Establishing the protocol validity of an electronic standardised measuring instrument

Sebastiaan Rothmann, B.Sc (Honours)

Dissertation submitted in partial fulfilment of the requirements for the degree Master of Science in Human Resource Management at the Potchefstroom Campus of the North-West University

Supervisors: Prof. L.T.B. Jackson
Prof. H.S. Steyn

Potchefstroom
2009

COMMENTS

The reader is reminded of the following:

- The editorial style as well as the references referred to in this dissertation follow the format prescribed by the Publication Manual (5th edition) of the American Psychological Association (APA). This practice is in line with the policy of the Programme in Human Resources Sciences of the North-West University (Potchefstroom) to use APA style in all scientific documents as from January 1999.

- The dissertation is submitted in the form of two research articles. The editorial style specified by the South African Journal of Industrial Psychology (which agrees largely with the APA style) is used, but the APA guidelines were followed in constructing tables.

ACKNOWLEDGEMENTS

In writing this dissertation, I was fortunate to have the advice and assistance of many people. I would hereby like to thank the following key individuals and organisations which assisted with and contributed to the completion of this dissertation:

- Prof. Leon Jackson for making this dream a reality.
- Prof. Faans Steyn for the brilliant statistical inputs.
- Prof. Ian Rothmann, the genius researcher for all the help, support, coaching and teaching me that research is structured art.
- Prof. Karina Mostert for all the advice, inputs and continuous support. Also thanks to her family, Frans Mostert for the psychological contract and little Christian Mostert for his roaring inputs.
- Prof. Jaco Pienaar for all the support and teaching me that the only way out is through.
- (Dr.) Ina (Rothmann) for teaching me how to fly.
- Afriforte (Pty) Ltd. for making the data available and supporting me in this endeavour.
- Willie Cloete and Mariëtte Postma for the language editing.
- The love of my life, Zinnley, thank you for being you. You are everything to me.
- Thinus Liebenberg for keeping up with my thousands of (sometimes incoherent) ideas and the rest of the Afriforte family, Lelani Brand-Labuschagne for the encouragement, Rika Barnard for keeping me at bay, Martin Noome and Douw Cronje for the hours of brainstorming.
- My dearest family, Ouma Susan (for the rooibos and love), Mikkie, Gerrie (Myburgh), Suzelle, Gerhard (Myburgh), Paul, Surika, Renier, Sandri, Gerhard (Rothmann), Naomi, Naomi Jr., Antonie and Noemie, thank you for shaping me through the years, you are all very close to my heart.
- My dearest friends, Paulette (my other mom), Leon (for all the schemes and dreams), Hendrè, Chucky, Ruan, Hester (Britney), Jeankia, Elzuette, Jannie, Leandi, Henry, Leandri, Vis, Glen, Jardus, Kerry-ann, Rudo and Jani for endeavouring with me on this journey called life.

TABLE OF CONTENTS

		Page
List of Figures		v
List of Tables		vi
Abstract		vii
Opsomming		ix
CHAPTER 1: INTRODUCTION		1
1.1	Problem statement	1
1.2	Research objectives	5
1.2.1	General objective	5
1.2.2	Specific objectives	5
1.3	Research method	5
1.3.1	Research design	6
1.3.2	Participants and procedure	6
1.3.3	Measuring instrument	6
1.3.4	Statistical analysis	7
1.4	Overview of chapters	9
1.5	Chapter Summary	9
	References	10
CHAPTER 2: RESEARCH ARTICLE 1		13
CHAPTER 3: RESEARCH ARTICLE 2		44
CHAPTER 4: CONCLUSIONS, LIMITATIONS AND RECOMMENDATIONS		70
4.1	Conclusions	70
4.2	Limitations of the present study	75
4.3	Recommendations	75
4.3.1	Recommendations for the organisation	75
4.3.2	Recommendations for future research	76
	References	78

LIST OF FIGURES

Table	Description	Page
	Research Article 1	
1	Structure of the Multilayer Perceptron Neural Network	21
2	Histogram of the non-random Exhaustion dimension in sample 1	30
3	Histogram of the non-random Exhaustion dimension in sample 2	30
4	Histogram of the random Exhaustion dimension in sample 1	30
5	Histogram of the random Exhaustion dimension in sample 2	30

LIST OF TABLES

Table	Description	Page
	Research Article 1	
1	Characteristics of the participants	23
2	Characteristics of the samples	28
3	Descriptive statistics of the Exhaustion items in the samples	28
4	Internal consistency of Exhaustion in the samples	29
5	Descriptive statistics of the Exhaustion dimension	29
6	Neural network classification results for Dataset 1	31
7	Neural network cross validation classification results	31
8	Cronbach alphas for the non-random classification groups	32
9	Component matrix for the correctly classified cases	32
10	Cronbach alpha for the random classification groups	33
11	Component matrix for the misclassified Random cases	34
	Research Article 2	
1	Interpretation of parameter-level mean-square fit statistics	51
2	Characteristics of the participants	53
3	Rasch model item fit statistics for Exhaustion	57
4	Descriptive statistics, Cronbach α and factor analysis for different outfit categories on non-random data	58
5	Tucker's Φ of the outfit mean-square categories	59
6	Descriptive statistics, Cronbach alpha and factor analysis for different outfit categories on random data	59
7	Structural equivalence between non-random and random outfit categories	60
8	Outfit descriptive statistics of the different neural network classifications	60
9	Cross tabulation of the neural network prediction versus the outfit categories	61
10	Internal consistency, factor analysis and structural equivalence for the neural network prediction and outfit categories cross tabulation	62

ABSTRACT

Title: Establishing the protocol validity of an electronic standardised measuring instrument

Key terms: Protocol validity, item response theory, neural networks, well-being instruments

Over the past few decades, the nature of work has undergone remarkable changes, resulting in a shift from manual demands to mental and emotional demands on employees. In order to manage these demands and optimise employee performance, organisations use well-being surveys to guide their interventions. Because these interventions have a drastic financial implication it is important to ensure the validity and reliability of the results. However, even if a validated measuring instrument is used, the problem remains that wellness audits might be reliable, valid and equivalent when the results of a group of people are analysed, but cannot be guaranteed for each individual. It is therefore important to determine the validity and reliability of individual measurements (i.e. protocol validity). However, little information exists concerning the efficiency of different methods to evaluate protocol validity.

The general objective of this study was to establish an efficient, real-time method/indicator for determining protocol validity in web-based instruments. The study sample consisted of 14 592 participants from several industries in South Africa and was extracted from a work-related well-being survey archive. A protocol validity indicator that detects random responses was developed and evaluated. It was also investigated whether Item Response Theory (IRT) fit statistics have the potential to serve as protocol validity indicators and this was compared to the newly developed protocol validity indicator.

The developed protocol validity indicator makes use of neural networks to predict whether cases have protocol validity. A neural network was trained on a large non-random sample and a computer-generated random sample. The neural network was then cross-validated to see whether posterior cases can be accurately classified as belonging to the random or non-random sample. The neural network proved to be effective in detecting 86,39% of the random responses and 85,85% of the non-random responses correctly. Analyses on the misclassified cases demonstrated that the neural network was accurate because non-random classified cases were in fact valid and reliable, while random classified cases showed a problematic factor

structure and low internal consistency. Neural networks proved to be an effective technique for the detection of potential invalid and unreliable cases in electronic well-being surveys.

Subsequently, the protocol validity detection capability of IRT fit statistics was investigated. The fit statistics were calculated for the study population and for random generated data with a uniform distribution. In both the study population and the random data, cases with higher outfit statistics showed problems with validity and reliability. When compared to the neural network technique, the fit statistics suggested that the neural network was more effective in classifying non-random cases than it was in classifying random cases. Overall, the fit statistics proved to be effective indicators of protocol invalidity (rather than validity) provided that some additional measures be imposed.

Recommendations were made for the organisation as well as with a view to future research.

OPSOMMING

Titel: Vasstelling van die protokolgeldigheid van 'n elektroniese gestandaardiseerde meetinstrument

Sleutelterme: Protokolgeldigheid, item-responsteorie, neurale netwerke, welstand instrumente

Die aard van werk het die afgelope paar dekades merkwaardige veranderings ondergaan, wat gelei het tot 'n verskuiwing van behoeftes van fisiese tot verstands- en emosionele eise aan werkers. Om hierdie eise te bestuur en werknemerprestasie te optimeer, maak organisasies gebruik van welstandsondersoeke om hulle intervensies te lei. Aangesien hierdie intervensies drastiese finansiële implikasies inhou, is dit belangrik om die geldigheid en betroubaarheid van die resultate te verseker. Indien 'n geldige meetinstrument egter gebruik word, is die probleem nog steeds dat welstandsoudits betroubaar, asook geldig en ekwivalent mag wees wanneer die resultate van 'n groep mense geanaliseer word, maar nie vir elke individu gewaarborg kan word nie. Daarom is dit belangrik om die geldigheid en betroubaarheid van individuele metings (d.i. protokol-geldigheid) te bepaal. Min inligting is egter beskikbaar oor die doeltreffendheid van verskillende metodes om protokolgeldigheid te evalueer.

Die algemene doelwit van hierdie studie was om 'n doeltreffende, intydse metode/aanduider vir die vasstelling van protokolgeldigheid in web-gebaseerde instrumente daar te stel. Die steekproef het bestaan uit 14 592 deelnemers vanuit verskeie industrieë in Suid-Afrika wat uit 'n werksverwante welwees ondersoekargief getrek is. 'n Protokolgeldigheidsaanduider wat lukraak response vasstel, is ontwikkel en geëvalueer. Ondersoek is ook ingestel na die Item Respons Teorie (IRT) en of passingstatistiek die potensiaal het om as protokol-geldigheidsaanduider te dien en dit is vergelyk met die nuutontwikkelde protokol-geldigheidsaanduider.

Die ontwikkelde protokolgeldigheidsaanduider maak van neurale netwerke gebruik om te voorspel watter gevalle geldig is of nie. 'n Neurale netwerk is op 'n groot nie-lukraak monster en 'n rekenaargegeneerde lukraak monster geoefen. Die neurale netwerk is daarna gekruis-

valideer om te sien watter later gevalle akkuraat geklassifiseer kan word as behorende aan die lukraak of nie-lukraak monster. Die neurale netwerk was doeltreffend met die korrekte opsporing van 86,39% van die lukraak response en 85,85% van die nie-lukraak response. Analises van die foutief-geklassifiseerde gevalle het aangetoon dat die neurale netwerk inderdaad akkuraat was, omdat nie-lukraak geklassifiseerde gevalle inderwaarheid geldig en betroubaar was, terwyl lukraak geklassifiseerde gevalle 'n problematiese faktorstruktuur getoon het asook lae interne konsekwentheid. Neurale netwerke het getoon dat dit 'n doeltreffende tegniek was vir die vasstelling van potensiële ongeldige en onbetroubare gevalle in elektroniese welstand ondersoeke.

