The development of an online neuropsychological test battery

The Amsterdam Cognition Scan

Feenstra, H.E.M.

Publication date
2018

Document Version
Other version

License
Other

Chapter 7

Summary and General Discussion

INTRODUCTION

Because of the need for more efficient research tools to assess cognitive functioning in oncology, the aim of this thesis was to establish a research tool for self-administered cognitive assessments. To accomplish this aim we (1) developed and (2) evaluated a new online neuropsychological test battery: the Amsterdam Cognition Scan (ACS). In the previous chapters, I have reported on the developmental process and evaluation studies. Here, I will first summarize our main findings ('Summary'). Thereafter, I will reflect on these findings and discuss implications and future directions ('General Discussion').

SUMMARY

ACS: development

At the initiation of this project no efficient tools were available for large-scale, international cognitive testing in the oncology setting; the few adequate (with acceptable psychometric properties and accompanying norm data) neuropsychological tests, suitable for efficient, unmonitored assessments, did not include a variety of cognitive domains, but only memory. Therefore, we developed and evaluated a new online neuropsychological test battery for self-administered assessments: the Amsterdam Cognition Scan.

At the start of this project, no specific guidelines were available to online self-administered cognitive assessments either. Therefore, we performed a literature study to identify prerequisites for proper development and use of online, neuropsychological tests in unmonitored settings. In **Chapter 2** we, first, discussed advantages and disadvantages of (unsupervised) online neuropsychological testing, and, second, outlined several technical, contextual, and psychological factors, as well as methods for quality assurance, that should be taken into account to facilitate reliable measurement of cognitive functioning in the unmonitored setting. Below I will summarize the main prerequisites that were identified.

Factors that allow for standardized data collection under a variety of technical conditions are: deliberate client-side/server-side processing, insensitivity to hardware and software differences, monitored computer configurations, and smart data recording. Contextual (testing context) factors are: test design suitable for patients with impairments or language constraints, and clear instructions on appropriate testing environment. Beneficial psychological factors are: a test design and test instructions that entail little cognitive load and optimize motivation; a trustworthy setting; securely recorded test scores; and avoidance of impeding psychological factors. In addition, for proper interpretation of tests scores from new online tests, test properties—such as reliability, validity, and norm scores—need to be determined. Finally, for long-term data quality, maintenance of the test itself, its norm scores, and data safety need to be ensured. We concluded that when properly addressing factors that could hamper reliable test results both during test development and during self-administered assessments, online neuropsychological testing could

aid large-scale data collection for clinical studies in the future.

Subsequently, we (web programming by NeuroTask B.V.) used these prerequisites to develop the ACS. We did so, in particular, by applying the following features:

1. Clear instructions
2. Practice sessions with feedback
3. Fixed and standardized breaks
4. Accessibility for all common types of browsers, operating systems, and hardware (e.g., input devices, screen types)
5. Independence from computer familiarity, language and motor abilities

This resulted in a test battery of seven neuropsychological tests: Connect the Dots (I & II), Worlist (Learning, Delayed Recall & Recognition), Reaction Speed, Place the Beads, Box Tapping, Fill the Grid and Digit Sequences (I & II).

ACS: psychometric properties & normative data

In **Chapter 3**, we reported on the psychometric properties of the ACS for assessments in clinical studies. Two studies were performed with non-CNS cancer patients to evaluate (study 1) test-retest reliability (with an interval of ± six weeks) and the influence of test setting (home or hospital) on ACS performance, and (study 2) concurrent validity (by comparing results from the ACS with results from traditional neuropsychological tests).

Test–retest reliability was studied in 96 cancer patients (57 female; mean age=51.8 years) who completed the ACS twice. Intraclass correlation coefficients (ICCs) were used to assess consistency over time; with a criterion of .60 to have a first indication of which ACS measures had acceptable reliability results and which didn't. We found ICCs to range from .29 to .76, with an ICC of .78 for the ACS total score. These correlations are generally comparable with the test–retest correlations of the traditional tests as reported in the literature. From the measures with reliability results below the .60 threshold, only two (Box tapping & Fill the Grid) had relatively low test-retest consistencies when comparing to those of traditional tests as reported in the literature.

The test setting was counterbalanced between home and hospital so that influence on test performance could be assessed by repeated measures analyses of variance. Results indicated that test setting can influence ACS performance; generally, patients performed somewhat better when the ACS was performed at the hospital than when performed at home. This difference was more pronounced for speed-based tests that depend on very precise time measurements (with significant results for Fill the Grid and Reaction Speed, and for the ACS total score). We argued that better performance from the hospital compared with the home setting is likely caused by the corresponding higher level of structure.

Concurrent validity was studied in 201 cancer patients (112 female; mean age=53.5 years) who completed both the ACS and an equivalent traditional neuropsychological test battery. Spearman or Pearson correlations were used to assess consistency between online and traditional measures. Correlations between these measures were interpreted (1) by using a .40 threshold, and (2) with respect to reliability, calculating a "ceiling consistency" (using results on the ACS and literature on traditional measures). We observed medium to large concurrent validity (r/ρ=.42 to .70; ACS total score r=.78). Although Box tapping had validity results below our .40 threshold (ρ=.36), the correlation with its traditional tests was a large proportion (75%) of the ceiling consistency. Correlations were affected—as expected—by design differences between online tests and their offline counterparts.

We concluded that, although development and optimization of the ACS remains an ongoing process, it can be used as a tool to obtain efficient (online) measures of various cognitive abilities.

In **Chapter 4**, we presented reference data to allow for initial interpretation of ACS performance. We studied psychometric properties and established normative data in a healthy reference group. More specifically, we (1) evaluated test-retest reliability of ACS measures; (2) explored the influences of computer familiarity; and (3) presented multiple regression-based formulas for calculating demographically (age, gender, and/or education) corrected norm scores. To do so, the ACS was self-administrated twice from home (with an interval of ± six weeks) by 248 healthy Dutch speaking adults (157 female; mean age=49.1 years). Virtually all participants completed the test battery from home without additional help.

