OULU 2015
D 1313

ACTA UNIVERSITATIS OULUENSIS

Eka Roivainen

VALIDITY IN PSYCHOLOGICAL MEASUREMENT: AN INVESTIGATION OF TEST NORMS

UNIVERSITY OF OULU GRADUATE SCHOOL;
UNIVERSITY OF OULU,
FACULTY OF MEDICINE;
MEDICAL RESEARCH CENTER OULU;
OULU UNIVERSITY HOSPITAL

ACTA UNIVERSITATIS OULUENSIS
D Medica 1313

EKA ROIVAINEN

VALIDITY IN PSYCHOLOGICAL MEASUREMENT

An investigation of test norms

Academic dissertation to be presented with the assent of the Doctoral Training Committee of Health and Biosciences of the University of Oulu for public defence in Auditorium F101 of the Faculty of Biochemistry and Molecular Medicine (Aapistie 7), on 30 October 2015, at 12 noon.

UNIVERSITY OF OULU, OULU 2015

Acta Univ. Oul. D 1313, 2015

Supervised by
Professor Jouko Miettunen
Professor Juha Veijola

Reviewed by
Professor Laura Hokkanen
Professor Jan-Erik Lönnqvist

ISBN 978-952-62-0942-5 (Paperback)
ISBN 978-952-62-0943-2 (PDF)

ISSN 0355-3221 (Printed)
ISSN 1796-2234 (Online)

Cover Design
Raimo Ahonen

JUVENES PRINT
TAMPERE 2015

Roivainen, Eka, Validity in psychological measurement. An investigation of test norms

University of Oulu Graduate School; University of Oulu, Faculty of Medicine; Medical Research Center Oulu; Oulu University Hospital

Acta Univ. Oul. D 1313, 2015

University of Oulu, P.O. Box 8000, FI-90014 University of Oulu, Finland

Abstract

A psychological test may be defined as an objective and standardized measure of a sample of behaviour. The interpretation of test results is usually based on comparing an individual's performance to norms based on a representative sample of the population.

The present study examined the norms of popular adult tests. The validity of the Wartegg drawing test (WZT) was studied using two rating scales, the Toronto Alexithymia Scale and the Beck Depression Inventory as criterion tests. Weak to moderate correlations were found. It is concluded that the WZT has some validity in the assessment of Alexithymia. Efforts to develop a psychometrically valid and reliable method of interpreting the WZT should be continued. Cross-national and historical analyses of the norms of Wechsler's adult intelligence scale (WAIS) were performed. The results show that the Finnish WAIS III test norms are distorted in the younger age groups. Significant cross-national and cross-generational differences in relative subtest scores, test profiles were also observed. Differences in general intelligence cannot explain such variations, and educational and cultural factors probably underlie the observed differences. It is suggested that the concept of a national IQ profile is useful for cross-national test validation studies. The validity of a validity scale, the Chapman Infrequency Scale, was studied in the context of a survey study. Results showed that careless responding is significantly more frequent among psychiatric patients relative to healthy respondents. The common procedure of excluding careless responders from final samples may affect the results of survey studies targeting individuals with psychiatric symptoms. Cut-off scores for exclusion should be flexible and chosen according to the demographic and health characteristics of the sample.

In conclusion, the results of this study underscore the need for up-to-date and representative test norms for valid test interpretation.

Keywords: careless responding, cohort study, IQ test, national IQ, projective test, test norms, vocabulary test, Wartegg test

Roivainen, Eka, Näkökulmia psykologisten arviointien luotettavuuteen. Tutkimus testinormeista
Oulun yliopiston tutkijakoulu; Oulun yliopisto, Lääketieteellinen tiedekunta; Medical Research Center Oulu; Oulun yliopistollinen sairaala
Acta Univ. Oul. D 1313, 2015
Oulun yliopisto, PL 8000, 90014 Oulun yliopisto

Tiivistelmä

Psykologiset testit voidaan ymmärtää otoksiksi tutkittavan käyttäytymisestä. Mittauksen tulosta tulkitaan yleensä vertaamalla sitä tavalliseen tai keskimääräiseen tulokseen eli testinormeihin.

Väitöskirjatutkimus tarkastelee suosittujen aikuistestien normien pätevyyttä. Warteggin piirrostestin validiteettia aleksitymian ja depression mittarina tutkittiin käyttämällä vertailukriteerinä kahta lomaketestiä, Toronton aleksitymia-asteikkoa ja Beckin depressioasteikkoa. Mitatut korrelaatiot olivat melko matalia. Tutkimuksen johtopäätöksenä oli, että Wartegg-testi saattaa olla hyödyllinen menetelmä aleksitymian toteamisessa ja että empiiriseen tutkimukseen perustuvaa tulkintamenetelmien kehittämistä pitäisi jatkaa. Tutkimuksessa selvitettiin myös Wechslerin aikuisten älykkyystestien (WAIS) eri versioiden osatestien kansallisten normien välisiä eroja ja eroja ikäkohorttien välillä. Tulokset osoittivat, että suomalaiset WAIS III testinormit ovat vinoutuneet nuorempien ikäryhmien osalta. Tutkimuksessa havaittiin merkitseviä eroja osatestien keskiarvojen suhteissa eli testiprofiileissa eri maiden ja ikäkohorttien välillä. Kyseisiä eroja ei voida selittää älykkyyden yleisellä faktorilla, vaan niiden taustalla on luultavasti koulutukseen ja kulttuuriin liittyviä tekijöitä. Osa eroista kansallisissa testiprofiileissa näyttää olevan luonteeltaan pysyviä, ja tätä tietoa voidaan käyttää hyväksi testinormien pätevyyttä arvioitaessa. Chapmanin vastaustapa-asteikon (CIS) validiteettia tutkittiin Pohjois-Suomen vuoden 1966 syntymäkohortin kyselytutkimusaineistolla. Psykiatrisista oireista kärsivät henkilöt saivat korkeampia pistemääriä kuin terveet vastaajat. Johtopäätöksenä oli, että vastaustapamittarit voivat karsia psykiatrisia potilaita liian herkästi ulos tutkimusjoukosta, mikä voi vääristää tutkimusten tuloksia. Kriteeripistemäärän pitäisi olla joustava ja sen määrityksessä pitäisi ottaa huomioon tutkimusjoukon ominaisuudet.

Tutkimukset osoittavat, että testituloksen luotettava tulkinta vaatii ajanmukaiset ja edustavaan otokseen perustuvat testinormit.

Asiasanat: kansallinen ÄO, kohorttitutkimus, projektiivinen testi, sanavarastokoe, testinormit, vastaustapa-asteikko, Wartegg-testi, älykkyystesti

Acknowledgements

I owe my deepest and warmest gratitude to my supervisor, Professor Jouko Miettunen, for guiding me with a positive attitude throughout the writing process. I feel privileged to have been able to be influenced by his excellent knowledge in psychiatric research. I am sincerely grateful to my other supervisor, Professor Juha Veijola, for his advice and support.

I wish to express my sincere thanks to the pre-examiners of this thesis, Professor Laura Hokkanen and Professor Jan-Erik Lönnqvist, whose advice substantially improved the quality of this thesis.

I thank Semantix for proofreading the summary part of the thesis.

I would like to extend my thanks to my former colleagues at Oulu Deaconess Institute, particularly Ms. Piritta Ruuska, who co-authored one of the original publications. I thank my co-workers at Verve Rehabilitation for their support and a pleasant working atmosphere.

My warmest appreciation goes to my parents and my sister for love and support in my life. My dearest gratitude and appreciation go to my family: Marina, Fred, and Isabella, for their love and understanding.

September 2015 Eka Roivainen

Abbreviations

16PF	Sixteen Personality Form test
BDI	Beck Depression Inventory
CIS	Chapman Infrequency Scale
FSIQ	Full scale intelligence quotient (WAIS I-IV)
GDP	Gross domestic product
GSS	General Social Survey
IQ	Intelligent Quotient
MMPI	Minnesota Multiphasic Personality Inventory
PAS	Physical anhedonia scale
PER	Perceptual aberration scale
PIQ	Performance intelligence quotient (WAIS I-III)
PISA	Programme for International Student Assessment
POI	Perceptual organization index (WAIS III)
PRF	Personality Research Form
PRI	Perceptual reasoning index (WAIS IV)
PSI	Processing speed index (WAIS III-IV)
RBDI	Raitasalo Beck Depression Inventory
SAS	Social anhedonia scale
SB	Stanford–Binet test
SDS	Self Directed Search (Holland's test)
SPM	Standard Progressive Matrices (Raven's matrices)
TAS	Toronto Alexithymia Scale
TAT	Thematic Apperception Test
TCI	Temperament and Character Inventory
VCI	Verbal comprehension index (WAIS III-IV)
VIQ	Verbal intelligence quotient (WAIS I-III)
WAIS	Wechsler Adult Intelligence Scale
WISC	Wechsler Intelligence Scale for Children
WJ	Woodcock–Johnson test
WMI	Wechsler memory index (WAIS III-IV)
WZT	Wartegg Zeichen Test, Wartegg Drawing test

Key concepts

Alexithymia	personality trait involving a lack of fantasy and difficulty in expressing feelings
Criterion validity	correlation between a test and an outside measure
Intelligence	capacity of the individual to act purposefully, to think rationally and to deal effectively with his environment
Performance subtest	nonverbal subtest of an intelligence test
Personality	a person's consistent patterns of feeling, thinking, and behaving
Projective test	a performance-based test of personality
Psychological test	a standardized measure of a sample of behaviour
Reliability	consistency, accuracy and stability of the result of the measurement
Standardized test	a test with norms and consistent scoring and administration procedures
Test norms	distribution of test scores in a representative sample of the population
Validity	the degree to which an instrument measures what it should measure
Validity scale	measure of a respondent's motivation and capacity to respond meaningfully to test items
Verbal subtest	a subtest of an intelligence test that measures verbal skills

List of original publications

This thesis is based on the following publications, which are referred to throughout the text by their Roman numerals:

I Roivainen E & Ruuska P (2005) The use of projective drawings to assess alexithymia: the validity of the Wartegg test. European Journal of Psychological Assessment 21: 199–201.

II Roivainen E (2010) European and American WAIS III norms: Cross-national differences in performance subtest scores. Intelligence 38: 187–191.

III Roivainen E (2013) Are cross-national differences in IQ profiles stable? A comparison of Finnish and US WAIS norms. International Journal of Testing 13: 140–151.

IV Roivainen E (2014) Changes in word usage frequency may hamper intergenerational comparisons of vocabulary skills: An Ngram analysis of WAIS, WISC and Wordsum test items. Journal of Psychoeducational Assessment 32: 83–87.

V Roivainen E, Veijola J & Miettunen J (2015) Careless responses in survey data and the validity of a screening instrument. Nordic psychology. DOI: 10.1080/19012276.2015.1071202.

Table of contents

1 A brief history of psychological testing

A psychological test may be defined as an objective and standardized measure of a sample of behaviour. The samples of behaviour measured by tests represent larger psychological constructs such as intelligence or personality (Anastasi & Urbina 1997). Table 1 shows the popularity ranking of adult tests in Finland.

Table 1. Proportion of Finnish psychologists using different psychological tests for adults (Kuuskorpi 2012).

Test	Popularity
1. WAIS-Wechsler adult intelligence scale	53%
2. WZT - Wartegg Zeichen Test	47%
3. RO – Rorschach Inkblot Test	46%
4. WMS - Wechsler Memory Scale	37%
5. BDI - Beck Depression Inventory (incl. R-BDI)	26%
6. PRF - Personality Research Form	21%
7. H-T-P - Draw-A-House-Tree-Person	15%
8. MMPI - Minnesota Multiphasic Personality Inventory	14%
9. TMT - Trail Making Test	13%
10. TAT - Thematic Apperception Test	12%

1.1 Intelligence tests

The birth of scientific psychology is often dated to 1879, when Wundt established the first psychological laboratory. The scientific, experimental method of study was applied to the human psyche. Previously, conceptions of psychology and mental phenomena had been based on introspection, folk psychology, and on philosophical and theological theories. Wundt used simple instruments such as a reaction time apparatus to measure mental processes (Gregory 2013).

Galton published *Inquiries into Human Faculty and Its Development* (Galton 1883), a series of articles on individual differences in perception, reasoning and other psychological faculties, regarded as the beginning of the "mental test movement" (Boring 1950). In addition to physical characteristics such as height and head length, Galton measured visual acuity, highest audible tone, and reaction time in his subjects. One of the first scientists to use the concept of "mental test" was McKeen-Cattell (1890), who had studied with Wundt and Galton. In his paper *Mental tests and measurements*, McKeen-Cattell described ten tests or instruments, including reaction time for sound, judgment of 10 seconds of time, time for naming

colours and weight differentiation–judging the relative weights of two boxes varying by one gram from 100 to 110 grams. The early psychometricians assumed that people who performed well on this type of simple physiological and sensory tests were more intelligent than others, but the results proved inconclusive. More useful measures of higher psychological processes were then introduced by Binet (Gregory 2013:2–28).

The progress of psychiatry in the 19th century was dominated by French psychiatrists. Esquirol (1838) proposed that mental illness (dementia) and mental retardation (idiocy) are two distinct psychiatric disorders, and he recognized three levels of mental retardation based on linguistic skills: 1) patients who are able to form short phrases, 2) patients who use only monosyllables, and 3) those with no speech, but an ability to scream or cry. According to Nicolas and others (2013), one of the factors underlying the emergence of psychometrics as a discipline in France at the turn of the century was rivalry between psychiatrists and psychologists on the matter of special education. Psychiatrists argued that abnormal children should be referred to special education classes in asylums, where psychiatrists were in charge, while Binet and others at the Société libre de l'étude psychologique de l'enfant founded in 1899 sought to keep children in schools and psychologists in charge in a previously psychiatric domain: identifying and treating "the abnormal".

In 1904 the Commission for Public Instruction in Paris concluded that medical and educational examinations should be used to identify those children who could not profit from regular instruction. Binet was appointed to develop a practical tool for this purpose. He had noticed that elementary processes such as reaction time and sensory acuity were not good measures of higher mental faculties. He also concluded on the basis of his experiments that attention was an important component of intelligence (Gregory 2013).

