Chapter 4
Mental Health Assessment via Internet: The Psychometrics in the Digital Era

Jéferson Ferraz Goularte and Adriane Ribeiro Rosa

Introduction

Mental health assessment is a critical step in the clinical practice and research guiding the treatment and follow-up of patients by clinicians. So far, much of the tools utilized for screening and diagnosis have been paper-and-pencil assessments to evaluate psychopathology of several mental disorders such as depression, anxiety disorders, bipolar disorder, and schizophrenia. As the need to easily handle information about the patient's psychiatric symptoms has increased over time, the paper-and-pencil instruments have been transformed into digital questionnaires and used in different digital formats to assess mental health. Recently, there was a high number of mobile phone mental health assessment applications (apps) available on platforms such as Google Play (Android) and iPlay (iOS) accessible to anyone with a smartphone or tablet. Mobile health (mHealth) is a promising field available to clinicians and patients from distinct areas of medicine including psychiatry. In this

J. F. Goularte
Laboratory of Molecular Psychiatry, Hospital Clinic of Porto Alegre, Porto Alegre, RS, Brazil

Postgraduate Program in Psychiatry and Behavioral Sciences, University Federal of Rio Grande do Sul, Porto Alegre, RS, Brazil
e-mail: jefgoularte@hcpa.edu.br

A. R. Rosa (✉)
Laboratory of Molecular Psychiatry, Hospital Clinic of Porto Alegre, Porto Alegre, RS, Brazil

Postgraduate Program in Psychiatry and Behavioral Sciences, University Federal of Rio Grande do Sul, Porto Alegre, RS, Brazil

Department of Pharmacology, Institute of Basic Health Sciences, University Federal of Rio Grande do Sul, Porto Alegre, RS, Brazil

I. C. Passos et al. (eds.), *Digital Mental Health*,
https://doi.org/10.1007/978-3-031-10698-9_4

chapter, we review the current literature regarding the psychometric properties of the self-reported digital instruments used for screening, diagnosis, symptoms, and treatment response of mental illness. When available, the paper-and-pencil questionnaires are compared to its transformed digital version. In addition, we discuss the potential and limitations of mHealth in the assessment of mental disorders.

Psychometrics: A Brief Overview

There are several psychological scales available that are able to assess aspects of human behavior such as personality traits, thoughts, memory, cognition, mood, and motivation. However, all scales used to measure these psychological characteristics must be meaningful and reliable. The science that analyzes the basic principles of psychological scales is known as psychometrics [1] and deals with the validity and reliability of instruments that measure some hypothetical construct (for example, depression, anxiety, self-esteem, intelligence, etc.).

When we say that a scale is valid, we are referring to the degree of an instrument that explains the behavior that is intended to be measured. According to the Standards for Educational and Psychological Testing, "validity refers to the degree to which evidence and theory support the interpretations of scale scores for proposed uses of scales. Validity is, therefore, the most fundamental consideration in developing and evaluating instruments" [2]. Furthermore, in case an already developed scale is transformed from paper-and-pencil format to digital format (web page, computer software, and mobile application), there are some steps necessary to assess whether the two formats are equivalent, such as some problems they may arise from differences in the visual presentation of the items and the environment in which the assessment was carried out [3]. Thus, studies that assess psychometric equivalence in different formats of a scale (e.g., paper vs. digital) are needed to ensure that both instruments measure the same construct.

In terms of psychometric properties, there are objective ways to analyze the validity and reliability of an instrument based on the contemporary view that *construct validity* is the essential concept of validity. In this sense, construct validity is the degree to which an instrument score represents or can be interpreted as reflecting a psychological construct (e.g., anxiety, depression, self-esteem, motivation, etc.). According to some authors, the validity of an instrument can be assessed by types of evidence, such as *content validity* (*face validity*), *the internal structure of the scale or reliability* (*internal consistency and test-rest*), *construct validity* (*convergent validity and discriminant validity*), and *criterion validity* (*concurrent and predictive validity*) [4]. Thus, validity is a unitary concept and those types of evidence taken together add information about the scale validity.

Content Validity

The content validity refers to the items or questions of a scale and the content that would be expected in this instrument to measure a specific construct. The items of an instrument must include all relevant facets of the construct; otherwise, this instrument may have irrelevant content of the construct (question or items) and reduce validity. For example, an instrument to measure occupational functioning includes some questions relating to the ability to work, or looking for a job, or the ability to take care of one's home on their own [5]. However, if the instrument had included questions about work preferences or house cleaning skills, they would likely be irrelevant items for measuring occupational functioning and would not reproduce the functioning construct. In addition, the construct facets should be composed of as many questions or items as possible that represent the construct to avoid reduced validity by under-representation of the construct.

Another important aspect in assessing the content validity of a scale, especially in the process of developing a new one or translating it into different languages, is face validity [4]. Face validity deals with how the respondent perceives the items of an instrument as relevant to measure the construct under study. For example, Mustafa et al. [6] translated and adapted the mHealth App Usability Questionnaire (M-MAUQ) into Malay, an app that aims to assess the usability of mobile apps and measure face validity by comparing expert scores and target user opinions on the understandability of the translated M-MAUQ items. In this example, all items had an excellent level of agreement (modified kappa >0.75) with a mean face validity index for 18 items (understandability = 0.961), indicating equivalence of face validity with the original version.

The Internal Structure of the Scale

The internal structure of the test is another important aspect while analyzing the validity of a new instrument. The internal structure refers to the items (or questions) in a scale and how they are related to each other to form one or more clusters that reflect the construct intended to be measured. Usually, items that strongly correlate with some items but weakly correlated with other items form clusters, indicating more than one domain is being measured. This is particularly useful to understand if the scale allows the assessment of a global measure or specific domains of the construct. Therefore, if a test was developed to have one dimension and the factor analysis shows a good correlation between items, there is good evidence that the internal structure predicted was achieved [7]. The internal structure of scales is commonly measured by means of factorial analysis (exploratory or confirmatory factor analysis) or principal component analysis.

