EHPS report

Measurement in health psychology: combining theory, qualitative, and quantitative methods to do it right.

6th Methods in Health Psychology Symposium

Gjalt-Jorn Ygram Peters
Open University
Alexandra Dima
University of Amsterdam
Anne Marie Plass
Measure Mind
Rik Crutzen
Maastricht University
Chris Gibbons
University of Cambridge
Frank Doyle
Royal College of Surgeons in Ireland

A recent debate in Health Psychology Review demonstrated the importance of careful attention to measurement and operationalisation of health psychology constructs (Beauchamp, 2016; Brewer, 2016; de Vries, 2016; Schwarzer & McAuley, 2016; Williams & Rhodes, 2016a, 2016b). This need is met by rapid developments in the theory and measurement of health psychology constructs as evidenced by recent publications and conference contributions (e.g. Dima et al., 2014). However, these enhanced methods have been slow to disseminate into research practice. One reason may be that the new perspectives afforded by these developments and the related tools were not part of the curricula of most researchers currently active in health psychology. This lack of familiarity may manifest itself as an obstacle that appears difficult to overcome, thereby obstructing wide-spread use of these methods in research.

The goal of the sixth Methods in Health Psychology symposium, held at the annual EHPS conference in Aberdeen in 2016, was to address this by increasing attendees' familiarity with several new developments in this field. The symposium brought together five contributions, combining theory and methods from qualitative and quantitative traditions to provide a broad overview of the state of the art, limitations of current practices, and options for improvement. Moreover, the symposium aimed to give its attendants practical suggestions to apply these insights, as well as facilitate access to their corresponding tools.

The symposium started with the presentation from Gjalt-Jorn Peters of a novel perspective on the nature and inter-relations of psychological variables and implications for their measurement. This perspective facilitates a flexible and theoretically promiscuous approach to operationalization and measurement, affording researchers more flexibility in the development and assessment of measurement instruments. This was followed by the presentation of Anne Marie Plass introducing tools to explore and improve operationalization in questionnaire development or adaptation using Cognitive Interviewing. Several problems with common assumptions about validity were pointed out and solutions provided for addressing these. Rik Crutzen provided an overview of the current practices regarding assessment of the quality of measurement instruments. Although these practices are strongly rooted in classical testing theory, important assumptions of the statistical models used were routinely violated. An accessible, freely-available procedure for improvement was introduced and explained. Alexandra Dima demonstrated stepwise procedures that leverage psychometric techniques to improve the understanding and operationalization of psychological constructs. Chris Gibbons introduced

computer adaptive testing using Concerto, an open source system based on the flexible R and mySQL platforms, and discussed its benefits for health psychology research. At the end of the symposium, Frank Doyle summarized the five previous contributions and proposed several directions regarding how these insights can be implemented in practice to improve the standard of measurement in health psychology.

The presentations and additional materials are available on the Open Science Framework through links on the Health Psychology Methods page on the EHPS website at http://ehps.net/content/health-psychology-methods. These materials are available under the Creative Commons Attribution license, unless indicated otherwise. Below, each contribution is briefly summarized from the perspective of this symposium.

Pragmatic Nihilism

Gjalt-Jorn Ygram Peters

Health psychology aims to explain and change a wide variety of behaviours, and to this end has developed a plethora of theories. Several attempts have been undertaken to build integrative theories, and some even strive for a Theory of Everything (also see Peters & Kok, 2016). We argue against these efforts; instead, adopting a stance that may be called 'pragmatic nihilism' is more fruitful.

The first tenet of pragmatic nihilism is that psychological variables, defined in our health psychology theories, are usefully considered as metaphors rather than referring to entities that exist in the mind. As a consequence, the second tenet emphasizes theories' definitions and guidelines for the operationalisation of those variables. The third tenet of pragmatic nihilism is that each operationalisation represents a cross-section of a variety of dimensions, such as behavioural specificity and duration of the behaviour, and most importantly, psychological aggregation level. Any operationalisation thus represents a number of implicit or explicit choices regarding these dimensions.

These three tenets of pragmatic nihilism have two implications. First, they provide a foundation that enables combining theories in a more flexible manner than made possible by integrative theories. Second, this perspective emphasizes the importance of operationalisations, underlining the importance of investing in the careful development of measurement instruments, and thorough and extensive reporting of the specifics and performance on those measurement instruments as well as disclosure of the instruments themselves.