The Binet-Simon (1905) scale from was composed of 30 tests including very simple "follows a moving object with eyes" to fairly difficult "define the difference between boredom and weariness". The more difficult tasks were verbal in nature, in opposition to the Galtonian sensory tasks. In the revised version of the test, the concept of mental level was introduced, based on a standardization study with 300 children between the ages of 3 and 13. The test items were ranked according to the age level at which they were passed by 80–90% of the children. In the 1911 version of the test, each age level had five tests, and the age range was extended to adults. Stern (1912) suggested that intelligence quotient may be calculated by dividing mental level by chronological age. So as to remove fractions, Terman suggested multiplying IQ by 100. Terman's revision of Binet's test, the Stanford–Binet

(Terman 1916) included 90 test items and was suitable for testing children, adults and individuals with mental retardation. The Stanford–Binet remained the prevailing IQ test up to the 1960s, when the Wechsler tests started to gain popularity (Gregory 2013).

1.1.1 The definition of intelligence

There is no scientific consensus definition of the concept of intelligence. Wechsler (1955) defines intelligence as "the global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment". Legg and Hutter (2007) list 70 different definitions for this concept. Boring (1950) defined intelligence as "what the intelligence tests test". While Boring's definition may seem jocular, it describes fairly well the prevailing theories of intelligence, such as the *Cattell–Horn–Carroll* (Carroll 2013) model that are based on factor analysis. The correlational pattern of tasks that are presumed to measure intelligence are modelled statistically with factor models. A general factor of intelligence is thought to underlie test scores and correlations between tests, and individual differences between subjects or groups of subjects.

Gardner (1983) has proposed in his *multiple intelligences theory* that there are several distinct cognitive abilities with only weak correlations between them. The evidence for Gardner's theory is weak (Sternberg & Grigorenko 2000), and there are no major intelligence tests based on Gardner's theory. Borsboom, Hellenberg and Van Heerden (2003) and Van der Maas and others (2014) have criticized the prevailing g-factor model of intelligence for its lack of theoretical coherence. The ontological nature of the latent variable, g-factor, remains unclear; is the relationship between the g-factor and IQ test scores really causal? As an alternative of the g-factor model, Van der Maas and others (2006) have proposed a "mutualism model". In the mutualism model, the correlations between test scores are not explained through the dependence on a common latent variable, but "*as a result of reciprocal positive interactions between abilities and processes that play key roles in cognitive development, such as memory, spatial ability, and language skills*". The impact of this theory on the development of future intelligence tests remains to be seen.

1.1.2 The Wechsler Adult Intelligence Scale

The Wechsler intelligence scale for children, WISC (Wechsler 1949) was released in 1949, and the adult intelligence scale, WAIS (Wechsler 1955) in 1955. The WAIS standardization sample consisted of 850 males and 850 females divided into seven age groups (16–17, 18–19, 20–24, 25–34, 35–44, 45–54, and 55–64 years) of 200 or 300 individuals each (half men, half women in each group). The sample was representative of the US adult population (based on the 1950 census) in terms of ethnic background, occupation, education, geographical region, and urban/rural residence. The scaled scores used to calculate IQs were based on a reference group of 500 cases, which included the subjects between the ages of 20 and 34. For each subtest, the distribution of raw scores of this reference group was converted to a scale with a mean score of 10 and a standard deviation of 3. However, the IQs were derived separately for each of the seven age groups by setting the mean sum of the scores for each age group equal to 100 and the standard deviation to 15. The sum of scaled scores on the six verbal subtests is used to calculate the verbal IQ, while the non-verbal IQ is based on the five performance tests, and the full scale IQ is based on all the 11 subtests.

WAIS subtests

The first version of WAIS consisted of 11 subtests that were also used in the later WAIS-R and WAIS III versions of the test. The WAIS III includes the following subtests (Wechsler 1997, Groth-Marnat 2003):

1. The *information* subtest consists of questions that average individuals should be able to answer, such as "What are the colours of Finland's flag?" This subtest is intended to measure intellectual curiosity, and alertness to the surrounding society and environment.
2. The *comprehension* subtest is based on an understanding of how to apply general knowledge in practical situations. For example: "Why do we need traffic safety regulations?". This subtest measures a knowledge of normal schemes and norms of behaviour, social maturity, reasoning and generalization.
3. In the *vocabulary* subtest, the respondent is requested to explain the meaning of words such as courageous or ossobuco. The vocabulary subtest measures verbal intelligence and fluency.

4. The *arithmetic* subtest consists of arithmetical problems such as "2+3" or "What is 5% out of 90?", and measures computational skill, auditory memory and the ability to concentrate.
5. In the *similarities* subtest, the respondent is presented with word pairs, such as "chair–table" and is requested to explain the common underlying concept (furniture).
6. The *digit span* subtest measures short-term verbal memory. The respondent is requested to repeat strings of digits read by the experimenter.
7. In the *block design* subtest, the respondent has to assemble different types of patterns shown in the test manual using nine red and white cubes. This task measures non-verbal reasoning, spatial visualization and visuomotor co-ordination.
8. In the *picture completion* subtest, the respondent is requested to point out a missing detail in a picture, for example, a door without a doorknob. This test measures visual alertness, and the ability to differentiate essential details from non-essential ones.
9. The *picture arrangement* subtest consists of short stories narrated in three to six pictures. The pictures are presented to the respondent in a jumbled order, and the respondent is asked to rearrange the pictures in the correct order that restores the story. This subtest measures visual alertness, the ability to anticipate consequences, and an understanding of cause-effect sequencing.
10. The *object assembly* subtest consists of four puzzles. The respondent receives extra points for fast performance. This is a test of visuomotor co-ordination and control, visual organization and non-verbal reasoning.
11. The *digit symbol coding* is a test of psychomotor processing speed and requires the respondent to quickly draw symbols based on a code where the numbers 1–9 are each substituted by a specific symbol. The respondent has 90 seconds in which to draw as many symbols as they can to form correct digit-symbol pairs.
12. The *letter-number sequencing* subtest measures auditory short-term memory, sequencing ability and concentration and attention.
13. The *matrix reasoning subtest* is a test of abstract reasoning and visual organization. It involves the analysis of wholes into smaller parts.
14. The *symbol search subtest* is a measure of visual-motor coordination, speed of information processing and learning ability.

The WAIS-R (Wechsler 1981) was published in the United States in 1981, and a sample of 1 880 subjects was used to standardize the test. In addition to the new norms, some items that appeared dated were changed in each subtest, and small changes were made in test instructions and scoring. The publication of the WAIS III (Wechsler 1997) represented a major revision of the original test. The main reason for the revision was to update the norms. Unlike its predecessors, WAIS III standard scores were calculated separately for each age group, and no separate reference group was created. Outdated items were replaced in the subtests. In addition to the Full scale (FSIQ), Performance (POI) and Verbal (VI) IQ scores, there are four index scores in the WAIS III that represent the four major factors of intelligence: 1) The Verbal comprehension index (VCI) composed of the Vocabulary, Similarities and Information subtest scores, 2) The Perceptual organization index (POI), composed of Block design, Matrix reasoning and Picture completion subtests, 3) The Working memory index (WMI) composed of Arithmetic and Digit span subtests, and 4) The Processing speed index (PSI) composed of Digit symbol coding and Symbol search.

WAIS IV (Wechsler 2008) was published in 2008. This test consists of 15 subtests, including nine of the original WAIS subtests with extensively changed and revised items, and six new subtests: Matrix reasoning, Symbol search, Letter–number sequencing, Visual puzzles, Figure weights, and Cancellation. The Picture arrangement and Object assembly subtests from the first edition of WAIS have been excluded.

1.1.3 The Finnish versions of the WAIS

The Finnish version of the original WAIS was published in 1971 (Wechsler 1971). The standardization was conducted between 1956 and 1968, with 68% of the sample comprising 1 000 subjects tested in 1967–68. The sample was representative of the Finnish population (1964 census) in terms of education, occupational status, and gender. The Finnish versions of the Digit span and Coding subtests were identical with the US versions, but there were some small differences in the order of presentation of the items in the other non-verbal subtests, and the scoring of the Object assembly was substantially modified. The verbal subtests were partly based on American WAIS items and partly developed by the Finnish test publisher. The Finnish WAIS-R was published in 1992 (Wechsler 1992). The standardization study was based on a sample of 1 052 subjects. The Block design, Coding, Digit span, Object assembly, and Picture arrangement subtests were

identical to the US WAIS-R subtests. The Picture completion subtest contained six items from the US WAIS-R, seven items from WISC-R, and nine items from the Finnish WAIS.

The Finnish WAIS III was published in 2005 (Wechsler 2005). The standardization sample was composed of 446 participants, and was divided into nine age groups. The testing equipment (cubes, pictures, puzzles), procedure, and test instructions of the non-verbal subtests were identical in the Finnish and US versions of the test. The verbal subtests were partly based on American WAIS items and partly developed by the Finnish test publisher.

The Finnish WAIS IV was published in 2012 (Wechsler 2012), and the standardization sample was composed of 657 participants divided into 11 age groups. The verbal subtests were based on Finnish WAIS III items, items from the American WAIS IV and some new national items. The non-verbal subtests are the same as in the American WAIS IV.

1.1.4 Other adult intelligence tests

Other adult intelligence tests composed of subtests that measure different factors of intelligence include the latest version of the Stanford–Binet (Roid 2003) and the Woodcock–Johnson (Woodcock, McGrew, Mather 2001) WJ tests. These tests have not been translated into Finnish, and a national test, the AVO intelligence test (Työministeriö 1995) developed by the Ministry of Labour career counselling services has been used in group testing.

1.2 Personality tests

1.2.1 Questionnaire measures

Personality can be defined as "*those characteristics that account for an individual's consistent patterns of feeling, thinking, and behaving*" (Pervin, Cervone, & John 2005: 6). One of the first personality tests of the questionnaire type was Woodworth's (1919) Personal Data Sheet that was developed for military purposes in detecting recruits who suffered from psychiatric disorders. This questionnaire consisted of 116 yes/no questions such as "Are you bothered by a feeling that things are not real?". The major shortcoming of this instrument was that it was based on the assumption that subjects would be honest when responding. For clever

individuals who wished to avoid service despite being psychologically healthy or those who wished to serve despite psychiatric disorders, faking was a possibility due to the fairly transparent type of questions (Gregory 2013).

Faking and socially desirable answering remain potential shortcomings of questionnaire measures (Bäckström & Björklund 2013). A more complex instrument still in use today, the Minnesota Multiphasic Personality Inventory (MMPI) was developed by Hathaway and McKinley (1940). The original MMPI had 13 standard scales; 3 were validity scales and 10 scales were related to psychiatric disorders. The test consisted of 506 statements that could be answered with "true" or "false". These items were presented to a control group comprised of 724 healthy individuals and a clinical group that was composed of patients being treated at hospitals. Items that differentiated, for example, between the control group and individuals suffering from depression, were included in the depression scale. Validity scales were added to the MMPI when test developers noticed that respondents could manipulate the impression they made on the test for various reasons (Hathaway & McKinley 1940). The MMPI Cannot say scale is the total number of unanswered questions. The Lie scale consisted of items that indicated a reluctance to admit even minor problems. The Infrequency scale (F) had items that were endorsed by fewer than 10% of respondents in the control group. Therefore, a high score on the F scale indicates that the respondent is endorsing deviant responses. The Defensiveness scale (K) consisted of items that were endorsed by psychiatric patients with a diagnosis who had normal scores on the clinical scales. In other words, these respondents consciously or unconsciously denied their symptoms. The K score was used as a correction factor for the other scales. It was hypothesized that the distorted scores of defensive patients on the clinical scales could be corrected by adding the K score, or a fraction of it to the clinical scale scores (Groth-Marnat 2003).

In Finland, the most popular questionnaire-type tests are the MMPI and the Personality research form PRF (Jackson 1974) based on Murray's theory of personality. Other general personality tests in Finnish include the factor test PK-5 (PKOY 2007), which measures the five major factors of personality and fifteen facet traits, and Cloninger's temperament and character inventory, TCI (Cloninger 1994, Miettunen *et al.* 2004). In addition to wide-scope personality tests, a large number of more specific clinical scales, such as the Beck depression inventory (Beck *et al.* 1961), the Beck anxiety inventory (Beck & Steer 1993), the Toronto Alexithymia Scale (TAS) (Bagby *et al.* 1986, Joukamaa *et al.* 2001) and Chapman's schizothypy scales (Chapman & Chapman & Kwapil 1995) have been developed

in order to screen and assess psychiatric symptoms (Jääskeläinen & Miettunen 2011, Miettunen *et al.* 2011).

1.2.2 Projective tests

Projective tests are performance-based tests of personality. The development of projective tests is related to the rise of psychoanalysis in the early 20th century (Wood, *et al.* 2003). Methods for studying the unconscious were needed. At the beginning of his career, Freud used hypnosis to access the unconscious of his patients, but soon moved to the methods of free association, dream analysis and word association. The method of word association is a simple projective test. The patient is instructed to say the first word that comes to their mind when the therapist gives a stimulus word. Other projective tests involving verbal stimuli include sentence completion tests such as the Rotter test (Rotter & Rafferty 1950) which continues to be used frequently in psychological assessment today.

Projective tests based on visual stimuli were developed by Rorschach (1921) and Wartegg (1939). Rorschach, who was working at a psychiatric hospital, noticed that his schizophrenic patients responded differently to inkblots in a children's creativity game called Blotto. He then tested 40 inkblots on a sample of 405 patients, and chose 15 inkblot cards for his test. For practical and commercial reasons, the test published in 1921 consisted of 10 inkblot cards (Wood *et al.* 2003). Since the 1940s the Rorschach has been one of the most frequently used tests in clinical psychology. The Thematic Apperception Test (Murray 1943) is another popular projective test that uses ambiguous pictures as stimuli.

1.2.3 The Wartegg test

The Wartegg drawing completion test (Wartegg Zeichen Test, WZT) is a projective drawing test developed in the 1920s and 1930s by Ehrig Wartegg, an Austro-German psychologist (Wartegg 1939). The A4-sized test form is shown in Figure 1. Each of the eight small (4x4 cm) squares on the test form is blank except for a small sign that is given as the starting point of a drawing. For example, there are two perpendicular lines in square 6. Subjects are instructed to complete the drawings, incorporating the given sign into the drawing. Like other projective drawing tests, Wartegg's test is based on the projective hypothesis, i.e. the assumption that the content and the qualitative aspects of the drawings reflect the

personality of the individual who is drawing. For example, drawings of clouds in the WZT protocol may be interpreted as a sign of depression (Gardziella 1985).