Gevolglik is die protokol geldigheidvasstellingskapasiteit van IRT-passingstatistiek ondersoek. Die passingstatistiek is bereken vir die studiebevolking en vir lukraak gegenereerde data met 'n uniforme verspreiding. In beide die studiebevolking en die lukraak data, het gevalle met hoër uitsetstatistiek probleme getoon met geldigheid en betroubaarheid. Wanneer dit met die neurale netwerktegniek vergelyk is, het die passingstatistiek aangedui dat die neurale netwerk meer doeltreffend was met die klassifikasie van nie-lukraak gevalle as wat die geval was in die klassifikasie van lukraak gevalle. Oorkoepelend beskou, het die passingstatistiek geblyk om meer doeltreffende aanduiders te wees van protokolongeldigheid (eerder as geldigheid) indien sekere addisionele maatreëls toegepas is.

Aanbevelings is vir organisasies asook met die oog op toekomstige navorsing gemaak.

CHAPTER 1

INTRODUCTION

This dissertation is concerned with whether measurements by a self-report instrument can be trusted as being valid and reliable on an individual level.

This chapter provides the background and the problem statement of this study. The research objectives and the significance of the study are also presented. Finally, the research method is explained and the division of chapters is provided.

1.1 PROBLEM STATEMENT

Over the past few decades, the nature of work has undergone remarkable changes. According to Schreuder and Coetzee (2006), these changes include the increased utilisation of information and communication technology, the expansion of the services sector, the globalisation of the economy, the changing structure of the workforce, the increasing flexibilisation of work, the creation of the 24-hour economy, and the utilisation of new production concepts. Barling (1999) points out that the nature of work has changed from manual demands to mental and emotional demands. In addition, job resources such as choice and control at work and organisational support are often lacking, which might affect the energy and motivation of employees (Nelson & Simmons, 2003; Schaufeli & Bakker, 2004; Turner, Barling, & Zacharatos, 2002). In order to survive and prosper in a continuously changing environment, organisations need energetic, healthy and motivated employees (Weinberg & Cooper, 2007).

As a first step to promote health and well-being in organisations, Rothmann and Cooper (2008) recommend that well-being audits, which focus on both positive and negative aspects of work-related well-being, should be implemented and feedback should be given at individual, group and organisational levels. Questionnaires are often used to assess psychological well-being dispositions and states in South Africa. It is believed that these instruments can contribute to the efficiency of management of human resources (Pieterse & Rothmann, 2009; Sieberhagen, Rothmann, & Pienaar, 2009). Huysamen (2002) stresses the

importance of responsible use of psychological assessment instruments. The responsible use of well-being audits implies that they should be reliable, valid, and equivalent for different demographical groups (Rothmann & Cooper, 2008; Van de Vijver & Rothmann, 2004).

Two psychological assessment instruments have been developed for the purpose of conducting well-being audits in South Africa, namely the South African Employee Health and Wellness Survey (SAEHWS) (Rothmann & Rothmann, 2006) and the South African Psychological Fitness Index (SAPFI) (Rothmann, 2008). The SAEHWS is used to assess the health and wellness of employees in South African organisations, while the SAPFI is used to assess the psychological fitness of employees. These instruments are administered electronically via the internet. Each participant receives an online personal feedback report after completion. Management also receives feedback at a group level. These instruments have been standardised for use in South Africa and have been proven to be internally consistent, valid and equivalent for different language, race and gender groups (Rothmann, 2008; Rothmann & Rothmann, 2006). This is especially important considering the following stipulation of the Employment Equity Act, 55 of 1998, Section 8 (South Africa, 1998): "Psychological testing and other similar assessments are prohibited unless the test or assessment being used – (a) has been scientifically shown to be valid and reliable, (b) can be applied fairly to all employees, and (c) is not biased against any employee or group."

However, the problem remains that wellness audits might be reliable, valid and unbiased when the results of a group of people are analysed, but cannot be guaranteed for each individual. The validity and reliability of an individual measurement is termed protocol validity (see Kurtz & Parrish, 2001). Protocol validity is an area of concern for any psychological measuring instrument (Johnson, 2004). Problems with protocol validity arise when the participant completes the instrument in such a way that the ability of the instrument to accurately measure the intended constructs is compromised (Ben-Porath, 2003).

There are several threats to protocol validity. A linguistically incompetent participant will be unable to produce a valid protocol even for a well-validated test. Reasons for linguistic incompetence includes limited vocabulary, poor verbal comprehension, a particular way of interpreting item meaning, and/or cultural differences in item interpretation. Negligence or inattentiveness may result in random responding or using the same response pattern

repeatedly. Participants might also deliberately attempt to respond uncharacteristically (Johnson, 2004).

The direct result of protocol invalidity is that the scores of the outcomes of the instrument are invalid. This has serious implications in the case of wellness audits, where decisions are based upon the outcomes of these instruments. Because wellness instruments are used as a basis for the referral of individuals for counselling and group interventions, valid and reliable results for each individual is important. The protocol validity should therefore be determined directly after the completion of the instrument to determine if the results can be trusted. If the outcomes are not trustworthy, individuals might be misdiagnosed and resources will be spent on ineffective and expensive interventions. It would also be beneficial to have information about the validity of individual cases during group analyses so that invalid cases may be discarded and kept from distorting the outcomes.

Quite a number of attempts have been made to determine protocol validity across a wide range of psychometric instruments. Goldberg and Kilkowski (1985) suggested a semantic antonym approach, where the instrument's items for a single construct are semantic opposites. The participant should then answer in opposite directions on the scale. The respondent's answers on the opposite items are correlated to determine if the person responded in the desired direction of the scale. The Minnesota Multiphasic Personality Inventory (MMPI) and the Revised NEO Personality Inventory (NEO-PI-R) are examples where these types of correlational indicators are used to determine protocol validity (Schinka, Kinder, & Kremer, 1997). Unfortunately, these indicators are flawed in the sense that they assume all items to be answered equally in order to be reliable.

The use of correlational indicators to determine protocol validity fits the paradigm of Classical Test Theory (CTT). However, the danger with using the CTT-based correlational approach is that inconsistent but valid protocols are routinely misdiagnosed (Johnson, 2005). During group analyses, CTT techniques also pose some problems. The typical CTT techniques that are used to determine validity and reliability are factor analysis and internal consistency tests like Cronbach's alpha (Allen & Yen, 2002). However, these statistics provide no information about individual cases. Although a large group of invalid cases is highly unlikely to provide acceptable group-level statistics, acceptable group-level statistics

cannot guarantee the validity of each and every protocol. Acceptable CTT statistics simply mean that a large part of the group of cases is acceptable.

The modern approach to test theory is Item Response Theory (IRT). In IRT, the identification of invalid protocols is potentially less of an issue because fit statistics are generated for each individual. These fit statistics indicate whether an individual's responses fit the chosen IRT model (Bond & Fox, 2007). If the fit statistics are unacceptable, it is an indication that the case is probably invalid. Literature also suggests that reasons for misfitting responses might be related to the threats of protocol validity (see Linacre, 2002; Smith, 1996). Furthermore, different items have different difficulty (intensity) and discrimination levels. Therefore, one may expect a valid protocol to have inconsistent responses depending on the items (Hambleton & Rogers, 1990). A possible problem with IRT is that scores are not calculated, but estimated with dedicated software implementing complex iterative algorithms like the Maximum Likelihood Estimation algorithm (Bond & Fox, 2007). This might complicate the calculation of fit statistics in real time on the internet.

It is clear that organisations in South Africa have to make responsible decisions regarding the health and wellness of their employees. Therefore a need exists not only for reliable, valid and equivalent measuring instruments, but also for proof of protocol validity. Currently, little information exists concerning the efficiency of different methods to evaluate protocol validity. If methods could be developed to assess protocol validity, individual responses on wellness audits could be analysed, which could improve human resource decisions. This research will make a contribution to the science of Industrial Psychology by contributing to a better understanding regarding the possibilities of more computationally advanced protocol validity indicators for use in electronic measuring instruments. This study will also contribute to the practice of Industrial Psychology in organisations by providing possible tools for determining the validity and reliability of individual measurements, promoting evidence-based practices and sound intervention investments.

From the above-mentioned description of the research problem, the following research questions arise:

- What are the major threats to protocol validity?

- Can a more advanced protocol validity indicator be developed for use in electronic wellbeing surveys?
- Can IRT fit statistics be used as protocol validity indicators?
- Can these measures be implemented programmatically for an online instrument?

1.2 RESEARCH OBJECTIVES

1.2.1 General objective

The general objective of this study is to establish an efficient, real-time method/indicator for determining protocol validity in web-based instruments.

1.2.2 Specific objectives

The specific objectives of this study are as follows:

- To study the major threats to protocol validity.
- To develop and evaluate a protocol validity indicator that detects random responses in electronic well-being surveys.
- To evaluate the IRT fit statistics for use as protocol validity indicators and to compare the IRT fit statistics with the developed protocol validity indicator.
- To discuss the practical implications of implementing the protocol validity indicators in an online wellness instrument.

1.3 RESEARCH METHOD

The research method for each of the two articles consists of a brief literature review and an empirical study. The reader should note that a literature review is conducted for the purposes of each article. This section focuses on aspects relevant to the empirical study that is conducted.

1.3.1 Research design

A survey design is used to reach the specific research objectives (Huysamen, 2001). In this type of research, data is collected by posing questions and recording people's responses.

1.3.2 Participants and procedure

The study sample consists of 14 592 participants from several industries in South Africa, including financial, engineering, mining, human resources and manufacturing. The data is gathered from a survey data archive (see Whitley, 2002, p. 383). The survey archive contains people's responses to survey questions in wellness audits and demographic data concerning the respondents. The data is kept on computer files. Survey archives are useful because they have been collected for research purposes; consequently, great care is taken to ensure the reliability and validity of the data. The following criteria are considered when evaluating archived survey data (Whitley, 2002):

- What was the purpose of the original study? Data collected for some purposes (e.g. influencing legislation) may be biased in ways that support the purposes.
- How valid was the data collection? There should be documentation that includes information such as how respondents are sampled and the validity and reliability of measures.
- What information was collected? The data set should include all the variables needed to test the research hypotheses.
- When was the data collected? Social attitudes and processes can change over time and responses in old data sets might not represent the ways in which responses are currently related.

1.3.3 Measuring instrument

One subscale of the South African Employee Health and Wellness Survey is used, namely Exhaustion (5 items, e.g. "I feel tired before I arrive at work"). A seven point rating scale is used, ranging from 0 (never) to 6 (always). The SAEHWS is a self-report instrument based on the dual-process model of work-related well-being (Rothmann & Rothmann, 2006) and is based on the assumption that employees' perceptions and experiences represent important

information regarding the wellness climate in the organisation. The SAEHWS measures organisational climate, wellness, health and lifestyle, organisational commitment, and personal variables (Rothmann & Rothmann, 2006).

