Computerized adaptive testing in primary care: CATja

van Bebber, Jan

Chapter 8

General Discussion

The main aim of the research presented in this thesis was to develop a number of computerized adaptive test (CATs) modules that jointly can be used as a screening device (named CATja) to assess mental well-being of GP clients. The purpose of this device was to facilitate mental health assistants (MHAs) in assessing their clients' strengths and weaknesses in order to better determine which level of care (within GP practices, generalist or specialist mental health care services) best suits their clients' needs. Importantly, CATja was designed to screen efficiently (i.e. adaptively) and developed in close collaboration with the MHAs that will use the device in the future. The individual chapters in this thesis comprise research that helped us to lay the scientific foundation for these adaptive test modules.

8.1 Main findings of this thesis

The modules of CATja comprise several domains of psychopathology (weaknesses) as well as constructs of positive psychology (strengths). In this thesis, I describe how CATs (including parameter specifications) were developed by our team for some of these domains, and how, for other domains, we investigated the applicability of already existing item pools (i.e. item banks plus parameter specifications) for use in the Netherlands. With respect to the latter domains, we collaborated with the Dutch Flemish PROMIS group (Terwee et al., 2014). The Patient-Reported Outcomes Measurement Information System (PROMIS) research group collected candidate items for various patient reported outcomes measures (Cella et al., 2007), and based on data that were representative of the 2000 U.S. census, final item banks were compiled and calibrated that can be used to construct computer adaptive tests (CAT). The aim of the PROMIS research group is that these item banks will be used worldwide so that results from studies conducted in different countries can be compared more easily. See chapter five for more information regarding the PROMIS initiative and the particular PROMIS item banks that are used in CATja. We developed adaptive tests for positive and negative symptoms of psychosis by means of item response theory modeling (chapter 2). The research presented in chapter 3 showed that GP clients and eHealth clients respond in a comparable way to the a-specific symptoms of stress that comprise the distress scale of the Four-Dimensional Symptom

Questionnaire (4DSQ). By that, we demonstrated that GP clients and eHealth clients experiencing equal levels of general distress tend to choose the same response options on the distress items of the 4DSQ. Thus, using the same item parameters and norm scores in both application modes is appropriate. In chapter 4, we found that the current practice of collapsing the three highest response options of the distress items of the 4DSQ prior to summation of item scores into total scale scores should be avoided because this practice decreases measurement precision for high levels of distress. In chapter 5, we showed that the original PROMIS item parameters for anxiety and depression that were estimated based on data collected in the U.S. did not show a good fit when applied to data collected in the Netherlands with the same item banks. In other words, Dutch and U.S. American people do not respond in the same way to symptoms of anxiety and depression. Simulations revealed, however, that using the official PROMIS item parameters instead of Dutch parameters (parameters estimated from the data collected in the Netherlands) did not lead to meaningful decrements in accuracy of predicting which individuals have been diagnosed to have an anxiety or mood disorder respectively. Therefore, we eventually recommended usage of the original American parameters, in order to facilitate international research collaborations. In chapter 6, we described a cohort study on symptomatic and functional recovery in individuals who experienced a first episode of psychosis (FEP) in the past. The research unveiled that levels of baseline negative symptoms were most important in predicting which individuals were going to relapse, followed by the length of the duration of untreated psychosis (DUP). Furthermore, also for predicting functional outcome, levels of baseline negative symptoms (and not number of relapses) were most important. Thus, we argued that solely focusing on relapse prevention in posttreatment of FEP patients may be insufficient, and that attention should be diverted to monitoring levels of negative symptoms instead. Note that this presumption (although at that time not tested yet) led us to incorporate a module for negative symptoms in CATja (chapter 2). Findings of this study are in line with the research described in chapter 2 which also revealed that baseline negative symptoms had the strongest relationship with social and occupational functioning at the end of the follow-up period of eighteen months. In chapter 7, we described the developmental approach that we took for CATja and the pilot study that accompanied the implementation of CATja. We reported that clients' levels of psychopathology as computed by CATja were generally lower than the levels of psychopathology estimated by the MHAs prior to testing. In addition, we also tentatively concluded that MHAs seem to lower the treatment level they advise their clients when they are provided with the score profiles generated by CATja. By means of the work described in this thesis, we contributed to the scientific literature, and also developed a practical tool that facilitates decision-making processes in clinical practice.