We found that test-retest reliability was moderate to high and comparable to that of equivalent traditional tests. ICCs (>.60) indicated sufficient consistency over time for 7 out of 12 outcome measures (ranging from .45 to .80, and .83 for the ACS total score). These results were also comparable to the test-retest results on the ACS as observed in a sample of non-CNS cancer patients (n=96; see Chapter 3), although we found higher reliability for Reaction Speed, Box tapping, and Fill the Grid in the reference sample. Since these three tests are highly dependent on mouse input, we argued that the use of similar computer set-ups over assessments improves test-retest reliability. Therefore, in future applications with repeated testing (e.g., before and after treatment), it will be important to pursue the consistency of computer configurations; preferably using a single computer and browser. In addition, since we found significant differences between the first and the second assessment on most of our tests (all measures except for Digit Sequences I & II, Reaction Speed, and Wordlist Recognition) it is important to take practice effects into account.

Multiple regression analyses indicated that (1) participants' age negatively influenced all (12) cognitive measures; (2) gender was associated with different performance on 6 measures; and (3) education level was positively associated with performance on 4 measures. In addition, we observed an influence of tested computer skills and of self-reported amount of computer use on cognitive performance. Demographic characteristics that proved to influence ACS test

performance were included in regression-based predictive formulas to establish demographically adjusted normative data. Measures of computer familiarity were not included yet, as we first need to learn more about the measurement constructs and their relation with both online tests and traditional tests. Preliminary data on correlations between measures of computer familiarity and the traditional equivalents of our online neuropsychological tests in a different sample of healthy adults (n=40; 29 female; mean age=40 years) did indicate that self-reported number of hours computer use is potentially most informative, as—unlike tested computer skills—it did not correlate with offline measures or online measures for which influence of computer familiarity were not to be expected.

Based on the results of these two studies, we concluded that the ACS has high usability and can give reliable measures of various generic cognitive ability areas for assessments with healthy Dutch adults. Combined with our normative data that describe which demographic characteristics influence performance, these results allow for initial interpretation of ACS performance. However, to improve the interpretation of the test scores, we will continue to collect reference data from (more) heterogeneous samples. We also aim to improve our MRA parameters by performing input device specific calculations and by studying the influence of computer familiarity and related factors (such as age and gender) more in depth.

In the context of future international use of the ACS, **Chapter 5** reported on the development and validation of the English version of the ACS for North American populations. In this chapter we assessed the usability and validity of ACS-EN. First, the Dutch ACS (ACS-NL) was translated and used as a basis to create the ACS-EN. Subsequently, concurrent validity was studied in 35 cancer patients (19 female; mean age=57.1 years) who completed both the ACS-EN and an equivalent traditional neuropsychological test battery at the Memorial Sloan Kettering Cancer Center (New York). Spearman and Pearson correlations were used to assess consistency between online and traditional tests. In addition, usability was assessed based on online debriefing and technical reports.

Concurrent validity was observed to be moderately-high to high (r/ρ=.51 to .81). Based on a matched (age and education) sample, validity results were found to be similar between the ACS-EN and the ACS-NL (r/ρ=.32 to .78). Usability of the ACS-EN was good: 34 out of 35 participants were able to complete the online assessment without supervision, and instructions and practice were rated as sufficiently clear.

These results indicated that the ACS-EN has good usability and that its tests provide valid measures of their cognitive measurement constructs in North American non-CNS cancer patients. However, norm data are required to attain a fully employable research tool. We argued that, a functional and validated ACS-EN will aid future international collaborations for cognitive research in the oncology and related fields, and can foster larger-scale data collection in these areas.

Detection of cognitive impairment

In **Chapter 6**, we used the ACS to evaluate current criteria for determining cognitive impairment in cancer patients. To do so, we applied ICCTF guidelines for impairment classification on results from the ACS and a traditional neuropsychological test battery in a sample of 200 adult non-CNS cancer patients (sample of the validity study described in Chapter 3).

Using the ICCTF guidelines (1.5 SD below the normative mean on two tests and/or 2 SDs below the normative mean on one test), the proportion of cognitively impaired patients was 40% (80/200) based on the traditional test battery, and 38% (76/200) based on the ACS. No difference between proportions was found ($\chi^2(1)=.17$, $p=.68$). Within-person agreement in impairment classification was "fair" (K=.35). Likewise, simulation-based analyses indicated limited average agreement for a 1.5 SD criterion (K=.36 (SD=.10)), and average agreement dropped for a 2 SDs criterion (K=.29 (SD=.15)).

These results indicated that using cut-off scores to classify patients can lead to a situation where two (neuropsychological) test batteries that are highly similar have only limited agreement on who is classified as impaired. Additional simulation results indicated that, with the use of binary cut-off scores, agreement between two classification methods is inherently low. Consequently, caution should be taken when applying cut-off scores for impairment detection in research and clinical practice. We argued that more nuanced methods for interpretation of neuropsychological test scores are needed to improve the diagnostic process of patients, as well as research efforts into the untoward cognitive effects of cancer and cancer therapies.

Conclusion

Taken together, ACS measures were generally found to have similar cognitive measurement constructs as traditional (supervised) neuropsychological tests; ACS-based performance was as could be expected considering applied adaptations and reliability levels. Notably, to gather reliable data, test setting, computer configuration, and computer familiarity should be taken into consideration. In addition, available reference data can be used for initial interpretations of ACS performance. Currently, the ACS appears suitable for research purposes. In this context, while in the process of optimization, expansion, and further establishment of norms, the ACS could facilitate efficient gathering of (large-scale/ international) data on cognitive functioning in the near future.