1.3.4 Statistical analysis

Statistical analyses are conducted with SPSS 16.0 (SPSS, 2008) and Winsteps 3.68 (Linacre, 2009). Descriptive statistics (e.g. means and standard deviations) are used. Pearson's product-moment correlation (Tabachnick & Fidell, 2001) is used to investigate the relationship between variables. Exploratory factor analyses, specifically principal component analyses (Kline, 1994), are conducted to determine the validity of the constructs that are measured in this study. Coefficient alpha (Cronbach, 1951) is used to assess reliability, as it contains important information regarding the proportion of variance of the total variance of a scale that consists of true variance.

The Multilayer Perceptron (MLP) neural network is used as a possible alternative for determining protocol validity. The MLP is a feed-forward neural network that can be trained to store knowledge, based on the relationship between the dependent and independent variables, and to predict values for posterior cases. The MLP is used for the following reasons:

- Neural networks can approximate either a linear or a non-linear relationship, depending on the relationship in the data (Haykin, 1998).
- A model does not have to be hypothesised in advance (Haykin, 1998).
- Minimal demands are made on assumptions (SPSS, 2008).

In addition, the Rasch IRT model is used with several of its statistics (Bond & Fox, 2007). First, Rasch reliability is used to provide an estimate of the reproducibility of measures. Rasch reliability is a more conservative estimate for the ratio of real person variance than Cronbach's alpha (Linacre, 2002). Second, item measures (indicated by δ) are used to assess the severity of items' measurement of the latent construct. Last, infit statistics are used to assess how accurately or predictably the items fit the Rasch model. Outfit statistics are used to assess person-fit for the purposes of protocol validity, because the outfit statistic is not adjusted for outliers (Bond & Fox, 2007).

Cross-validation (Tabachnick & Fidell, 2001) is used to ensure repeatability by testing the model against an unknown sample. If the protocol validity indicator is based on one sample and tested against an unknown sample of cases, the efficiency of the indicator can be determined with more confidence. Tucker's coefficient of congruence phi (ϕ) is used to compute structural equivalence between factors for different samples (Tucker, 1951). Structural equivalence can be used to prove differences in factor structures for non-random and random predicted cases, confirming the validity of the neural network prediction. Tucker's ϕ is defined by the following formula:

$$\phi = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$

In this formula, x_i and y_i represent the respective component loadings. Tucker's ϕ ranges from -1,00 to +1,00 (perfect similarity). Values above 0,95 can be taken to indicate factorial similarities, while values below 0,85 show unavoidable incongruencies (Van de Vijver & Leung, 1997).

The better-than-chance effect size index I is used to determine the success of the neural network (Huberty & Lowman, 2000). This index adjusts the observed hit rate of a category for incidental correct classification of cases. In other words, it indicates if the classification was correct by chance or not. The better-than-chance index is calculated by the following formula:

$$I = \frac{H_o - H_e}{1 - H_e}$$

In this formula, H_o represents the observed hit rate (correct classifications divided by total cases), while H_e represents the chance rate, which is the proportional prior probabilities of classification. Huberty and Lowman (2000) provides guidelines for the interpretation of I. Values below 0,10 are seen as a small effect, while values above 0,35 represent a large effect.

1.4 OVERVIEW OF CHAPTERS

In Chapter 2, a potential protocol validity indicator is developed and evaluated. The protocol validity indicator is based on a specific predictive modelling technique called neural networks. The neural network is trained to distinguish between non-random and random data and then cross-validated against a second sample. Validity, reliability and structural equivalence tests are used to evaluate the effectiveness of the neural network's classification. In Chapter 3, it is investigated whether the Rasch IRT model fit statistics can be used as protocol validity indicators. The fit statistics are also compared to the neural network technique from Chapter 2. Conclusions, recommendations and limitations of the study follow in Chapter 4.

1.5 CHAPTER SUMMARY

This chapter discussed the problem statement and research objectives. The measuring instruments and research method that are used in this research were explained, followed by a brief overview of the chapters that follow.

REFERENCES

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory.* Long Grove, IL: Waveland Press.

Barling, J. (1999). Changing employment relations: Empirical data, social perspectives and policy options. In D. B. Knight & A. Joseph (Eds.), *Restructuring societies: Insights from the social sciences* (pp. 59-82). Ottawa: Carlton University Press.

Ben-Porath, Y. S. (2003). Self-report inventories: Assessing personality and psycho-pathology. In J. R. Graham & J. Naglieri (Eds.) Vol. X: *Handbook of assessment psychology* (pp. 554-575). New York: Wiley.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Goldberg, L. R., & Kilkowski, J. M. (1985). The prediction of semantic consistency in self-descriptions: Characteristics of persons and of terms that affect the consistency of responses to synonym and antonym pairs. *Journal of Personality and Social Psychology, 48*, 82-98.

Hambleton, R. K., & Rogers, J. H. (1990). Using item response models in educational assessments. In W. Schreiber & K. Ingenkamp (Eds.), *International developments in large-scale assessment* (pp. 155-184). England: NFER-Nelson.

Haykin, S. (1998). *Neural networks: A comprehensive foundation* (2nd ed.). New York: Macmillan College Publishing.

Huberty, C.J. & Lowman, L.L. (2000). Group overlap as a basis for effect size. *Educational and Psychological Measurement, 60*, 543-563.

Huysamen, G. K. (2001). *Methodology for the social and behavioural sciences.* Cape Town: Oxford University Press.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality, 39*, 103-129.

Kline, P. (1994). *An easy guide to factor analysis.* London: Routledge.

Kurtz, J. E., & Parrish, C. L. (2001). Semantic response consistency and protocol validity in structured personality assessment: The case of the NEO–PI–R. *Journal of Personality Assessment, 76*, 315-332.

Linacre, J. M. (2002). Cronbach alpha or Rasch reliability: Which tells the "truth"? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2009). *WINSTEPS®: Rasch measurement computer program.* Beaverton, OR: Winsteps.com

Nelson, D. L., & Simmons, B. L. (2003). Health psychology and work stress: A more positive approach. In J. C. Quick & L. E. Tetrick (Eds.), *Handbook of occupational health psychology* (pp. 97-119). Washington, DC: American Psychological Association.

Pieterse, H., & Rothmann, S. (2009). Perceptions of the role and contribution of human resource practitioners in a global petrochemical company. *South African Journal of Economic and Management Sciences, 12*, 370-384.

Rothmann, S. (2008, April). *Psychological fitness: Concept and measurement.* Paper presented at the SASOM Conference, Pretoria.

Rothmann, S., & Cooper, C. L. (2008). *Organizational and work psychology.* London: Hodder Education.

Rothmann, J. C., & Rothmann, S. (2006). *The South African Employee Health and Wellness Survey: User manual.* Potchefstroom: Afriforte (Pty) Ltd.

Schaufeli, W. B., & Bakker, A. B. (2004). Job demands, job resources and their relationship with burnout and engagement: A multi-sample study. *Journal of Organizational Behavior, 25*, 293-315.

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

Schreuder, A. M. G., & Coetzee, M. (2006). *Careers: An organisational perspective* (3rd ed.). Johannesburg: Juta Academic.

Sieberhagen, C., Rothmann, S., & Pienaar, J. (2009). Employee health and wellness in South Africa: The role of legislation and management standards. *SA Journal of Human Resource Management, 7*(1), 1-9.

Smith, R. M. (1996). Polytomous mean-square fit statistics. *Rasch Measurement Transactions, 10*(3), 516-517.

South Africa. (1998). *Government Gazette , 400 , 19370.* Cape Town: Government Printers.

SPSS Inc. (2008). *SPSS 16.0 for Windows.* Chicago, IL: SPSS Inc.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Needham Heights, MA: Allyn & Bacon.

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies.* Personnel Research Section Report No. 984. Washington, DC: Department of the Army.

Turner, N., Barling, J., & Zacharatos, A. (2002). Positive psychology at work. In C. R. Snyder & S. J. Lopez (Eds.), *Handbook of positive psychology* (pp. 715-728). Oxford: Oxford University Press.

Van de Vijver F. J. R., & Leung, K. (1997). *Method and data analysis for cross-cultural research*. Beverly Hills, CA: Sage.

Weinberg, A., & Cooper, C. (2007). *Surviving the workplace: A guide to emotional well-being*. London: Thomson.

Whitley, B. E. (2002). *Principles of research in behavioral science* (2nd ed.). Boston, MA: McGraw-Hill.

CHAPTER 2

RESEARCH ARTICLE 1

THE DEVELOPMENT AND EVALUATION OF A PROTOCOL VALIDITY INDICATOR

ABSTRACT

The aim of this study was to develop and evaluate a protocol validity indicator that detects random responses in electronic well-being surveys. The study sample consisted of 14 592 participants from several industries in South Africa. A literature review indicated that neural networks could be used to evaluate protocol validity. A neural network was trained on a large non-random sample and a computer-generated random sample. The neural network was then cross-validated to see whether posterior cases can be accurately classified as belonging to the random or non-random sample. The neural network proved to be effective in detecting 86,39% of the random protocols and 85,85% of the non-random protocols correctly. Analyses on the misclassified cases demonstrated that the neural network was accurate because non-random classified cases were in fact valid and reliable, whereas random classified cases showed a problematic factor structure and low internal consistency. Neural networks proved to be an effective technique for the detection of potential invalid and unreliable cases in electronic well-being surveys.

OPSOMMING

Die doel van hierdie studie was om 'n indikator van protokolgeldigheid te ontwikkel en te evalueer wat lukraak response in elektroniese welstandopnames kan identifiseer. 'n Literatuuroorsig het aangedui dat neurale netwerke gebruik kan word om protokol-geldigheid te evalueer. Die steekproef het bestaan uit 14 592 deelnemers uit verskeie industrieë in Suid-Afrika. 'n Neurale netwerk is opgelei op 'n groot nie-ewekansige steekproef en 'n ewekansige steekproef wat met behulp van 'n rekenaar gegenereer is. Die neurale netwerk is met behulp van kruisvalidering getoets om te bepaal of protokolle akkuraat geklassifiseer kan word as behorende tot die ewekansige of die nie-ewekansige steekproef. Die neurale netwerk was effektief in die opsporing van 86,39% van die lukraak protokolle en 85,85% van die nie-lukraak protokolle. Ontleding van die gevalle wat verkeerd geklassifiseer is, het aangetoon dat die neurale netwerk akkuraat was omdat nie-ewekansige geklassifiseerde gevalle geldig en betroubaar was, terwyl ewekansige geklassifiseerde gevalle 'n problematiese faktorstruktuur en lae interne konsekwentheid getoon het. Neurale netwerke blyk 'n effektiewe tegniek te wees om potensieel ongeldige en onbetroubare gevalle in elektroniese welstandopnames te identifiseer.