8.2 Generalizability of research findings

To what degree may the research findings of this thesis be generalized beyond the situational specifics (i.e., the specific instruments used in specific contexts) in which they were found? Note that we are not referring to generalizability in strict statistical sense – or more precisely – in the strict sense of inferential statistics. In order to provide the reader with a better understanding of what is meant here, the following two examples might be helpful.

8.2.1 Measurement invariance between application modes: GP clients/paper and pencil versus eHealth clients

When instruments possess the properties of structural equivalence (i.e., the property of collections of items to have the same meaning for subgroups) and scalar invariance (i.e. the property of collections of items that the same observed scores indicate the same position on the latent continuum for individuals belonging to different subgroups) for certain application modes, measurements may be considered to be invariant between application modes (see chapter 4 for a more in depth introduction to the topic). Based on our findings of structural equivalence and scalar invariance for the distress scale of the 4DSQ, the question rises whether we can assume that GP- and eHealth clients who are seeking professional help for psychological complaints respond in the same way to all different kinds of psychopathology items.

Although many studies that investigated measurement invariance of paper and pencil and online administrations have been published, most of these studies consider either various measures of personality dimensions (e.g. Fouladi, Mccarthy, & Moller, 2002; Hays & McCallum, 2005), or measures of (neuro-)cognitive functioning for educational and developmental testing purposes (e.g. Silverstein et al., 2007; Kern, Green, Nuechterlein, & Deng, 2004). A few exceptions to this general rule are studies that investigate equivalence of administration modes for specific instruments. Schulenberg et al. (2001) found the Beck Depression Inventory (Beck, Steer, & Brown, 1996) to be invariant with respect to paper and pencil versus online administration, but Coles et al. (2007) found the Obsessive Compulsive Inventory (Foa et al., 2002) not be invariant with respect to these differences in application. Because different results have been found with different instruments, we cannot state that all evidence collected on this matter so far suggests either presence or absence of measurement invariance for measures of psychopathology. In this case, we first have to rely on the frequently stated advice that equivalence has to be established for each measure between all application modes of interest (American Psychological Association. Board of Scientific Affairs. Committee on Psychological Tests, & Assessment, 1986; van Bebber et al., 2017).

Although only some patient reported outcomes (PROs) are indicators of psychopathology, like anxiety and depression, an informative review and meta-analysis on the psychometric

equivalence of paper and pencil and online testing of PROs is provided by Muehlhausen et al. (2015). The 435 extracted Interclass correlation coefficients between application modes varied between .65 and .99, with a pooled correlation coefficient of .88. Most researchers would agree that an R^2 of .65*.65*100 = 42.25% demonstrates absence of measurement invariance, while most researchers would also agree that an R^2 from .95*.95*100 = 90.25% does demonstrate measurement invariance between application modes. With respect to PROs, no general conclusion can be formulated either. With respect to measurement invariance of PROs, the recommendations given by Coons et al. (2009) are worth mentioning. These researchers categorized modifications from so called *migrations* of paper and pencil versions to computerized versions as minor (e.g. simply placing item content into a text screen format and/or visualizing one item on each page instead of multiple), moderate (e.g., splitting one item onto multiple screens), or substantial (e.g., changes in item stem wordings and/or changes in response options). For minor changes, equivalence may be simply assessed by cognitive interviewing techniques (i.e. techniques that explicitly focus on the cognitive processes that respondents use to answer survey questions). For moderate changes, Coons et al. (2009) advised quantitative equivalence testing, as we performed for the distress scale of the 4DSQ. For substantial changes, Coons et al. (2009) advised to handle the electronic version as a new instrument, which requires full psychometric testing. Although measures of psychopathology differ from PROs, the same guidelines that Coons et al. (2009) provided for PROs apply here as well. Thus, we recommend readers to logically and critically evaluate in which way(s) online testing situations differ from taking the paper and pencil versions of instruments, and how these differences might influence the response behavior of subjects. Additionally, researchers should carefully question themselves in which way(s) online sampling might additionally influence research results, because this data collection approach leads to different sample compositions in terms of demographic variables (e.g. gender, age, socio-economic status) that are related to the constructs of interests.