GENERAL DISCUSSION

Reflections on development and use of the ACS

In the following paragraphs I will reflect on the developmental process and on the use of the ACS in unmonitored settings.

Development

The development of the ACS was a considerable part of this PhD project. Before arriving at the current state of the ACS, the following steps were (repeatedly) taken: literature research, design, (web) programming, testing, and adapting. In order to go through all these steps properly, a multidisciplinary team—providing knowledge on neuropsychological testing and web-programming—and a sufficient amount of time were required.

As a starting point for the development of the online tests, traditional neuropsychological tests were selected considering a combination of the following factors: (1) sensitivity for measuring cognitive problems related to non-CNS cancer and cancer treatment (Wefel et al., 2011); (2) frequency of use by clinical neuropsychologists; (3) quality of psychometric properties; and (4) suitability for conversion to (online) computerized versions. This resulted in a selection of seven tests meeting at least a subset of our criteria. For example, a selected traditional test could have suboptimal reliability (based on available literature), but this could be compensated by proven sensitivity to problems with attention and processing, frequent use, and suitability for computerized assessments (e.g. Reaction Speed). The seven tests were converted into computerized tests, together forming the ACS.

The ACS was designed for the unmonitored setting. Previous studies have indicated that it is feasible to elicit high quality data from online cognitive tests that are self-administered by individuals in their own homes (Assmann et al., 2016; Cromer et al., 2015). However, since this means that there are no test leaders involved for structure and individual support (Troyer et al., 2014), additional effort should be put into creating structure from a distance. Note that online test results—similar to all computerized test—may be influenced by varying computer skills, and—in addition—by variance in computer configuration and test setting. To facilitate reliable online, self-administrated assessments, developers can focus on understandability, technical suitability (e.g. software and hardware requirements) and accessibility (e.g. for people with colorblindness or varying computer skills) of the interface. To this end, we first searched the literature for useful features that may help to limit noise form the unmonitored setting and, subsequently, incorporated these features (e.g. by using video instructions) while developing the ACS. This resulted in all ACS elements being self-explanatory and test performance being influenced as little as possible by technical variance, or variance in language and computer skills. But did this result in a fully functional online neuropsychological test battery? Technical and usability studies are crucial for assessing functionality and developing a user-friendly measurement tool. While developing the ACS, results from cross-system and cross-browser tests

served to assess technical functionality to indicate the effect of technical variation on test results, and to determine technical minimum requirements. Results from our usability studies will be discussed in the next paragraph.

Usability

To optimize test development, it is crucial to assess usability: can the tests be used to measure cognition effectively, efficiently, and in a user-friendly fashion (http://www.usabilitynet.org/tools/r_international.htm#9241-11)?

We performed three usability studies during ACS development. In these studies, I was present during self-administration with the ACS, asking (open) questions after every element. The first study focused on elderly participants (n=8) to indicate feasible computer actions for participants with varying levels of computer skills and motor abilities. A main lesson from this study was that we needed to limit the need for drag-and drop actions, as these are relatively difficult for people with limited computer skills. Generally, usability tests are most informative when performed with heterogeneous and relatively small groups of people, representative for the target population (Nielsen, 2012). Therefore, the succeeding two studies were performed with seven participants each (six non-CNS cancer patients and one neuropsychological professional), varying in age, education, and gender. From these studies we gained vital feedback on ease of navigation (reduce the number of buttons, and clarify when to press where), instructions (short and clear, create more distinction between instructions and practice), provided feedback (clarification), and technical issues (increase loudness of sound; and fix inconsistencies, typos, and programming bugs).

In addition, we assessed usability in the main studies of this thesis (as described in Chapters 3, 4, and 6) by means of unmonitored, online debriefing (questionnaires). Overall, the results of studies indicated good usability of the ACS: almost all participants were able to complete the ACS without help of a test leader and rated both the instructions and the practice sessions as sufficiently clear. Based on the debriefing, we did learn that it is important—particularly in the home setting—to strongly emphasize in pre-test instructions to have a functional sound system, to check functionality of the internet connection, not to use paper-and-pencil, not to use tablets or smartphones, and to strive not to be disrupted during the assessment. Furthermore, usability could be further increased by reducing the number of required click-actions and shortening instructions where possible.

Psychometric properties and normative data

Online adaptations of existing neuropsychological tests should be regarded as new tests (Bauer et al., 2012). Therefore, psychometric properties for the use of the ACS in (clinical) research studies had to be newly determined. To enable appropriate use and interpretation of the ACS, results from studies on reliability, validity, and norm scores are described in Chapter 3, 4, 5, and 6 of this thesis. In the following paragraphs, I will reflect on the results from these studies.

Reliability studies

Our reliability studies consisted of a study with non-CNS cancer patients (n=96) and a study with people who had never had cancer ("healthy controls"; (n=248), enabling us to look into: (1) test-retest reliability; (2) influence of test setting (patient study only); and (3) practice effects. I will discuss these topics in this order.

Test-retest reliability

Test-retest reliability was assessed with the ICC as main reliability parameter. The ICC is recommended for use with continuous variables since it is a critical measure that can detect both random and systematic error (de Vet et al., 2011). It provides a population-specific measure using between-person as well as within-person variance estimates (*ICC = variance (between-person)/[variance(between-person) + variance(time) + variance(error)]* (Weir, 2005)). Pearson *r* and Spearman ρ (depending on the distribution of scores on the particular measurements) were also calculated to allow for comparisons with test-retest reliabilities of traditional tests as reported in the literature. Table 1 gives an overview of reliability results (ICC and Spearman/ Pearson) for both reliability studies.

Table 1. Overview of reliability results for both the patient and reference reliability studies.