Questionnaires are increasingly used to assess psychological well-being dispositions and states in South Africa. These questionnaires are used by managers to understand the strengths and weaknesses within the organisation before implementing expensive organisational development interventions (Rothmann & Cooper, 2008). It is believed that these instruments can contribute to the efficiency of management of human resources (Pieterse & Rothmann, 2009; Sieberhagen, Rothmann, & Pienaar, 2009). Huysamen (2002) stresses the importance of responsible use of psychological assessment instruments. The responsible use of well-being audits implies that they should be reliable, valid, and equivalent for different demographical groups (Rothmann & Cooper, 2008; Van de Vijver & Rothmann, 2004).

Two psychological assessment instruments have been developed for the purpose of conducting well-being audits in South Africa, namely the South African Employee Health and Wellness Questionnaire (SAEHWS) (Rothmann & Rothmann, 2006) and the South African Psychological Fitness Index (SAPFI) (Rothmann, 2008). The SAEHWS is used to assess the health and wellness of employees in South African organisations, whereas the SAPFI is used to assess the psychological fitness of employees. These instruments have been standardised for use in South Africa and have been shown to yield reliable, valid and unbiased scores for different language, race and gender groups (Rothmann, 2008; Rothmann & Rothmann, 2006). This is important considering the following stipulation of the Employment Equity Act 55 of 1998, Section 8 (South Africa, 1998): "Psychological testing and other similar assessments are prohibited unless the test or assessment being used - (a) has been scientifically shown to be valid and reliable, (b) can be applied fairly to all employees; and (c) is not biased against any employee or group."

Both the SAEHWS and the SAPFI are self-report inventories (SRIs) which are administered online, providing the employee with an immediate feedback report upon completion. However, Ben-Porath (2003) points out that intentional or unintentional distortion is the primary limitation to SRIs. He further explains that even if the SRI is psychometrically sound, individuals might approach the assessment in a manner that compromises the ability to respond accurately on the item measuring the construct. Thus, in these cases, a reliable and valid psychometric instrument might yield invalid test results. This is referred to as protocol validity.

The reality of protocol validity for well-being audits is that scores might be reliable, valid and equivalent across groups when a group of people are analysed, but that the validity of individual assessments cannot be guaranteed. These individual assessments are often used as the basis for decisions regarding the health, well-being and/or fitness of respondents, creating an immediate concern. The user of the well-being audits needs to consider the validity of the responses before using the results. If a decision is made utilising invalid information, the decision might have harmful effects on the employee and/or the organisation which could result in labour issues given the rights of employees (South Africa, 1995). Therefore, a need exists for efficient protocol validity indicators on these instruments.

To develop protocol validity indicators, it is necessary to understand the different threats to protocol validity. Ben-Porath (2003) classifies the threats into two broad categories, namely non-content-based invalid responding and content-based invalid responding. These categories reflect the role of the instrument item content in invalid responding. Non-content-based invalid responding refers to responding without reading, processing or comprehending the items. This has adverse effects on the protocol validity of the measurement, because the individual did not portray an answer related to the item or construct. Content-based invalid responding occurs when a respondent reads and comprehends the item content, but distorts answers (intentionally or unintentionally) to create a misleading impression (social desirability and acquiescence).

Non-content-based invalid responding is categorised into three modes, namely non-responding, random responding and fixed responding (Ben-Porath, 2003). These modes are all different behaviours to the same threat, i.e. that participants did not evaluate the content of items before responding. Non-responding occurs where a participant fails to respond to a certain number of items. Random responding takes place when an individual provides a random answer without considering the content of the item. Fixed responding occurs when a participant adopts a systematic response approach by providing the same answer to multiple items in the SRI, thereby creating a response pattern.

Content-based invalid responding is organised into two main categories, namely over-reporting and under-reporting (Ben-Porath, 2003). These categories are defined by an individual providing an answer that is more (over-reporting) or less (under-reporting) severe

than the actual situation. Both of these categories of threats might occur intentionally or unintentionally.

In the context of the electronic well-being surveys, certain threats are more problematic than others. Non-responding is dealt with by forcing participants to answer a question before continuing to the next one. The risk of this approach is that participants might provide a random answer because they are unable to non-respond. Thus, to a certain extent, non-responding is replaced with random responding. Fixed responding is also less of an issue because the surveys consist of multiple pages with a limited number of items on a single page. If a participant should provide a fixed response pattern, that exact pattern will in all probability not be repeated continuously, because the participant starts on a new page every few items. This, to a certain extent, also substitutes random responding for fixed responding. Furthermore, fixed responding can easily be identified by investigating an algorithm that detects patterns in the responses.

These arguments stress the importance of the random response threat in electronic well-being surveys. When a decision is made or money invested based upon the outcome of such a survey, it is important to have confidence in the outcome of the survey. Knowing if random response was evident during the completion of the survey will provide more confidence in the decisions made. Therefore, a need exists to develop and evaluate a protocol validity indicator that can be used to detect random responding in an electronic well-being survey.

The aim of this study was thefore to develop and evaluate a protocol validity indicator that detects random responses in electronic well-being surveys.

Random responding

Ben-Porath (2003) defines random responding as an unsystematic response approach that occurs when an individual provides a random answer without reading or comprehending a test item. It is described as not being dichotomous, i.e. it presents itself in varying intensities throughout the instrument. This non-content-based protocol validity threat can be divided into three categories, namely intentional random responding, unintentional random responding and response recording errors.

Intentional random responding comes about when a respondent has the capacity to respond appropriately to an item, but chooses to respond in an unsystematic way (Ben-Porath, 2003). A typical example of this would be an uncooperative individual who would respond randomly just to complete the instrument, thereby avoiding conflict with third parties. Unintentional random responding occurs when an individual does not have the capacity to provide an answer to a specific item (Ben-Porath, 2003). Instead of non-responding, the individual provides an answer without having an understanding of the item. Reasons for unintentional random responding might include reading difficulties or comprehension deficits.

The final category of random responding is response recording errors. This is related to the user-friendliness of the instrument presentation (Ben-Porath, 2003). Some instruments are presented in a booklet and answer sheet format, others in a booklet-only format, and others are electronic. If the respondent makes a mistake by marking the answer in the wrong position, the response is essentially random. A well constructed electronic instrument should be less prone to response recording errors than conventional methods, because there is little room for error if only one question is displayed at a time.

Currently, random responding is detected predominantly with inconsistency scales and examples can be found in the Minnesota Multiphasic Personality Inventory (MMPI; e.g. Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989; Butcher et al., 1992) and the Revised NEO Personality Inventory (NEO-PI-R; Goldberg & Kilkowski, 1985). Scores on these scales are fairly simple to compute, which makes it possible for psychologists to calculate the scores without computers. The MMPI utilises the Variable Response Inconsistency Scale (VRIN) and the NEO-PI-R makes use of the INC inconsistency scale (Schinka, Kinder, & Kremer, 1997). The inconsistency scales focus on comparing scores on items from a test with scores on other items from the same test. Highly correlated (similar and opposite) items within the test are selected, and the expectation is that respondents should provide similar responses on all these items (Kurtz & Parrish, 2001). Confidence intervals are then created based on deviations from the normative means or dissimilarity between distributions from a known random and non-random sample.

There are, however, problems with these inconsistency scales. Costa and McCrae (1997) found that most participants who score high on the inconsistency scale of the NEO-PI-R are

in fact not responding randomly. Piedmont, McCrae, Riemann and Angleitner (2000) indicated that the NEO-PI-R's inconsistency scales lack utility. Kurtz and Parrish (2001) found that cases identified as invalid by NEO-PI-R inconsistency scales, were in fact psychometrically valid and reliable. According to Archer, Handel, Lynch and Elkins (2002), the MMPI-A's inconsistency scale is limited in detecting partially random responses.

Another approach to random response is to compare an individual case to a normative sample of other cases. Item Response Theory (IRT; Reise & Widaman, 1999) uses person-fit statistics to calculate how a participant's responses fit to theoretical expectations based on a normative sample. The items are scaled according to how they are rated in the larger group of cases, and for a respondent to have a good person-fit they should rate items according to their estimated level for the construct (Johnson, 2005).

Neural networks (De Ville, 2001) present another alternative for evaluating the protocol validity of a measure. Neural networks are discussed in the next section.

The use of neural networks to evaluate protocol validity

Predictive classification techniques can model and infer trends from a large database and apply them to individual, posterior cases (SPSS, 2008). These techniques are widely used in data mining applications for creating business intelligence (De Ville, 2001). It would be possible to create a random response classification model based on a large training sample, and apply it to individual posterior cases. The power of predictive modelling is that the model is only created once. This model is then used for posterior classification with little computational effort. If a predictive model were built to understand what a valid and reliable response is, it would be able to identify a similar case with a certain probability. In essence, these predictive classification techniques can model and infer response styles from a large group of protocols and apply them to individual protocols for calculating individual protocol validity and reliability.

Several classification techniques can serve as candidates for this purpose, namely discriminant analysis, special cases of regression, decision trees, or neural networks. Neural networks are very sophisticated modelling techniques capable of modelling extremely complex functions. They have been found to outperform discriminant analysis and logistic

regression especially where the latter's assumptions are violated (Sommer, Olbrich, & Arendasy, 2004). This study utilises neural networks for the following reasons:

- Neural networks can approximate a linear or non-linear relationship, depending on the relationship in the data (Haykin, 1998).
- A model does not have to be hypothesised in advance (Haykin, 1998).
- Minimal demands are made on assumptions (SPSS, 2008).

Haykin (1998) defines a neural network as a massively parallel distributed processor that has a natural propensity for storing experiential knowledge and for making it available for use. It resembles the brain in two respects. First, knowledge acquisition is done by training the neural network and second, interneuron connection strengths known as "synaptic weights" are used to store the knowledge.

Neural networks, similar to regression analysis, learn through a series of independent and dependent variables. Just as regression analysis acquires knowledge through the least-squares method, and saves the knowledge in regression coefficients, a neural network acquires knowledge through minimising the prediction error in the dependent variable (training), and saves it as synaptic weights (SPSS, 2008).

The Multilayer Perceptron is an important class of neural networks that have been used widely for forecasting, prediction and classification across several disciplines of science (see Reifman & Feldman, 2002). Multilayer Perceptron networks are also commonly available in modern statistical packages. Haykin (1998) explains that Multilayer Perceptron networks consist of a series of sensory units (independent variables) that constitute the input layer, one or more hidden layers consisting of computational nodes, and an output layer (dependent variables). The input signal propagates forward through the different layers of the network (usually referred to as feed-forward).