8.2.2 Relevance of baseline negative symptoms for relapse prevention and long-term functional outcome

Relapse prevention

In chapter six, we reported that levels of negative symptoms assessed with the positive and negative syndrome scale (PANSS; Kay, Flszbein, & Opfer, 1987) at ´baseline´ are predictive of both relapse risk and functional outcome (i.e. social and occupational functioning). May we assume to find the same relationships in case we would use another instrument to assess levels of negative symptom experiences at baseline (BNS)? For example, when we would use the CAT-NEG (Bebber et al., 2017)

based on the Prodromal Questionnaire (Loewy, Bearden, Johnson, Raine, & Cannon, 2005), or the Scale for the Assessment of Negative Symptoms (SANS; Andreasen, 1989)? To formulate the question in a slightly different way, to what extend is the idea that levels of BNS experiences (i.e. the domain, irrespective of which specific symptoms are utilized and how these are phrased) are predictive of relapse risk and functional outcome justified?

With respect to the relationship between BNS and relapse prevention, earlier studies reported mixed results. Probably the most cited (> 1200 citations according to Google scholar in February 2018) study on this matter is the study conducted by Robinson et al. (1999) who used the SANS as an indicator of BNS. The relationship was found to be non-significant. But there were important methodological differences between the study conducted by Robinson et al. and our study. First, Robinson et al. compared dose maintenance against dose discontinuation, while we compared dose maintenance against either dose reduction or dose discontinuation. The difference between treatment arms in the study of Robinson et al. was thus greater than the difference between treatment arms in our study. Related to this difference is the fact that Robison et al. entered medication strategy as first predictor in their model on relapse prevention, because it had the strongest (odds ratio of 5) relationship with relapse risk. Second, and even more important according to the view of the author of this chapter, Robinson et al. did enter an indicator of premorbid adaptation to school and premorbid social withdrawal as predictor to the model before testing the effect of BNS. The effect of premorbid adaptation and social withdrawal was significant (odds ratio of 1.6), even when controlling for different medication strategies. In my opinion, the construct of poor premorbid adaptation and social withdrawal may in fact be conceived as direct consequences of BNS, or even as alternative indicators of BNS. So it does not surprise that Robinson et al. did not find BNS being related to relapse rates when already controlling for differences in premorbid adaptation to school and social withdrawal. In a recent systematic review and meta-analysis conducted by Alvarez-Jimenez et al., the authors found that in only two out of eight studies, BNS were related to relapse risk (Alvarez-Jimenez et al., 2012). With respect to the methodological quality of the studies included in their meta-analysis, the authors note that “Statistical methods and description of methodology and results were poor in many studies. (...) and potentially important predictors of outcome such as premorbid adjustment, diagnosis, sex, age or negative symptoms were rarely included in the multivariate models” (Alvarez-Jimenez et al., 2012, p.117).

To conclude this section, until now, not enough high-quality studies have been conducted as to provide a definitive answer to the question whether the relationship between BNS and relapse risk may or may not be generalized beyond the situational specifics (using the PANSS as indicator of BNS) of our study.

Functional outcome

In chapter 2, we reported a substantial relationship (at least when considering restriction of range in the predictor induced by the data collection design) between levels of BNS assessed with the CAT-NEG and long-term social and occupational functioning in prodromal subjects (Bebber et al., 2017). In fact, many research teams have documented the predictive value of negative symptoms (either as the only predictor or in combination with measures of neurocognitive functioning) for later social and occupational functioning. See for example (Rabinowitz et al., 2012) and (Norman et al., 2000) for the PANSS, and (Milev, Ho, Arndt, & Andreasen, 2005) for the SANS. Note that the author was unable to retrieve any study that reported absence (or inverse direction) of this relationship. So, in my opinion, the generalization that levels of BNS, irrespective of the way in which these were assessed, are predictive of long-term functioning in individuals that are either prodromal or psychotic seems justified.

8.3 Lessons learned

8.3.1 Evaluation of model fit and Differential Item Functioning

The evaluation of model fit is usually done using some kind of test statistic (e.g. summed score chi-square as proposed by Orlando & Thissen, 2000 or Lagrange multiplier as proposed by Glas, 1999), where often the differences between observed and model-implied item responses are compared. Comparable test statistics are employed to investigate differential item functioning (DIF), a popular technique to assess measurement invariance across groups, an important aspect of item quality. When testing for DIF, response functions between groups are either compared directly to one another (the typical two group scenario), or group-specific response functions are compared to general response functions (the multigroup scenario). Regardless of whether testing model fit or DIF effects, and regardless of which test statistic is implemented, large sample sizes quickly lead to significant test results even though deviations between model and data may be small and negligible for practical testing purposes. An alternative is to investigate the magnitudes of differences between observed and expected item score frequencies, as was done in chapters 2 and 3, or to quantify the detrimental effect of using estimated parameters under a misfitting model on coefficients expressing criterion- or predictive validity, as was done in chapter 5. In my opinion, test statistics should be used as a first step, that is to 'flag' items that are most problematic, either with respect to model fit or DIF.