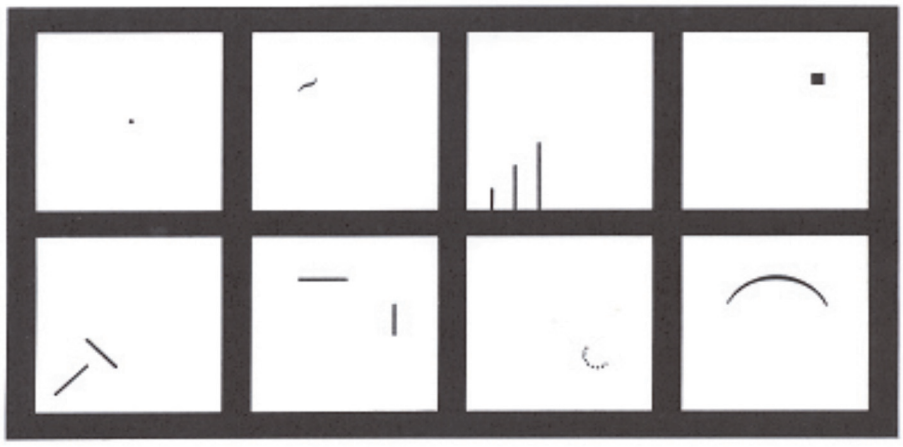

Fig. 1. The WZT form. (Roivainen 2009. Reprinted with permission by Hogrefe).

The WZT is practically unknown in English-speaking countries, but is widely used in continental Europe and Latin America. Mystical philosophies, modern art, psychoanalysis and Gestalt psychology may be considered to be the roots of the WZT (Roivainen 2006). Wartegg was introduced to these doctrines in the 1920s, while his academic work on the WZT at the University of Leipzig during the 1930s was mainly based on the theory of Ganzheit psychology. The rise of Nazism, the Second World War and the post-war division of Germany significantly affected Wartegg's work. Wartegg had limited opportunities to develop his test and communicate with western researchers (Petzold 2000).

The WZT became popular in Finland in the 1960s in the clinical setting, vocational guidance, and in personnel selection. Gardziella (1985) developed a new interpretation method at the Institute of Occupational Health during the 1970s which was taught on wide scale in courses organized by the Ministry of Labour and Psykologien Kustannusosakeyhtiö, the Finnish test publisher. Since Gardziella's manual was published, his method of Wartegg interpretation has been dominant in the different sectors of applied psychology in Finland.

2 Reliability, and validity of psychological tests

2.1 Reliability

The reliability of a psychological test refers to its consistency and accuracy and the stability of the result of the measurement (Groth-Marnat 2003). In a reliable test, the error of measurement, the range of random fluctuation in an individual's test score, is small. In an unreliable test, the test result of an individual re-examined on the same test is different on different occasions. Reliability of a test may be determined in three different ways (Groth-Marnat 2003: 12–13):

1. Test-retest reliability refers to the correlation between the scores obtained by the same respondent on two different administrations of the same test.
2. Split-half reliability refers to the internal consistency of the test. The test items are divided into two parts, and the two halves are correlated.
3. Inter-scorer reliability refers to the degree of agreement between two judges in tests where scoring is based on the judgment of the examiner. For example, drawing tests involve qualitative aspects that cannot objectively be scored.

In general, intelligence tests are considered to be somewhat more reliable than personality tests, because personality traits are more dependent on quickly varying factors such as the mood of the respondent. A reliability of 0.80 may be considered fairly high for psychological tests, while a reliability of 0.50 or lower means that the error of measurement is large and the test result is not very reliable (Hershen 2004: 7).

2.2 Validity

Test validity may be defined as *the degree to which a test measures what it is intended to measure*. A test can be reliable without being valid, but not the other way round (Groth-Marnat 2003). For example, tests of processing speed that involve naming of alphabets or digits are usually highly reliable, but they may actually measure reading skills and not processing speed (Roivainen 2011). Most psychological constructs, such as processing speed, intelligence or personality are abstract by nature and cannot be directly observed. They must be studied by indirect means. However, psychological theories and the definition of psychological constructs change over time. Therefore, the first concern of a test constructor is to

consider whether the test items they have selected are representative of the construct being measured. This aspect of validity is usually referred to as content validity. Content validity is subjective by nature, as it is based on the judgment of the test developers. Criterion validity refers to the correlation between a test and an outside measure. For example, IQ scores are highly correlated with school grades. If performance on a test item has zero correlation with education, then that item is probably a poor measure of intelligence. Predictive validity involves outside measurements that are performed some time after the psychological test, for example, the correlation of school grades with IQ scores from a test taken a year earlier (Blacker & Endicott 2000, Morgan *et al.* 2001).

2.2.1 Test norms

Valid norms are an essential aspect of test validity. Test results are usually interpreted by comparing the performance of an individual to norms based on a representative sample of the population. The more closely the standardization sample resembles the respondent in terms of age, education, sex and health, the more reliable the comparison is between test norms and the respondent's test score. Typical problems in test standardization involving norms include (Groth-Marnat 2003: 11):

1. Sampling bias. Random sampling based on population registers is the ideal. However, convenience sampling is less time-consuming and costly. Norms based on patients at a hospital or students at a university are easier to establish. Unfortunately, convenience sampling often leads to distorted test norms because the sample is not representative of the general population.
2. Sample size is too small. Age, education and gender are three basic factors that affect most psychological test scores. In a random adult sample of 100 individuals, theoretically 50 are men, and seven men are between the ages of 40 and 50, and one of these men has an academic degree. The interpretation of the test results of a 40–50 year-old man with an academic degree is not very reliable with norms from this sample, if age, sex, or education affects the test score of the construct measured.
3. Too rigid or too loose exclusion criteria. For example, individual's with *"a current treatment for alcohol or drug dependency; those who consume more than three alcoholic beverages more than two nights per week; and those currently taking antidepressants, anti-anxiety medication, or anti-psychotic*

drugs" (Wechsler 1997: 42) were excluded from the American WAIS III standardization sample. Of course, these disorders affect cognitive performance, but on the other hand, they are fairly common.

Test standardization also refers to establishing consistent administration and scoring procedures (Groth-Marnat 2003). A well-constructed test can be scored and administered in a consistent manner by different examiners at different test sessions.

2.3 Studies of test norms of intelligence tests

The validity of test norms of an intelligence test may be analysed by comparing them to a) norms of the previous editions of the test, b) norms of the foreign editions of the test, and c) norms of other valid cognitive tests including scholastic aptitude tests.

The standardization study of the revised edition of Wechsler's test, WAIS-R included a validity study where a subset (n = 72) of the standardization sample also took the earlier version of the test. In a similar fashion, 192 subjects from the WAIS III standardization sample were tested on the WAIS-R and WAIS, and 200 individuals from the WAIS IV standardization sample took the WAIS III and WAIS IV. These studies, as well as similar validity studies based on comparing different editions of other intelligence tests such as Wechsler's children's intelligence scales, WISC (Wechsler 1949), WISC-R (Wechsler 1974), and WISC III (Wechsler 1991) show high correlation between the different test editions in general, but also a general rise in scores over time (Flynn 2012). The new standardization samples tend to have higher mean scores on the tests than the original standardization samples. James Flynn has estimated the average gain in IQ to be approximately 0.3 IQ points per year (Flynn 1984). The phenomenon of secular gains in IQ scores has been named "the Flynn effect". Table 2 shows the gains for the Full scale IQ and for selected subtest scores.

Table 2. Demonstration of the Flynn effect in U.S. WAIS scores (Flynn 2010, 2012).

Index/Subtest	WAIS(1955)/ WAIS-R(1981)	WAIS-R(1981)/ WAIS III (1997)	WAIS III (1997) WAIS IV (2008)
Full scale IQ	104/111	102/106	100/103
Block design	100/101	100/101	100/100
Vocabulary	100/102	100/101	100/101
Coding	100/102	100/101	100/100

Vanhanen (2008, Vanhanen & Laulumaa 2011) compared Finnish WAIS III and WAIS-R scores in a group of patients from Kuopio University Hospital. Table 3 shows some of the results of this study. Vanhanen hypothesized that the WAIS III norms, based on a smaller sample than the WAIS-R standardization, were distorted.

Table 3. Comparison of Finnish WAIS III and WAIS-R norms (Vanhanen & Laulumaa 2011).

WAIS scale	Age group	n	WAIS III	WAIS-R
Full Scale IQ	18–19	21	110	102
Full Scale IQ	20–24	25	103	96
Full Scale IQ	45–64	25	96	102

Test norms may also be compared to norms of other tests that measure the same construct. For example, the correlation between WAIS III FSIQ and the Stanford-Binet IV global score is 0.88 (Wechsler 1997), and the correlations between SB Verbal reasoning/WAIS VCI and SBI Visual reasoning/WAIS POI are 0.87 and 0.78, respectively (Wechsler 1997). In a study by Cheramie and others (2012), the correlation between the WAIS IV VCI score and WJ III Crystallized Intelligence was 0.82, and between the WAIS IV PRI and WJ III Fluid intelligence scores it was 0.65.

Raymond Cattell developed a “culture fair” intelligence test in order to study intelligence across nations (Cattell 1949). His goal was to produce a test that would measure intelligence irrespective of environmental factors. In essence, the culture fair test is based on non-verbal tasks, while verbal subtests such as those used in the WAIS and the WJ, are excluded. However, the analysis of test scores across nations was a fairly marginal topic before Lynn and Vanhanen launched the concept of national IQ (Lynn & Vanhanen 2002). They calculated national IQs based on results from the standardization studies of Raven's matrices (Raven *et al.* 1996), a non-verbal reasoning test highly similar to the WAIS III matrix reasoning, and supposedly “culture-free”. For example, Lynn and Vanhanen calculated the British

IQ to be slightly higher than the US IQ (100 vs. 98) based on the 1992 and 1993 standardization studies of the SPM (Raven *et al.* 1996), and the German national IQ to be 99, based on German standardization studies (Kratzmeier & Horn 1980).

Lynn and Vanhanen (2012) have shown that national IQ is highly correlated with a wide range of economic, social, and political phenomena, such as the Gross domestic product (GDP), educational output, economic and political freedom, health, and religiosity of the nation. They conclude that genetic factors underlie national differences in g, the general factor of intelligence. Rindermann (2007) found that national means from the Programme for International Student Assessment (PISA), for scholastic aptitude among 15-year old pupils have a roughly 0.90 correlation with national IQ. This result is keeping with national studies that show a high correlation between education and IQ. For example, in a study by Longman, Saklofske, and Fung (2007), the mean IQ of Americans with eight-years of education or less was 86 while the respective figure for those with at least 16 years of schooling was 112. Roivainen (2010a) found the mean Performance IQs of the younger age groups of the Finnish WAIS III sample to be inferior to those of the US and Spanish standardization samples, while Finnish students have outperformed American and Spanish students in the PISA assessments (OECD 2015). The results were interpreted as supporting Vanhanen's (2008) findings, and further studies involving other nations and a wider selection of scales were proposed.

2.4 Validity studies of projective tests

Despite the widespread popularity of the Rorschach test in clinical settings starting from the 1940s, first empirical studies of the test found that it performed poorly in predicting vocational success, as well as in differentiating individuals with psychiatric problems from psychologically healthy individuals (Guilford 1948, Cronbach 1949, Holtzman & Sells 1954). The only exception to poor validity was shown for differentiating schizophrenics from other patients (Armitage & Pearl 1957). The conclusion of leading academic psychometricians by the end of the 1950s was that the Rorschach failed to meet most of the criteria for psychological tests such as internal consistency, interrater reliability, and validity (Eysenck 1959). The leading Rorschach practitioners such as Beck (1959) and Klopfer (Klopfer & Davidson 1962) dismissed such criticism based on the concept of "clinical validation": *"Let it be said at once and unequivocally that validation as is sought in a laboratory experiment is not at present to be expected for whole personality*

findings, whether by the Rorschach test or by any other. The measure of validity must be limited to indicating direction. It cannot be a number such as a correlation or other coefficient" (Beck 1959: 275). The clinical validation argument roughly states that, based on experience from a large number of patient cases, clinicians are able to notice and interpret details of test responses, and behaviour that empirical analysis may overlook and integrate these into a coherent picture of the individual's personality (Wood *et al.* 2003).

One major shortcoming of the Rorschach test was that there were divergent systems of administration and scoring. The two major competing methods of interpretation were those of Klopfer (Klopfer & Davidson 1962) and Beck (1952). Others were developed by Hertz (1959) and Piotrowski (1957). Exner (1974) aimed at fixing this problem by introducing his comprehensive system. Exner reviewed the published Rorschach research and conducted new studies to choose from each Rorschach system those scales and interpretation principles that had the strongest scientific support. Exner also presented norms for the Rorschach that by 1986 included 1 332 test protocols of individuals with no psychiatric hospitalizations or long-term psychotherapy (Exner 1993). Studies by Exner and other researchers showed the Comprehensive scoring system to be reliable (interscorer agreement > 0.75), and high test-retest reliability (> 0.80) was also reported (Exner 1993). However, subsequent research questioned the validity of the norms collected by Exner and associates at the Rorschach Workshop Society as well as the high reliability figures reported (Shaffer, Erdberg & Haroian 1999). Based on a review of 32 studies, Wood, Nezworski, Garb & Lilienfeld (2001) concluded that Exner's norms tended to overpathologize subjects. The scientific debate, at times intensive, and research on the validity of the Rorschach method and on Exner's interpretation system has continued in Finland (Nummenmaa & Hyönä 2005, Koistinen 2005) and abroad. For example, the Google Scholar index lists 32 studies involving "Rorschach norms" that have been published since 2010 (Google 2015).

The history of validity studies on other projective techniques, such as the Thematic Apperception Test (Murray 1943, Niitamo 1999, Meyer 2004) tends to mirror that of the Rorschach. For example, in a study by Fineman (1977), the correlation between TAT and different self-report questionnaire scales on achievement was 0.15. Niitamo (1999) lists three potential explanations for the lack of convergent validity between personality measures: 1) One of the measures is invalid, 2) Poor validity of all personality dispositions, and 3) Different types of measures (projective vs. questionnaire type tests) tap different aspects of the personality constructs: "underlying and implicit" aspects versus "surface-level,

explicit and conscious" aspects. Niitamo attributes the poor convergent validity shown in his own studies between TAT scales, self-report scales and interview ratings to the difference between "implicit motives" seen in TAT responses and "explicit motives" measured by questionnaires and interviews. This phenomenon, the general lack of relationship between free response and self-report methods, has been called the *heteromethod convergence problem* (Bornstein 2009).