The Multilayer Perceptron is trained to create the structure of the neural network. During training, the weights between the input, hidden and output layers are optimised. Different training algorithms could be applied to optimise these weights, but the backpropagation algorithm is most widely used (Agirre-Basurko, Ibarra-Berastegi, & Madariaga, 2006). It

might happen that too much detail is included in the neural network, causing it to lose its ability to generalise (Haykin, 1998). This is referred to as over-training. To solve over-training, a testing sample can be assigned to track errors made during training and guarantee the generalisation of the network (SPSS, 2008). After the neural network has been trained, it can be used for classification of unknown cases.

Figure 1 depicts the structure of the Multilayer Perceptron neural network. When classifying unknown cases, a series of independent variables ($x_1...x_p$) is provided to the input layer. The input layer then distributes the values in the variables to the neurons in the hidden layer. The neuron in the hidden layer multiplies the value by a weight (w_{ji}), and the resulting weighted values are summed to produce a combined value u_j. This weighted sum (u_j) is then provided to a transfer function, σ, which outputs a value h_j. These outputs are distributed to the output layer where the value from each hidden layer neuron is again multiplied by a weight (w_{kj}). The resulting weighted values are summed to produce a combined value v_j. The weighted sum (v_j) is provided to a transfer function, σ, which outputs a value y_k. The y values are the outputs of the network.

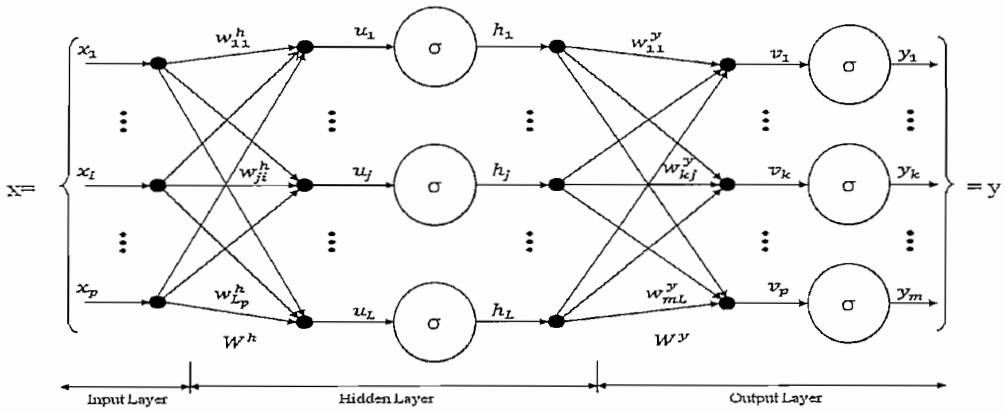

Figure 1. Structure of the Multilayer Perceptron Neural Network

When the Multilayer Perceptron classifies posterior cases, it also generates pseudo-probabilities for each classification. These pseudo-probabilities estimate the level of certainty that the case belongs to the predicted group (SPSS, 2008).

If non-random and random samples are created, the construct's items are used as inputs of the neural network, and the familiar source of the cases (random or non-random) as outputs. The neural network will model neurons and synaptic weights based on the relationship between

the items of the construct for the applicable sample. A posterior case can then be classified as belonging to either the random or non-random sample of data.

METHOD

Research design

This research follows the quantitative research tradition. A cross-sectional survey design was used (Huysamen, 2001). In this type of research, data is collected by posing questions and recording people's responses. A correlational approach was followed where each individual in the sample was measured on variables (i.e. items of a scale) at the same point in time and the relationship between these variables were analysed.

Participants

The study sample consisted of 14 592 participants from several industries in South Africa, including financial, engineering, mining, human resources and manufacturing. Descriptive information of the sample is given in Table 1. The mean age of the participants was 40,24 (SD = 9,98). Slightly more males (62,49%) than females (37,51%) were represented in the study population. In terms of race, 27,42% of participants were black and 36,64% white. The race values were missing for 4 288 (29,39%) participants, due to the sensitivity of posing questions relating to racial differences in South Africa. Almost half of the study population (49,75%) had a qualification of grade 12 or lower, 13,38% a certificate, 15,48% a diploma or a degree and 10,38% a postgraduate qualification.

Table 1
Characteristics of the Participants

Item	Category	Frequency	Percentage
Gender	Male	9119	62,49
	Female	5473	37,51
Race	Black	4001	27,42
	White	5346	36,64
	Coloured	586	4,02
	Indian	359	2,46
	Other	12	0,08
	Missing	4288	29,39
Qualification	Up to grade 12	7260	49,75
	Certificate	1953	13,38
	Diploma or degree	2259	15,48
	Postgraduate qualification	1515	10,38
	Missing	1605	11,00

Measuring instrument

Two qualitative questions measuring helping and restraining factors at work were used for selecting the cases where respondents took care in completing the survey. One subscale of the South African Employee Health and Wellness Survey (SAEHWS), namely Exhaustion, was used to reach the objective of this study. The SAEHWS is a self-report instrument based on the dual-process model of work-related well-being and is based on the assumption that employees' perceptions and experiences represent important information regarding the wellness climate in the organisation (Rothmann & Rothmann, 2006). The SAEHWS instrument measures an employee's health and wellness status, relates the data to the organisational climate and compares the results to the South African norm (Rothmann & Rothmann, 2006). The factor structures of all the subscales in the SAEHWS support the validity of the scales and are equivalent for different ethnic groups and organisations. The internal consistencies are also acceptable and above the cut-off point of 0,70 (Rothmann & Rothmann, 2006).

The exhaustion scale was used because the items are posed amongst the first 15 items in the 239-item survey. Research has shown that consistent responding is more likely to occur in the

beginning of the test before participants get bored, tired or impatient (see Berry et al., 1991). Exhaustion was measured with 5 items (e.g. "I feel tired before I arrive at work") on a 7-point scale varying from 0 (*never*) to 6 (*always*). Helping and restraining factors were measured with two items (e.g. "Which factors are helping you to be motivated and effective in your current job and organisation?") where participants provided unrestrained and spontaneous answers. Exploratory and confirmatory analyses showed that the factor structure of the exhaustion scale is valid and equivalent for different ethnic groups and organisations.

Statistical analyses

Statistical analyses were conducted with the SPSS 16.0 program (SPSS, 2008). Descriptive statistics (e.g. means and standard deviations) were used. Histograms, skewness and kurtosis were used as measures of spread (Tabachnick & Fidell, 2001). Pearson's product moment correlations were used to assess the relationship between variables (Tabachnick & Fidell, 2001). Exploratory factor analyses, specifically principal component analyses (Kline, 1994), were conducted to determine the validity of the construct that was measured in this study. Coefficient alpha (Cronbach, 1951) was used to assess reliability as it contains important information regarding the ratio of true variance to observed variance explained by the particular scale.

As discussed earlier, a multilayer perceptron neural network was used for the predictive classification of data. In addition, cross-validation (Tabachnick & Fidell, 2001) was used to ensure repeatability by testing the model against an unknown sample. If the model is trained on one sample and tested against an unknown sample of cases, the efficiency of the model can be determined for classifying posterior unknown cases.

Tucker's coefficient of congruence phi (ϕ) was used to compute structural equivalence between factors for different samples (Tucker, 1951). Structural equivalence analysis can be used to detect differences in factor structures for non-random and random predicted cases, supporting the validity of the neural network prediction. Tucker's ϕ is defined by the following formula:

$$\phi = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$

In this formula, x_i and y_i represent the respective component loadings. Tucker's ϕ ranges from -1,00 via 0 to +1,00 (perfect similarity). Values above 0,95 can be taken to indicate factorial similarities, while values below 0,85 show non-avoidable incongruencies (Van de Vijver & Leung, 1997).

The better-than-chance effect size index I was used to determine the success of the neural network (Huberty & Lowman, 2000). This index adjusts the observed hit rate of a category for incidental correct classification of cases. In other words, it indicates if the classification was correct by chance or not. The better-than-chance index is calculated by the following formula:

$$I = \frac{H_o - H_e}{1 - H_e}$$

In this formula, H_o represents the observed hit rate (correct classifications divided by total cases), while H_e represents the chance rate, which is the proportional prior probabilities of classification. Huberty and Lowman (2000) provides guidelines for the interpretation of I. Values below 0,10 could be seen as a small effect, while values above 0,35 represent a large effect.

Research procedure

The data was gathered from a survey data archive (see Whitley, 2002). The survey archive contains responses to survey questions in the well-being audits and demographic data concerning the respondents. This data is kept on computer databases. Survey archives are useful because they have been collected for research purposes; consequently, great care was taken to ensure the reliability and validity of the data.

In order to build a predictive model that classifies non-random responses, certain assumptions had to be made regarding the definition of non-random responses. A non-random response set was defined as a response set that belongs to a group of cases that were found to be valid and reliable. To minimise the effect of potential individual unreliable responses in the training sample, some cases were filtered out based on sufficient time spent answering the survey and an adequate amount of qualitative data provided in items measuring helping and restraining factors at work. A total of 3 496 cases were discarded based on the criteria, resulting in a

sample of 11 097. Subsequently, the sample was split equally by means of random sampling for cross-validation purposes. The sample sizes were 5 549 for the training sample and 5 548 for the cross-validation sample. The factor structures (in support of validity), reliability and structural equivalence were computed for both samples.

Next, random data was generated for each sample to serve as rejection samples. The purpose of the rejection sample is to train the neural network in what a case should not look like. To ascertain a prior probability of 50,00%, random data was generated to match the amount of non-random data in each sample. The property of the random number generator was to assign an equal probability to each element (uniform distribution). The random and non-random data were marked appropriately, and would be used to train the predictive model. Descriptive statistics were computed on the items and the exhaustion construct (mean of the items) for both the random and non-random data. These statistics were used to analyse the comprehensiveness of the samples.

The neural network was then trained on the first sample and cross-validated against the second sample. The cross-validation was done by comparing the known classification with the predicted classification. In this comparison, known random and non-random cases can either be correctly classified or misclassified by the neural network. To more precisely assess the performance of the neural network, the factor structures (in support of validity), reliability and structural equivalence were calculated for the correctly classified and misclassified non-random and random cases.

RESULTS

Minimising the effect of unreliable response in the sample

On average, 65 minutes were spent completing the 239-item questionnaire. Twenty percent of individuals (2 638) took less than 25 minutes to complete the questionnaire. It was assumed that these individuals have answered the questions excessively fast (less than 6 seconds per question). This data was discarded, resulting in a sample of 11 955 valid cases.

Subsequently, the data of the two qualitative questions was analysed. Of the 11 955 cases, the average length of the qualitative data (the sum of both items) was 137 characters. After inspecting the data, it was concluded that responses longer than nine characters provided meaningful responses. There were 766 participants (6,41%) who provided less than nine characters on both of the qualitative questions. These cases were also discarded, resulting in a sample of 11 097 usable cases.