Then, the alternatives described above should be implemented for the 'flagged' items to investigate whether deviations are meaningful for test practice. Furthermore, in chapter 3, using the response data of GP patients and eHealth clients, we found that considering all respondents to be

random draws from the same population in which distress is standard normally distributed would have been incorrect. Instead, a multigroup IRT model was used to correctly estimate item parameters for the distress items of the 4DSQ. The effect of erroneously assuming equal prior distributions on item- and person parameter estimates is a topic that until now has not received proper attention in the scientific literature on estimation strategies.

8.3.2 Never underestimate people's resistance to change when implementing new tools

On a more practical level, CATja was developed taking a 'bottom up' approach, as was stated in the general introduction of this thesis. That is, we organized regular meetings with our envisioned end users (GPs and MHAs) in which we inventoried their ideas and wishes, and in which we checked whether our plans found support. The first reason for these meetings was that we wanted to incorporate the knowledge and expertise of the people who would be using CATja in daily practice in developing CATja. The second reason was that we wanted to avoid, or at least minimize, MHAs' resistance to adopt their working routines when asked to implement CATja. Still, our first result showed that it is not easy to get CATja used in practice.

Interesting in this context is that on December 1, 2017 the search term *resistance to change with IT innovations* yielded 1,130,000 hits on Google Scholar. The concept is of such importance within organizational theory that even measures have been developed to measure this construct (Oreg, 2003). Furthermore, resistance to change in the context of IT innovations is no longer seen as inherently negative. Ford et al. (2008) pointed out that resistance may actually also have functional aspects. Individuals may be against change as a result of their rational tendencies to pursue their own strategies and objectives. Furthermore, sometimes change agents are advised to embrace resistance to change because authentic resistance might contribute to successful implementation of change.

To what degree was our approach of implementing CATja in line with these new findings? Although the version that we piloted in 2017 was only the alfa version of the instrument that will be improved based on experience, there was the problem of the use of various so-called GP information systems. These are electronic systems in which caregivers may store all kinds of information that they consider relevant for a good understanding of clients digitally. During all meetings held with the MHAs, they stated that they would only use our instrument in case it would be integrated within the GP information system they were using. The problem was that there are numerous different systems, and that integrating CATja into all these different systems was too expensive in the first phase of the project. Our argument that we would integrate CATja in case GPs and MHAs would be

positive or enthusiastic about the application in general did help to convince some MHAs, but others did not collaborate in the implementation of CATja probably for this reason.

8.4 Limitations

A limitation of the research described in this thesis was that we did not further investigate or validate the value of *CATja* for triaging GP patients with psychological complaints. There are also more specific limitations to the individual studies. In the study described in chapter 2 on the applicability of CAT for positive and negative symptoms experiences of psychosis, we could not provide a conclusive answer to the question whether the CATs based on positive and negative symptom experiences (Bebber et al., 2017) or the PQ-16 (i.e. the brief version that is implemented in the Netherlands; Ising et al., 2012) are to be preferred in clinical practice. Therefore, a definitive recommendation on the best questionnaire to use could not be provided. Furthermore, the study on the optimal number of response options for the items of the distress scale of the 4DSQ in chapter 4 suffers from the shortcoming that parts of the data have not been gathered with a response scale consisting of three response options in the first place. Therefore, it was difficult to determine whether the use of three response options would have had better psychometric properties than the existing scales in which the three highest response options are collapsed *after* administration. Conclusions from this study should therefore be interpreted with caution.

8.5 Future Research

8.5.1 The incremental value of CATja: Improved placement and recovery?

Although the first results of our pilot study discussed in chapter 7 of this thesis were promising, more research is needed before CATja can be implemented on a large scale in practice. Ideally, a randomized controlled treatment (RCT) design is applied, where clients are randomly assigned to either a treatment (triage based on screening with *CATja*) or a control (triage without CATja) condition. For all cases in which clients are referred to either generalist or specialist health care services, a proper criterion for judging whether placement improves through the use of *CATja* would be to ask caregivers to rate the appropriateness of the referrals. If CATja works, the referrals in the treatment condition should, on average, be judged as more appropriate than those in the control condition. Also, clients in both experimental conditions could be requested to judge the degree to which they think their condition did improve since baseline. Baseline would be defined as the moment the clients approached their GPs for reasons of psychological complaints, i.e. intake.