Awareness of the dimensions of the tesseract, of which each operationalization represents a slice, can aid researchers in scrutinizing the exact items (elements) of both newly developed operationalisations and operationalisations that have been in use for decades. For example, when using questionnaires, it is important to pay close attention to the questions used. A very easy, fast, and affordable method of identifying potential problems related to item content and interpretation was provided by Plass in the following talk.

Valid or not valid that's the question: the limited validity of 'proven valid' measurement instruments

Anne Marie Plass

The results of the 2015 landmark study of Nosek and colleagues suggested that the vast majority of recent psychology studies cannot be replicated, and it thus became clear that evidence for the most published findings is not as strong as claimed (Open Science Collaboration, 2015). It was argued that replication bias might be due to the different research methods used, publication bias, or the so-called 'statistical jackpot', which indicates that a

study result may be sheer luck, or the result of endlessly trying various analytic methods until something pans out. Yet, the quality of the measurement instruments, used in many social scientific studies, was never questionned in relation to this. Whereas, almost every individual that ever completed a questionnaire has experienced the unclear nature of this task, giving answers to questions that were difficult to understand. A large body of evidence demonstrates that items researchers thought to be perfectly clear are often vague and hard to understand (Markhous, Siksma, & Plass, 2014; Van Kessel, Hendriks, van der Hoek, & Plass, 2015; Willis & Artino, 2013). We hardly know how our respondents interpret and understand our questions.

Researchers often make use of existing measurement instruments that have proven valid through the statistical testing of its psychometric qualities. While this seems an excellent approach at first glance, there are serious risks that are being overlooked, in particular regarding the validity assumed. Validity is the extent to which a measurement-instrument (scale, or questionnaire) measures what it claims to measure. There are three conditions to achieve adequate conceptual coverage of the relevant construct. First, every element of a measurement instrument must measure a part of the construct as defined by the relevant theory. Second, no elements may be included that do not measure that construct. And third, every element must be processed as intended by research participants. The big question is: Is this the case?

With regard to the first and second condition, recent studies, using modern statistical techniques, e.g. Item Response Theory (IRT) and Rasch analyses, the validity of the assumed validated measurement instruments (Markhous et al., 2014; Van Kessel et al., 2015). They revealed substantial weaknesses of questionnaires that previously were proven 'valid' using traditional validation methods, and made clear that the 'quality guarantee' implied when a measurement instrument is validated is in fact largely unsatisfactory. Apart from this, the third and utmost critical condition for validity: verifying the interpretation of the items for a given target population, is even a largely unknown step, and extremely rare. However, if the elements of a measurement instrument are interpreted differently by a sample than what was intended when the instrument was developed, none of the previously gathered data and indicators of validity and reliability still apply. Thus, none of the three necessary conditions for construct validity are being met. Yet, we draw our conclusions based on these data.

One way to establish (better) content validity, and at an earlier stage, is through applying cognitive testing (Holch et al., 2016; Markhous et al., 2014; Willis, 2005; Willis & Artino, 2013). Cognitive interviewing involves the study of how survey questions are interpreted, how information is recalled, and how respondents make decisions to provide a particular response. Cognitive interviewing is conducted using two key procedures that are combined: 1. Think Aloud, requesting the survey respondents to actively verbalize their thoughts as they attempt to answer the survey questions (Willis, 2005; Willis & Artino, 2013), revealing how they interpret and understand the questions and answer options, and 2. Probing, a form of data collection in which the cognitive interviewer administers a series of probe questions to elicit detailed information to give researchers a better idea about the completeness of the survey and its fit to the target group. Cognitive Interviewing is an iterative process, in which usually two to three rounds of six to ten interviews, with in-between carefully structured analyses and adjustment of the items, are sufficient to optimize the survey and to understand what our respondents think we are asking.

Various studies that made use of cognitive interviewing, testing content validity of well established measurement-instruments, showed that

the majority of the items were not well understood by the target population, even though well-thought out by researchers and other stakeholders (Holch et al., 2016; Markhous et al., 2014; Van Kessel et al., 2015; Willis & Artino, 2013). Items are often phrased in a way which is common to researchers and stakeholders, but largely uncommon to the target population, and far from being representative to the way they would express themselves. Therefore, there is an urgent need to look deeper into the (content and construct) validity of measurement-instruments used, before drawing our conclusions.