2.5 Validity studies of the Wartegg test

The first empirical studies of the WZT were performed in Finland by Takala and Hakkarainen (1953), who developed a quantifiable scoring system and administered the WZT to over 1 000 subjects. The results showed that the test could serve as a possible predictor of vocational success, and it was valid in differentiating occupational groups. In a later study, Takala (1964) found that test scores correlate with intelligence and occupational interests, but a correlation with personality traits was not found.

According to Gardziella (1985), his method is not based on any personality theory, but on his clinical experience involving "thousands of cases". For example, human drawings are considered to be a sign of sociability; ambitious people draw long lines in Box 3 of the test form (depicting ascending stairs, graphs showing growth, etc.), while inactive or depressed individuals tend to draw shorter lines (descending stairs, downhill, etc.). According to Gardziella, impulsive people may begin drawing in any of the eight boxes, while controlled people proceed in numerical order, beginning from box one. The interpretation of the qualitative aspects of the drawings such as drawing size, pencil pressure, and number of details resemble those concerning other projective drawing tests (Machover 1949).

A major shortcoming of Gardziella's WZT manual is that it does not include any empirical evidence concerning the validity of the method. This fact raised increasing criticism in the 1980s, and 90s from academic psychologists. According to Tamminen and Lindeman (2000), the only validity study up to year 2000 in which other personality tests had been used as a criterion for WZT scores based on Gardziella's interpretation method was an unpublished study by Roivainen (1997).

In the study by Roivainen (1997), Wartegg drawings by 246 clients of vocational counselling at Kemijärvi Job centre were analysed. The median age of the subjects was 20 years. Sixty-seven subjects were men and 179 were female. The majority of the subjects were students close to completing their compulsory education or high school and planning their future studies. A minority of subjects

were adults who were planning to change their job. The Wartegg drawings were compared to scores on Holland's (1985) vocational SDS test ($n = 201$) and an abbreviated version of Cattell's 16PF personality test ($n = 65$) (Cattell, Cattell & Cattell 1993). It was found that the mean time for completing the Wartegg form was 12 minutes, with 13 subjects finishing in less than five minutes while three subjects had used more than 30 minutes to complete the test. Drawing time had a negative correlation (-0.37) with Cattell factor I (Sensitivity) ($p < 0.05$) and a weak positive correlation with SDS score on factor I (Intelligence) ($r = 0.16$, $p < 0.05$). Age and gender did not have a significant correlation with drawing time.

Gardziella's hypotheses concerning the drawing sequence were not confirmed in the study. For example, the WZT manual suggests that warm and emotional individuals prefer Box 2 in the Wartegg form and tend to draw in this box first, while emotionally cold and schizoid individuals often find it difficult to complete this box and leave it until last to be completed. However, the mean score on Cattell factor A (warmth) was practically the same (11.9) for those who started drawing from Box 2 as compared to the mean of other respondents (12.0). Other analyses showed that age and gender did not have a significant effect on drawing sequence.

The Wartegg drawings were divided into three groups based on drawing content: human drawings, animals and plants, and inanimate objects. Age and Gender had an effect on drawing content. Table 4 shows the correlation between drawing content and scores on the 16PF and Holland's SDS in 179 female subjects. The results show that interest in technical, artistic, theoretical or social vocations and hobbies has a weak positive correlation with the number of drawings in the WZT protocol depicting humans and nature, and a moderate negative correlation with drawings depicting inanimate objects. The correlations with the 16PF scores were low.

Table 4. Correlation between personality test scores and the number of human, nature and inanimate object drawings in WZT protocols (Roivainen 1997).

	Human drawings	Nature	Inanimate objects
SDS scales (n = 144)			
Realistic	0.24[1]	0.19[2]	-0.32[1]
Investigative	0.28[1]	0.28[1]	-0.42[1]
Artistic	0.22[1]	0.25[1]	-0.35[1]
Social	0.20[2]	0.09	-0.19[2]
Enterprising	-0.03	0.18[2]	-0.27[1]
Conventional	-0.06	0.10	-0.04
Cattell 16PF (n = 52)			
A Warmth	0.21	0.06	-0.17
C Emotional stability	0.07	0.05	-0.09
E Dominance	0.30[1]	0.03	-0.17
G Consciousness	-0.03	0.04	-0.03
H Social boldness	0.20	0.10	-0.20
I Sensitivity	-0.06	0.27[2]	0.20
N Privateness	-0.25[2]	-0.09	0.24
O Apprehension	0.16	0.00	-0.07
Q2 Self-reliance	-0.10	-0.11	0.16
Q4 Tension	0.06	0.16	-0.16

[1] Significance = 0.005, [2] Significance = 0.025, [3] Significance = 0.05

The major shortcoming of this study is that the drawings were analysed by one person only, and thus the reliability of the analysis may be questioned.

Tamminen and Lindeman (2000) studied the criterion validity of Gardziella's method using the Personality Research Form (Jackson 1974) and the State-Trait Anxiety test (Spielberger *et al*. 1977) as criterion tests. They found that drawing contents were not correlated with criterion measures of anxiety, need for affiliation, need for achievement or attachment styles, as suggested by the WZT interpretation manual (Gardziella 1985). They concluded that much of Gardziella's method was based on magical thinking and that the method had poor validity in personality assessment.

Mattlar, Lindholm, Haasiosalo and Vesala (1991), who investigated whether the WZT may be used to detect alexithymia, present a more favourable view of the WZT. They hypothesized that alexithymic subjects would favour non-creative solutions in their drawings and prefer to draw inanimate objects. In this study, the interrater agreement between assessments of alexithymia based on WZT drawings was as high as 0.77 for four independent judges. However, WZT scores were not

compared to other measures of alexithymia. Thus, the study by Mattlar *et al.* suggests that WZT scores may be reliable, but says little about the validity of this test.

2.6 The validity of validity scales

One of the main reasons for the popularity of projective tests in the clinical setting is that questionnaire-type methods based on self-report are not useful in cases where the respondent has poor introspective skills (Groth-Marnat 2003). One of the criteria of psychotic illness is a distorted sense of reality, and many other less severe psychiatric disorders may also involve a distorted self-image. One way to solve this problem has been to use response-style and validity scales in addition to the scales that measure psychiatric symptoms or other psychological constructs.

Validity scales may be defined as measures of an individual's motivation and their capacity to respond meaningfully to psychological test items (Ben-Porath 2012).

Studies of validity scales show that they can effectively detect faking (Piedmont *et al*. 2000, Schinka *et al*. 1997, Sellbom & Bagby 2008).The MMPI infrequency scales are valid in the detection of over-reporting of pathology (Baer *et al*. 1999) and the K scale for under-reporting of symptoms (Baer *et al*. 1998, Putzke *et al*. 1999). However, poor specificity constitutes a major problem for validity scales. For example, Lim and Butcher (1996) found that a screening scale that was 100% accurate in detecting faking in a student sample produced a 30% false positive rate in a sample of honest psychiatric patients.

2.6.1 The Chapman Infrequency Scale

Validity scales are also frequently used in survey studies to screen out careless respondents and invalid protocols. The Chapman Infrequency Scale, CIS, developed by Chapman and Chapman (1986) is one of the most widely used screening instruments in psychiatric survey studies. The CIS consists of 13 questionnaire items that are unusual or nonsensical in content, such as "*I cannot remember a time when I talked with someone who wore glasses*". The CIS is modelled on Jackson's infrequency scale (Jackson 1974), used in his Personality Research Form (PRF). In their original study, Chapman and Chapman used three points as a cut-off score for careless responding, and this has become the standard in psychiatric research. The CIS is a non-standardized test and, regardless of its

popularity, there have been no validity studies investigating its psychometric qualities. In light of the nonsensical content of the items, careless responding rather than impression management or over-reporting of symptoms is assumed to underlie high scores on this scale (Peltier & Walsh 1990). Respondent motivation is a major factor affecting the CIS score. For example, Fonseca-Pedrero and colleagues (2009) report 53 careless respondents in a psychiatric survey sample of 737 Spanish university students from different faculties, while in another psychiatric survey sample, of 325 American undergraduate psychology students, no one scored more than two points (Bonogofsky 2007). In a study by Merckelbach and colleagues (2010), nine student volunteers in a sample of 306 admitted that they often intentionally gave wrong answers in the survey.

On the basis of their study comparing several methods for identifying careless responding in an online survey of a student sample – a) outlier analysis, b) response consistency indices, c) response time, and d) nonsensical items – Meade and Craig (2012) endorse the use of nonsense items of the type used in the CIS (e.g., "Respond with 'strongly agree' for this item"). However, due to the potential problem of poor specificity discussed above, the cut-off score of three points recommended by Chapman and Chapman may not be universally valid across survey target populations.

3 The aims and objectives of the study

The overall aim of the study was to examine the validity of test norms of widely used psychological tests. The aims of the original publications were:

I To study the validity of the Wartegg test in the assessment of alexithymia.
II To study cross-national variation in WAIS test profiles and the validity of national test norms.
III To study the stability of cross-national differences in WAIS test profiles.
IV To study the cross-generational validity of norms of vocabulary tests.
V To study the validity of the Chapman infrequency scale norms.

4 Methods and subjects

4.1 Subjects

The subjects in original study I were 83 (35 males) patients of the Oulu Deaconess Institute, who participated in a group test as a part of a health assessment requested by social insurance and employment offices on behalf of their clients.

Studies II, III and IV are based on the test standardization samples of Wechsler's intelligence scales WAIS and WISC in USA (Wechsler 1949, 1955, 1974, 1981, 1991, 1997), Finland (1971, 1992, 2005), France (2006), Germany (2006), and Spain (1999), and on the United States General Social Survey (2009) samples.

In study V, the sample consisted of 5 024 (2 264 males) participants from the Northern Finland Birth Cohort 1966 study who had completed psychiatric questionnaires as a part of their follow-up clinical examination in 1997 (average age 31 years).

4.2 Methods

4.2.1 Original study I

Original study I was a convergent validity study. All subjects were administered the WZT and the criterion tests: the RBDI, a Finnish version of the Beck depression inventory (Beck 1961, Raitasalo 1995), and the Toronto Alexithymia scale (Bagby *et al.* 1986, Joukamaa *et al.* 2001). Standard test instructions were used for all the tests. The Wartegg drawings were assessed by two independent judges (the authors). Clear-cut criteria for human drawings were set: all drawings depicting human beings or body parts, including strawmen were accepted. Snowmen, ghosts, robots and masks were excluded.

4.2.2 Original study II

Original study II is based on the data from the standardization studies of the Wechsler adult intelligent scale III. In study II, the data from national standardization studies of the WAIS III in the USA, United Kingdom, Germany, France, Spain and Finland was analysed. The mean raw scores from national

standardization studies that are reported in national test manuals were compared across nations. Only non-verbal subtests were included in the analysis, because verbal subtests such as Vocabulary have different test items in different countries. The WAIS index scores were calculated by converting national raw scores into USA standard scores. For example, the age group 20–24 in the Finnish WAIS III standardization sample had a mean raw score of 18 on the Picture completion test, which gives 8 standard points in the American test. The Performance IQ and the index scores POI and PSI are calculated on the basis of the sum of the standard points from the respective subtests. For example, the age group 20–24 of the Finnish standardization sample scored 8 standard points on the Picture completion subtest, 10 points on the Block design subtest and 11 points on the Matrices subtest. The sum of 29 points gives a Perceptual organization index score of 97. The mean standard scores of this group on the Coding and Symbol search subtests were 7 and 10, respectively, and the Processing speed index score based on the sum of the standard points (= 17) is 91.

4.2.1 Original study III

In original study III, a combined longitudinal and cross-national analysis of WAIS norms was performed. In this study, US and Finnish WAIS, WAIS-R and WAIS III non-verbal and Digit span subtest norms for the age group 20–34 were compared. Longitudinal analysis of intra-national IQ gains was possible for the Coding and Digit span subtests only. Concerning other subtests, test editions differed significantly in the testing materials and procedures. The time limit for the Coding subtest was changed from 90 seconds (WAIS and WAIS-R) to 120 seconds in the WAIS III. To compare the three versions, the WAIS and WAIS-R coding scores were multiplied by 1.33.

4.2.4 Original study IV

Original study IV is based on data from the US standardization studies of different editions of the WAIS and WISC. In study IV, IQ gains on the Vocabulary subtest of the different editions of the WAIS and WISC were compared with the frequency of the use of the words used as test items. Word frequency was studied using the Google Books database (Michel *et al*. 2011) that shows the annual frequency of the use of words in English language books starting from year 1800. Another vocabulary test, the General Social Survey Wordsum (GSS 2009) was also included

in the study. Correlation between item frequency and item difficulty was analysed. The order of presentation of test items in the WAIS and WISC vocabulary subtests is based on item difficulty with the easiest words presented first. Thus, they present a rank-order scale of difficulty of the words.

4.2.5 Original study V

Original study V is based on data from the Northern Finland Birth Cohort 1966 Study, where a collection of self-rating-based psychological scales were applied (Miettunen *et al*. 2011) as a part of a clinical examination in 1997 (at the age of 31 years). Careless responding was studied by using a 12-item version of the Chapman Infrequency Scale (CIS). The item that was excluded states *I often notice that I am limping, which is a consequence of an old parachute jumping accident.* Other scales included in the study were Wisconsin schizotypy scales–the Social and Physical Anhedonia Scales (Chapman *et al.* 1976) and the Perceptual Aberration Scale (Chapman *et al.* 1978) that were used to study differences in schizotypal traits between careless respondents (three or more CIS points) and other respondents. These four scales are composed of 30 to 61 items, and they measure, perceptual aberration, social anhedonia, and physical anhedonia The number of careless respondents among subjects with a history of psychiatric hospitalization ("the psychiatric group") was compared to that among healthy subjects ("the healthy group") with no psychiatric hospitalizations registered in the nationwide Finnish Hospital Discharge Register (FHDR) between the years 1982–1997. The endorsement for individual items of the CIS in the two respondent groups was also analysed, and the effects of education and gender on CIS and other test scores were studied.

4.3 Statistical methods

In original study I, the subjects were divided into three groups: alexithymics (TAS score > 59), depressed non-alexithymics (TAS < 60 and RBDI score > 4) and individuals with no diagnosis (TAS < 60; RBDI < 5). The number of individuals that had drawn no human drawings was compared across groups using the Chi-squared test. The mean scores on the TAS were calculated for subjects with no human drawings in their WZT protocol, and for subjects with at least one human drawing in the WZT. The difference was statistically analysed with the t-test.

Interrater agreement was analysed by calculating the correlation between the two judges' independent ratings (human drawing = 1, non-human = 0) of drawings.