Internal Consistency Reliability

Internal consistency reliability assesses the degree to which questions on an instrument measure the same underlying concept. It can be used to determine the consistency of instrument score when it is applied at once or across replications of the same test. When the test–retest approach has been applied the analysis of scores in distinct periods of time may be assessed by correlation analysis, while coefficient alpha or Cronbach's alpha may be used when the instrument was applied once [1, 2, 8]. Furthermore, the reliability of a score can be estimated empirically by its reliability coefficient, generalizability coefficient, item response theory (IRT) information functions, standard errors, error/tolerance ratios, or various indices of classification consistency [2]. Based on the classical test theory (CCT), the reliability coefficients are estimated by statistical analysis of internal consistency.

In general, reliability can be considered as strong or weak as there is no score that represents a 100% reliablility. Keeping that in mind, and according to CTT, the reliability coefficient of a score ranges from 0 to 1, with 0 indicating no evidence of reliability and 1 a perfect measure of reliability. As the CTT takes into account observed scores, true scores, and measurement error, a score with a reliability coefficient of 0.70 would indicate that 70% of the score is actually measuring a true score of a construct and 30% of the score is a measuring error of any source [2]. According to some authors, a reliability coefficient >0.70 means a satisfactory level of reliability [8].

Construct Validity (Convergent Validity and Discriminant Validity)

Convergent Validity

Convergent validity refers to a construct measured in different ways that produce similar results. Specifically, it is the degree to which scores on a studied instrument are related to measures of other constructs that can be expected on theoretical grounds to be close to the one tapped into by this instrument. Evidence of convergent validity of a construct can be provided by the extent to which the newly developed scale correlates highly with other variables designed to measure the same construct. Therefore, if the score of the newly developed scale is highly correlated with another scale that measures the same construct, we conclude there is some level of convergent validity [4].

Discriminant Validity

Discriminant validity refers to a measure that is novel and not simply a reflection of some other construct [4]. In other words, it is the degree to which the scores of a studied instrument are differentiated from the behavioral manifestations of other constructs, which, from a theoretical point of view, cannot be related to the underlying construct of the investigated instrument [4]. For instance, González-Robles et al. [9] studied the psychometric properties of the online version of the Overall Anxiety Severity and Impairment Scale (OASIS) among Spanish patients with anxiety and depressive disorders, including discriminant validity. In this study, correlation of OASIS with the Positive and Negative Affect Schedule-Positive Affect was not as high ($r = -0.40$, $p < 0.01$) as for Beck Anxiety Inventory (BAI, $r = 0.61$, $p < 0.01$), suggesting OASIS maintained the property to evaluate symptom of anxiety and not positive affect.

Criterion Validity (Concurrent and Predictive Validity)

In addition to what has been mentioned so far, the other relevant aspects of validity are concurrent validity and predictive validity [4].

C*oncurrent validity* refers to the relationship between the scores of two instruments measuring the same construct taken at the same time, usually the new instrument compared to another "gold standard" for the construct of interest. For example, the BDI score of depression delivered through the ReMAP app showed good correlation with "gold standard" clinician-rated depression severity using the HDRS in a subset of the sample ($r = 0.78$), suggesting evidence of concurrent validity [10].

Contrary to concurrent validity, *predictive validity* is the extent to which a measure predicts the answers to some other question or a result to which it ought to be related with, i.e., the scale should be able to predict a behavior in the future [4]. For instance, the online version of the Dutch Penn State Worry Questionnaire (PSWQ), a self-reported assessment of pathological worry, had their predictive validity estimated by relationship with worry frequency and worry duration variables [11]. In this study, score of PSWQ was significantly associated with the total time spent worrying during the day ($r(187) = 0.446$, $p < 0.001$) and during the night time ($r(187) = 0.324$, $p < 0.001$), as well as with the frequency of worry episodes during the day ($r(187) = 0.418$, $p < 0.001$), and during the night time ($r(187) = 0.310$, $p < 0.001$), suggesting that worry frequency and worry duration were predicted by PSWQ scores.

The Psychological Process Used in the Scale Responses

The psychological process used in the test responses deals with the cognitive process that a respondent uses while answering a test and the cognitive process they should use to answer the test [1, 2]. This is an important step in assessing the validation degree of a measure as any deviation of expected process to answer a test can affect the test score beyond the intended purpose of the test. Some authors exemplified the issue of process used in a test when test takers used more than cognitive attentional resources to answer a word task [1]. In this example, the scores were inflated because one group did not follow the rules and the scores did not show strong evidence of validity.

Consequences of Using Test

The consequences of using the test deal with sources of bias and useful application of scores when making decisions, affecting the degree of validity of the construct measure and their intended use. For example, men who take the test score higher than women on a screening for depression and, for that result, are referred to see a psychiatrist. However, there are some concerns that the test items were not truly gender balanced and therefore male was given priority in the consultation. In this hypothetical scenario, construct validity is impaired, as scores can be biased and result in adverse consequences for test participants. Typically, most instrument comparisons in clinical practice do not assess this aspect of construct validity.

In sum, we must give an overview of the main components of the psychometric properties commonly used by researchers to assess the validity and reliability of scales when they are developed or for existing instruments that are transformed into digital format, mainly for the purpose of helping the reader in the following sections. However, it is beyond the scope of this section to discuss Item Response Theory (IRT) as another method for evaluating measurement at scale. For this, we suggest readers to read [12] as a starting point. Finally, we have chosen examples to clarify most definitions of validity, although we cannot guarantee that the results given in the examples are in fact a confirmation of validity, as validity is a matter of degree rather than a matter of yes/no.

Psychometric of Mental Health Instruments: Paper-and-Pencil Versus Digital Formats

With the widespread use of the internet in the 1990s, the assessment of mental disorders started a new era of digital assessments through computer-based assessment, internet web page assessments, and more recently by mobile apps through

smartphone or tablets [7, 13, 14]. While in the previous section we discussed the main steps to consider when assessing the validation of new instruments in psychiatry, here we describe the process that must be followed for those instruments to be transformed from paper-and-pencil format to digital versions.

While the instruments available in digital format cover a broad range of mental illnesses [15], there are some concerns that psychometrics of the digital format may not be the same as the original paper-and-pencil format and can affect, to some extent, the validity and reliability of the scores measured [16]. For instance, the assessment of mild cognitive impairment by the Cambridge University Pen to Digital Equivalence assessment (CUPDE) showed significant differences in reliability and validity of scores to its paper-and-pencil version Saint Louis University Mental Status examination (SLUMS) [17], even after change from web-based to app-based interface/layout [18]. In addition, the assessment of anxiety in patients with panic disorder by the internet-based BDI questionnaire showed significant difference in means scores, with lower scores observed in the internet version compared to pen and paper assessment [19]. Furthermore, not all studies assessing psychological symptoms by mobile apps have been validated suggesting that more studies are needed to analyze the equivalence between instruments [16].