Coefficient alpha, omega & factor-analytic evidence

Rik Crutzen

Cronbach's alpha is a commonly reported estimate to assess scale quality in health psychology and related disciplines. To illustrate this, we have screened all articles published in Psychology & Health in 2015 (see: http://osf.io/v7jxe). A total of 288 scales were reported in 88 articles. For 233 of these scales (80.9%), an estimate of scale quality was reported, which was alpha for 210 scales (90.1%). These figures demonstrate that reporting alpha is a widespread habit in health psychology. In this paper (Crutzen & Peters, 2016), we argued that alpha is an inadequate estimate for both validity and reliability – two key elements of scale quality – and that one of the readily available alternatives should be used. More importantly, we argued that also for these alternatives, factor-analytic evidence should be presented first when assessing scale quality.

Analyses of internal scale structure can indicate the degree to which the relationships among measurement items conform to the construct on which the proposed interpretation of scale scores is based. For example, the degree to which self-efficacy items used in a certain study reflect an underlying construct – in this case self-efficacy. Alpha, despite being frequently reported as such, is unrelated to a scale's internal structure. A recent critical review of application of Cronbach's alpha in research shows that 'both very low and very high alpha values can go either with unidimensionality or multidimensionality of the data' (Sijtsma, 2009). Therefore, in line with many others, we have previously argued to abandon alpha (Peters, 2014). Instead, we recommend reporting alternative estimates such as omega, which provides a more accurate approximation of a scale's internal structure (Revelle & Zinbarg, 2009).

Before reporting omega, however, researchers should verify if for their sample (and by implication, their population), their measurement instrument retained its intended structure. In other words, we need to know whether a single latent variable is being measured in case of a unidimensional construct (Revelle & Zinbarg, 2009), or in the case of a multidimensional construct, whether the construct's dimensions are consistent with the exhibited factor structure. Subsequently, omega is reported per subscale. Hence, dimensionality should first be verified in order to know whether the measurement instrument retained its intended structure, because if not, the measurement instrument's validity is compromised, relegating reliability assessment to a secondary concern. In order to do so, a set of analysis techniques known as exploratory factor analysis (EFA) is available. Despite the availability of methods to verify dimensionality, such analyses rarely seem to accompany reports of alpha. Of the 288 scales we surveyed in our state-of-the art review, authors assessed dimensionality for only 10 scales (3.4%). Therefore, in the vast majority of cases, readers (and likely, reviewers) have no information on the performance of the scales used. This means that the validity of these operationalisations cannot be verified. Of course,

unexpectedly discovering a multidimensional scale structure can have implications for the interpretation of the data. This is why it is so important to conduct and report these analyses. If a supposedly unidimensional scale turns out to have a two-dimensional structure in a given study, then this affects the interpretation of the scale's internal structure. Therefore, we recommend that factor-analytic evidence should be presented first when assessing the internal structure of a scale.

In the next talk, Dima extended this idea of providing factor-analytic evidence and introduced a 6-step psychometric analysis for health psychology research.

R-based 6-step psychometric analysis for health psychology research

Alexandra L. Dima

Measurement accuracy is an essential requirement for valid inferences in health psychology research and needs to be explicitly demostrated irrespective of whether concepts are measured via validated, adapted, or new tools. For multi-item scales, Crutzen and Peters (2016) showed that researchers usually rely on limited (if any) psychometric testing; to facilitate reporting of scale properties, they provided an accessible R-based tool that reports automatically item descriptives, exploratory factor analysis results, and several reliability indices. These statistics are an informative and an indispensable first glimpse of scale quality, but they can only provide a partial (and sometimes puzzling) view on the concepts under investigation. In my experience, once we get this far, we need to investigate further; luckily, R gives easy access to a whole range of tests and solutions once we become familiar with a few basic psychometric concepts and the related R packages. I introduced a 6-step analysis protocol that condenses the possibilities R offers into an analysis template that can be adapted relatively quickly for various purposes.

Why investigate scale properties further? First, we can diagnose any inconsistencies and thus correct them before they might bias substantive results. Second, factor analysis is not appropriate for all types of questionnaires and concepts, and can give misleading results in certain conditions, for example when items have different probabilities of being endorsed by respondents (van Schuur, 2003). And third (and most important), a comprehensive psychometric analysis is an opportunity to understand the concept better and thus improve theory not only in terms of statements about relationships between concepts, but also regarding measurement issues; concept and theory development are best performed in sync (Nunally & Bernstein, 1994). In essence, by skipping scale analysis in our rush to run multiple regression models using total scores we might deprive ourselves of a large part of the wisdom stored in our hard-earned data.