Although the caregivers' opinions would be a proxy for the – hard to operationalize –criteria of "recovery", the main problem with conducting an experimental design is that it will be very difficult to randomly assign patients to both experimental conditions (treatment and control). We do not have direct access to clients, but only indirect, via their MHAs.

8.5.2 Future research and improvements 4DSQ

The research presented in chapter 3 only tested whether the principles of structural equivalence and scalar invariance would hold for the distress scale of the 4DSQ. In future research, one should investigate whether these principles also apply to the other three symptom dimensions of the 4DSQ: anxiety, depression, and somatization. As for the distress scale, cut-off values for classifying eHealth clients' levels of anxiety, depression, or somatization as low, moderate, and high were based on total scores of GP clients. Furthermore, our study presented in chapter 4 revealed that the current practice of recoding 4DSQ distress item scores prior to computing scale scores should be avoided because it leads to decreased measurement precision for high levels of distress. In case this finding would generalize to the other three dimensions of the 4DSQ, the current practice of recoding item scores should be discouraged. In addition, according to the current cut-off values for the distress scale, approximately half of all GP-clients and about two-thirds of all eHealth clients are experiencing high levels of distress. Although these cut-off values were based on clinical expertise in the first place (instead of having certain percentages of respondents in each category), it is questionable whether this classification scheme (and those for the other three 4DSQ dimensions) is still up to date. Research on these new cut-off values should be based on scale totals that are computed from raw item scores (the original five response options weighted as 0,1,2,3,4) as to preserve measurement precision for high levels of distress (and possible high levels of anxiety, depression, and somatization).

8.5.3 The intercorrelations of CATja's domains of psychopathology and constructs of positive psychology

In future research, it may be investigated how strongly the seven domains and constructs that currently can be assessed with CATja are intercorrelated in the population of individuals who contact their GP for psychological complaints. First, this would increase our understanding of our patients' score profiles. How much of the variance is shared by items with similar content, and how much variance is unique to particular constructs? Second, the answers to these questions could also improve test practice. In case, say, constructs A and B do correlate at least moderately, knowledge about the standing of an individual on construct A may be used to accelerate adaptive testing of

construct B (Paap et al., 2017). To provide the reader with an example, in case of full item bank administration, the correlation between the PROMIS domains companionship (6 items) and emotional support (16 items) equals r = .78, which means that both domains have approximately 61% common variance. In this case, the score that individual A received on Companionship could be used as prior information for testing individual A on Emotional support by using this score as preliminary input for item selection and/or score estimation (instead of using the group average for these purposes). In case many items of both domains provide information on the relative standings of individuals on both domains, multidimensional IRT models may be fitted to combinations of domains with similar content (e.g., anxiety and depression). Both approaches, the 'empirical prior' and the multidimensional approach would further decrease the number of items respondents would have to respond to in order to reach reliable score estimates, thus making the time to complete the screening instrument even shorter.

8.6 The future of CATja

The development of information technology tools is an ongoing process. For *CATja*, a next step is to expand the number of domains from which MHAs can choose. Kessler et al. (2003) found that problems concerning alcohol- and/or substance abuse/dependency are frequently encountered in primary health care. For the development of this module, we are currently in close collaboration with Addiction Care Northern Netherlands (Verslavingszorg Noord Nederland, VNN). In order to reduce the influence of social desirability response bias, we will combine the items on alcohol and drug taking habits and behavioral effects of substances with questions concerning the dieting style, physical exercise habits, and sleeping patterns of respondents. In addition to possibly reducing response bias, we consider the aforementioned aspects as important parts of an holistic perspective on clients. Furthermore, we are currently working on adding a module for the residual effects of traumatic experiences, another for autism spectrum disorders, and one for attention deficit hyperactivity disorders. More domains will be added in the more distant future.