TEST	n	ICC	Spearman ρ/ Pearson *r* (P)	n	ICC	Spearman ρ/ Pearson *r* (P)	Literature
	Patient study			Reference study			
Connect the Dots I	90	.66	.58	240	.67	.75	.55 - .73
Connect the Dots II	92	.71	.79	246	.71	.74	.56 - .79
Wordlist Learning	95	.60	.77	241	.59	.75	.80
Wordlist Delayed recall	95	.49	.72	241	.50	.64	.83
Wordlist Recognition	95	.76	.53	242	.70	.54	.48
Reaction Speed	95	.52	.35	241	.67	.74	.20 - .82
Place the Beads	96	.40	.48	143	.45	.50	.38 - .70
Box Tapping	89	.29	.29 (P)	232	.49	.46	.42 - .69
Fill the Grid	93	.33	.36	241	.80	.81	.72 - .86
Digit Sequences I	96	.66	.66 (P)	245	.54	.54 (P)	.61 - .78
Digit Sequences II	96	.64	.64 (P)	242	.64	.64 (P)	.46 - .71
Online sum score	83	.78	.78 (P)	206	.83	.83 (P)	

In the first reliability study, based on assessments with non-CNS cancer patients, we found acceptable reliability results (ICC≥.60) for 6 of the 11 ACS measures (not counting the results for Wordlist Recognition because of ceiling effects). In the second reliability study, based on assessments with healthy adults, we found the same number of measures to have acceptable reliability results, but ICCs were found to be higher, particularly for the measures with lowest reliabilities in the patient study.

First, these reliability scores appeared to reflect the relatively low inherent test–retest reliability of the original tests on which we based our online versions, which inherently limits the reliability of certain tests. Note, however, that for research purposes, ICCs somewhat below .60 might be acceptable as long as the sample sizes are sufficiently large (based on power calculations) (Weir, 2005). Second, both reliability studies resulted in high consistency over time for the ACS as a whole (total score). Third, the ACS did not underperform in its field as similar results are reported on few other self-administrated cognitive test batteries with studied reliability (evaluated in healthy participants with similar test-retest intervals). Nevertheless, it is important to explore factors that may have hampered ACS's reliability results.

Possible explanations for the differences in the magnitude of reliability results between patients and healthy controls are:

(1) Instability of the patient group. Even though we included a sample of "stable" patients with non-active disease, ongoing recovery or treatment after-effects may have increased within-person variability in health and psychological factors, and, subsequently, affected test-retest variability.

(2) Heterogeneity of the reference group. The ICC normalizes measurement error relative to the participant heterogeneity (de Vet et al., 2011), meaning that ICCs are higher in the case of more between-person variability (Weir, 2005). However, even though our reference group did vary slightly more in education level than our patient group, we did not find higher variability (SDs) for reference scores.

(3) Influence of test setting. Study design related variability in setting and computer configuration may have added systematic error to the patient data. While the reference group performed the ACS twice from home, likely on the same computer, the patient group performed the ACS over two settings (home and hospital) and consequently on two different computers. This study design enabled us to assess influences of setting on test performance (see 'Test setting', next paragraph), but may have affected reliability in the patient sample. Note that effects of setting could not be filtered out by our counterbalanced design, since reliability calculations are partly based on within-person variability. To adjust for the effects of setting, we included additional variance estimates (variance by setting from mixed-effect models for repeated testing) in our reliability calculations (Weir, 2005). However, this correction may not have been strong enough to explain all setting-related variance for tests that strongly rely on precise measurements from the input device as for two of such tests (Reaction Speed and Fill the Grid) we found (1) a significant effect of setting on tests performance (see 'Test setting'), and (2) much lower reliability results in the patient compared to the reference group.

Test setting

Patients generally performed better from the hospital than from home (a difference that was

significant for 2 specific measures and the ACS total score). This influence of test setting may be explained by two factors. First, in the home setting there is less environmental structure. Although we designed the ACS for unmonitored assessments (see 'Development'), unmonitored home assessments are susceptible to factors that may add noise to cognitive data such as unscheduled breaks, or distractions from phones, TV, or persons entering the room. Second, each home setting will have different computer configurations. People will use all kinds of computer input devices, screen sizes (and zoom factors), and processors with a varying amount of processor speed to take the ACS. Some of these may have been suboptimal for testing purposes. As the current studies did not provide enough data to investigate the influence of computer configuration (required variability was available from home assessments only), future research should investigate this. In the meantime, for comparative cognitive studies and for the collection of reference data, we allow the use of a variety of computer settings (within the basic requirements, ensuring accessibility of the ACS), but advise to use strictly one constant test setting: either home or lab based.

Practice effects

It is common for people to improve in cognitive performance with repeated testing (Mathews et al., 2013; Wesnes & Pincock, 2002). As expected, such "practice effects" were found for several of the ACS tests in our test-retest studies (particularly on measures of memory and executive functioning). As practice effects produce a type of systematic error variance, critical test-retest evaluations require taking practice effects into account (de Vet et al., 2011). To do so we used the ICC and applied a correction for practice effects to the SDCs. Alternatively, to reduce practice effects and possibly improve reliability analyses, assessments could be repeated three to four times, focusing on performance form the second assessment on.

The reference study allowed us to look into practice effects more closely (the patient study was hampered by its design with counterbalanced settings). In line with the additional error detection related to the ICC measure, ICCs were found to be smaller than Pearson/ Spearmen correlations (as there is more error detected). For future analyses, to indicate the magnitude of practice effects one can look into variance estimates (from mixed-effect models for repeated measures/ ANOVA). However, note that, as variance estimates represent systematic error as a whole (effects of fatigue, etcetera), it is impossible to quantify practice effects specifically.