Performing psychometrics analyses within substantive research is therefore preferable. But is it possible? Until recently, it used to be a daunting task: more advanced techniques required dedicated proprietary software, psychometrics theory was less accessible to non-statisticians, and gathering results of different analyses into formatted reports took a long time. But nowadays most relevant statistical tools are available for free in R, together with worked examples and suggestions of relevant and accessible theoretical literature. Moreover, R provides several options for automatic report generation such as Sweave (Friedrich Leisch, 2002) and R markdown (Allaire, Horner, Marti, & Porte, 2015). In this new context, it becomes possible to streamline psychometric and substantive analyses in one analysis report that takes full advantage of the data available.

The 6-step analysis protocol is designed to facilitate this for scales with binary or ordinal

response options (an example script is accessible at https://github.com/alexadima/6-steps-protocol). Step 1 includes data preparation and descriptive statistics (package psych). Step 2 examines item fit with non-parametric and parametric item response theory (IRT) requirements (packages mokken, ltm, eRm, mirt). Step 3 tests scale structure according to exploratory or confirmatory factor analysis (psych, lavaan). Step 4 calculates reliability (classical test theory) for item (sub)-sets that show unidimensionality (psych, CTT, MBESS). Step 5 examines possible clustering of respondents via cluster analyses (stats, cluster). After each step, decisions for item exclusion can be taken and recorded in the script. Finally, step 6 computes total scores and score statistics (psych). The 6-step protocol and related script can be extended with further analyses of total scores (depending on the study hypotheses), and can be intergrated into automated reporting tools.

The benefits of integrating psychometric and substantive analyses in one data analysis protocol are manifold. For individual studies, the psychometric findings can lead to using modified scales with best performing items in sensitivity analyses to assess the influence of measurement quality on substantive results. It can also trigger a process of scale adaptation for specific populations, or of regular scale updates to keep up with changes in the phenomenon they measure. More broadly, using such R-based protocols facilitates transparency and replicability of both psychometric and substantive findings, and a more efficient and complete use of the available data. Thus, it can be part of the answer to the recent calls for increasing research quality and efficiency.

Introducing Concerto, an open-source platform designed to realise the potential of modern measurement theories

Chris Gibbons

Item response theory (IRT) models and algorithms for computer adaptive testing (CAT) were originally developed in the 1960s (Rasch, 1960). However, their widespread use was restricted by available computer processing power, lack of suitable software for conducting IRT analyses and, until recently; the absence of any accessible tools for administering questionnaires within an IRT framework. In 2011, the open-source Concerto platform (http://concertoplatform.com) was released to allow psychologists to develop and administer questionnaires and create flexible computer adaptive tests which include automatic scoring and tailored feedback. The talk introduced CAT principles, described the features of Concerto, and presented three recent implementations of Concerto for health assessment.

The main advantage of CAT compared to traditional survey administration tools (paper-based or electronic) is that it allows assessments to be better targeted, more efficient (shorter) and more accurate (reliable) (Gershon, 2005). These improvements are the result of an item selection process while a participant is taking a test: after a first item administration, the CAT selects from a larger item bank the next most informative item that matches the response pattern of that participant. Test administration stops when a pre-defined reliability threshold is reached for that particular assessment; if the test is well designed and the respondent is engaged with the task, this threshold is reached long before the item bank is exhausted. This process requires complex dedicated software that is not implemented in common survey tools and, until recently, was implemented only in proprietary tools. The development of

Concerto changed all this.

Concerto allows users to develop psychological assessments within the freely-available, fully flexible R-based environment. The open-source accessibility of Concerto means that CATs are readily available for any researcher in a relatively easy-to-use system, which still maintains the capacity to apply advanced measurement theories. CAT can be conducted in Concerto using a wide variety of pre-installed IRT models for item selection, score estimation, and prediction (Gibbons, 2016; Magis & Raîche, 2011). Concerto also offers flexibility in assessment presentation and layout using JavaScript, HTML and CSS. In addition to adaptive assessments, Concerto is capable of supporting R-based machine learning and statistical inference algorithms for automated classification of new data over the internet (opentextanalysis.com). The system can be installed on a range of locations ('cloud' or local servers) and devices running Linux or Windows operating systems.