In this sense, the equivalence of different formats of instruments used in psychiatry has been reviewed by some studies [7, 13, 14] considering some aspects of validity and reliability. According to van Ballegooijen et al. [13] the equivalence between distinct formats should be assessed by the same steps used in the validity and reliability studies of newly developed scale. Therefore, the following tests should be considered in order to examine equivalence between formats: internal consistency, test-retest reliability, measurement error, internal structure and model fit or explained variance, correlation between the two instruments, difference in mean scores between online and paper versions and criterion validity in terms of sensitivity and specificity (for the optimal cut-off point). Likewise, another systematic review [7] highlighted the importance of performing test-retest reliability, internal consistency, and mean differences between instruments, including the effect size test.

Furthermore, there is some evidence that respondent's perception of the questions delivered should be taken into account and may produce evidence of face validity. For instance, participants reported preference for single items instead of multiple items per web page when they answered instruments such as Beck Depression Inventory (BDI), Beck Anxiety Inventory (BAI), Quality of Life Index (QOLI), and Montgomery-Åsberg Depression Rating Scale (MADRS) [20]. It is also important to consider the respondent perception of the digital layout along with functionality, navigation, personalization, and appearance of a mobile app [21].

In sum, all those aspects might influence the way that respondents answer questions, thus affecting the validity and reliability of the instrument. Thus, instruments that assess psychological symptoms need further validation study when the original format is adapted to digital devices, including original paper-and-pencil versions transformed to computer-based instruments, web page instruments, and mobile applications. In the tables below, we summarize psychometric properties (i.e., face validity, discriminant validity, concurrent validity, internal consistency, intraclass

correlation coefficients (ICC), correlation, and mean scores comparisons) of some digital instruments based on pen and paper scales commonly used to measure symptoms in the field of psychiatry.

Online Web Page Self-Reported Questionnaires

The online web page includes any platform accessed over the Internet using a browser. This digital format requires an Internet connection and a mouse, keyboard, or fingertip as devices to navigate and select web page content. In the field of mental health, few studies have compared the equivalence of an online web page with pen and paper [7, 13]. Overall, online and pen and paper versions have been compared in terms of correlation between scores, comparing score's mean, effect size of differences, internal consistency, convergent validity and criterion validity [13] (see Table 4.1). For example, instruments that assess symptoms of anxiety have shown

Table 4.1 Psychometric properties of instruments to assess self-reported symptoms of depression and anxiety in online web pages

Study	Instrument	Format correlation[a]	Means (SD)		Format difference	Test–retest reliability[b]	Internal consistency[b] (Cronbach's alpha)
			PnP	Digital			
Depression							
Brock et al. [3]	CES-D	N/a	12.00 (7.09)	12.17 (7.75)	N.s.	ICC = 0.84	N/a
Bush et al. [22]	PHQ-9	ICC = 0.92	5.9 (5.6)	5.1 (4.9)	N.s.[c]	N/a	0.85
Carlbring et al. [19]	BDI-II	$r = 0.94$	17.52	18.01	$F = 6.3$, $p < 0.05$, $d = 0.27$[d]	N/a	0.88/0.89
	MADRS-S	$r = 0.91$	16.69 (7.4)	16.42 (7.1)	N.s.	N/a	0.82/0.83
			17.11 (9.4)	16.79 (8.3)			
Fortson et al. [23]	CES-D	N/a	13.81 (8.89)[e]	12.34 (8.59)[e]	N.s.	N/a	0.88/0.89
Fortson et al. [23]	CES-D	N/a	13.81 (8.89)[e]	12.34 (8.59)[e]	N.s.	N/a	0.88/0.89
Holländare et al. [29]	BDI-II	$r = 0.89$	30.55 (10.72)	29.68 (10.07)	N.s.	N/a	0.87/0.89
	MADRS-S	$r = 0.84$	24.43 (6.97)	23.79 (7.98)	N.s.	N/a	0.73/0.81
Herrero and Meneses [30]	CESD-7	N/a	11.85 (3.78)	11.57 (3.79)	N.s.	N/a	0.82

Table 4.1 (continued)

Study	Instrument	Format correlation[a]	Means (SD)		Format difference	Test–retest reliability[b]	Internal consistency[b] (Cronbach's alpha)
			PnP	Digital			
Thorén et al. [31]	HADS	$r = 0.67$	7.3 (5.9)	6.6 (5.4)	N.s. subscales.	N/a	0.85
Whitehead [27]	HADS-depression	N/a	3.24 (3.05)	3.52 (3.04)	N.s.	N/a	0.76
Yu and Yu [32]	CES-D	N/a	12.14 (8.02)	11.03 (7.87)	$t = 2.39$, $p = 0.02$[c]	N/a	N/a
Zimmerman and Martinez [33]	CUDOS	ICC = 0.96	20.0 (14.6)	20.6 (13.9)	N.s.	N/a	0.93
Anxiety							
Brock et al. [3]	BAI	N/a	8.55 (6.87)	6.21 (6.42)	N.s.	ICC = 0.84	>0.70
			9.08 (8.72)	9.43 (6.96)			
Carlbring et al. [19]	BAI	$r = 0.84$	22.62	19.63	$F = 82.2$, $p < 0.01$, $d = 0.98$[d]	N/a	0.88/0.91
Hirai et al. [34]	SIAS	N/a	20.5 (12.39)	20.0 (13.23)	N.s.	N/a	0.93
	SPS	N/a	15.6 (10.68)	16.4 (12.66)	N.s.	N/a	0.93
Whitehead [27]	HADS-anxiety	N/a	6.31 (3.72)	6.39 (3.68)	N.s.	N/a	0.80