Validity studies

From the Dutch validity study (n=201; Chapter 3), we found all but one of ACS measures to have above threshold (Pearson/ Spearman ≥.40) validity results, similar to other concurrent validity studies comparing self-administered online cognitive test batteries with traditional tests. The highest validity was found for the ACS as a whole. Although validation is best performed on separate measures, not on full assessment batteries (de Vet et al., 2011), these results indicate that relations between the ACS measures and the traditional measures were as expected. For 6 out of 11 measures, correlations were found to be medium (<.50; (Cohen, 1988a)). Importantly though, concurrent validity is bound by the tests' test–retest reliabilities, which in some cases were suboptimal (see 'Reliability'). When also taking reliability into account, we see that validity

results are generally not far from what can maximally be expected (ceiling consistencies).

From the US validity study on the ACS-EN (n=35; Chapter 6), we found moderately-high to high validity results. Unlike for the ACS-NL, concurrent validity exceeded the .40 criterion for all ACS-EN measures. Potentially this is due to the homogeneity of the US sample (almost exclusively highly educated and frequent users of computers).

In addition to the restricting quality of the ceiling consistencies, to understand why not all correlations between our online measures and their offline counterparts were large, it is important to realize that perfect correlations were not expected because: (1) we adapted tests to an online mode (computerized and unmonitored setting); and (2) we applied several additional adaptations to enhance test usability and validity compared to the traditional tests. This means that we developed new tests with some noteworthy differences compared to the originals. Indeed, lower concurrent validity results were found for tests in which we introduced relatively large adaptations. For example, Place the Beads—a test for which we found relatively weak validity results—was adapted in several ways, of which problem structure (type and sequence) and response mode (drag-and-drop instead of manual) were most impactful. In addition, tests with only a minimum amount of changes from its original test were found to have highest validity results. We saw that Reaction Speed—the only tests from the battery that was based on a computerized traditional test, meaning that the online mode and the type of input system (keyboard for traditional assessments vs. mouse/ touchpad for online assessments) should be the only differences with the traditional test—was the only measure for which validity exceeded ceiling consistency.

Based on the Dutch data, we aimed to investigate the validity of the ACS tests using an alternative approach also by assessing agreement in the proportion of patients classified as cognitively “impaired”. Even though we found similar proportions of impaired patients with the ACS and traditional assessments, there was only limited agreement on who was classified as impaired. At first these results seem to indicate poor external validity of the ACS tests, however, additional simulation results indicate that, with the use of binary cut-off scores, agreement between two classification methods is inherently low. Therefore, methods for impairment detection should improve (e.g. by using methods such as Multivariate Normative Comparison (van Rentergem et al., 2017)), before impairment proportions are used to look further into the validity of the ACS.

Finally, some possible limitations of our validity studies have not been discussed in the previous chapters. First, in both validity studies (ACS-NL and ACS-EN) we corrected for order (adjusting all second assessments), using an in-house developed correction method. In some cases, such as Place the Beads, this correction was relatively large because of large practice effects. Importantly though, results did not differ substantially between corrected and uncorrected scores. Second, we presented single correlations (between ACS tests and their traditional counterparts) only and left out correlation matrices, since our online tests were based on one specific traditional test each. Correlation matrices may be applied to study the measurement constructs of new ACS

tests after expanding the battery.

Reference study

In our reference study, we collected data from 248 healthy Dutch speaking adults (aged 18 to 81 years). We used multiple regression analyses (MRA) to establish norms and to enable interpretation of ACS scores (by comparing newly obtained scores with reference data from people with similar demographic characteristics). To establish regression-based norms that account for as much variance in test performance as possible, we investigated predominantly the influences of age, gender, and education as these characteristics are often found to be associated with cognitive performance (Strauss et al., 2006). Sensitivity to these demographic characteristics was largely consistent with what is known from traditional neuropsychological tests. Because of the computerized nature of the ACS assessments, we also looked into sensitivity of the measures to participants' computer skills and experience, but so far, based on the current measures, results remain too ambiguous to include them, as an adjustment factor, into ACS's norms.

For each ACS measure, demographic characteristics that proved to influence performance were used to establish demographically adjusted norms. For research with Dutch speaking adults, the regression-based predictive formulas described in Chapter 4 allow for initial interpretation of ACS results. However, since reference data were collected from home assessments, one should bear in mind that current norms could be slightly less sensitive when applied in a more structured lab or hospital setting.

In addition, there are some attention points for improving interpretation of test scores:

- A larger reference group will enable more precise regression parameters, allowing for more reliable predictions.
- Additional data will have to be collected from a more heterogeneous sample, as our current sample did not include participants with a low education level.
- Newly collected data should come from a larger variety of computer configurations; in particular input devices, as response time latencies can differ substantially depending on input device type. Subsequently, we can establish device-specific norm data.
- To account for possible fatigue effects the order of the ACS tests may need to be altered (alternating instead of fixed) when collecting additional reference data. This will especially be of importance as soon as ACS test order becomes customizable.
- Tests that show ceiling effects (e.g., Wordlist Recognition) should be re-evaluated to increase variance (and subsequently interpretation of test performance) and to better fit the cognitive level of future participants.
- In the future, quantile regression might be applied, since this type of regression is more suitable for skewed data—common for reaction time measures—than mean regression (Sherwood, Zhou, Weintraub, & Wang, 2016).
- After expanding the reference data and improving MRA parameters, availability and usability of the norm data needs to be considered. We plan to develop scoring software

and make this available via an online tool. Such a tool will provide up-to-date calculations of demographically corrected scores by simply entering the raw scores and demographic characteristics.

Populations and computer familiarity

In this section, I will discuss two additional factors that may have influenced interpretation of ACS results and the results of this thesis. First, I will reflect on the population characteristics of the above described evaluation studies, and, second, I will reflect on the role of computer familiarity.