Concerto is increasingly used as an assessment platform in health science research. For example, it hosts a computer adaptive version of the World Health Organisation Quality of Life -100 scale, which is significantly shorter than the paper-based version and provides tailored graphical and text feedback (Gibbons, Bower, Lovell, Valderas, & Skevington, 2016). US researchers have recently created the Movement Ability Measure, an adaptive test which assesses the disparity between people's current and ideal functional capacity, with clear feedback (Scalise & Allen, 2015). In higher-stakes assessment, Concerto is being developed for patient-reported outcome measures based clinical intervention that combines standard and adaptive assessment with feedback linked to clinical practice guidelines. The Concerto developers are strong supporters of open-source, accessible, and user-friendly measurement software for non-experts, and keen to provide support for researchers interested in implementing CAT for research or clinical assessment.

Reflections on the symposium and the future

Frank Doyle

To situate the previous five contributions in the wider context of health psychology measurement and start exploring future research possibilities, it is important to first reflect on the relative value of psychometrics and theory in health psychology research and practice. In my talk, I therefore began by highlighting some alternative perspectives on the limitations of psychometrics for both psychologists and non-psychologists.

The limited success of sustained efforts to improve psychometric quality of many commonly-used scales suggests that perhaps we should not exclude the possibility that psychologists are always going to be limited by the inherent inaccuracy of psychological scales. For example, depression is surely one of the most-studied latent traits, yet questionnaires for identifying major depressive disorder are not really very accurate. Thumbs et al. (2008) conducted a systematic review of sensitivity and specificity of depression scales for identifying major depression in people with coronary heart disease. They reported that, when adopting the median sensitivity (84%), specificity (79%) and depression prevalence (15%) levels, less than half of those who screened positive according to a scale will actually have major depression. Other systematic reviews report similar findings (Meader, Moe-Byrne, Llewellyn, & Mitchell, 2014; Mitchell, Vaze, & Rao, 2009). Furthermore, there is always going to be substantial sample variability which drives individual study psychometric results, differences in predictive validity, or even temporal issues with items (Cosco, Doyle, Ward, & McGee, 2012; Doyle, Conroy, & McGee, 2012; Freedland et al., 2016). There can be age-related, condition-

related and cross-cultural issues preventing scales from performing as expected across samples, despite undergoing rigorous psychometric development. Popular scales, such as the HADS, have questionable content validity (Doyle, Conroy, & McGee, 2007; Maters, Sanderman, Kim, & Coyne, 2013). Attempts to improve scales, such as using reverse-coding (van Sonderen, Sanderman, & Coyne, 2013), or adopting restrictive measurement assumptions (Meijer & Egberink, 2012) do not always yield better outcomes. These, and other issues, are summarised in Table 1. In essence, there is a large gap between what we might want from psychometric scales and what they can actually offer, and filling this gap completely might be unachievable even with the most sophisticated methods.

Overall, these findings suggest that we have to be cognisant of quite a degree of inaccuracy in psychometric scales. Against this background, the true value of adopting a pragmatic nihilistic approach, as outlined by Peters and Crutzen, can be seen. In addition to what the authors propose, this approach may allow for exploration of important issues such as sample variability and non-performing items within an individual study. A potential drawback of this approach is that it allows for subset constructs, which are difficult to analyse in current conventional approaches, and may require more sophisticated network analyses (Hevey, Collins, & Brogan, 2013).

This issue also links with the presentation from Plass – if content validity is questionable, then sample variability and non-performing items are inevitable. There is always the potential for the operationalisation of theory to be suboptimal, but adopting a cognitive interviewing technique may go some way towards alleviating such discrepancies. Indeed, it is difficult to envisage how talking to the people you are studying about these constructs or scales could be a bad idea. However, one can also question the validity of think-aloud or qualitative

Table 1

Psychometric ideals versus current reality?

What psychologists want from psychometric scales (a selection of ideals)	What psychometric scales usually comprise (the current reality)
Few robust items, that:	Many items, that:
1. Adequately cover the domain/ trait of interest	1. Do not always have good content validity
2. Are pure measures of the theoretical construct	2. Therefore are not pure measures of the construct, or overlap
3. Each discriminate along the latent trait and have predictive validity	3. Sometimes do not discriminate well or have variable predictive validity
4. Are unaffected by sample or temporal variability	4. Are variable, or even developed for individual groups
5. Can be generalised across contexts	5. Are not always generalizable
6. Have psychometric properties that are understood widely	6. Are often used by people with little psychometric training

methods – is what is verbalised a 'true' reflection of a person's emotional or cognitive state?