 Note: Results are shown for total sample in studies with many groups. *PnP* pen and paper, *N.s.* non-significant, *N/a* not available, *CES-D* Center for Epidemiological Studies Depression, *PHQ-9* Patient Health Questionnaire-9, *BDI-II* Beck Depression Inventory-II, *MADRS-S* Montgomery–Asberg Depression Rating Scale-Self-report, *CESD-7* Center for Epidemiologic Studies Depression scale-7, *HADS* Hospital Anxiety and Depression Scale, *HADS-Depression* Hospital Anxiety and Depression Scale-Depression, *CUDOS* Clinically Useful Depression Outcome Scale, *BAI* Beck Anxiety Inventory, *SIAS* Social Interaction Anxiety Scale, *SPS* Social Phobia Scale, *HADS-Anxiety* Hospital Anxiety and Depression Scale-Anxiety

[a] ICC = Intraclass Correlation and r = Pearson's r

[b] Digital version

[c] t-tests conducted and interpreted by ref. [7] based on values from original article

[d] Effect sizes calculated by ref. [7] based on values from the original article

[e] Mean score calculated and standard deviation estimated by ref. [7] based on values from the original article

good degrees of reliability (Table 4.1). However, the Beck Anxiety Inventory (BAI) assessed online showed a remarkable difference in terms of average compared to the paper-and-pencil version.

In general, instruments that assess post-traumatic stress disorder had a good level of reliability when delivered in web page format compared to their pen and paper instrument counterpart. For example, means scores of the PTSD Check List–Civilian Version (PCL-C), Trauma Symptom Screen Frequency (TSS Frequency), Trauma Symptom Screen Distress (TSS Distress), and Traumatic Life Events Questionnaire (TLEQ) were similar to pen and paper version [7]. In addition, all showed format correlation (ICC and/or $r > 0.65$) with pen and paper and internal consistency >0.80 in the web page format [22–24].

As for other measures summarized so far, questionnaires that assessed self-reported symptoms of panic disorders (Body Sensations Questionnaire, BSQ; Agoraphobic Cognitions Questionnaire, ACQ; Mobile Inventory Accompanied, MI Accompanied; Mobile Inventory Alone, MI Alone) showed a good reliability, with format correlation (ICC or $r > 0.90$) with pen and paper and high internal consistency (Cronbach's alpha >0.9) [19, 25]. However, assessment of web page means scores showed that BSQ, ACQ, and MI alone might slightly differ from pen and paper score [7]. Even though the results are informative, researchers have to consider such differences when transforming the pen and paper format to web page format of those instruments.

The instruments used to measure perceived physical and mental health had not performed well in web page format. For instance, there were some differences in scores on subscales of General Health Questionnaire-28 (GHQ-28) and Symptom Checklist 90 Revised (SCL-90-R) [7], indicating the scores of subscales might not be consistent with the pen and paper versions. However, format correlation (GHQ-28 $r = 0.49–0.92$; SCL-90-R $r = 0.74–0.96$) and internal consistency (Cronbach's alpha >0.90) showed some evidence of validity [26]. Other scales such as the Short Form [12] Health Survey Version Two (SF12V2) had similar scores compared to paper-and-pencil version [7] with moderate to good internal consistency (Cronbach's alpha of 0.68) [27]. Thus, researchers should use GHQ-28 and SCL-90-R with caution regarding scores of subscales, while SF12V2 might be a good alternative to assess the physical and mental health construct.

The instruments to assess self-reported drug abuse had shown a good level of evidence of reliability. For example, the Alcohol Dependence Scale (ADS), the Alcohol Use Disorder Identification Test (AUDIT), the Rutgers Alcohol Problem Index 1 month (RAPI 1 month), the Rutgers Alcohol Problem Index 6 months (RAPI 6 months), and the Rutgers Alcohol Problem Index 1 year (RAPI 1 year), all showed equivalence of means scores to pen and paper versions [7]. In addition, all performed very well regarding test-retest reliability ($r = >0.78$) [28].

The only instrument analyzed by Alfonsson et al. [7] to assess symptoms of insomnia, the ISI, showed a good reliability compared to the pen and paper version. For instance, analysis showed format correlation of 0.99/98 and internal consistency of 0.61/0.88 [20], with identical mean scores in paper and pen (15.86 ± 3.80) compared to online version (16.00 ± 3.87) when compared by statistical analysis [7].

Altogether, there is a good level of evidence that instruments that assess a wide range of psychological symptoms by online web pages maintain equivalence with pen and paper measurements, except for a few subscales that assess panic symptoms (BSQ and ACQ, with lower and upper marginal scores, respectively, compared to pen and paper format) and physical and mental health (SCL-90-R and GHQ-28) which showed some differences in mean scores.

Computer-Based Instruments

Computerized self-report instruments are digital versions of pen and paper questionnaires delivered through desktop software [35] instead of a web page accessed through the internet. For instance, the PHQ-9 and BDI-II were part of a computer-based therapy design to improve symptoms of depression delivered by a flash drive on a designated computer onsite in an outpatient clinic [36]. The assessment of mental health by computer-based instruments also covers a broad range of self-reported symptoms, including depression and anxiety (Table 4.2).

In the assessment of depression, the BDI was studied by four independent authors [35, 37–39], with a good reliability (Table 4.2). For the assessment of reliability

Table 4.2 Psychometric properties of instruments to assess self-reported symptoms of depression and anxiety in computer-based studies

Study	Instrument	Format correlation[a]	Means (SD)		Format difference	Test–retest reliability[b]	Internal consistency[b] (Cronbach's alpha)
			PnP	Digital			
Depression							
George et al. [37]	BDI	N/a	6.02 (5.17)	8.21 (4.69)	$t = 2.18$, $p < 0.05$, $d = 0.44$[c]	N/a	N/a
Glaze and Cox [43]	EPDS	$r = 0.98$	13.34 (7.60)[d]	13.59 (7.75)[d]	N.s.[e]	N/a	N/a
Kurt et al. [44]	GDS-15	$r = 0.72/0.83$	17.68 (2.48)[d]	17.59 (2.38)[d]	N.s.[e]	$r = 0.70$	N/a
	CESD-R 20	$r = 0.61/0.74$	10.19 (14.11)[d]	10.59 (10.85)[d]	N.s.[e]	$r = 0.85$	N/a
Lankford et al. [38]	BDI	N/a	5.72 (3.83)	6.32 (4.34)	N.s. effect of format	N/a	N/a
Lukin et al. [39]	BDI	N/a	7.68 (5.88)	7.67 (5.84)	N.s.	N.s. effect of time.	N/a
Murrelle et al. [42]	CES-D	$r = 0.54$	N/a	N/a	N/a	N/a	N/a
Ogles et al. [45]	CES-D	$r = 0.96$	N/a	N/a	N/a	N/a	0.91