Splitting the dataset for cross-validation purposes

The large sample (n = 11097) was divided into two smaller samples by means of random sampling. Sample 1 was used to train the neural network, and sample 2 for cross-validation. The sample sizes were 5 549 for sample 1 and 5 548 for sample 2. Table 2 shows the characteristics of the samples. The mean age of the participants was 40,71 (SD = 9,86) in sample 1 and 40,81 (SD = 10,02) in sample 2. Sample 1 contains 3 436 (61,92%) male respondents, while sample 2 has 3 394 (61,18%). In total, 25,57% of the respondents in sample 1 were black, and 38,78% white, which is in line with the 25,34% black and 39,26% white respondents in sample 2. In terms of qualification, sample 1 included 2 703 (48,71%) respondents with grade 12 or lower, similar to the 2 729 (49,19%) respondents in sample 2.

Table 2
Characteristics of the Samples

Item	Category	Sample 1 (n=5549)		Sample 2 (n=5548)	
		Frequency	*Percentage*	*Frequency*	*Percentage*
Gender	Male	3436	61,92	3394	61,18
	Female	2113	38,08	2154	38,83
Race	Black	1419	25,57	1406	25,34
	White	2152	38,78	2178	39,26
	Coloured	193	3,48	230	4,15
	Indian	126	2,27	117	2,11
	Other	5	0,09	6	0,11
	Missing	1654	29,83	1611	29,00
Qualification	Up to grade 12	2703	48,71	2729	49,19
	Certificate	715	12,89	695	12,53
	Diploma or degree	916	16,51	838	15,10
	Postgraduate qualification	574	10,34	622	11,21
	Missing	641	11,55	664	11,97

Table 3 shows the descriptive statistics for the Exhaustion items in both samples. Item 4 has the highest mean of 3,09 (*SD* = 1,67) in sample 1 and 3,13 (*SD* = 1,67) in sample 2. Item 3 has the lowest mean of 1,69 (*SD* = 1,56) in sample 1 and 1,76 (*SD* = 1,58) in sample 2. It is apparent that the means and standard deviations are quite similar in both samples.

Table 3
Descriptive Statistics of the Exhaustion Items in the Samples

Item	Mean		*SD*	
	Sample 1	*Sample 2*	*Sample 1*	*Sample 2*
Item 1	2,93	2,95	1,55	1,56
Item 2	2,79	2,84	1,63	1,64
Item 3	1,69	1,76	1,56	1,58
Item 4	3,09	3,13	1,67	1,65
Item 5	2,57	2,60	1,65	1,65

Table 4 shows the Cronbach alpha coefficients of the samples. The internal consistencies of the samples are both acceptable ($\alpha \geq 0{,}70$). Principle components analyses extract a one-factor model for both samples, with 61,75% of the variance explained in sample 1 and 60,67% of the variance explained in sample 2. The component loadings of the samples are structurally equivalent with a Tucker's ϕ of 1,00.

Table 4

Internal Consistency of Exhaustion in the Samples

Sample	α
Sample 1	0,84
Sample 2	0,84

Next, an Exhaustion dimension score was created by calculating the mean of the items for both samples. Table 5 contains the descriptive statistics of the Exhaustion dimension. The non-random data is skewed in both samples ($z \geq 2{,}58$). Non-random sample 1 has a mean of 2,61 ($SD = 1{,}27$) and sample 2 a mean of 2,66 ($SD = 1{,}26$). The random data has a mean of 2,99 ($SD = 0{,}89$) in sample 1 and 3,01 ($SD = 0{,}89$) in sample 2.

Table 5

Descriptive Statistics of the Exhaustion Dimension

Sample	**Random**	**Minimum**	**Maximum**	**Mean**	***SD***	**Skewness**	**Kurtosis**
						z	*z*
Sample 1	Non-random	0,00	6,00	2,61	1,27	4,69	-5,48
Sample 1	Random	0,20	5,80	2,99	0,89	0,20	-4,07
Sample 2	Non-random	0,00	6,00	2,66	1,26	4,18	-4,99
Sample 2	Random	0,20	5,80	3,01	0,89	-0,19	-4,70

Figure 2 to Figure 5 depict histograms to show the distribution of the data. One can easily slight difference in distributions between the random and non-random data.

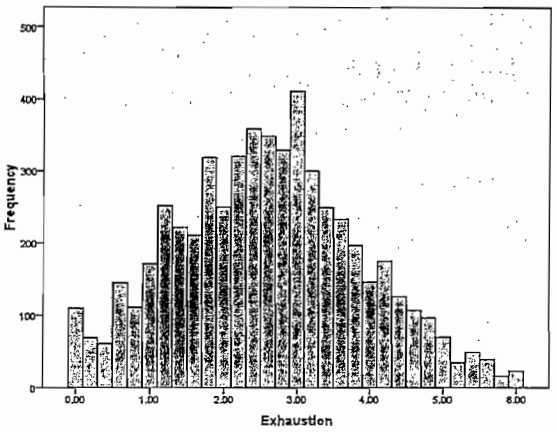

Figure 2. Histogram of the non-random Exhaustion dimension in sample 1

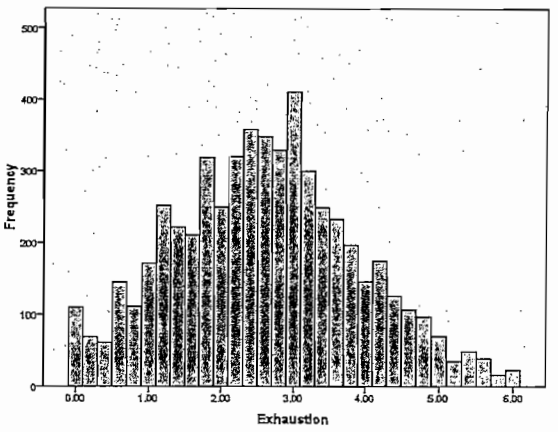

Figure 3. Histogram of the non-random Exhaustion dimension in sample 2

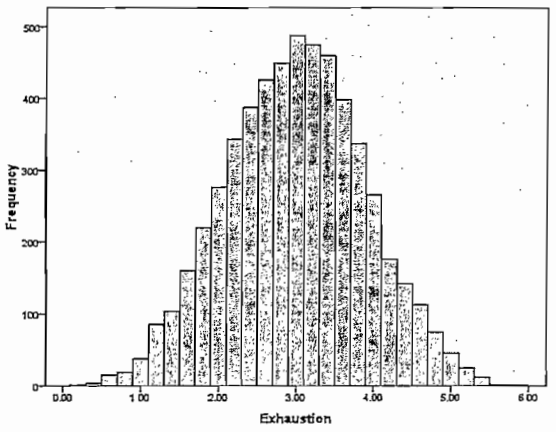

Figure 4. Histogram of the random Exhaustion dimension in sample 1

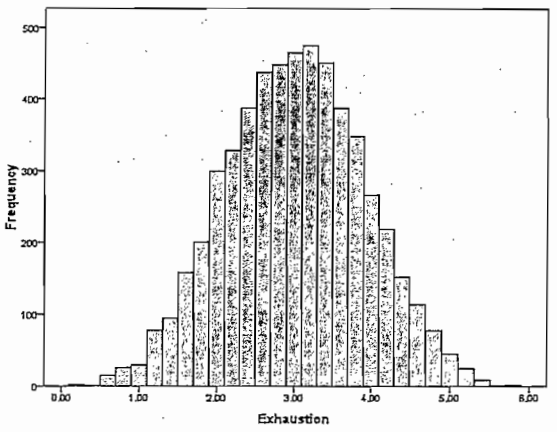

Figure 5. Histogram of the random Exhaustion dimension in sample 2

Training the neural network

The neural network was trained on the first dataset. To prevent overtraining, 70,00% of the cases were used for training the neural network and 30,00% were assigned to a testing sample. The exhaustion items were used as independent variables in the input layer, and the random/non-random classification as the dependent variable in the output layer.

Table 6 shows the *post hoc* classification results of the neural network. It is apparent that the neural network performed quite well in predicting both the random and non-random cases. Approximately 87,00% of the non-random cases were predicted correctly in both the training and testing samples, while 86,82% and 85,41% of the random cases were predicted correctly in the training and testing samples respectively.

Table 6

Neural Network Classification Results for Dataset 1

Sample	Observed	*n*	Predicted		
			Non-random	*Random*	*Percent Correct*
Training (*n*=7802)	Non-random	3857	3354	503	86,96
	Random	3945	520	3425	86,82
Testing (*n*=3296)	Non-random	1692	1472	220	87,00
	Random	1604	234	1370	85,41

Subsequently, the neural network was cross-validated against the second dataset.

Table 7 shows the results of the cross-validation. The neural network performed equally well in the prediction of unfamiliar cases, predicting 85,85% and 86,39% of the respective non-random and random cases correctly. The better-than-chance index *I* shows a value of 0,72, which implies a large effect ($I \geq 0{,}35$), meaning that the correct predictions did not occur by chance. *I* was calculated with a prior probability of 0,50 for both samples. This provides an acceptable level of confidence in the performance of the neural network when classifying unfamiliar cases.

Table 7

Neural Network Cross Validation Classification Results

Observed	*n*	Predicted		
		Non-random	*Random*	*Percent Correct*
Non-random	5548	4763	785	85,85
Random	5548	755	4793	86,39

Reliability and validity of the predicted outcomes

A total of 785 misclassified non-random cases were identified during cross-validation. To investigate whether these cases were misclassified because of the incompetence of the neural

network or because they were in fact random responses by individuals, reliability, validity and construct equivalence analysis procedures were performed.

Cronbach's alpha statistic was used to determine the reliability of the correctly classified and misclassified cases (see Table 8). The correctly classified cases have an acceptable level of reliability ($\alpha \geq 0{,}70$), whereas the misclassified cases present with a Cronbach alpha of 0,35, indicating a low level of internal consistency in that group of cases.

Table 8

Cronbach Alphas for the Non-Random Classification Groups

Classification	α
Correctly classified non-random cases	0,89
Misclassified non-random cases	0,35

In a principal components factor analysis, the correctly classified cases yield a one-factor model for the Exhaustion dimension (Eigenvalues > 1), as would be expected, which explains 69,38% of the variance. The component loadings are all sufficiently high (indicated in Table 9), with item 3 showing the lowest component loading of 0,75. In addition, all the items have practically significant (large effect) correlations with each other ($r > 0{,}50$). This provides sufficient evidence about the reliability and validity of the correctly classified non-random cases.