Furthermore, the report section will be expanded. First, the table that contains the client's scores on the selected domains of psychopathology and constructs of positive psychology is now just a snapshot of the strengths and weaknesses of a client. In the future, we want to add the option to visualize differences between measurement waves (e.g., baseline, post-treatment, and various follow-up measures) in graphical form. Second, in line with existing evidence of the effectiveness of interventions, certain therapeutic options may be connected to specific score profiles (Pilling, Whittington, Taylor, Kendrick, & Guideline Development Group, 2011). A particular promising, but just upcoming intervention is the experience sampling method (ESM), which has already proven its

value as a personalized measure in the treatment of depression (Kramer et al., 2014). The core assumption is that patients are capable of influencing their mind states (e.g. positive affect) if they are empowered by providing them insight in personal and contextual factors that have impact on their mind states. However, in line with our guiding principles (chapter 7), MHAs will still be in charge to decide what they think the best choices would be for their patients.

8.7 References

Alvarez-Jimenez, M., Priede, A., Hetrick, S., Bendall, S., Killackey, E., Parker, A., . . . Gleeson, J. (2012). Risk factors for relapse following treatment for first episode psychosis: A systematic review and meta-analysis of longitudinal studies. *Schizophrenia Research, 139*(1), 116-128.

American Psychological Association. Committee on Professional Standards, American Psychological Association. Board of Scientific Affairs. Committee on Psychological Tests, & Assessment. (1986). *Guidelines for computer-based tests and interpretations* The Association.

Andreasen, N. C. (1989). The scale for the assessment of negative symptoms (SANS): Conceptual and theoretical foundations. *The British Journal of Psychiatry.*

Bebber, J., Wigman, J. T., Meijer, R. R., Ising, H. K., Berg, D., Rietdijk, J., . . . Jonge, P. (2017). The prodromal questionnaire: A case for IRT-based adaptive testing of psychotic experiences? *International Journal of Methods in Psychiatric Research, 26*(2).

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). Beck depression inventory-II. *San Antonio, 78*(2), 490-498.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., . . . PROMIS Cooperative Group. (2007). The patient-reported outcomes measurement information system (PROMIS): Progress of an NIH roadmap cooperative group during its first two years. *Medical Care, 45*(5 Suppl 1), S3-S11.

Coles, M. E., Cook, L. M., & Blake, T. R. (2007). Assessing obsessive compulsive symptoms and cognitions on the internet: Evidence for the comparability of paper and internet administration. *Behaviour Research and Therapy, 45*(9), 2232-2240.

Coons, S. J., Gwaltney, C. J., Hays, R. D., Lundy, J. J., Sloan, J. A., Revicki, D. A., . . . Basch, E. (2009). Recommendations on evidence needed to support measurement equivalence between electronic and paper-based patient-reported outcome (PRO) measures: ISPOR ePRO good research practices task force report. *Value in Health, 12*(4), 419-429.

Foa, E. B., Huppert, J. D., Leiberg, S., Langner, R., Kichic, R., Hajcak, G., & Salkovskis, P. M. (2002). The obsessive-compulsive inventory: Development and validation of a short version. *Psychological Assessment, 14*(4), 485.

Ford, J. D., Ford, L. W., & D'Amelio, A. (2008). Resistance to change: The rest of the story. *Academy of Management Review, 33*(2), 362-377.

Fouladi, R. T., Mccarthy, C. J., & Moller, N. (2002). And-pencil or online? evaluating mode effects on measures of emotional functioning and attachment. *Assessment, 9*(2), 204-215.

Glas, C. A. (1998). Detection of differential item functioning using lagrange multiplier tests. *Statistica Sinica, 8*(3), 647-667.

Glas, C. A. (1999). Modification indices for the 2-PL and the nominal response model. *Psychometrika, 64*(3), 273-294.

Hays, S., & McCallum, R. S. (2005). A comparison of the pencil-and-paper and computer-administered Minnesota multiphasic personality Inventory–Adolescent. *Psychology in the Schools, 42*(6), 605-613.

Ising, H. K., Veling, W., Loewy, R. L., Rietveld, M. W., Rietdijk, J., Dragt, S., . . . van der Gaag, M. (2012). The validity of the 16-item version of the prodromal questionnaire (PQ-16) to screen for ultra high risk of developing psychosis in the general help-seeking population. *Schizophrenia Bulletin, 38*(6), 1288-1296.

Kay, S. R., Flszbein, A., & Opfer, L. A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophrenia Bulletin, 13*(2), 261.

Kern, R. S., Green, M. F., Nuechterlein, K. H., & Deng, B. H. (2004). NIMH-MATRICS survey on assessment of neurocognition in schizophrenia. *Schizophrenia Research, 72*(1), 11-19.