(continued)

Table 4.2 (continued)

Study	Instrument	Format correlation[a]	Means (SD)		Format difference	Test–retest reliability[b]	Internal consistency[b] (Cronbach's alpha)
			PnP	Digital			
Schulenberg and Yutrzenka [46]	BDI-II	$r = 0.98$	8.83 (6.80)	10.09 (9.08)	N.s.	N/a	0.91
Anxiety							
George et al. [37]	STAI-S	N/a	34.88 (7.03)	38.69 (9.61)	$t = 2.23$, $p < 0.05$, $d = 0.45$[c]	N/a	N/a
Lukin et al. [39]	STAI-T	N/a	46.35 (6.77)	46.06 (8.23)	N.s.	N.s. effect of time	N/a
Murrelle et al. [42]	STAI	$r = 0.35$	N/a	N/a	N/a	N/a	N/a

This table has been adapted from ©Sven Alfonsson, Pernilla Maathz, Timo Hursti. Originally published in the Journal of Medical Internet Research (http://www.jmir.org), 03.12.2014. This is an open-access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. Note: Results are shown for total sample in studies with many groups. *PnP* pen and paper, *N.s.* non-significant, *N/a* not available, *BDI* beck depression inventory, *EPDS* Edinburgh Postnatal Depression Scale, *GDS-15* Geriatric Depression Scale-15, *CESD-R 20* Center for Epidemiologic Studies Depression scale -R-20, *CES-D* Center for Epidemiologic Studies Depression scale, *BDI-II* Beck depression Inventory-II, *STAI-S* State-Trait Anxiety Inventory-State, *STAI-T* State-Trait Anxiety Inventory-Trait, *STAI* State-Trait Anxiety Inventory

[a] ICC = Intraclass Correlation and r = Pearson's r

[b] Digital version

[c] Effect sizes calculated by ref. [7] based on values from the original article

[d] Mean score calculated and standard deviation estimated by ref. [7] based on values from the original article

[e] t-tests conducted and interpreted by ref. [7] based on values from original article

between pen and paper to computer-based format, studies performed with anxiety instruments showed few data to allow a full analysis, including few with scores and intraclass correlation (Table 4.2).

The study of Schmitz et al. [40] reported comparison of pen and paper and computerized versions of the SCL-90-R to assess perceived mental health. In this study, there was high internal consistency (Cronbach's alpha =0.98), but there was no information regarding interformat correlation. In addition, there was no statistical difference in mean scores between formats (pen and paper: 1.20 ± 0.66 vs. computerized version: 1.29 ± 0.66) [7].

The studies performed by Chan-Pensley [41] and Murrelle et al. [42] assessed the psychometrics of instruments delivered by computer to measure alcohol and tobacco dependence or misuse. The instruments AUDIT (mentioned in item 3.1), Michigan Alcohol Screening Test (MAST), CAGE Substance Abuse Screening Tool

(CAGE), Drug Abuse Screen Test (DAST), and Fagerstrom Tolerance Questionnaire (FTQ) showed format correlation $r = >0.65$. More recent analysis showed that the computerized version of AUDIT had very close scores compared to the paper-and-pencil version, while the other instruments did not report mean scores for comparison studies [7]. However, not all studies have assessed format correlation and internal consistency between instruments, which may limit the interpretation of results.

The paper-and-pencil scales transformed to computer-based instruments were the earliest digital format used to assess psychological symptoms. In general, most scales delivered through computer software showed some evidence of equivalence to pen and paper format, except for BDI (depression) and STAI-S (anxiety) that had higher means scores in the pen and paper version [7].

Mobile Application (App) Format

The number of health apps available for download can be as high as 325,000 according to estimates published in 2017 [47], with >10,000 related to mental health [48]. The use of mHealth technologies in severe mental disorders such as bipolar disorder, schizophrenia, and major depressive disorder has been systematically reviewed yielding valuable results regarding the psychometric properties of some apps [15]. Most studies in the area of mental health assessment through mobile apps were published after 2013 [13], probably as a result of the widespread use of smartphones. Thus, in this section, we summarize some findings in the field published in recent years.

The Mobile Screener was an app developed in an iOS platform (iPhone) to assess symptoms of PTSD (PTSD Checklist, PCL-C), depression (Patient Health Questionnaire-9, PHQ-9), suicidal ideation (Revised Suicidal Ideation Scale, R-SIS), anger (Dimensions of Anger 5, DAR5), common sleep difficulties and daytime tiredness (Sleep Evaluation Scale), and clinical symptoms (BI Self-Report of Symptoms) in health volunteer soldiers [22]. All measures were analyzed by internal consistency and intraclass correlation between app and pen and paper formats. In general, digital scores in all instruments were close to the original format and with intraclass correlation ranging from 0.62 (DAR5) to 0.95 (Sleep Evaluation Scale). In addition, these apps were satisfactorily qualified by the respondents as easy to submit answers, navigation through pages, sections, and questions. Indeed, more than 70% of them prefer digital app format rather than other formats of questionnaires [22]. However, the limitations of the study included the assessment of symptoms in healthy volunteers and sample ($N = 46$) meaning the results may not be generalized to patients.

Another study developed a mobile tablet app to measure psychosocial functioning in patients with schizophrenia based on the pen and paper full version of University of California San Diego Performance-Based Skills Assessment (UPSA) [49]. The mobile app (UPSA-M) retained 4 out of 5 subsets (planning recreational

activities, finance, communication, and transportation) of the original version. The UPSA-M app showed feasibility and 80% sensitivity to differentiate health subjects from patients with schizophrenia, and the app scores significantly correlated with UPSA pen and paper version ($r = 0.61$). However, in the health controls the correlation did not reach significance ($r = 0.24$). The authors stated that the USPA-M may possess the same psychometric properties of full UPSA and further studies are needed to validate for use in clinical practice [49].