In original study II, differences between nations in the WAIS index scores were statistically analysed using the Student's t-test, and the standard deviation value of 15 was used for all calculations.

In study III, differences in the mean raw scores between the US and Finnish standardization samples in WAIS subtests were analysed statistically using the t-test.

In study IV, correlation between word frequency and word difficulty was calculated using Spearman's rank correlation test. The correlation between gain in the vocabulary score from one test standardization sample to the other and median change in word usage was calculated. Correlation between change in usage frequency and change in difficulty for each test item separately was calculated for the Wordsum items.

In study V, logistic regression analysis was performed in order to compare the number of careless respondents in the "psychiatric" and "healthy" groups. The scores on the psychiatric scales for the psychiatric and healthy groups were compared using t-test for statistical analysis. The endorsement for individual items of the CIS in the two respondent groups was analysed statistically by using the using the Chi-square test.

4.4 Ethical considerations and personal involvement

The author is the sole author of original studies II, III and IV. The data used in studies II, III and IV already existed and had been published, and no specific ethical considerations were involved. The author planned study I and collected the data. The analysis of the drawings was planned and executed together with the co-author, Ms. Ruuska. The report was written by the first author. The study design was approved by the research director of the Oulu Deaconess Institute. The participants were informed of the study when they were given feedback on their test results, and oral consent was obtained to use the WZT, TAS and RBDI scores for research purposes. The Northern Finland Birth Cohort 1966 study design has been approved and is under continuous review by the Ethical Committee of the Northern Ostrobothnia Hospital District. After complete description of the study to the participants, written informed consent was obtained. Study V was planned by the author with the co-authors Professor Miettunen and Professor Veijola. The

statistical data analysis was performed by Professor Miettunen. The report was written by the first author.

5 Results

5.1 Original study I

In original study I, twenty respondents (24%) out of a total of 83 scored 60 or more points on the TAS, thus meeting the criterion for alexithymia. Eleven out of the twenty alexithymics had no human drawings in their WZT protocol, while only one subject out of 25 among subjects with TAS score lower than 40 did not draw humans. The mean score on the Toronto Alexithymia Scale was 56.0 for subjects with no human drawings as compared to 45.4 for those subjects with at least one human drawing in their WZT form. The difference is significant, $t(81) = 3.71$, $p < 0.001$. The correlation between the total number of human drawings in the WZT and the TAS score was -0.33.

Correlation between the Depression inventory RBDI score and the number of human drawings in the WZT protocol was -0.10. In all, 42 subjects scored 5 points or more on the RBDI, the cut-off score for mild depression. There was no significant difference in the number of subjects with no human drawings between the depressed (RBDI > 4, n = 6) and non-depressed groups (RBDI < 5, n = 5), $\chi^2(1) = 0.30$, $p = 0.586$. It was also found that age had a weak positive correlation with the Alexithymia score (0.17) and a weak to moderate negative correlation (-0.39) with human drawings. Inter-rater reliability, the agreement between two judges, was 0.94.

Table 5. The number of alexithymic and depressed subjects drawing human drawings (HD) in the Wartegg test (From Original study I).

Drawings	Alexithymic	Depressed nonalexithymic	No diagnosis
HD	9	23	29
No HD	11	6	5
Total	20	29	34

5.2 Original study II

In study II, significant cross-national differences were found in WAIS III scores. Compared to the US norms, the WAIS perceptual organization index score POI was significantly higher in France and Germany for the age groups 16–55, and in Spain for the age groups 16–34. The American sample had higher scores in the age groups 35–69, compared to Spain. The Finnish age groups 15–17 and 35–44 had

significantly lower POI scores than the respective groups in the American sample, while the oldest Finnish age group, 65–74 had higher mean POI scores than the American sample

The results also showed cross-Atlantic differences in the ratio between the POI score and the Processing speed index score (PSI). The POI mean was higher than the PSI mean in all age groups in each European sample. For example, eight age groups in the German sample had higher mean POI scores and lower PSI scores than the US sample (Table 6).

Table 6. WAIS III IQ and Index means for four European countries; US norms (From Original study II).

Country	Age	PIQ	POI	PSI	n
Finland	15–17	90[2]	94[1]	87[2]	40
	18–19	94[1]	96	88[2]	27
	20–24	93[1]	97	91[2]	44
	25–34	98	100	96[1]	91
	35–44	96[2]	96[2]	93[2]	85
	45–54	98	101	93[2]	59
	55–64	98	101	99	71
	65–74	99	103[1]	95[2]	57
Spain	16–19	101	103[1]	101	163
	20–24	102	103[1]	99	153
	25–34	103[2]	104[2]	101	272
	35–54	94[2]	95[2]	93[2]	408
	55–69	88[2]	91[2]	86[2]	237
France	16–17	106[2]	107[2]	105[2]	84
	18–19	110[2]	114[2]	101	78
	20–24	111[2]	116[2]	103	101
	25–29	109[2]	111[2]	105[2]	99
	30–34	107[2]	111[2]	100	103
	35–44	106[2]	109[2]	100	102
	45–54	106[2]	109[2]	100	87
	55–64	105[2]	107[2]	100	86
	65–69	108[2]	109[2]	105[2]	94
	70–74	104[3]	106[2]	95[2]	101
	75–79	95[2]	97	91[2]	93
	80–89	95[2]	97	95[2]	76

Country	Age	PIQ	POI	PSI	n
Germany	16–17	100	103[1]	96[2]	139
	18–19	102	105[2]	96[2]	142
	20–24	106[2]	109[2]	99	168
	25–29	109[2]	114[2]	99	171
	30–34	110[2]	116[2]	103[1]	146
	35–44	107[2]	111[2]	99	140
	45–54	103	106[2]	96[2]	141
	55–64	98	101	96[2]	140
	65–69	99	101	96[2]	140
	70–74	102	103[1]	93[2]	139
	75–79	98	99	93[2]	142
	80–84	100	100	96[2]	138
	85–89	98	96[2]	96[2]	140

[1] Statistical significance $p < 0.05$, [2] Statistically significant difference $p < 0.01$ from USA value of 100 ($n = 200$). SD = 15, [3] performance IQ, [4] perceptual organization index, [5] processing speed index

5.3 Original study III

In study III, the results showed significantly higher mean scores for the Coding and Digit span subtests for young adults (16–34) in the American WAIS, WAIS-R and WAIS III samples, as compared to the respective Finnish samples. The Finnish WAIS-R and WAIS III samples had significantly higher scores on the Block design subtest. The Finnish WAIS sample also had a higher mean on the Block design test than the US sample; however, the difference was not statistically significant, $t(926) = 1.518$, $p = 0.13$. The American WAIS and WAIS III samples also had higher scores on the Picture completion subtest.

A longitudinal analysis was possible for the Coding and Digit span subtests only, because test editions differ somewhat in the testing materials and procedures for the other subtests. A rise in test scores over time and across test editions was evident in the extrapolated Coding and Digit span scores. The Finnish–US gap had grown in the Digit span test and decreased in size in the Coding subtest.

5.4 Original study IV

In original study IV, the results showed that the difficulty of vocabulary test items is dependent on their frequency of use. Usage frequency, as estimated by Google search and the Google Ngram viewer, had a 0.69 to 0.81 correlation with item difficulty in the WAIS and WISC vocabulary subtests and a 0.38 to 0.52 correlation

in the Wordsum vocabulary test. Changes in word usage frequency between test standardizations were fairly small. For single words, the largest decrease and increase were close to 50%, but the median change in usage frequency was only -17% for WISC words, and -8% for Wordsum words. There was little or no change in the usage frequencies of the WISC-R (-5%), WAIS (-4%) and WAIS-R (0%) vocabulary items. However, the results showed that the vocabulary test items tended to become less rather than more popular over time.

Because there was no item-level longitudinal data available for the Wechsler tests, the correlation between the change in usage frequency and the change in difficulty over single test items could be calculated for the 10 Wordsum items only. The correlation was fairly weak (-0.21). For all the seven tests, the correlation between vocabulary score gain from one test standardization sample to the other, and the median change in word usage was 0.33.

5.5 Original study V

In study V, logistic regression analyses showed that the odds ratio for careless responding was 2.1 (95% CI: 0.9–5.0) for respondents with a psychiatric diagnosis relative to healthy respondents (Table 7). Only for one test item was the frequency of careless responding significantly lower among psychiatric patients *("On some occasions I have noticed that some people are better dressed than I")*. Basic education (OR 1.9; 95% CI: 1.0–3.6) and male sex (OR 2.1; 95% CI: 1.4–3.1) were also found to correlate with careless responding. The difference between careless respondents and other respondents was statistically significant (Student t test, $p < 0.001$) for all the three psychiatric rating scales, the Chapman Revised Physical (PAS), and Social Anhedonia (SAS) scales, and on the Perceptual Aberration (PER) scale. The ranking of CIS items among careless responders (CIS score > 2) differed somewhat from the whole sample. Compared to the whole sample, careless responders were eight times more likely not to endorse the item *"I have never combed my hair before going out in the morning"* but only twice as likely to endorse the item *"I believe that most light bulbs are powered by electricity"*.

Table 7. Chapman's Infrequency Scale (CIS) score by sex, education, and health (From original study V).

Variable	0–2	3 or above	OR[1] (95% CI[2])
Diagnosis			
No psychiatric diagnosis	4 784	102	reference
Psychiatric diagnosis	131	6	2.1 (0.9–5.0)
Gender			
Female	2 719	40	reference
Male	2 196	68	2.1 (1.4–3.1)
Education			
Tertiary	1 319	26	reference
Secondary	3 875	66	1.1 (0.7–1.7)
Basic	421	16	1.9 (1.0–3.6)

[1] odds ratio, [2] confidence interval.

6 Discussion

6.1 Human drawings in the WZT as a measure of Alexithymia (I)

The results of Study I show that human drawings are slightly less often found in the WZT test protocols of alexithymic individuals as compared to non-alexithymic individuals. No correlation was found between the number of human drawings in the WZT protocol and depression score in the depression inventory.

Definitions of Alexithymia include as central features of this syndrome difficulty in identifying and expressing feelings and a lack of fantasy. Obviously, both of these features may lead to a lack of human content in WZT drawings. Completing the symbols of the Wartegg test form to depict squares, circles or other abstract forms requires less imagination and emotional involvement than drawing humans. The absence of human drawings should not by itself be considered a sufficient criterion for alexithymia. However, the results of the study show that the projective hypothesis should not be discarded. The results show that projective drawings may correlate with psychological constructs such as alexithymia.

Soilevuo-Grönneröd & Grönneröd (2012) collected all available studies, published and unpublished on the Wartegg test for a meta-analysis. They found 507 studies in all, 230 were journal articles, 113 were books, 40 were dissertations, and 124 were other types of publications. They concluded with an overall positive evaluation of the potential of the WZT. The interscorer agreement correlation of 0.94 observed in original study I is one of the highest reliability figures reported in the 37 empirical studies reported by Soilevuo-Grönneröd and Grönneröd. The interscorer reliability was reported in 15 of the studies included in the meta-analysis, and the average correlation calculated by Soilevuo-Grönneröd and Grönneröd was 0.74. The figure of 0.94 seen in study I is likely to be due to the fairly narrow scope of the study. The variable of human content can be defined in an explicit way as was done in study I. Aspects such as pencil pressure, “maturity” or “compatibility with the given initial sign” that affect WZT interpretation in the Gardziella and other systems are obviously harder to evaluate objectively. The -0.33 correlation found between TAS and WZT is not very high, but among the studies analysed by Soilevuo-Grönneröd and Grönneröd this result is one of the strongest empirical findings that speaks for the validity of the Wartegg test. We may conclude that efforts to develop an empirically validated interpretation method for the WZT should be continued.

6.2 Cross-national variation in intelligence test norms and the validity of national WAIS norms (II, III)

The results of original study II show a clear cross-Atlantic difference in intelligence test profiles. Europeans have higher scores on the Perceptual reasoning subtests, while Americans perform better on the Processing speed subtests. The Finnish index scores were low compared to those of other nations. The results of original study III show that the differences in test profiles between the US and Finland seen in study II are fairly stable. Thus, they are not specific to the WAIS III test or the standardization samples.

The differences between Spanish, American, German and French WAIS III POI scores found in study II are correlated with the differences in mean PISA scores of these countries. As noted above, education correlates with the POI score within national samples (Wechsler 2005, Longman *et al.* 2007), and thus countries with high mean scores in the PISA assessments should also have high national POI scores. However, differences in general intelligence cannot explain the fact that the European PSI scores are in many cases lower than the US scores. The PSI subtests have a fairly high loading on the g-factor.

The results of study II seem to verify the hypothesis that the Finnish WAIS III norms are distorted for the younger age groups (Vanhanen & Laulumaa 2011, Roivainen 2010a). In study II, the mean POI score was lower in the Finnish sample as compared to the American sample, while the Finnish PISA scores are higher than the American scores. It is probable that the Finnish WAIS III standardization sample is not representative of the Finnish general population, because the sample for the younger age groups was recruited from clients of labour offices. The Finnish test publisher, PKOY has also acknowledged (Heiskari 2010) the problems involving the norms of the younger age groups reported in study II and published a new scoring table that combines the age groups 15–24. This manoeuvre improves the validity of the Finnish norms somewhat, because the largest cross-national discrepancy was observed for the youngest age group 15–17.

It can be hypothesized that one of the non-g related factors that raises American scores on the PSI index is related to test-taking attitudes. Test-taking attitudes have been shown to have a significant effect on test performance. For example, the increase in guessing behaviour over generations has been shown to be one factor underlying the Flynn effect (Must & Must 2013). Anecdotal evidence suggests that fast performance and speed are more highly valued in the US as compared to Europe, where accuracy and avoiding mistakes is a relatively higher priority

(Rosselli & Ardila 2003). In cross-national studies comparing norms of neuropsychological tests, American samples tend to perform well on processing speed measures (Agranovich & Puente 2007).

The results of original study II have been used in other studies on the Flynn effect and on the cross-national differences in IQ. Based on the results of original study II, Flynn (2012) has estimated the French national IQ to be 106.2, Germany at 100.5, Spain at 97.1 and Finland at 93.1, when the reference value of USA = 100. For his analysis, Flynn reduced 0.3 IQ points for each year from the US standardization to the year of national standardization of the WAIS III in the respective nations in order to adjust for the Flynn effect. Lynn and Vanhanen (2013) have also used the figures reported in original study II and calculated the national IQs for France, Germany and Spain to be 101, 101 and 94, respectively. Lynn and Vanhanen use the national IQ for the United Kingdom ("Greenwich IQ" = 100) as their reference value.