Study populations

Our study populations consisted of non-CNS cancer patients (Dutch: Chapters 3 and 5, North American: Chapter 6) and adults who never had cancer (reference group; Chapter 4).

Response rates and reasons for not participating are provided in Table 2. Dropouts were not found to have different demographic characteristics from participants who completed the assessments.

Table 2. Reasons for not participating or dropping out from the ACS-NL validation studies.

Non-participants % (n)	Reliability study 54.5 (115)	Validity study 60 (304)	Reference 30 (105)
Exclusion	22 (25)	20 (61)	27.5 (29)
Unreachable	12 (14)	11.5 (35)	25 (26)
Decline	64.5 (74)	68.5 (208)	44.5 (47)
Technical	1.5 (2)	-	3 (3)

Patients

The ACS was developed for cognitive research on non-CNS cancer patients. Therefore, we aimed to include heterogeneous samples of non-CNS cancer patients, enabling generalizability of our study results to other non-CNS cancer patients. Eligible patients (based on screening of electronic patient files from the Netherlands Cancer Institute; treated between 2010 and 2013) were stratified by age, gender, and tumor type (one third breast; one third testis/ prostate; one third other). We expected patients to be of stable health as we applied the following exclusion criteria: (a) tumor or metastases in the central nervous system; (b) distant metastasis; (c) disease progression; and (d) psychiatric/neurologic symptoms hampering test completion. These criteria were chosen to ensure feasibility of participation and to allow for stable test-retest analyses. We expected mild cognitive problems to be present in a subset of the patient participants (Wefel, Kesler, Noll, & Schagen, 2015). As patients received initial treatment between 1 and 5 years prior to testing, we expected these potential cognitive problems to be relatively stable over the test-retest interval (6 weeks), allowing us to see if it is feasible for mildly affected patients to self-administer the ACS. We did not include patients with CNS cancers, likely presenting with more severe cognitive problems, as we aimed to ensure feasibility in higher functioning individuals first. However, it will be highly valuable to study ACS performance in patients with CNS cancer

and other patients with more severe cognitive and medical symptoms in the future.

Reference group

The reference sample was matched on the demographic characteristics of the patient sample. Generally, we succeeded in including heterogeneous samples, but when expanding reference data, it could be valuable to include additional male participants and participants at the extremes of the age spectrum. For one of the ACS tests—Place the Beads—we had to collect additional reference data by means of a respondent panel because this test was still under development during data collection. This means that, compared to the main reference sample, the Place the Beads sample received fewer personal instructions, had a longer mean test interval, and received financial compensation (€10). It is possible that these differences contributed to Place the Beads having relatively low reliability results.

General limitations

A first limitation of our studies is that people with lower education levels were underrepresented in our participants. All our study samples (the US sample the most, and the reference sample the least) represent mainly highly educated patients. Although this is a common phenomenon in cognitive research, it limits generalizability of results to lower educated groups. For our norm data, the underrepresentation of lower education levels has hampered the assessment of the influence of (the full range of) level of education on ACS performance. In addition, as ceiling effects—generally resulting in high test–retest coefficients—are more likely to take place in highly educated participants, we could have overestimated reliability and validity. However, only for one ACS measure (Wordlist Recognition) a ceiling effect was found.

A second limitation could be that it is possible that patients with relative good computer skills were more likely to participate in the studies of this thesis. Patients with poor computer skills or computer anxiety may have been prone to decline participation in an online study. This may affect the generalizability of the study results to less skilled participants. I will discuss this topic in the next paragraph.

Computer familiarity

Familiarity with the use of computers has been found to influence performance on computerized tests (Bauer et al., 2012; Parsey & Schmitter-Edgecombe, 2013). Subsequently, individual differences in computer familiarity may influence reliability and validity of test scores. Even though research populations are becoming more and more familiar with the use of computers (Chuah et al., 2006), it is still likely that participants will have varying computer skills. Therefore, it is important to consider computer familiarity when interpreting ACS results.

We studied the influence of both tested computer skills and self-reported computer experience on ACS performance; partly because we were interested in both aspects of computer familiarity and partly because—frankly—we were not sure how to optimally assess the concept of

computer familiarity. In the patients, as well as the reference sample, computer skills were found to predict ACS performance to certain extent, mainly on time-based measures. However, interestingly, we also found associations between computer skills and scores on several traditional neuropsychological tests; again, mainly on time-based measures. This could indicate that (1) computer skills overlap with certain cognitive functions; and (2) that there are certain confounding factors at play (e.g. age). Self-reported computer experience, indicated by "mean number of hours computer use per week", did not correlate with offline measures or online measures for which influence of computer skills was not to be expected; it was associated with several online tests that require relatively demanding use of computer input devices (on the domains of processing speed and motor coordination), but not with traditional (offline) measures. Therefore, self-reported computer experience may be most useful as a measure of computer familiarity at the moment. However, notably, number of hours of computer use only covers a limited part of computer familiarity, not providing any information on type and level of computer actions.

Overall, further research on these measures should result in determining an optimal measure of computer familiarity: a sensitive measure that represents relevant computer skills, while independent of general cognitive functions. The question remains, however, how to use this information when interpreting cognitive performance? One option is to adjust for computer familiarity by including the appropriate measure in the regression-based norms. In that case it would be important to prevent "over-correction" to make sure that the measure does not tap into the cognitive constructs of interest and to take confounding variables such as age and education carefully into account (Hansen et al., 2015). As a second option, results on computer familiarity could function as a selection criterion: only participants within a certain cut-off can continue with the online assessment. Alternatively, to avoid the complexity of the issue as much as possible, the ACS could move (even more) towards only including tests that merely require very basic computer skills. For example, drag-and drop actions could be left out completely.