Kessler, R. C., Barker, P. R., Colpe, L. J., Epstein, J. F., Gfroerer, J. C., Hiripi, E., . . . Zaslavsky, A. M. (2003). Screening for serious mental illness in the general population. *Archives of General Psychiatry, 60*(2), 184-189.

Kramer, I., Simons, C. J., Hartmann, J. A., Menne-Lothmann, C., Viechtbauer, W., Peeters, F., . . . Delespaul, P. (2014). A therapeutic application of the experience sampling method in the treatment of depression: A randomized controlled trial. *World Psychiatry, 13*(1), 68-77.

Loewy, R. L., Bearden, C. E., Johnson, J. K., Raine, A., & Cannon, T. D. (2005). The prodromal questionnaire (PQ): Preliminary validation of a self-report screening measure for prodromal and psychotic syndromes. *Schizophrenia Research, 79*(1), 117-125.

Milev, P., Ho, B., Arndt, S., & Andreasen, N. C. (2005). Predictive values of neurocognition and negative symptoms on functional outcome in schizophrenia: A longitudinal first-episode study with 7-year follow-up. *American Journal of Psychiatry, 162*(3), 495-506.

Muehlhausen, W., Doll, H., Quadri, N., Fordham, B., O'Donohoe, P., Dogar, N., & Wild, D. J. (2015). Equivalence of electronic and paper administration of patient-reported outcome measures: A

systematic review and meta-analysis of studies conducted between 2007 and 2013. *Health and Quality of Life Outcomes, 13*(1), 167.

Norman, R. M., Malla, A. K., McLean, T., Voruganti, L. P. N., Cortese, L., McIntosh, E., . . . Rickwood, A. (2000). The relationship of symptoms and level of functioning in schizophrenia to general wellbeing and the quality of life scale. *Acta Psychiatrica Scandinavica, 102*(4), 303-309.

Oreg, S. (2003). Resistance to change: Developing an individual differences measure. *Journal of Applied Psychology, 88*(4), 680.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50-64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S-X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement, 27*(4), 289-298.

Paap, M. C., Kroeze, K. A., Glas, C. A., Terwee, C. B., van der Palen, J., & Veldkamp, B. P. (2017). Measuring patient-reported outcomes adaptively: Multidimensionality matters! *Applied Psychological Measurement*.

Pilling, S., Whittington, C., Taylor, C., Kendrick, T., & Guideline Development Group. (2011). Identification and care pathways for common mental health disorders: Summary of NICE guidance. *BMJ (Clinical Research Ed.), 342*.

Rabinowitz, J., Levine, S. Z., Garibaldi, G., Bugarski-Kirola, D., Berardo, C. G., & Kapur, S. (2012). Negative symptoms have greater impact on functioning than positive symptoms in schizophrenia: Analysis of CATIE data. *Schizophrenia Research, 137*(1-3), 147-150.

Robinson, D., Woerner, M. G., Alvir, J. M. J., Bilder, R., Goldman, R., Geisler, S., . . . Mayerhoff, D. (1999). Predictors of relapse following response from a first episode of schizophrenia or schizoaffective disorder. *Archives of General Psychiatry, 56*(3), 241-247.

Schulenberg, S. E., & Yutrzenka, B. A. (2001). Equivalence of computerized and conventional versions of the beck depression inventory-II (BDI-II). *Current Psychology, 20*(3), 216-230.

Silverstein, S. M., Berten, S., Olson, P., Paul, R., Williams, L. M., Cooper, N., & Gordon, E. (2007). Development and validation of a world-wide-web-based neurocognitive assessment battery: WebNeuro. *Behavior Research Methods, 39*(4), 940-949.

Terwee, C., Roorda, L., De Vet, H., Dekker, J., Westhovens, R., Van Leeuwen, J., . . . Perez, B. (2014). Dutch–Flemish translation of 17 item banks from the patient-reported outcomes measurement information system (PROMIS). Quality of Life Research, 23(6), 1733-1741.

van Bebber, J., Wigman, J. T., Wunderink, L., Tendeiro, J. N., Wichers, M., Broeksteeg, J., . . . Meijer, R. R. (2017). Identifying levels of general distress in first line mental health services: Can GP-and eHealth clients' scores be meaningfully compared? *BMC Psychiatry, 17*(1), 382.