In original study III, the observed differences between the Finnish and US norms were by and large those that were expected. The American samples had higher means on the Coding subtest that measures processing speed, while the Finnish samples had higher mean scores on the Block Design subtest. These results may be explained by the cultural and educational factors discussed above. However, the fact that among non-verbal subtests only Block design showed a consistent Finnish advantage requires an explanation. In national WAIS samples, Block design tends to have a moderate to high (0.40–0.60) positive correlation with the other non-verbal subtests (Picture arrangement, Picture completion) that are also highly g-loaded and correlate with PISA scores (Rindermann 2007). In the case of WAIS III, sampling bias may also have had some effect on the results. The Block design subtest is somewhat less affected by an examinee's educational background than the other non-verbal subtests (Groth-Marnat 2003) and there was an overrepresentation of individuals with only primary education in the Finnish WAIS III sample.

The observation that the mean score on the Picture arrangement subtest was higher in the oldest US WAIS sample (Wechsler 1955) compared to the Finnish WAIS sample (Wechsler 1971) might be explained by different rates of TV ownership and comic book reading. Approximately 50% of US homes had a TV set in 1953–1954 (TV History 2011), the year of the American WAIS standardization, while in Finland this percentage was reached only by the end of the 1960s (Ilmonen 1996). By the 1980s (WAIS-R standardization), TV ownership was equally common in the two countries. It may be that the plots of the Picture arrangement

stories are easier to understand for experienced consumers of TV entertainment. The American WAIS and WAIS III samples were also superior on the Picture completion subtest. The WAIS-R picture completion subtests cannot be compared due to differences in test items. The US/Finland difference is difficult to explain. The pictures seem culturally fair, and they do not depict things that would be clearly more familiar to Americans than to Finns. Actually, logs covered with snow (Picture 25), rowing boats (18), fireplaces (13) and footprints on soft surfaces (sand or snow) (12) might even be more common in Finland.

The US/Finland difference on the working memory Digit span subtest might be explained by the phonological loop hypothesis. Cross-cultural studies on working memory show that differences in digit articulation time across linguistic borders affect digit span, such that the longer the articulation time, the shorter the digit span (Hoosain & Salili 1988, Naveh-Benjamin & Ayres 1986). Finnish numerals (nolla, yksi, kaksi, kolme, neljä, viisi, kuusi, seitsemän, kahdeksan, yhdeksän) contain more phonemes than English numerals; therefore articulation is probably slower.

On the basis of the results of study III, it was predicted that raw Coding scores would be lower and Block Design scores would be higher in the then ongoing Finnish WAIS IV standardization study compared to the US standardization scores. A recent study that compares the US and Finnish WAIS IV matrices (Dutton & Kierkegaard 2014) subtest actually shows a higher mean (103 vs. 100 IQ points) for the Finnish sample. Table 8 shows the means for other subtests by age group in the Finnish and US samples. The Finnish sample has higher mean scores on the Perceptual reasoning subtests (Block design, Matrix, Visual Puzzle), while the American sample has higher means on the processing speed subtests (Coding and Symbol search) and on the Digit span subtest. Therefore, the factors that underlie the differences between the US and Finnish national IQ profiles have not disappeared.

Table 8. Finnish and US WAIS IV mean raw scores by age groups (Wechsler 2008, 2012).

Age group	Block design (PRI)	Matrix reasoning (PRI)	Visual puzzles (PRI)	Coding (PSI)	Symbol search (PSI)	Letter-number (WMI)	Digit span (WMI)
20–24							
USA	46.5	19	16.5	73	34	20.5	28.5
Fin	50	20	19	69	33.5	19	26.5
25–30							
USA	46	19	15.5	73	34	20.5	28.5
Fin	51	20	20	72.5	32.5	20	27
30–35							
USA	45	18.5	15.5	72	33.5	20.5	28.5
Fin	50	21	19.5	71	32	19.5	27.5

6.3 The cross-generational validity of vocabulary test norms (IV)

In original study IV, the results showed that the difficulty of vocabulary test items is dependent on their frequency of use. However, changes in word usage frequency between test standardizations were fairly small.

The results of study IV imply that the WAIS, WISC, WISC-R, and GSS vocabulary tests may have become somewhat more difficult due to the test words becoming less popular over time. It can be hypothesized that old-fashioned words function well as vocabulary test items in an IQ test, because such words appear in books after they have become obsolete in the spoken language (Curzan 2009). Book-reading is strongly correlated with general intelligence. The standardization samples of the newer test versions are more educated, but old-fashioned words favour older cohorts.

Recent studies (Dorius *et al*. 2014, Meisenberg 2015, Woodley *et al*. 2015) seem to confirm some of the findings of study IV. Dorius and others used a "exposure to word frequency" method based on the Google Ngrams database to measure cohort differences in exposure to word frequency. In this study, a "window of exposure" was defined for each birth cohort based on analysing the frequencies of the WORDSUM words during the school-age years of each cohort. The effect of the changes in word popularity on word knowledge was examined. Dorius and others conclude that *"Our results establish a strong basis for the conclusion that intercohort differences in WORDSUM, across all levels of conceptual difficulty, can be explained by variations over time in cohort-specific exposures to the test*

items, thereby pointing to a "cohort experience" interpretation of intercohort trends in WORDSUM".

In conclusion, while the comparison of vocabulary skills over generations is not an absurd task like that of comparing verbal skills across nations, the same type of "cultural fairness" problems arise. Birth cohorts constitute mini-cultures and people belonging to the same generation share experiences that are different from those of younger and older cohorts (Schaie 2005). The General information subtest may also be affected by this effect. Politicians, athletes or artists well known to one generation may be less known in other cohorts. The results of study IV show that item obsolesce is one factor that needs to be controlled when analysing trends of rising or declining IQ scores.

6.4 Validity of the CIS validity scale (V)

The results of original study V indicate that that individuals with a psychiatric diagnosis score higher on the CIS than healthy individuals do and are, therefore, prone to be excluded more often from studies as "careless respondents."

The results of study V suggest that a more flexible approach may be warranted in the use of validity scales in surveys. High scores on the CIS and on similar scales may be related to characteristics that are actually under study; accordingly, the results for these validity scales should be interpreted and analysed in consideration of this fact. In large-scale studies, the cut-off score for careless responding should be adjusted for the demographic and health characteristics of the sample, and in accordance with the research hypotheses. A slightly higher CIS criterion score might be appropriate with samples of psychiatric patients, in order to avoid loss of data due to a high exclusion rate. In samples of healthy individuals with a high level of education, a lower cut-off score might be used.

Only six individuals with a psychiatric disorder were excluded from the cohort study because of careless responding. In a sample of 5 024 this is, of course, an insignificant number. However, we should bear in mind that these individuals were already a selected group: there were 3 449 non-responders (Haapea *et al.* 2008); cohort members who chose not to participate in the study. We may hypothesize that, in studies where respondents are paid for their participation or encouraged through social and psychological means to take part, careless responding may be much more frequent among passive, reluctant, or ill respondents.

6.5 Limitations of the study

6.5.1 Human drawings in the WZT as a measure of Alexithymia (I)

Study I is based on a convenience sample, and while there is no specific reason to assume that the results of the study are not generalizable and would not apply to the general population, this is of course always possible when a convenience sample is used. Roughly half of the subjects had mild depression based on the depression inventory, a proportion greater that was than in the general population. This may or may not have affected the results. The criterion test that was used in the study, the Toronto Alexithymia Scale is a valid test (Bagby *et al.* 1986), but not a perfect measure of the alexithymia construct. A third measure of alexithymia such as the Rorschach Alexithymia scale (Porcelli & Mihura 2010) also based on a projective test might be used in an optimal study design. Another shortcoming of Study I was that, due to the small sample, age was not controlled for in the analysis of the relationship between alexithymia and human drawings. Age had a negative correlation with human drawings and a positive correlation with alexithymia. Thus, controlling for age, the correlation between human drawings and TAS score would likely be lower than the -0.33 figure.

6.5.2 Cross-national variation in test profiles (II, III)

While the number of studies of differences in national IQs is in the hundreds studies of cross-national differences in test profiles are few. The concept of national IQ profile presented in study III may be considered to be a hypothesis at this moment. The number of nations and test versions analysed in studies II and III is quite limited, and the conclusions warrant some caution. The discussion of the cultural factors underlying the cross-Atlantic differences in speed test is obviously speculative by nature. Additional more detailed studies involving other nations and test versions are needed to analyse whether these hypotheses are valid or not.

6.5.3 The cross-generational validity of vocabulary test norms (IV)

The results and conclusions of study IV are based on English-language tests only. Of course, we may presume that the correlation between word popularity and difficulty as test items is negative for other languages as well. The Google books database includes books in other major languages. Studies, for example, of French

and German vocabulary tests should be initiated so as to validate the hypotheses proposed in study IV. In such studies, the method of estimating "exposure to word frequency" by Dorius *et al.* (2014) should be used instead of the rudimentary method employed in study IV. The WAIS and WISC manuals do not report means and standard deviations for single test items, and therefore, only an ordinal scale of item difficulty is available. Other tests that report more detailed item-level information might be used to improve the accuracy of analysis.

6.5.4 Validity of the CIS validity scale (V)

While the results of the study show that careless responding is more frequent among psychiatric patients than non-patients, the data analysed in this study do not actually indicate the reason underlying this phenomenon. On the basis of previous studies, we know that the lack of concentration and cognitive skills is more prevalent in psychiatric patients and that these factors probably play a part in careless responding. However, a more conclusive study would require actual measuring of these factors. Thus, in study V, the conclusions are partly based on indirect evidence. A minor shortcoming of this study is that a modified 12 item version of the CIS scale was used instead of the original 13 item version.

7 General Discussion

The results of the study underline the contextual, relative nature of test scores, which calls for continued efforts in developing valid test norms. The observations support the suggestion by Vanhanen and Laulumaa (2011) that the publication of new Finnish editions of psychological tests involving costly translation and test adaptation work should perhaps be avoided and that we should instead focus on collecting valid norms for the existing tests.

The nature of psychological testing is easily misunderstood. Numerical values based on the application of a measuring instrument somehow seem more reliable, professional or scientific than data based on, for example, interviews. The Rorschach was nicknamed the "x-ray" of the mind, and unfortunately such analogies have strong appeal. Psychological tests are mistakenly seen as analogous to weight scales or blood pressure monitors (Anastasi 1985).

Modern management theories emphasize the importance of measurement to monitor processes. In the health care and rehabilitation setting, one application of this ideology has been the increasing use of questionnaires in order to estimate the impact of health interventions on psychological well-being. For example, clients of Finnish rehabilitation centres take the Beck Depression inventory at the beginning and end of their rehabilitation course (Kela 2015). While a fall or rise in the score of the valid and reliable BDI inventory on the average means an increase or decrease in the level of depression, the possibility of false positive and false negative results in the assessment of individuals and small groups is easily overlooked (Roivainen 2008).

In the United States, there has been a long debate among psychologists, lawyers and politicians on the validity of IQ scores in capital punishment cases (Young 2012). For example, the state of Florida rigidly requires an IQ of 70 or below to demonstrate mental retardation, with no allowance for the test's margin of error. Some states have been reluctant to acknowledge the Flynn effect, and regard the test norms reported in test manuals as "official". In a recent case, a death row petitioner, who had scored 71 on the WAIS III was exempted from the penalty by a Supreme Court decision (5 votes to 4) which concluded that, since IQ scores contain a margin of error, states must generally consider other factors in determining intellectual disability as an exemption from the death penalty (Huffington Post 2014).

Cross-national studies of intellectual skills continue to include verbal tests in the analyses. For example, Armstrong and others (2014) report that the non-verbal

IQ of the Sami people is higher and their verbal IQ is lower than that of Finns. Georgas and colleagues (2003) compared WISC III scores from 12 national standardization studies, and based their analysis on the non-adapted international items of the verbal subtests. It was assumed that these items are, on average, equally difficult in different countries. However, as the results of study IV show, the difficulty of test items in vocabulary subtests changes over generations speaking the same language, and the comparison of verbal skills across borders is probably highly unreliable. It is impossible to answer the question of the type: do Finnish or Swedish children have a richer average vocabulary?

The paradigm of cross-national and cross-generational comparison of test norms has great potential for the study of intelligence and cognitive processes in general. For example, the effects of schooling, language and cultural factors on cognitive skills can be analysed. In a recent review article, Mingroni (2014) lists the use of subtest profile data as one of six main methods in future research on intelligence. However, the use of test norm data to rank countries or ethnic groups to prove one's political views, as well the censorship of such data based on political correctness, seem equally unfruitful pursuits.

The use of "Big data", large electronic databases, has radically increased in psychological research in recent years. Psychological theories have traditionally been based on experimentation with rats and undergraduate students. Obviously, conclusions based on data from large samples such as those used in studies IV and V, the Northern Finland birth cohort and the Google books database, are less affected by sampling-related problems.

Efforts to develop an empirically valid interpretation method for the Wartegg test should be continued. In cases where the respondent has poor introspective skills, performance-based tests are potentially more appropriate than questionnaires. Compared to the multitude of self-report test and questionnaires, the present-day selection of valid projective tests is small. New tests and valid interpretation methods for the old tests are needed.

8 Conclusions

1. WZT drawings may correlate with personality constructs such as alexithymia. Efforts to develop an empirically valid interpretation method for the WZT should be continued.
2. The Finnish WAIS III norms are distorted in the younger age groups due to a non-representative sample.
3. There are stable cross-national differences in WAIS subtest norms that cannot be explained by differences in the general factor of intelligence. Educational, cultural and linguistic factors may underlie differences in national IQ profiles.
4. The difficulty of vocabulary test items depends on the frequency of the use of the words and varies across birth cohorts. Test norms become outdated over time, and the magnitude of the Flynn effect may vary across subtests.
5. Data on cross-national and cross-generational differences in IQ profiles may aid the development of valid of test norms in small countries with few resources for large standardization studies.
6. The Chapman Infrequency Scale cut-off score for excluding careless respondents should be flexible according to the sample studied. Validity scales may measure different things in different populations.

References

Agranovich AV, & Puente AE (2007) Do Russian and American normal adults perform similarly on neuropsychological tests? Preliminary findings on the relationship between culture and test performance. Archives of Clinical Neuropsychology 22: 273–282.