Table 9

Component Matrix for the Correctly Classified Cases

Items	**Component 1**
Item 1	0,85
Item 2	0,84
Item 3	0,75
Item 4	0,84
Item 5	0,89

Interestingly, the misclassified cases yield a two-factor model in a principal components analysis (Eigenvalues $\geq$ 1), explaining 28,16% and 22,81% of the variance respectively. In addition, these cases fail to provide a practically significant correlation between any of the

items ($r \leq 0{,}30$). The Exhaustion dimension should theoretically yield a one-factor model (Rothmann & Rothmann, 2006). This indicates instability in the factor structure for the misclassified cases. Given this instability and the low internal consistency of the misclassified cases, one can assume that these cases are problematic.

Analysis of misclassified random data

Table 7 shows that 755 random cases were misclassified as non-random by the neural network. Again, it should be investigated whether these cases were misclassified because of the incompetence of the neural network or whether the random data might actually have been similar to the non-random data.

As would be expected, the correctly classified random cases generate an undesirable Cronbach alpha (see Table 10), indicating that there is indeed no internal consistency in the random data. The misclassified cases produce an alpha of 0,77, indicating an acceptable level of internal consistency ($\alpha \geq 0{,}70$).

Table 10

Cronbach Alpha for the Random Classification Groups

Classification	α
Misclassified random cases	0,77
Correctly classified random cases	-0,25

Confirming the inherent nature of the random data, the correctly classified random cases yield a totally unrelated four-factor model (Eigenvalues >1) explaining 21,56%, 20,95%, 20,71% and 20,21% of the variance respectively. These cases also fail to provide a practically significant correlation between any of the items ($r \leq 0{,}30$).

The misclassified random cases yield a one-factor model (Eigenvalues > 1) explaining 52,85% of the variance. The component loadings are sufficiently high (see Table 11), with item 3 having the lowest component loading of 0,62. All the items have practically significant (medium effect) correlations with each other ($r > 0{,}30$).

Table 11

Component Matrix for the Misclassified Random Cases

Items	Component 1
Item 1	0,80
Item 2	0,63
Item 3	0,62
Item 4	0,76
Item 5	0,80

In order to calculate structural equivalence between the correctly classified non-random cases and the misclassified random cases, their component loadings (see Table 9 and Table 11 respectively) were compared using Tucker's ϕ. The Tucker's ϕ proportionality coefficient of the misclassified random cases and the correctly classified non-random cases is 1,00. This suggests that the components are largely equivalent ($\phi \geq 0{,}95$), keeping in mind that with a small number of items Tucker's ϕ is often misleadingly high.

DISCUSSION

The aim of this study was to develop and evaluate a protocol validity indicator that detects random responses in electronic well-being surveys. A predictive model that classifies an individual case as being random or non-random was used to achieve this. The results showed that 14,15% of participants could be regarded as random responders according to the neural network. These cases showed a low internal consistency and instability in the factor structure, while the non-random classified cases showed acceptable internal consistency and extracted the expected factor structure.

In total, 13,61% of the computer-generated random data was classified as non-random by the neural network. These cases proved high internal consistency, an acceptable factor structure and structural equivalence with the non-random natural sample. The structural equivalence implies that the factor structure extracted from these misclassified random cases corresponds to the expected factor structure of the non-random cases provided by participants in the natural sample.

The misclassified random cases serve as evidence that the non-random classification of the neural network cannot guarantee the absence of the random response threat. Though it does

particularly well in detecting random responses (86,39% of the random cases), it is not infallible. A random responder might coincidentally display a response pattern similar to that of a non-random responder. In such as case, the random response would be regarded as valid and reliable.

Although the neural network is inadequate to classify the random response threat perfectly, these findings do provide evidence that when the neural network classifies cases as non-random they can be trusted to be valid and reliable on group level. The opposite is also true: when the neural network classifies cases as random, problems with validity and reliability should be expected. This is potentially powerful, because as it is quite difficult to put a single response set in relation to a certain factor structure or reliability level. The assumptions of factor analysis and Cronbach's alpha both include an adequate sample size, and it would not be possible to calculate these statistics on only one case (see Kline, 1994; Cronbach, 1951). The only other technique that potentially provides for the identification of invalid individual cases is IRT (Bond & Fox, 2007).

Exploring reasons why the neural network performs well in detecting valid and reliable cases requires an understanding of exactly what the neural network does when classifying the data. When the neural network is trained with a construct's items as inputs, it infers a model of which responses tend to correspond with different values for the other items. Such a model is inferred for each of the input items and stored in neurons and paths with synaptic weights (the hidden layer). It merely provides a model of which responses tend to "go together" in a large training sample. If the neural network classifies a case as not corresponding to the training sample, it means that the specific response pattern would in all probability not exist in the training sample, and the neural network will reject the case. Thus, if the training sample is valid and reliable, the neural network will classify similarly valid and reliable cases on an individual level.

When a response does not correspond with the training sample, it either means that certain threats to protocol validity were evident or that the training sample was not comprehensive enough. The threats might be a combination of content- or non-content-based invalid responding. Fixed response, random response or response recording errors could have been evident or the participant could have over- or underreported inconsistently. Any threat (or

combination thereof) that distorts the response pattern would cause rejection by the neural network.

The nature of the training sample is fundamentally important as it has a direct influence on the effectiveness of the neural network. The neural network can, at best, only perform as well as its training sample (Haykin, 1998). The training sample consists of a natural sample and a rejection sample. The natural sample trains the neural network in what a valid and reliable case should look like, while the rejection sample does exactly the opposite. During post hoc classification, the neural network decides to which sample an individual case belongs.

If the natural sample is too small, the neural network will not correctly classify all possible valid and reliable cases. If the natural sample contains too many invalid responses, the neural network will again misclassify posterior invalid responses as valid. If the distribution of the items is heavily skewed on the Likert scale (i.e. individuals tend to respond very high or very low on the scale), the neural network will only classify individual cases correctly if they are similarly skewed. However, an interesting argument arises regarding the skewness of the natural data. If a very large group of participants provided skewed responses for a specific construct, one could argue that individuals tend to embrace or reject the construct easily. This trend should be expected from a posterior case. On the other hand, it might be risky to restrain the level of the construct measured by limiting the valid responses on the scale. This should be thoroughly tested and investigated when implementing a neural network in this context.

Furthermore, the importance of the composition of the rejection sample should be stressed. The same criteria apply for this sample as for the natural sample. The difference between the inherent properties of the natural and rejection sample also plays an important role in the success of the neural network classification. In this article, the rejection sample was populated with computer-generated random data because it was used to test for random responding and also provided an adequate picture of what responses should not look like.

Another critical factor is the number of items used as inputs. Using a limited number of items restricts the range of possible response distribution, contributing to the coincidence of having random data (in the rejection sample) correspond to non-random data (in the natural sample). Items worded semantically opposite could also influence the effectiveness of the neural

network. Semantic opposites refer to wording certain items as antonyms of other items (Goldberg & Kilkowski, 1985). The expectation is that the antonyms should be answered in the opposite direction of the other items on the scale (Johnson, 2005). When using more items (including semantic opposites) as inputs for the neural network, a more precise nature of response would be inferred from the natural sample. More precision in the response patterns of the network structure should minimise overlap between the natural and rejection samples.

In conclusion, the neural network learns from a large group of cases and predicts whether a posterior individual case would also belong to that group of cases. The neural network does not identify protocol validity threats, but gives an indication if a collection of threats have corrupted the case to the extent that it would not belong to the group of valid and reliable cases. This serves as valuable information for a facilitator who should make a decision based on the outcome of the SAEHWS. It also creates the opportunity for data analysts to exclude single cases from a group analysis because of validity and reliability issues.

There are several limitations to this study: Firstly, only one construct with five items was used as input for the neural network. Moreover, these items did not include semantically opposite items. The neural network might respond differently to other constructs and items (or combinations thereof). Secondly, it is indefinite whether the natural training sample included all possible valid and reliable combinations of responses on the five items. The neural network might classify potentially valid and reliable responses as invalid because such a response did not feature strongly in the natural sample. Thirdly, the inherent property of the computer-generated random data in the rejection sample is that every element has an equal probability to be selected (uniform distribution). This might not be the case when human beings are providing the data randomly. Different rejection samples were not tested in this study. Finally, the neural network was validated with group reliability and validity techniques (i.e. factor analysis and Cronbach α) that do not provide fit statistics for individual cases.

RECOMMENDATIONS

Predictive modelling has proved effective in detecting possible invalid and unreliable cases on an individual level. However, it does not indicate the applicable threats that were apparent in the individual case. It is recommended that a neural network be implemented in the SAEHWS to provide facilitators with a validity and reliability indicator. Facilitators should evaluate the indicator before making decisions based on the results of the survey. The indicator should be used with caution and if a problem is evident, the facilitator should conduct an interview with the participant to investigate possible threats to the validity of the protocol. This interview can be structured within Ben-Porath's (2003) framework for protocol validity threats by asking questions such as "Did you read all the questions before providing an answer?", "Did you understand all the questions?", and "Do you feel you were totally honest in answering all the questions?".

The neural networks can be used when performing group analyses on the SAEHWS data. Although it is straightforward to determine reliability and validity on a group level, it is difficult to identify cases that contribute to the invalidity or unreliability of the group. Cases that prove to be invalid and unreliable on an individual level can be excluded from the group analyses to fortify the case for organisations to invest in expensive interventions.

Implementing the neural network in the SAEHWS requires the model structure to be exported after training. This model structure includes information about the trained neurons and synaptic weights. A neural network prediction algorithm should then be used to classify posterior cases based on the model structure when each individual SAEHWS report is generated.

Future research should be conducted to further validate the neural network with techniques beyond factor analysis and Cronbach's α, such as IRT. Also, interviews should be conducted with individuals identified as random responders to establish if they had indeed responded randomly. Reasons for random responding should also be investigated. The effect of predicting validity and reliability with more than 5 items as inputs in the neural network should be investigated. It would be beneficial to include semantically opposite items in this research, as more items would minimise the probability of random data being valid and reliable by coincidence. The relationship between the size of the training sample, the number

of independent variables and the prediction success of the neural network should be investigated. This should include an attempt to determine the optimal number of cases that would prevent overtraining, yet provide the neural network with enough information to make accurate predictions.

In addition, future research should observe whether it is more effective to include items for several constructs into one neural network for use as a validity and reliability score, or to use multiple neural networks for each construct for creating an aggregated validity and reliability score. It is proposed that the relationship between the validity, reliability and pseudo-probabilities be investigated. If a linear relationship is found, the pseudo-probabilities could potentially be used to indicate a severity score for protocol validity. Norms, benchmarks and cut-off points should also be considered for the pseudo-probabilities. These norms, benchmarks and cut-off points can then be used by facilitators to decide about the validity and reliability of cases in different situations.

REFERENCES

Agirre-Basurko, E., Ibarra-Berastegi, G., & Madariaga, I. (2006). Regression and multilayer perceptron-based models to forecast hourly O3 and NO2 levels in the Bilbao area. *Environmental Modelling & Software, 21*, 430-446.