Conclusions

We found that it is feasible to have participants of clinical studies (cancer patients and healthy controls) self-administer the ACS from home. Furthermore, we found encouraging levels of reliability and validity, especially for the ACS as a whole. To favor reliability of cognitive data, researchers should take test setting, computer configuration, and computer skills into consideration (e.g., clear guidelines on test setting and monitoring computer configuration). Reference data that is currently available can be used for initial interpretations of ACS performance, but will be expanded to establish improved norms. In addition, reference data for languages other than Dutch will be collected. Taken together, the results from this thesis indicate that the ACS could contribute to large-scale cognitive data collection in the oncology setting and beyond. While working on current limitations, we will continue with further development and implementation of the ACS.

Implications

Considering the positive user experiences and promising psychometric properties, the ACS seems suitable for implementation in research studies. As long as optimal conditions for collecting reliable data are taken into consideration, the ACS could be used to efficiently measure cognitive functioning and, subsequently, facilitate large-scale data collection. The need for large cognitive data sets in oncology has formed the starting point for the development of the ACS. The ACS could especially be valuable in clinical research studies as home assessments may improve user-friendliness for patients with limited mobility or who have had many treatments. Nevertheless, since the ACS covers a variety of cognitive functions, it can be used in other research areas also.

In addition to allowing for collecting large(r) cognitive data sets, online ACS assessments provide more cognitive data points than traditional assessments and therefore enable looking at cognitive data in more detail. For example, results from a divided attention test such as Connect the Dots do not only include success/error rates and total completion times, but also time-stamps per response, providing information on performance over time and enabling to look into specific responses. Furthermore, online cognitive data are in line with the general increase of electronic research data and with the digitalization of healthcare (Kane & Parsons, 2017), allowing for efficient data storage and integration of data when multiple sources of information are used.

When applying online neuropsychological assessments, it is important to monitor factors that might affect test performance (Troyer et al., 2014). This can be done by automatic registration of technical characteristics (e.g. processor or browser type) and by (online) debriefing on test circumstances. Obtained information may be used to understand performance and to identify outliers. In our studies we have excluded test results based on debriefing only a few times (because reports of a disrupted internet connection or the use of paper and pencil).

In addition, measures must be taken to ensure online data safety. To securely transfer and save data for the studies of this thesis we (1) stored personal data and performance data on two separate servers, both ISO 27001 certified, and (2) used encrypted data transfer (of passwords and user “tokens”) via SSL (2048 bit RSA key). Safety measures should be updated regularly and always comply with the latest privacy legislation. Finally, it is important to upfront think about circumstances in which an online unmonitored assessment will lead to poor reliability or is merely not feasible. For example, people with very poor computer skills or strongly disoriented people (e.g. dementia) will not be able to perform a self-administered online test. Moreover—although outside the scope of the intended use of the ACS—, assessments with high-stakes (e.g. with diagnostic purposes) are currently not suitable for the online setting, as this would require in-person observations or person identification checks. Use for screening might be considered as the ACS could serve as an accessible test to provide first indications of cognitive problems. To extend the options of applying the ACS beyond research and to optimize its use, we could use and improve the potential of online cognitive testing. In the next section, I will elaborate on possible improvements.

Future directions

The future of the ACS

As discussed above, several adaptations will move the ACS forward as a fully functional, user-friendly tool for gathering reliable cognitive research data. Currently, we are working on the following adaptations:

- Several new tests will be developed and added to the ACS in order to increase the coverage of cognitive domains.
- New language versions of the ACS are being developed (e.g. French, German, Spanish, and Swedish).
- Findings from this thesis are being used to optimize ACS measures. Sub-threshold reliability results will be used as main indications for need of improvement.
- Data export is being made more efficient and user-friendly.
- Options for visualization of test results are being developed (for research purposes).
- The influence of computer familiarity on cognitive test performance is further studied, with the aim to improve interpretation of test results.
- Additional outcome measures (from existing tests) are being developed in order to create more sensitive measures.

Furthermore, we plan on working on the following adaptations in the more distant future:

- Reference data will be expanded (in heterogeneous samples) in order to optimize norm scores.
- Parallel versions will be developed in order to enable repeated (longitudinal) testing. After developing parallel versions psychometric properties will be reevaluated.
- Validation of ACS versions in new language versions (apart from Dutch and English) and gathering of reference data for these versions will take place in order to increase possibilities for international data gathering.
- The ACS will be made customizable to enable researchers to combine tests into personalized test batteries.
- The ACS will be made suitable for tablets.
- Options for adaptive testing will be explored in order to make assessments (more) efficient and to facilitate the development of parallel versions (by creating an item bank to select from).
- An end-user manual will be finalized, providing precise instructions on when (and when not!) and how to use the ACS.

The future of cognitive assessments

Over the years, cognitive assessments have been through limited changes. Despite widespread developments in technology and unlike many other fields in science and healthcare, neuropsychological testing continues to rely primarily on traditional paper-and-pencil tasks (Kane

& Parsons, 2017). According to a survey from McMinn and colleagues (2011), in 2009, only 47% of clinical neuropsychologists had used computerized tests (Schatz, 2017). In addition, based on a survey with 512 North American clinical neuropsychologists, only 4% (40/693) of all utilized tests were computerized (Rabin et al., 2014). Main explanations for sticking to traditional tests are financial costs associated with new tests, lack of appropriate normative data, and concerns about utility or validity (Kane & Parsons, 2017; Rabin et al., 2014). In addition, many publishers continue to sell minimally revised versions of long-standing tests, using the same stimuli and materials that were developed in non-clinical setting many years ago (J. B. Miller & Barr, 2017). For example, the RAVLT wordlist still consist out of words selected in 1919.