The critique of alpha, and the 6-step process for psychometric evaluation, are important contributions to the literature. While it is difficult to defend the current, unquestioning, widespread adoption of alpha, the alternative – omega – is not available in all statistical packages. Furthermore, while the widespread adoption of R would perhaps alleviate this practice, and allow for further appropriate psychometric investigations, R can seem daunting to master, in comparison to the popular SPSS, or indeed other statistical packages. However, it probably will not be too long before most of these procedures are available in other applications (e.g. Stata already has most of these options). However, one potential drawback of the recommendations from Dima and Crutzen is that, again due to sample variability, but also the other issues outlined above, there is always going to be non-performing items/subscales. This could potentially lead to an endless cycle of psychometric assessment and evaluation. For example, requiring authors to report the factor analytic results along with alpha values could lead to 'rotation hacking', where researchers are simply trying all possible rotation options until one leads to the findings that they believe reviewers and editors are most likely to want. It seems that to expect reviewers to understand the strengths and weaknesses of all rotation options is unreasonable. Such a cycle of psychometric evalution may also undermine psychology to other audiences, as most scales are in fact used by non-psychologists.

A final issue is that factor analysis itself can lead to spurious results (Cosco et al., 2012), and item response theory (IRT) is generally accepted to be superior (Embretson & Reise, 2000). However, IRT requires very large sample sizes that are typically not seen in health psychology research. This highlights the value of the open-source Concerto platform, described by Gibbons, which leverages computer adaptive testing, IRT and large samples to provide greater accuracy of measurement. Of note, however, is that findings from Concerto suggest that 4 items per construct are needed for good reliability – it is often the case that operationalisation of health psychology theories can have only 2-3 items per construct. Increasing the number of construct items will increase respondent burden, and potentially limit the amount of other constructs (e.g. health behaviours, health outcomes) that can be measured.

So, where does this leave us? I suggest that to improve measurement and theory, we should encourage, where possible

-the use of scales with appropriately-tested content validity

-the use of items tested in large IRT-based studies, such as Concerto

-adoption of psychometric meta-analytic techniques (e.g. Norton et al, 2013), given the issues around (small) sample variability

-consider further adoption of network analysis (Hevey et al., 2013), as per pragmatic nihilism

-the pooling of data for individual patient data network meta-analysis (Debray et al., 2016) –which should provide robust theory testing and refinement and address issues with sample variability.

-the reporting of sensitivity analyses, with and without non-performing items

-the submission of (fully anonymised) data with journal articles

While these recommendations might not take us all the way to reaching our psychometric ideals, they may give us better opportunities to understand the complex health care realities we study.

References

Allaire, J. J., Horner, J., Marti, V., & Porte, N. (2015). markdown: "Markdown" Rendering for R.

manual. Retrieved from https://cran.r-project.org/package=markdown

Beauchamp, M. R. (2016). Disentangling motivation from self-efficacy: implications for measurement, theory-development, and intervention. Health Psychology Review, 7199(April), 1–4. https://doi.org/10.1080/17437199.2016.1162666

Brewer, N. T. (2016). Building better boxes for theories of health behavior: a comment on Williams and Rhodes (2016). Health Psychology Review, 7199(April), 1–4. http://dx.doi.org/10.1080/17437199.2016.1162668

Cosco, T. D., Doyle, F., Ward, M., & McGee, H. (2012). Latent structure of the Hospital Anxiety And Depression Scale: A 10-year systematic review. Journal of Psychosomatic Research, 72(3), 180–184. https://doi.org/10.1016/j.jpsychores.2011.06.008

Crutzen, R., & Peters, G.-J. Y. (2016). Scale quality: alpha is an inadequate estimate and factor-analytic evidence is needed first of all. Health Psychology Review. https://doi.org/10.1080/17437199.2015.1124240

de Vries, H. (2016). Self-efficacy: skip the main factor paradigm! A comment on Williams and Rhodes (2016). Health Psychology Review, 7199(April), 1–4. https://doi.org/10.1080/17437199.2016.1163234

