doi.org/10.1177/2053168015590681

[10] Al Baghal, T. The Effect of Online and Mixed-Mode Measurement of Cognitive Ability. Social Science Computer Review 2019;37:89-103. https://doi.org/10.1177/0894439317746328

[11] McClain, C.A., Ofstedal, M.B., Couper, M.P. Measuring Cognition in a Multi-mode Context. Survey Research Center, Institute for Social Research, University of Michigan 2018; https://hrs.isr.umich.edu/publications/biblio/9606

[12] Murray, A. L., Obsuth, I., Eisner, M., Ribeaud, D. Evaluating longitudinal invariance in dimensions of mental health across adolescence: An analysis of the Social Behavior Questionnaire. Assessment, 2019;26:1234-1245. https://doi.org/10.1177/1073191117721741

[13] Liu, Y., West, S. G. Longitudinal measurement non-invariance with ordered-categorical indicators: How are the parameters in second-order latent linear growth models affected?. Structural Equation Modeling: A Multidisciplinary Journal, 2018, 25:762-777. https://doi.org/10.1080/10705511.2017.1419353

[14] Ferrer, E., Balluerka, N., Widaman, K. F. Factorial invariance and the specification of second-order latent growth models. Methodology, 2008, 4:22-36. https://doi.org/10.1027/1614-2241.4.1.22

[15] Zygouris, S., & Tsolaki, M. (2015). Computerized Cognitive Testing for Older Adults: A Review. *American Journal of Alzheimer's Disease & Other Dementias®*, *30*(1), 13–28. https://doi.org/10.1177/1533317514522852

[16] Shiffman, S., Stone, A.A., Hufford, M.R. Ecological Momentary Assessment. Annual Review of Clinical Psychology, 2008, 4:1-32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415