Anastasi A & Urbina S (1997) Psychological testing (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Anastasi A (1985) Psychological testing: Basic concepts and common misconceptions. In Rogers AM.; Scheirer C J. (eds). The G. Stanley Hall lecture series, 5. Washington, DC, US: American Psychological Association: 87–120.

Armitage SG & Pearl D (1957) Unsuccessful differential diagnosis from the Rorschach. Journal of Consulting Psychology 21: 479–484.

Armstrong E, Woodley M & Lynn R (2014) Cognitive abilities among the Sami population. Intelligence 46: 35–39.

Backström M. & Björklund F (2013). Social desirability in personality inventories: Symptoms, diagnosis and prescribed cure. Scandinavian Journal of Psychology 54: 152–159.

Baer RA, Ballenger J, & Kroll LS (1998) Detection of underreporting on the MMPI-A in clinical and community samples. Journal of Personality Assessment 71: 98–113.

Baer RA, Kroll LS., Rinaldo J & Ballenger J (1999) Detecting and discriminating between random responding and overreporting on the MMPI-A. Journal of Personality Assessment 72: 308–320.

Bagby RM, Taylor GJ & Ryan D (1986) Toronto Alexithymia Scale: relationship with personality and psychopathology measures. Psychotherapy & Psychosomatics 45: 207–215.

Beck AT, Ward CH, Mendelson M, Mock J & Erbaugh J (1961) An inventory for measuring depression. Archives of General Psychiatry 4: 561–71.

Beck AT & Steer, RA (1993) Beck Anxiety Inventory Manual. San Antonio: Harcourt Brace and Company.

Beck SJ (1952) Rorschach's test. Vol. 3. Advances in interpretation. New York: Grune & Stratton.

Beck SJ (1959) Review of the Rorschach inkblot test. In O.K. Buros (ed.). The fifth mental measurements handbook. Highland Park, NJ:Gryphon Press.

Ben-Porath YS (2012) Interpreting the MMPI-2-RF. Minneapolis, MN: University of Minnesota Press.

Binet A & Simon T (1905) Méthodes nouvelles pour le diagnostic du niveau intellectual des anormaux. L'Année psychologique 11: 191–336.

Blacker D & Endicott J (2000) Psychometric properties: concepts of reliability and validity. In: Handbook of psychological measures. American Psychiatric Association, Washington, DC.

Bonogofsky AN (2007) Self-Report Measures of Psychopathic and Schizotypal Personality Characteristics: A Confirmatory Factor Analysis of Hypothetical Antisocial Behavior and Hypothetical Psychosis-Proneness in a College Sample. MA thesis, Leland Stanford Junior University (Palo Alto, California).

Borsboom D, Mellenbergh GJ & Van Heerden J (2003) The theoretical status of latent variables. Psychological Review 110: 203–219

Boring EG (1950) A History of Experimental Psychology (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Bornstein RF (2009) Heisenberg, Kandinsky, and the heteromethod convergence problem: Lessons from within and beyond psychology. Journal of Personality Assessment 91: 1–8.

Burke HR (1985) Raven's Progressive Matrices (1938). More on norms, reliability, and validity. Journal of Clinical Psychology, 41: 231–235.

Carroll JB (2013) Human cognitive abilities: A survey of factor-analytic studies. Cambridge: Cambridge University Press.

Cattell McKeen J (1890). Mental tests and measurements. Mind, 15: 373–381.

Cattell RB (1949) Culture Free Intelligence Test, Scale 1, Handbook. Champaign: Institute of Personality and Ability.

Cattell RB, Cattell AK, & Cattell HEP (1993) 16PF Fifth Edition Questionnaire. Champaign, IL: Institute for Personality and Ability Testing.

Chapman LJ & Chapman JP (1986) Infrequency scale for personality measures. Available from TR Kwapil, Department of Psychology, University of North Carolina at Greensboro, P.O. Box 26164, Greensboro, NC 27402.

Chapman JP, Chapman LJ, & Kwapil TR (1995) Scales for the measurement of schizotypy. In A Raine, T Lencz, & SA Mednick (Eds.), Schizotypal personality. New York: Cambridge University Press.

Chapman LJ, Chapman JP, & Raulin ML (1976). Scales for physical and social anhedonia. Journal of Abnormal Psychology 85: 374–382.

Chapman LJ, Chapman JP, & Raulin ML (1978) Body-image aberration in schizophrenia. Journal of Abnormal Psychology 87: 399–407.

Cheramie GM, Stafford ME, Boysen C, Moore J & Prade C (2012) Relationship between the Wechsler Adult Intelligence Scale – Fourth Edition (WAIS-IV) and Woodcock-Johnson-III normative Update (NU): Tests of Cognitive Abilities (WJ-III COG). Journal of Education and Human Development 5: 1–7.

Cloninger CR (1994) The temperament and character inventory (TCI): A guide to its development and use. St. Louis, MO: Center for Psychobiology of Personality, Washington University.

Cronbach L (1949) Statistical methods applied to Rorschach scores: A Review. Psychological Bulletin 46: 393–429.

Curzan A (2009) Historical corpus linguistics and evidence of language change. In A Luedeling & M Kytö (Eds.), Corpus linguistics. Berlin, Germany: Gruyter:1091–1102.

Dorius S, Alwin DF & Pacheco J (2014) Cohort Differences in Verbal Ability: Testing the Word Obsolescence Hypothesis. Paper presented at the meeting of the American Sociological Association, August 17, 2014.
Dutton E & Kierkegaard E (2014) Fluid g in Scandinavia and Finland: Comparing results from PISA Creative Problem Solving and the WAIS IV matrices subtest. Open Differential Psychology.
Esquirol E (1838). Des maladies mentales considérées sous le rapport médical, hygiénique, et médico-légal. Paris: Bailliére.
Eysenck HJ (1959) Review of the Rorschach inkblot test. Teoksessa O.K.Buros (Ed.). The fifth mental measurements handbook. Highland Park, NJ:Gryphon Press: 276–278.
Exner JE (1974) Rorschach: A Comprehensive System. Vol.1. New York: Wiley.
Exner JE (1993) The Rorschach: A comprehensive system: Vol. I. Basic foundations (3rd ed.). New York: Wiley.
Fineman S (1977) The achievement motive construct and its measurement: Where are we now? British Journal of Psychology 68: 1–22.
Flynn JR (1984) The Mean IQ of Americans: Massive Gains 1932 to 1978. Psychological Bulletin 95: 29–51.
Flynn JR (2010) Problems with IQ gains: The huge vocabulary gap. Journal of Psychoeducational Assessment 28: 412–433.
Flynn JR (2012) Are we getting smarter? Rising IQ in the 21st century. Cambridge, UK: Cambridge University Press.
Fonseca-Pedrero E, Paíno-Piñeiro M, Lemos-Giráldez S, García-Cueto E, Villazón-García U, & Muñiz, J (2009) Psychometric properties of the Perceptual Aberration Scale and the Magical Ideation Scale in Spanish college students. International Journal of Clinical and Health Psychology 9: 299–312.
Galton F (1883) Inquiries into Human Faculty and Its Development. London: J.M. Dent & Co.
Gardner H (1983) Frames of Mind: The Theory of Multiple Intelligences. New York:: Basic Books.
Gardziella M (1985) Wartegg-piirustustesti: Käsikirja [The Wartegg Drawing Test: A handbook]. Jyvaskyla, Finland: Psykologien Kustannus Oy.
Georgas J, Weiss LR, Van de Vijver FJR, & Saklofske DJ (2003) Culture and children's intelligence: Cross-cultural analysis of the WISC III. New York, NY: Academic Press.
Google (2015) http://scholar.google.fi/ Retrieved 18.2.2015.
Gregory RJ (2013) Psychological testing: history, principles and applications. NJ:Pearson.
Groth-Marnat, G (2003) Handbook of psychological assessment. Hoboken, NJ: John Wiley.
GSS (2009) General Social Survey. Cumulative file for Wordsum 1972–2006. Chicago, IL: National Opinion Research Center.
Guilford JP (1948) Some lessons from aviation psychology. American Psychologist 3: 3–11.
Haapea M, Miettunen J, Läärä E, Joukamaa M, Järvelin MR, Isohanni M, & Veijola JM (2008) Non-participation in a field survey with respect to psychiatric disorders. Scandinavian Journal of Public Health 36: 728–736.

Hathaway SR & McKinley JC (1940) A Multiphasic Personality Schedule (Minnesota). I. Construction of the schedule. Journal of Psychology 10: 249–254.
Heiskari P (2010) Kommentti Eka Roivaisen artikkeliin suomalaisten WAIS-III normien arvioinnista[A response to Eka Roivainen's article on Finnish WAIS III norms]. Psykologia 45: 90–92.
Hershen M (2004) Comprehensive Handbook of Psychological Assessment. New Jersey: Wiley.
Hertz M (1959) The use and misuse of the Rorschach method. I. Variations in the Rorschach procedure. Journal of Projective Techniques 23: 33–48.
Holland JL (1985) Making vocational choices. A theory of vocational personalities and work environments. New Jersey: Prentice Hall.
Holtzman W & Sells SB (1954) Prediction of flying success by clinical analysis of test protocols. Journal of Abnormal and Social Psychology 49: 485–490.
Hoosain R, & Salili F (1988) Language differences, working memory and mathematical ability. In MM Gruneberg, PE Morris, & RN Sykes (Eds.), Practical aspects of memory: Current research and issues (Vol. II). Chichester, UK: John Wiley: 512–517.
Huffington Post (2014) Supreme Court Rules In Favor Of Death Row Inmates Who Have Low IQs http://www.huffingtonpost.com/2014/05/27 (Retrieved 17.02.2015)
Ilmonen K (1996) Tekniikka, kaiken perusta. Yleisradion historia 1926–96. Osa 3 [History of the Finnish Broadcasting Corporation 1926–1996, volume 3]. Helsinki, Finland: Yleisradio.
Jackson DN (1974) Personality Research Form: Manual. Port Huron, MI:Research Psychologists Press.
Joukamaa M, Miettunen J, Kokkonen P, Koskinen M, Julkunen J, Kauhanen J & Jokelainen J (2001). Psychometric properties of the Finnish 20-item Toronto Alexithymia Scale. Nordic Journal of Psychiatry 55: 123–7.
Jääskeläinen E & Miettunen J (2011) Psykiatriset arviointiasteikot kliinisessä työssä [Psychiatric rating scales in clinical work]. Duodecim 127: 1719–25.
Kela (2015) Standardit. [rehabilitation standards/The social insurance institution of Finland] http://www.kela.fi/standardit. Retrieved 18.2.2015.
Klopfer B & Davidson HH (1962) The Rorschach Technique. An Introductory manual. Orlando: Harcourt Brace.
Koistinen P (2005) Arvostelijan pitää tuntea aiheensa: projektiiviset testit. Psykologi 2005, 4: 28–29.
Kratzmeier H, & Horn, R (1980) RAVEN–Matritzen-Test Advanced Progressive Matrices — Manual, Deutsche Bearbeitung. Weinheim: Beltz Test.
Kuuskorpi T (2012) Psykologisten testien käyttö suomessa. [Psychological test usage in Finland]. PhD thesis, University of Turku.
Legg S & Hutter M (2007) A collection of definitions of intelligence. Frontiers in Artificial Intelligence and Applications 157: 17–24.
Lim J & Butcher J (1996). Detection of faking on the MMPI-2: Differentiation among faking-bad, denial, and claiming extreme virtue. Journal of Personality Assessment 67: 1–25.

Longman RS, Saklofske DH, & Fung TS (2007) WAIS-III percentile scores by education and sex for U.S. and Canadian populations. Assessment 14: 426–432.

Lynn R & Vanhanen T (2002) IQ and the Wealth of Nations. Westport, CT: Praeger.

Lynn R & Vanhanen T (2012) National IQ's: A review of their educational, cognitive, economic, political, demographic, sociological, epidemiological, geographic and climactic correlates. Intelligence 40: 226–234.

Lynn R & Vanhanen T (2013) Intelligence: A Unifying Concept for the Social Sciences. Ulster Institute for Social Research.

Maas HLJ van der, Dolan CV, Grasman RPPP, Wicherts JM, Huizenga HM, & Raijmakers MEJ (2006) A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. Psychological Review 113: 842–861.

Maas HLJ, Kan K-J & Borsboom D (2014) Intelligence Is What the Intelligence Test Measures. Seriously. Journal of Intelligence 2: 12–15.

Machover K (1949) Personality projection in the drawings of the human figure. Springfield: Thomas.

Mattlar CE, Lindholm T, Haasiosalo A & Vesala P (1991) Interrater agreement when assessing alexithymia using the Drawing Completion Test (Wartegg Zeichentest). Psychotherapy & Psychosomatics 56: 98–101.

Meade AW & Craig B (2012) Identifying careless responses in survey data. Psychological Methods 17: 437–455.

Meehl PE (1992). Needs (Murray, 1938) and state-variables (Skinner, 1938). Psychological reports 70: 407–450.

Meisenberg G (2015) Verbal ability as a predictor of political preferences in the United States, 1974–2012. Intelligence 50: 135–143.

Merckelbach H, Giesbrecht T, Jelicic M, & Smeets T (2010) The problem of careless respondents in surveys. Tijdschrift voor Psychiatrie 52: 663–669.

Meyer GJ (2004) The reliability and validity of the Rorschach and Thematic Apperception Test (TAT) compared with other psychological and medical procedures: An analysis of systematically gathered evidence. In MJ Hilsenroth & DL. Segal (Eds.), Comprehensive Handbook of Psychological Assessment: Vol. 2. Personality assessment. Hoboken, NJ: Wiley: 315–342.

Michel JB, Shen YK, Aiden AP, Veres A, Gray MK, The Google Books Team, & Aiden EL (2011). Quantitative analysis of culture using millions of digitized books. Science 331: 176–182.

Miettunen J, Kantojärvi L, Ekelund J, Veijola J, Karvonen JT, Peltonen L, Järvelin M-R, Freimer N, Lichtermann D, Joukamaa M. (2004) A large population cohort provides normative data for investigation of temperament. Acta Psychiatrica Scandinavica 110:150–7.

Miettunen J, Veijola J, Isohanni M, Paunio T, Freimer N, Jääskeläinen E, Taanila A, Ekelund J, Järvelin M-R, Peltonen L, Joukamaa M, Lichtermann D (2011) Identifying schizophrenia and other psychoses with psychological scales in the general population. Journal of Nervous and Mental Disorders 199: 230–8.