Debray, T., Schuit, E., Efthimiou, O., Reitsma, J., Ioannidis, J., Salanti, G., … GetReal Workpackage. (2016). An overview of methods for network meta-analysis using individual participant data: when do benefits arise? Statistical Methods Medical Research, (Aug 2), 962280216660741. https://doi.org/10.1177/0962280216660741

Dima, A., C., G., Kleppe, M., Byrka, K., de Bruin, M., & Johnston, M. (2014). The opportunities Item Response Theory (IRT) offers to health psychologists Methods in Health Psychology Symposium IV. European Health Psychologist, 16(6), 249–259. Retrieved from http://www.ehps.net/ehp/index.php/contents/article/viewFile/68/pdf_18

Doyle, F., Conroy, R., & McGee, H. (2007). Challenges in reducing depression-related mortality in cardiac populations: cognition, emotion, fatigue or personality? Health Psychology Review, 1(March), 137–172. https://doi.org/10.1080/17437190802046322

Doyle, F., Conroy, R., & McGee, H. (2012). Differential predictive value of depressive versus anxiety symptoms in the prediction of 8-year mortality after acute coronary syndrome. Psychosomatic Medicine, 74(7), 711–6. https://doi.org/10.1097/PSY.0b013e318268978e

Embretson, S. E., & Reise, S. P. (2000). Item Response Theory for Psychologists. Mahwah, N. J.: Lawrence Erlbaum Associates, Inc.

Freedland, K. E., Lemos, M., Doyle, F., Steinmeyer, B. C., Csik, I., & Carney, R. M. (2016). The Techniques for Overcoming Depression Questionnaire: Mokken Scale Analysis, Reliability, and Concurrent Validity in Depressed Cardiac Patients. Cognitive Therapy and Research. https://doi.org/10.1007/s10608-016-9797-6

Friedrich Leisch. (2002). Sweave: Dynamic generation of statistical reports using literate data analysis. Compstat 2002 - Proceedings in Computational Statistics, (69), 575–580. https://doi.org/10.1.1.20.2737

Gershon, R. (2005). Computer adaptive testing. Journal of Applied Measurement.

Gibbons, C. (2016). All CATS are grey in the dark: a novel approach to evaluating computer adaptive tests (CATs) in the real world. Quality of Life Research, 25(1), 48.

Gibbons, C., Bower, P., Lovell, K., Valderas, J., & Skevington, S. (2016). Electronic quality of life assessment using computer-adaptive testing. Journal of Medical Internet Research, 18(9), e240.

Hevey, D., Collins, A., & Brogan, A. (2013). Network Analysis. The Psychologist, 26(6), 430–431. https://doi.org/10.1007/978-1-4613-9458-7_5

Holch, P., Warrington, L., Potrata, B., Ziegler, L., Hector, C., Keding, A., … Velikova, G. (2016). Asking the right questions to get the right answers: using cognitive interviews to review the acceptability, comprehension and clinical meaningfulness of patient self-report adverse event items in oncology patients. Acta Oncologica (Stockholm, Sweden), 0(0), 1–7. https://doi.org/10.1080/0284186X.2016.1213878

Magis, D., & Raîche, G. (2011). catR An R Package for Computerized Adaptive Testing. Applied Psychological Measurement.

Markhous, E., Siksma, H., & Plass, A. (2014). Cognitive validation of the VascuQoL Questionnaire [In Dutch: Cognitieve validatie van de VascuQoL]. Utrecht, the Netherlands.

Maters, G. A., Sanderman, R., Kim, A. Y., & Coyne, J. C. (2013). Problems in Cross-Cultural Use of the Hospital Anxiety and Depression Scale: "No Butterflies in the Desert." PLoS ONE, 8(8). https://doi.org/10.1371/journal.pone.0070975

Meader, N., Moe-Byrne, T., Llewellyn, A., & Mitchell, A. (2014). Screening for poststroke major depression: a meta-analysis of diagnostic validity studies. Journal of Neurological Neurosurgery Psychiatry, 85(2), 198–206. https://doi.org/10.1136/jnnp-2012-304194

Meijer, R. R., & Egberink, I. J. L. (2012). Investigating Invariant Item Ordering in Personality and Clinical Scales: Some Empirical Findings and a Discussion. Educational and Psychological Measurement, 72, 589–607. https://doi.org/10.1177/0013164411429344