The app ClinTouch was developed to assess daily self-reported psychosis compared to face-to-face Positive and Negative Syndrome Scale (PANSS) and Calgary Depression Scale (CDS) interview [50]. The app was developed in the Android platform and contained two sets of questions based on PANSS and CDS. Set 1 consisted of questions to assess guilt, hopelessness, depression, social withdrawal, conceptual disorganization, excitement, and hallucinations, while set 2 assesses anxiety, grandiosity, hostility, somatic concern, guilty ideas of reference, paranoia, and delusions. The validity of ClinTouch was evaluated in remitted patients, acutely psychotic patients, and those with ultra-high risk of developed psychosis. The patients showed good compliance with the study procedure and only those who had negative symptoms were likely to show greater reactivity to the app (i.e., changing thoughts or mood by answering the questions). In addition, alpha scores showed satisfactory internal consistency (Cronbach's alpha >0.76). In general, there were significant correlations with PANSS positive and affective symptoms while no correlation between the passive and apathetic social withdrawal, hostility, excitement, and cognitive disorganization subscales with the PANSS subscales, suggesting there are some limitations in self-reported assessment in this group of patients.

Apps that allow patients to assess daily measures of mania and depression are extremely useful to provide data on mood changes over time and be used as a guide to prevent relapse in individuals with bipolar disorder. The "Monitoring, treatment and prediction of bipolar disorder episodes" (MONARCA) is a specific app developed to assess mood symptoms in bipolar disorder. This app asked participants to assess every evening (during 3 months) items regarding subjective mood, sleep duration, medicine intake, irritability, activity level, mixed mood, cognitive problems, alcohol consumption, stress, and individual warning signs. In addition, objective measurements were automatically taken regarding social activity, physical activity, speech duration, and cell tower ID. The MONARCA validity study showed 88% adherence to self-report measures using the app and significant correlation between depressive symptoms measured by the app and the Hamilton 17-item Depression Rating Scale interview. However, no correlation was found between Young Mania Rating Scale and self-reported manic symptoms, which was explained by the low prevalence of mania in the sample subpopulation (YMRS score = 2.7) [51].

The Mindful Moods app was developed to assess real-time symptoms of depression in real life in a sample of adult patients with major depressive disorder ($N = 13$) using a smartphone version of the PHQ-9 three times a day for 29–30 days [52]. Respondents received survey notifications via the app with three random PHQ-9

pen and paper questions to answer throughout the day on a Likert scale. In addition, patients attended personal visits to respond to a PHQ-9 pen and paper at the beginning and end of the study. The analysis showed good scoring correlation between the two formats ($r = 0.84$), although the app's scores were on average 3.02 points higher than the pen and paper version. Furthermore, suicide at levels 2 and 3 was reported only in the PHQ-9 app version, suggesting the scenario and may have influenced responses and scores. In addition, adherence to the study protocol was 77% for 30 days, suggesting the feasibility of a long-term protocol to assess symptoms of depression in real time.

Another study developed the Remote Monitoring Application in Psychiatry (ReMAP) app to collect ecological momentary assessment (EMA) symptoms of depression in a sample of healthy controls, patients with Major Depressive Disorder (MDD), bipolar disorder, social anxiety disorder (SAD), MDD with comorbid SAD, or specific phobia (SP) with spider subtype [10]. The study app was the digital format of the BDI and the concordance of scores with the paper-and-pen versions of the BDI, BDI-II, and HDRS assessed by the physician was compared by correlation of the intraclass coefficient and internal consistency (Cronbach's alpha). The overall agreement of the BDI between formats was high (ICC = 0.92), but lower for healthy controls (ICC = 0.63) and patients with anxiety disorders (ICC = 0.72). The internal consistency of ReMAP BDI (Cronbach's $\alpha = 0.944$) was similar to pen and paper BDI (BDI-I: $\alpha = 0.945$; BDI-II: $\alpha = 0.944$). In addition, concurrent validation was established for the ReMAP BDI, which was correlated with clinician-rated depression severity using the HDRS in a subset of the sample ($r = 0.78$) that was comparable to the association between the HDRS score and the score of the pen and paper BDI ($r = 0.68$), suggesting ReMAP showed evidence of equivalence with pen and paper BDI in bipolar patients.

Lastly, a recent systematic review determined the feasibility and evidence of validity of mobile apps developed to monitor episodic symptoms and course of symptoms over time in bipolar disorder patients [14]. The review included 13 studies, but only eight studies assessed the equivalence of the scores obtained in the digital version with clinician-rated assessment or pen and paper self-reported scales. In general, the authors concluded that there is some evidence of concurrent validity for the app Monsenso system (compared to clinician-rated HDRS and YMRS) and MONARCA (compared to clinician-rated HDRS-17 and YMRS), while a mood chart scale app did not show concurrent validity compared to pen and paper mood chart, MADRS and YMRS. In addition, there was convergent validity between the app MONARCA self-reported mixed symptoms and Cohen Perceived Stress Scale (PSS), but not with the abbreviated World Health Organization Quality of Life scale (WHOQoL-BREF) scores. Furthermore, the app MONARCA showed convergent validity for both irritability and mood instability with the Functional Assessment Short Test (FAST), PSS, and WHOQoL-BREF. These findings suggest that mobile app-based self-report tools are valid in the assessment of symptoms of mania and depression in euthymic patients with bipolar disorder.

Conclusion

The evaluation of psychometric properties of instruments transformed into digital format has great potential in psychiatry. When developing a scale, the researcher must carefully examine all types of validity evidence when developing new scale formats based on previous excellent mental health instruments. First, the selection of gold standard instruments is suggested, ideally those that were studied in the target population (general population or clinical sample). In the case of developing new instruments in digital format from scratch, it would be extremely important to choose the appropriate construction and content of the instrument, usually based on previous instruments and the opinions of experts in the field. Second, another key aspect of developing a digital assessment is testing whether the target population is able to use the format, especially considering their ability to use mobile devices. Ideally, pilot studies with the target population would improve face validity before establishing a new scale in digital format. Third, after collecting data in a pilot study, check the agreement of the internal structure of the digital instrument with the original paper-and-pen version, usually by internal consistency and factor analysis. If not fully adhered to, consider the extent to which the differences might impair the accuracy of the construct being measured. Fourth, it is very important to compare the scores of newly developed digital format instruments with other instruments that measure the same construct to confirm concurrent validity, preferably with a gold standard instrument. Finally, and most importantly, researchers must plan carefully before starting research on scale validation in a new format, as digital health is continually evolving in the way data is collected.