This does not necessarily mean that traditional neuropsychological tests are invalid and out-of-date, (in fact, many neuropsychological tests perform well in detecting cognitive problems; (Schmand et al., 2011)). However, one may wonder whether the potential to augment cognitive assessments is used optimally. As early as in 1987, and throughout the years, neuropsychologists addressed the potentials of computerized testing and advocated integrating new technologies in neuropsychology (Bilder, 2011; Dodrill, 1997; Kane & Parsons, 2017; Lezak et al., 2004; Miller & Barr, 2017; Parsey & Schmitter-Edgecombe, 2013; Parsons, McMahan, & Kane, 2018). Since then –mainly in research and sports and military psychology– some computerized measures have been adopted for measuring cognition, but the potential to augment cognitive assessment is barely used (Kane & Parsons, 2017; Parsey & Schmitter-Edgecombe, 2013), while current technology and knowledge may add to obtaining useful (additional) cognitive and behavioral information. Note that we have been conservative in the development of the ACS as well; we stuck to the original test designs as much as possible. This approach was chosen in order to create a test tool that provides data comparable to existing neuropsychological data, and to increase acceptance among cognitive researchers. However, we do plan on making use of additional options related to computerized and online testing in the future; see 'Future of the ACS'. Concerning the future of cognitive assessments I agree with Kane & Parsons (both clinical neuropsychologists): 'We should exploit the potential offered by technology to either make the assessment process more efficient or to develop new capabilities that augment the assessment of cognition' (Kane & Parsons, 2017).

Some of these technology-driven options are:

- Additional cognitive data points. The most obvious additional metric is precise measures of processing speed and/or response latency (e.g. time-stamped latencies for TOL moves). Other additional metrics are: speed and consistency of output over time (behavioral patterns), approaches to the test (e.g. organization and planning), pressure (in case of electronic pencils as input device or touchscreen use), evaluation of pauses (number, length, and position), perseverations, etcetera (J. B. Miller & Barr, 2017).
- Additional options for user responses: drawing or written responses via tablet/ electronic pencil, or speech recognition.
- Computer Adaptive Testing. CAT makes use of algorithms that select future test items based on prior performance. This reduces test administration time (discontinuing tests

after minimum criteria have been reached) and increases precision of measurement of cognitive limits (floor or ceiling) (J. B. Miller & Barr, 2017; Parsey & Schmitter-Edgecombe, 2013).

- Integration of (online) computerized testing with other technologies, such as online questionnaires, structural and functional neuroimaging, electro-encephalography (EEG), transcranial direct-current stimulation (tDCS), etcetera (J. B. Miller & Barr, 2017; Rabin et al., 2014).
- Continual behavioral data capture. Wearable sensors, environmental sensors/ systems (e.g. "smart homes"), and constant monitoring of the use of (portable) devices (e.g. typing speed on smartphone) provide the ability to capture behavioral patterns in everyday life (Schultheis & Doiron, 2017). These real-time behavioral data can be used for health monitoring, such as detecting age-related cognitive decline (e.g. (Stringer et al., 2018)).
- Virtual reality (VR). VR allows for replicating real-world activities (such as cooking or driving a car) in a controlled and objectively measured way (Schultheis & Doiron, 2017). VR does require much more equipment than just a computer, but it can closely resemble daily behavior and is often perceived as interesting/ enjoyable by participants.
- Gamification. Cognitive assessments can be integrated into computer games. Game scenarios can be designed to engage certain cognitive processes, which are measured continuously and therefore less obviously for the participant. This type of testing is likely to be most suitable for testing relatively computer-skilled individuals.

As mentioned above, people have varying computer skills. One of the most prominent challenges for the use of technology/computers in neuropsychological testing is that unfamiliarity with technology (often associated with older age) will affect (reliability of) cognitive performance. To what extent this plays a role in the ACS measures needs to be determined in more detail (see paragraph 'Computer familiarity'). Nevertheless, for people who interact with computers daily, behavior from computerized tests may actually resemble everyday behavior more closely than behavior from traditional test. The ease with which people interact with computers is related to the breadth of activities that people carry out in digital environments (Helsper & Eynon, 2010). Generally, these activities are increasing rapidly (in number and breadth) throughout the world. This means that, in addition to the benefits of standardization, efficiency, and increased measurement options, the use of technology in cognitive assessments may be beneficial because it provides a good fit with current daily human behavior. Therefore, in my opinion we should no longer question if cognitive assessments should be developed into making more use of digital technology, but rather how.

It will be important to first have a critical look at the measurement tools that are currently available, including the ACS. What works well (leading to reliable measures) and what could be improved? How could technology be used for possible improvements and for keeping cognitive assessments up-to-date? Technology can be implemented into cognitive testing via collaborations between neuropsychologists, (web)programmers, and human-interaction designers. Main points of attention should be: improving psychometric properties and collecting representative norm

data. Furthermore, it will be important to find ways to finance these developments and to make them sustainable (this means that maintenance, hosting, data management, and technical support need to be financed long term as well).

Altogether, I feel that it would be valuable for neuropsychology to think of innovative ways to augment cognitive measurements. Historically, there has been some resistance towards adapting digital technology and using online testing in neuropsychology (see General Introduction). It seems about time to be less afraid of technology; time to take chances to augment cognitive assessment while at the same time keeping up with developments around us.

Concluding remarks

In this thesis I have provided an overview of what is needed to develop online neuropsychological tests and presented results on the evaluation of our own neuropsychological test battery, the ACS. Results of this evaluation confirm that it is feasible to assess cognitive functioning with self-administrated online tests. In the current chapter I have discussed what is required to further optimize reliability and utility of ACS measures. Fortunately, we have the opportunity to continue working on these issues and to develop and optimize the ACS further. I truly hope that these efforts will enable the ACS to contribute to cognitive research.