Mitchell, A., Vaze, A., & Rao, S. (2009). Clinical diagnosis of depression in primary care: a meta-analysis. Lancet, 374(9690), 609–619. https://doi.org/10.1016/S0140-6736(09)60879-5

Norton, S., Cosco, T., Doyle, F., Done, J., & Sacker, A. (2013). The Hospital Anxiety and Depression Scale: A meta confirmatory factor analysis. Journal of Psychosomatic Research, 74–81. https://doi.org/10.1016/j.jpsychores.2012.10.010

Nunally, J. C., & Bernstein, I. H. (1994). Psychometric Theory (3rd ed.). New York: McGraw-Hill.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716. https://doi.org/10.1126/science.aac4716

Peters, G.-J. Y. (2014). The alpha and the omega of scale reliability and validity: why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality. The European Health Psychologist, 16, 56–69.

Peters, G.-J. Y., & Kok, G. (2016). All models are wrong, but some are useful: a comment on Ogden (2016). Health Psychology Review, 10(3). https://doi.org/10.1080/17437199.2016.1190658

Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma. Psychometrika, 74(1), 145–154. https://doi.org/10.1007/s11336-008-9102-z

Scalise, K., & Allen, D. D. (2015). Use of open-source software for adaptive measurement: Concerto as an R-based computer adaptive development and delivery platform. British Journal of Mathematical and Statistical Psychology, 68(3), 478–496. https://doi.org/10.1111/bmsp.12057

Schwarzer, R., & McAuley, E. (2016). The world is confounded: a comment on Williams and Rhodes (2016). Health Psychology Review, 7199(April), 1–3. https://doi.org/10.1080/17437199.2016.1162667

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. Psychometrika, 74, 107–120.

Thombs, B. D., de Jonge, P., Coyne, J. C., Whooley, M. a, Frasure-Smith, N., Mitchell, A. J., … Ziegelstein, R. C. (2008). CLINICIAN ' S CORNER Depression Screening and Patient Outcomes in Cardiovascular Care A Systematic Review, 300(18).

Van Kessel, P., Hendriks, M., van der Hoek, L., & Plass, A. M. (2015). Development of the CaReQoL Chronic Haert Failure: a questionnaire to measuring patient reported outcomes of care. [In Dutch: Ontwikkeling van de CaReQoL Chronisch Hartfalen: Een vragenlijst voor het meten van de ervaren uitkomsten van de zorg.]. Utrecht, the Netherlands.

van Schuur, W. H. (2003). Mokken Scale Analysis: Between the Guttman Scale and Parametric Item Response Theory. Political Analysis, 11(2), 139–163. https://doi.org/10.1093/pan/mpg002

van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: let's learn from cows in the rain. PloS One, 8(7), e68967. https://doi.org/10.1371/journal.pone.0068967

Williams, D. M., & Rhodes, R. E. (2016a). Reviving the critical distinction between perceived capability and motivation: A response to commentaries. Health Psychology Review, 7199(April), 1–7. https://doi.org/10.1080/17437199.2016.1171729

Williams, D. M., & Rhodes, R. E. (2016b). The confounded self-efficacy construct: review, conceptual analysis, and recommendations for future research. Health Psychology Review, 10(2), 113–128. https://doi.org/10.1080/17437199.2014.941998

Willis, G. B. (2005). Cognitive interviewing: a tool for improving questionnaire design. Thousand Oaks, CA, US: Sage Publishing.

Willis, G. B., & Artino, A. R. (2013). What Do Our Respondents Think We're Asking? Using Cognitive Interviewing to Improve Medical Education Surveys. Journal of Graduate Medical Education, 5(3), 353–6. https://doi.org/10.4300/JGME-D-13-00154.1

Gjalt-Jorn Ygram Peters
Open University, the Netherlands
gjalt-jorn@behaviorchange.eu

Alexandra Dima
University of Amsterdam, The Netherlands
a.l.dima@uva.nl

Anne Marie Plass
Measure Mind, The Netherlands
anne.marie.plass@kpnmail.nl

Rik Crutzen
Maastricht University, The Netherlands
rik.crutzen@maastrichtuniversity.nl

Chris Gibbons
University of Cambridge, United Kingdom
cg598@cam.ac.uk

Frank Doyle
Royal College of Surgeons in Ireland, Ireland
fdoyle4@rcsi.ie