In conclusion, the field of digital mental health assessment has evolved over the past 25 years from computer-based instruments to today's use of mobile apps to measure symptoms across a wide range of mental conditions. Although mobile assessment psychometrics has been studied for some recognized instruments, it is imperative that more psychometric studies be carried out in patients with symptoms of anxiety, post-traumatic stress disorder, and dependence or misuse of alcohol and tobacco. In addition, the respondent's perception of digital layout in mobile apps, along with their judgment on navigation, safety, and ease of use, should be addressed in future studies that compare the psychometric properties of mobile app questionnaires with their paper-and-pencil versions.

References

1. Furr RM, Bacharach VR. Psychometrics: an introduction. 2nd ed. SAGE; 2014.
2. American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational and Psychological Testing. Standards for educational and psychological testing, 1st ed. American Educational Research Association, American Psychological Association & NC on M in E, editor. Washington, DC: American Educational Research Association; 2014. 1–230 p.

3. Brock RL, Barry RA, Lawrence E, Dey J, Rolffs J. Internet administration of paper-and-pencil questionnaires used in couple research: assessing psychometric equivalence. Assessment. 2012;19(2):226–42.
4. Boateng GO, Neilands TB, Frongillo EA, Melgar-Quiñonez HR, Young SL. Best practices for developing and validating scales for health, social, and behavioral research: a primer. Front Public Heal. 2018;6:149.
5. Rosa AR, Sánchez-Moreno J, Martínez-Aran A, Salamero M, Torrent C, Reinares M, et al. Validity and reliability of the Functioning Assessment Short Test (FAST) in bipolar disorder. Clin Pract Epidemiol Ment Heal [Internet]. 2007;3(1):5. Available from: http://www.cpementalhealth.com/content/3/1/5.
6. Mustafa N, Safii NS, Jaffar A, Sani NS, Mohamad MI, Abd Rahman AH, et al. Malay version of the mHealth app usability questionnaire (M-MAUQ): translation, adaptation, and validation study. JMIR Mhealth Uhealth. 2021;9(2):e24457.
7. Alfonsson S, Maathz P, Hursti T. Interformat reliability of digital psychiatric self-report questionnaires: a systematic review. J Med Internet Res. 2014;16(12):e268.
8. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. 2006;119:66.e7–16.
9. González-Robles A, Mira A, Miguel C, Molinari G, Díaz-García A, García-Palacios A, et al. A brief online transdiagnostic measure: psychometric properties of the Overall Anxiety Severity and Impairment Scale (OASIS) among Spanish patients with emotional disorders. PLoS One. 2018;13(11):1–18.
10. Goltermann J, Emden D, Leehr EJ, Dohm K, Redlich R, Dannlowski U, et al. Smartphone-based self-reports of depressive symptoms using the remote monitoring application in psychiatry (ReMAP): interformat validation study. JMIR Ment Heal. 2021;8(1):e24333.
11. Verkuil B, Brosschot JF. The online version of the Dutch Penn State Worry Questionnaire: factor structure, predictive validity and reliability. J Anxiety Disord [Internet]. 2012;26(8):844–8. https://doi.org/10.1016/j.janxdis.2012.08.002.
12. Jabrayilov R, Emons WHM, Sijtsma K. Comparison of classical test theory and item response theory in individual change assessment. Appl Psychol Meas. 2016;40(8):559–72.
13. van Ballegooijen W, Riper H, Cuijpers P, van Oppen P, Smit JH. Validation of online psychometric instruments for common mental health disorders: a systematic review. BMC Psychiatry. 2016;16(1):45.
14. Chan EC, Sun Y, Aitchison KJ, Bch BM, Sivapalan S. Mobile app – based self-report questionnaires for the assessment and monitoring of bipolar disorder: systematic review. JMIR Form Res. 2021;5:1–14.
15. Batra S, Baker RA, Wang T, Forma F, DiBiasi F, Peters-Strickland T. Digital health technology for use in patients with serious mental illness: a systematic review of the literature. Medical Devices. 2017;10:237–51.
16. Van Ameringen M, Turna J, Khalesi Z, Pullia K, Patterson B. There is an app for that! The current state of mobile applications (apps) for DSM-5 obsessive-compulsive disorder, posttraumatic stress disorder, anxiety and mood disorders. Depress Anxiety. 2017;34:526–39.
17. Ruggeri K, Maguire Á, Andrews JL, Martin E, Menon S. Are we there yet? Exploring the impact of translating cognitive tests for dementia using mobile technology in an aging population. Front Aging Neurosci. 2016;8:1–7.
18. Maguire Á, Martin J, Jarke H, Ruggeri K. Getting closer? Differences remain in neuropsychological assessments converted to mobile devices. Psychol Serv. 2019;16(2):221–6.
19. Carlbring P, Brunt S, Bohman S, Austin D, Richards J, Öst LG, et al. Internet vs. paper and pencil administration of questionnaires commonly used in panic/agoraphobia research. Comput Hum Behav. 2007;23(3):1421–34.
20. Thorndike FP, Carlbring P, Smyth FL, Magee JC, Gonder-Frederick L, Ost LG, et al. Web-based measurement: effect of completing single or multiple items per webpage. Comput Hum Behav. 2009;25(2):393–401.