Mingroni M (2014) Future Efforts in Flynn Effect Research: Balancing Reductionism with Holism. Journal of Intelligence, 2:, 122–155.
Morgan GA, Gliner JA & Harmon RJ (2001) Measurement validity. Journal of the American Academy of Child & Adolescent Psychiatry 40: 729–731.
Murray HA (1943) The Thematic Apperception Test Manual. Cambridge, MA: The Harvard University Press.
Must O & Must A (2013). Changes in test-taking patterns over time. Intelligence 41: 780–790.
Naveh-Benjamin M & Ayres TJ (1986) Digit span, reading rate and linguistic relativity. Quarterly Journal of Experimental Psychology 38: 739–751.
Nicolas S, Andrieu B, Croizet J-C, Sanitioso RB, & Burman JT (2013). Sick? Or slow? On the origins of intelligence as a psychological object. Intelligence 41: 699–711.
Niitamo P (1999) "Surface" and "Depth" in human personality: Relations between explicit and implicit motives. PhD thesis, University of Helsinki. People and Work Research Reports, 27. Finnish Institute of Occupational Health.
OECD (2015) Programme for International Student Assessment, homepage. http:/www.pisa.oecd.org. Cited 01.06.2015.
Nummenmaa L & Hyönä J (2005) Mitä sinä näet tässä kuvassa? Voiko projektiivisiin testeihin luottaa. Psykologi 3: 14–16.
Peltier BD & Walsh JA (1990) An investigation into response bias in the Chapman scales. Educational and Psychological Measurement 50: 803–815.
Pervin LA, Cervone D & John O (2005) Personality: theory and research. NY: Wiley.
Petzold H (2000): Warteggs Zeichentest WZT [Wartegg's drawing test WZT], in Bernhardt, H & Lockot, R (eds.): Mit Ohne Freud. Zur Geschichte der Psychoanalyse in Ostdeutschland [With Without Freud. The history of psychoanalysis in Eastern Germany. Giessen: Psychosozial-Verlag: 128–131.
Piedmont RL, McCrae RR, Riemann R, & Angleitner A (2000) On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. Journal of Personality and Social Psychology 78: 582–593.
Piotrowski Z (1957) Perceptanalysis: A fundamentally reworked, expanded and systematized Rorschach method. New York: MacMillan.
PKOY (2007) PK5-persoonallisuustestin käsikirja [PK-5test manual] Helsinki: Psykologien Kustannus Oy.
Porcelli P & Mihura J L (2010) Assessment of alexithymia with the Rorschach Comprehensive System:The Rorschach Alexithymia Scale (RAS). Journal of Personality Assessment 92: 128–136.
Putzke JD, Williams MA, Daniel FJ, & Boll TJ (1999) The utility of K-correction to adjust for a defensive response set on the MMPI. Assessment 6: 61–70.
Raitasalo R (1995) RBDI Mielialakysely. Suomen oloihin Beckin lyhyen depressiokyselyn pohjalta kehitetty masennusoireilun ja itsetunnon kysely. Sosiaali- ja terveysturvan tutkimuksia 86. Helsinki: Kela.
Raven JC, Court JH, & Raven J (1996) Manual for Raven's Standard Progressive Matrices. Oxford: Oxford Psychologists Press.

Rindermann H (2007) The g-Factor of International Cognitive Ability Comparisons: The Homogeneity of Results in PISA, TIMSS, PIRLS and IQ-Tests Across Nations. European Journal of Personality 21: 667–706.

Roid GH (2003) Stanford-Binet Intelligence Scales, Fifth Edition. Itasca, IL: Riverside Publishing.

Roivainen E (1997) Onko Wartegg-piirustustesti validi? [Is the Wartegg test valid?]. Unpublished manuscript, Kemijärven työvoimatoimisto.

Roivainen E (2006). Ehrig Wartegg ja Wartegg-testin varhaisvaiheet [Ehrig Wartegg and the early history of Wartegg's Drawing Test]. Psykologia 41: 260–268.

Roivainen E (2008) Beckin depressiokyselyn tulkinta [Interpretation of BDI scores]. Duodecim 124: 2467 –2470.

Roivainen E (2009). A brief history of the Wartegg Drawing Test. Gestalt Theory 31: 55–71.

Roivainen E (2010a) Suomalaisten WAIS III normien arviointia [An examination of Finnish WAIS III norms]. Psykologia 4: 86–89.

Roivainen E (2011) Gender differences in processing speed: a review of recent research. Learning and Individual Differences 21: 145–149.

Rorschach H (1921) Psychodiagnostik. Tafeln. Bern: Hans Huber; 1921.

Rosselli M, & Ardila A (2003). The impact of culture and education on nonverbal neuropsychological measurements: A critical review. Brain and Cognition 52: 326–333.

Rotter JB., & Rafferty JE (1950) The Rotter Incomplete Sentences Blank manual: College form. New York: Psychological Corporation.

Schaie KW (2005) What can we learn from longitudinal studies of adult development? Research in Human Development 2: 133–158.

Schinka JA, Kinder BN, & Kremer T (1997) Research validity scales for the NEO-PI-R: Development and initial validation. Journal of Personality Assessment 68: 127–138.

Shaffer TW, Erdberg P & Haroian J (1999) Current nonpatient data for the Rorschach, WAIS-R and MMPI 2. Journal of Personality Assessment 73: 305–316.

Sellbom M & Bagby RM (2008) Validity of the MMPI-2-RF (Restructured Form) L-r and K-r scales in detecting underreporting in clinical and nonclinical samples. Psychological Assessment 20: 370–376.

Soilevuo-Grönneröd J & Grönneröd C (2012) The Wartegg Zeichen Test: A Literature Overview and a Meta-Analysis of Reliability and Validity. Psychological Assessment 24: 476–489.

Spielberger CD, Gorsuch RL, Lushene R, Vagg PR & Jacobs G (1977) Manual for the State-Trait Anxiety Inventory (Form Y). PaloAlto, CA: Consulting Psychologists Press.

Stern W (1912) Die psychologischen Methoden der Intelligenzprüfung und deren Anwendung an Schulkindern. Leipzig: Verlag von Johann Ambrosius Barth.

Sternberg RJ, & Grigorenko EL (2000) Teaching for successful intelligence. Arlington Heights, IL:Skylight.

Takala M (1964) Studies of the Wartegg drawing completion test: Studies of psychomotor personality tests II. Annales Academiae Scientiarum Fennicae, Serie B, 131: 1–112. Helsinki, Finland:Suomalainen Tiedeakatemia.

Takala M, & Hakkarainen M (1953) Ueber Faktorenstruktur und Validität des Wartegg-Zeichen-testes [Factor analysis and validity of the Wartegg Drawing test]. Annales Academiae Scientiarum Fennicae, SerieB: 81–95.

Tamminen S & Lindeman M (2000). Wartegg—luotettava persoonallisuustesti.

vai maagista ajattelua? [The Wartegg—A valid personality test of magical thinking?]. Psykologia 35: 325–331.

Taylor GJ, Ryan D & Bagby RM (1986) Toward the development of a new self-report alexithymia scale. Psychotherapy and Psychosomatics 44: 191–199.

Terman LM (1916) The measurement of intelligence: An explanation of and a complete guide forthe use of the Stanford revision and extension of the Binet-Simon Intelligence Scale. Boston: Houghton Mifflin.

TV History (2011) Television history, the first 25 years. Retrieved from www.tvhistory.tv

Työministeriö (1995) AVO 9 kykytestistö. Helsinki:Psykologien Kustannus OY.

Vanhanen M & Laulumaa R (2011) WAIS-R ja WAIS-III testistöjen vertailututkimus: normiongelma ja ratkaisuehdotus. Psykologia 46: 346–53.

Vanhanen M (2008) WAIS-R ja WAIS III älykkyystestien tulosten vastaavuus suomalaisilla. [Comparing Finnish WAIS-R and WAIS III norms]. Paper presented at the Psykologia 2008 conference, Helsinki.

Wartegg E (1939) Gestaltung und Charakter [Formation of gestalts and personality]. Zeitschrift für Angewandte Psychologie und Charakterkunde 84, Beiheft 2.

Wechsler D (1949) Wechsler intelligence scale for children. New York, NY: Psychological Corporation.

Wechsler D (1955) Wechsler adult intelligence scale: Manual. New York, NY: Psychological Corporation.

Wechsler D (1971). Wechslerin aikuisten älykkyysasteikko. Helsinki, Finland: Psykologien Kustannus OY.

Wechsler D (1974) Manual for the Wechsler Intelligence Scale for Children—Revised. New York: Psychological Corporation.

Wechsler D (1981) Wechsler adult intelligence scale–Revised: Manual. New York, NY: Psychological Corporation.

Wechsler D (1991) The Wechsler intelligence scale for children—Third edition. San Antonio, TX: Psychological Corporation.

Wechsler D (1992) Wechslerin aikuisten älykkyysasteikko. WAIS-R käsikirja [WAIS-R manual]. Helsinki, Finland: Psykologien Kustannus OY.

Wechsler D (1997) Wechsler adult intelligence scale–Third edition: Manual. San Antonio, TX:Psychological Corporation.

Wechsler D (1999a) WAIS-III: Escala de inteligencia de Wechsler para Adultos. Madrid: TEA.

Wechsler D (2000) Echelle d'intelligence de Wechsler pour adultes (WAISIII). Paris: ECPA.

Wechsler D (2005) WAIS III Käsikirja[WAIS III manual]. Helsinki: Psykologien kustannus OY.

Wechsler D (2006) Wechsler Intelligenztest fuer Erwachsene WIE III. Frankfurt: Pearson.

Wechsler D (2008) Wechsler Adult Intelligence Scale IV. San Antonio: Pearson.

Wechsler D (2012) WAIS IV esitys- ja pisteytyskäsikirja [WAIS IV manual]. Helsinki: Hogrefe Psykologien Kustannus Oy.

Wood JM, Nezworski MT, Garb HN & Lilienfeld SO (2001) The misperception of psychopathology. Problems with the norms of the Comprehensive System for the Rorschach. Clinical Psychology: Science and Practice 8: 350–373.

Wood JM, Nezworski MT, Garb HN & Lilienfeld SO (2003) What's Wrong With The Rorschach: Science Confronts the Controversial Inkblot Test. Wiley & Sons.

Woodcock RW, McGrew KS, & Mather N (2001) Woodcock-Johnson III. Itasca, IL: Riverside.

Woodley of Menie MA, Fernandes HBF, Figueredo AJ & Meisenberg G (2015) By their words ye shall know them: Evidence of genetic selection against general intelligence and concurrent environmental enrichment in vocabulary usage since the mid 19th century. Frontiers in Psychology 6:361.

Woodworth RS (1919) Examination of emotional fitness for warfare. Psychological Bulletin 15: 59–60.

Young G (2012) A More Intelligent and Just Atkins: Adjusting for the Flynn Effect Vanderbildt Law Review 65:615.

List of original publications

This thesis is based on the following publications, which are referred to throughout the text by their Roman numerals:

I Roivainen E & Ruuska P (2005) The use of projective drawings to assess alexithymia: the validity of the Wartegg test. European Journal of Psychological Assessment 21: 199-201.

II Roivainen E (2010) European and American WAIS III norms: Cross-national differences in performance subtest scores. Intelligence 38: 187–191.

III Roivainen E (2013) Are cross-national differences in IQ profiles stable? A comparison of Finnish and US WAIS norms. International Journal of Testing 13: 140–151.

IV Roivainen E (2014) Changes in word usage frequency may hamper intergenerational comparisons of vocabulary skills: An Ngram analysis of WAIS, WISC and Wordsum test items. Journal of Psychoeducational Assessment 32: 83–87.

V Roivainen E, Veijola J & Miettunen J (2015) Careless responses in survey data and the validity of a screening instrument. Nordic psychology. DOI: 10.1080/19012276.2015.1071202..

Reprinted with permissions from Hogrefe (I), Elsevier (II), Taylor and Francis (III, V) and SAGE (IV). Original publications are not included in the electronic version of the dissertation.

1297. Aatsinki, Sanna-Mari (2015) Regulation of hepatic glucose homeostasis and Cytochrome P450 enzymes by energy-sensing coactivator PGC-1?

1298. Rissanen, Ina (2015) Nervous system medications and suicidal ideation and behaviour : the Northern Finland Birth Cohort 1966

1299. Puurunen, Johanna (2015) Androgen secretion and cardiovascular risk factors in women with and without PCOS : studies on age-related changes and medical intervention

1300. Pakanen, Lasse (2015) Thrombomodulin and catecholamines as post-mortem indicators of hypothermia

1301. Mäkelä, Mailis (2015) Hoitoon ja kohteluun kohdistuva tyytymättömyys : potilaslain mukaiset muistutukset

1302. Nordström, Tanja (2015) Predisposing factors and consequences of adolescent ADHD and DBD : a longitudinal study in the Northern Finland Birth Cohort 1986

1303. Tanner, Tarja (2015) Healthy young adults' oral health and associated factors : cross-sectional epidemiological study

1304. Ijäs, Hilkka (2015) Gestational diabetes : metformin treatment, maternal overweight and long-term outcome

1305. Leskinen, Riitta (2015) Late-life functional capacity and health among Finnish war veterans : Veteran Project 1992 and 2004 surveys

1306. Kujala, Tiia (2015) Acute otitis media in young children : randomized controlled trials of antimicrobial treatment, prevention and quality of life

1307. Kämppi, Antti (2015) Identifying dental restorative treatment need in healthy young adults at individual and population level

1308. Myllymäki, Satu-Marja (2015) Specific roles of epithelial integrins in chemical and physical sensing of the extracellular matrix to regulate cell shape and polarity

1309. Antonoglou, Georgios (2015) Vitamin D and periodontal infection

1310. Valtokari, Maria (2015) Hoitoon pääsyn moniulotteisuus erikoissairaanhoidossa

1311. Toljamo, Päivi (2015) Dual-energy digital radiography in the assessment of bone characteristics

1312. Kallio-Pulkkinen, Soili (2015) Effect of display type and room illuminance in viewing digital dental radiography : display performance in panoramic and intraoral radiography

Book orders:
Granum: Virtual book store
http://granum.uta.fi/granum/

UNIVERSITY OF OULU P.O. Box 8000 FI-90014 UNIVERSITY OF OULU FINLAND

ACTA UNIVERSITATIS OULUENSIS

UNIVERSITY of OULU
OULUN YLIOPISTO

ISBN 978-952-62-0942-5 (Paperback)
ISBN 978-952-62-0943-2 (PDF)
ISSN 0355-3221 (Print)
ISSN 1796-2234 (Online)