21. Buitenweg DC, Bongers IL, van de Mheen D, van Oers HA, van Nieuwenhuizen C. Cocreative development of the QoL-ME: a visual and personalized quality of life assessment app for people with severe mental health problems. JMIR Ment Heal. 2019;6(3):e12378.
22. Bush NE, Skopp N, Smolenski D, Crumpton R, Fairall J. Behavioral screening measures delivered with a smartphone app psychometric properties and user preference. J Nerv Ment Dis. 2013;201(11):991–5.
23. Fortson BL, Scotti JR, Del Ben KS, Chen Y-C. Reliability and validity of an internet traumatic stress survey with a college student sample. J Trauma Stress [Internet]. 2006;19(5):709–20. Available from: http://doi.wiley.com/10.1002/jts.20165.
24. Read JP, Farrow SM, Jaanimägi U, Ouimette P. Assessing trauma and traumatic stress via the internet: measurement. Traumatology (Tallahass Fla). 2011;15(1):94–102.
25. Austin DW, Carlbring P, Richards JC, Andersson G. Internet administration of three commonly used questionnaires in panic research: equivalence to paper administration in Australian and Swedish samples of people with panic disorder. Int J Test. 2006;6(1):25–39.
26. Vallejo MA, Mañanes G, Isabel Comeche M, Díaz MI. Comparison between administration via Internet and paper-and-pencil administration of two clinical instruments: SCL-90-R and GHQ-28. J Behav Ther Exp Psychiatry. 2008;39(3):201–8.
27. Whitehead L. Methodological issues in internet-mediated research: a randomized comparison of internet versus mailed questionnaires. J Med Internet Res. 2011;13(4):1–6.
28. Miller ET, Neal DJ, Roberts LJ, Baer JS, Cressler SO, Metrik J, et al. Test-retest reliability of alcohol measures: is there a difference between internet-based assessment and traditional methods? Psychol Addict Behav. 2002;16(1):56–63.
29. Holländare F, Andersson G, Engström I. A comparison of psychometric properties between internet and paper versions of two depression instruments (BDI-II and MADRS-S) administered to clinic patients. J Med Internet Res. 2010;12(5):1–9.
30. Herrero J, Meneses J. Short web-based versions of the perceived stress (PSS) and Center for Epidemiological Studies-Depression (CESD) scales: a comparison to pencil and paper responses among internet users. Comput Hum Behav. 2006;22(5):830–46.
31. Thorén ES, Andersson G, Lunner T. The use of research questionnaires with hearing impaired adults: online vs. paper-and-pencil administration. BMC Ear, Nose Throat Disord. 2012;12(1):12.
32. Yu SC, Yu MN. Comparison of internet-based and paper-based questionnaires in Taiwan using multisample invariance approach. Cyberpsychol Behav. 2007;10(4):501–7.
33. Zimmerman M, Martinez JH. Web-based assessment of depression in patients treated in clinical practice. J Clin Psychiatry. 2012;73(03):333–8.
34. Hirai M, Vernon LL, Clum GA, Skidmore ST. Psychometric properties and administration measurement invariance of social phobia symptom measures: paper-pencil vs. internet administrations. J Psychopathol Behav Assess. 2011;33(4):470–9.
35. Tests C. Equivalence of computerized and conventional versions of the beck depression inventory-II (BDI-II). Curr Psychol. 2001;20(3):216–30.
36. Sandoval LR, Buckey JC, Ainslie R, Tombari M. HHS public access. 2018;48(3):413–25.
37. George CE, Lankford JS, Wilson SE. The effects of computerized versus paper-and-pencil administration on measures of negative affect. Comput Hum Behav. 1992;8(2–3):203–9.
38. Lankford JS, Bell RW, Elias JW. Computerized versus standard personality measures: equivalency, computer anxiety, and gender differences. Comput Hum Behav. 1994;10(4):497–510.
39. Lukin ME, Dowd ET, Plake BS, Kraft RG. Comparing computerized versus traditional psychological assessment. Comput Hum Behav. 1985;1(1):49–58.
40. Schmitz N, Hartkamp N, Brinschwitz C, Michalek S, Tress W. Comparison of the standard and the computerized versions of the Symptom Check List (SCL-90-R): a ramdomized trial. Acta Psychiatr Scand. 2000;102(2):147–52.
41. Chan-Pensley E. Alcohol-use disorders identification test: a comparison between paper and pencil and computerized versions. Alcohol Alcohol [Internet]. 1999;34(6):882–5. Available from: https://academic.oup.com/alcalc/article-lookup/doi/10.1093/alcalc/34.6.882.

42. Murrelle L, Ainsworth BE, Bulger JD, Holliman SC, Bulger DW. Computerized mental health risk appraisal for college students: user acceptability and correlation with standard pencil-and-paper questionnaires. Am J Heal Promot [Internet]. 1992;7(2):90–2. Available from: http://journals.sagepub.com/doi/10.4278/0890-1171-7.2.90.
43. Glaze R, Cox JL. Validation of a computerized version of the 10-item (self-rating) Edinburgh postnatal depression scale. J Affect Disord. 1991;22(1–2):73–7.
44. Kurt R, Bogner HR, Straton JB, Tien AY, Gallo JJ. Computer-assisted assessment of depression and function in older primary care patients. Comput Methods Prog Biomed. 2004;73(2):165–71.
45. Ogles B, France C, Lunnen K, Bell M, Goldfarb M. Computerized depression screening and awareness. Community Ment Health J. 1998;34(1):27–38.
46. Schulenberg SE, Yutrzenka BA. Equivalence of computerized and conventional versions of the Beck Depression Inventory-II (BDI-II). Curr Psychol. 2001;20(3):216–30.
47. Research2Guidance. mHealth economics 2017 – current status and future trends in mobile health [Internet]. 2017 [cited 2021 June 15]. Available from: https://research2guidance.com/325000-mobile-health-apps-available-in-2017/.
48. Torous J, Roberts LW. Needed innovation in digital health and smartphone applications for mental health transparency and trust. JAMA Psychiat. 2017;74(5):437–8.
49. Moore RC, Fazeli PL, Patterson TL, Depp CA, Moore DJ, Granholm E, et al. UPSA-M: feasibility and initial validity of a mobile application of the UCSD Performance-Based Skills Assessment. Schizophr Res [Internet]. 2015 [cited 2021 Apr 9];164(1–3):187–92. https://doi.org/10.1016/j.schres.2015.02.014.
50. Palmier-Claus JE, Ainsworth J, Machin M, Barrowclough C, Dunn G, Barkus E, et al. The feasibility and validity of ambulatory self-report of psychotic symptoms using a smartphone software application. BMC Psychiatry [Internet]. 2012 [cited 2021 Apr 9];12(1):172. Available from: http://bmcpsychiatry.biomedcentral.com/articles/10.1186/1471-244X-12-172.
51. Faurholt-Jepsen M, Frost M, Vinberg M, Christensen EM, Bardram JE, Kessing LV. Smartphone data as objective measures of bipolar disorder symptoms. Psychiatry Res [Internet]. 2014;217(1–2):124–7. https://doi.org/10.1016/j.psychres.2014.03.009.
52. Torous J, Staples P, Shanahan M, Lin C, Peck P, Keshavan M, et al. Utilizing a personal smartphone custom app to assess the patient health questionnaire-9 (PHQ-9) depressive symptoms in patients with major depressive disorder. JMIR Ment Heal. 2015;2(1):e8.