Archer, R. P., Handel, R. W., Lynch, K. D. & Elkins, D. E. (2002). MMPI–A Validity Scale Uses and Limitations in Detecting Varying Levels of Random Responding. *Journal of Personality Assessment, 78*(3), 417–431.

Arendasy M., Sommer, M., & Hergovich, A. (2007). Statistical judgment formation in personnel selection: A study in military aviation psychology. *Military Psychology, 19*(2), 119-136.

Barling, J. (1999). Changing employment relations: Empirical data, social perspectives and policy options. In D. B. Knight & A. Joseph (Eds.), *Restructuring societies: Insights from the social sciences* (pp. 59-82). Ottawa: Carlton University Press.

Ben-Porath, Y. S. (2003). Self-report inventories: Assessing personality and psycho-pathology. In J. R. Graham & J. Naglieri (Eds.) Vol. X: *Handbook of assessment psychology* (pp. 554-575). New York: Wiley.

Berry, D. T. R., Wetter, A. W., Baer, R. A., Widiger, T. A., Sumpter, J. C., Reynolds, S. K., & Hallam, R. A. (1991). Detection of random responding on the MMPI-2: Utility of F, Back F, and VRIN scales. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3,* 418-423.

Bond, T. G., & Fox, C. M., 2007. *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *MMPI–2: Minnesota Multiphasic Personality Inventory–2: Manual for administration and scoring.* Minneapolis, MN: University of Minnesota Press.

Butcher, J. N., Williams, C. L., Graham, J. R., Archer, R. P., Tellegen, A., & Ben-Porath, Y. S. (1992). *MMPI–A: Minnesota Multiphasic Personality Inventory–Adolescent: Manual for administration, scoring, and interpretation.* Minneapolis, MN: University of Minnesota Press.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*, 309-319.

Costa P. T., & McCrae R. R. (1997). Stability and change in personality assessment: The Revised NEO Personality Inventory in the Year 2000. *Journal of Personality Assessment, 68*, 86-94.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

De Ville, B. (2001). *Microsoft data mining: Integrated business intelligence for e-commerce and knowledge management.* Woburn, MA: Butterworth-Heinemann.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4,* 26-42.

Goldberg, L. R., & Kilkowski, J. M. (1985). The prediction of semantic consistency in self-descriptions: Characteristics of persons and of terms that affect the consistency of responses to synonym and antonym pairs. *Journal of Personality and Social Psychology, 48*, 82-98.

Greenholtz, J. F. (2005). Does intercultural sensitivity cross cultures? Validity issues in porting instruments across languages and cultures. *International Journal of Intercultural Relations, 29*(1), 73-89.

Haykin, S. (1998). *Neural networks: A comprehensive foundation* (2nd ed.). New York: Macmillan College Publishing.

Huberty, C.J. & Lowman, L.L. (2000). Group overlap as a basis for effect size. *Educational and Psychological Measurement*, 60, 543-563.

Huysamen, G. K. (2001). *Methodology for the social and behavioural sciences.* Cape Town: Oxford University Press.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality, 39*, 103-129.

Kline, P. (1994). *An easy guide to factor analysis.* London: Routledge.

Kurtz, J. E., & Parrish, C. L. (2001). Semantic response consistency and protocol validity in structured personality assessment: The case of the NEO–PI–R. *Journal of Personality Assessment, 76*, 315-332.

Nelson, D. L., & Simmons, B. L. (2003). Health psychology and work stress: A more positive approach. In J. C. Quick & L. E. Tetrick (Eds.), *Handbook of occupational health psychology* (pp. 97-119). Washington, DC: American Psychological Association.

Piedmont, R. L., McCrae, R. R., 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(3), 582-93.

Pieterse, H., & Rothmann, S. (2009). Perceptions of the role and contribution of human resource practitioners in a global petrochemical company. *South African Journal of Economic and Management Sciences, 12*, 370-384.

Reifman J., & Feldman, E. E. (2002). Multilayer perceptron for nonlinear programming. *Computers & Operations Research, 29*, 1237-1250.

Reise, R. P., & Widaman, K. F. (1999). Assessing the fit of measurement models at the individual level: A comparison of item response theory and covariance structure approaches. *Psychological Methods, 4*(1), 3-21.

Rothmann, S., & Cooper, C. L. (2008). *Organizational and work psychology.* London: Hodder Education.

Rothmann, J. C., & Rothmann, S. (2006). *The South African Employee Health and Wellness Survey: User manual.* Potchefstroom: Afriforte (Pty) Ltd.

Rothmann, S. (2008, April). *Psychological fitness: Concept and measurement.* Paper presented at the SASOM Conference, Pretoria.

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

Schaufeli, W. B., & Bakker, A. B. (2004). Job demands, job resources and their relationship with burnout and engagement: A multi-sample study. *Journal of Organizational Behavior, 25*, 293-315.

Sieberhagen, C., Rothmann, S., & Pienaar, J. (2009). Employee health and wellness in South Africa: The role of legislation and management standards. *SA Journal of Human Resource Management, 7*(1), 1-9.

Sommer, M., Olbrich, A., & Arendasy M. (2004). Improvements in personnel selection with neural networks: A pilot study in the field of aviation psychology. *International Journal of Aviation Psychology, 14*(1), 103-115.

South Africa. (1998). *Government Gazette , 400 , 19370.* Cape Town: Government Printers.

South Africa. (1995). *Labour Relations Act (nr. 66 of 1995).* Pretoria: Government Printers.

SPSS Inc. (2008). *SPSS 16.0 for Windows.* Chicago, IL: SPSS Inc.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Needham Heights, MA: Allyn & Bacon.

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies.* Personnel Research Section Report No. 984. Washington, DC: Department of the Army.

Turner, N., Barling, J., & Zacharatos, A. (2002). Positive psychology at work. In C. R. Snyder & S. J. Lopez (Eds.), *Handbook of positive psychology* (pp. 715-728). Oxford, UK: Oxford University Press.

Van de Vijver F. J. R., & Leung, K. (1997). *Method and data analysis for cross-cultural research*. Beverly Hills, CA: Sage.

Weinberg, A., & Cooper, C. (2007). *Surviving the workplace: A guide to emotional well-being*. London: Thomson.

Whitley, B. E. (2002). *Principles of research in behavioral science* (2nd ed.). Boston, MA: McGraw-Hill.

CHAPTER 3

RESEARCH ARTICLE 2

APPLYING THE RASCH MODEL TO ASSESS PROTOCOL VALIDITY

ABSTRACT

The aim of this study was to evaluate Rasch model fit statistics for use as protocol validity indicators and to compare these fit statistics to a newly developed neural network technique. Literature suggests that reasons for unacceptable fit statistics are in line with the threats to protocol validity. Rasch model fit statistics were calculated for a study population of 5548 participants who completed a wellness survey in several industries in South Africa. Fit statistics were also calculated for random generated data with a uniform distribution. In both the study population and the random data, cases with higher outfit statistics showed problems with validity and reliability. When compared to the neural network technique, the fit statistics suggested that the neural network was more effective in classifying non-random cases than it was in classifying random cases. Reasons for this are given and discussed in this article. Overall, the fit statistics proved to be effective indicators of protocol invalidity (rather than validity) provided that some additional measures be imposed.

OPSOMMING

Die doel van hierdie studie was om te evalueer of Rasch-model passtatistiek geskik is vir gebruik as protokolgeldigheidsaanduiders en om hierdie passtatistiek met 'n nuutontwikkelde neurale netwerktegniek te vergelyk. Volgens die literatuur is die redes vir onaanvaarbare geskiktheidstatistiek in lyn met bedreigings vir protokol-geldigheid. Rasch-model geskiktheidstatistiek is bereken vir 'n studiepopulasie van 5548 deelnemers wat 'n welweesondersoek by verskeie industrieë in Suid-Afrika voltooi het. Passtatistiek is ook bereken vir lukraak gegenereerde data met 'n eenvormige verspreiding. In beide studiepopulasie en die lukraak data, het gevalle met hoër uitsetstatistiek probleme getoon met geldigheid en betroubaarheid. Wanneer dit met die neurale netwerktegniek vergelyk word, het passtatistiek aangedui dat die neurale netwerk meer doeltreffend was om nie-lukraak gevalle te klassifiseer as wat dit was met die klassifikasie van lukraak gevalle. Redes is hiervoor gegee en bespreek in hierdie artikel. Oorkoepelend beskou, het die passtatistiek aangedui dat dit meer doeltreffende aanduiders was van protokolongeldigheid (eerder as geldigheid) indien sekere addisionele maatreëls ingestel is.

Keywords: Protocol validity, Rasch model, fit statistics, neural networks

Self-report questionnaires are often used in research in organisational behaviour. Research based on self-report questionnaires is used by managers to understand the strengths and weaknesses within the organisation before implementing expensive organisational development interventions (Rothmann & Cooper, 2008). The South African Employee Health and Wellness Survey (SAEHWS) is a self-report questionnaire used for well-being audits in South Africa (Rothmann & Rothmann, 2006). The SAEHWS has been standardised for use in South Africa and has been proven to be internally consistent, valid and equivalent for different language, race and gender groups (Rothmann, 2008; Rothmann & Rothmann, 2006).

When important decisions are made based upon the outcomes of an instrument, it is crucial to have confidence in the validity and reliability of the measurement. However, a valid and reliable instrument does not guarantee the reliability and validity of measurement in every instance (Ben-Porath, 2003). Reliability and validity tests are usually conducted on group level, and even if these tests are conducted before feedback is provided, reliability and validity are still not guaranteed for each individual. Protocol validity is thus concerned with the validity and reliability of a single case.

Protocol validity can be compromised by several threats. These threats fall into two broad categories (Ben-Porath, 2003), namely non-content-based invalid responding and content-based invalid responding. These categories reflect the role of item content when invalid response was evident. Non-content-based invalid responding refers to threats where participants responded without reading, processing or comprehending the items. This has obvious consequences for the protocol validity of the measurement, because the individual did not provide an answer related to the item or construct. Content-based invalid responding occurs when a respondent reads and comprehends the item content, but tries to create a misleading impression by distorting the answers (intentionally or unintentionally).

Non-content-based invalid responding is categorised into three specific behaviours, namely non-responding, random responding and fixed responding (Ben-Porath, 2003). These behaviours are all related to the same threat, i.e. that participants did not evaluate or understand the content of items before responding. Non-responding occurs where a participant fails to respond to a certain number of items. Random responding is when an individual provides a random answer without considering the content of the item. Fixed responding occurs when a participant adopts a systematic response approach by providing the

same answer to multiple items, following some kind of pattern. Content-based invalid responding is characterised by two main behaviours, namely over-reporting and under-reporting (Ben-Porath, 2003). These behaviours are defined by an individual providing an answer that is more (over-reporting) or less (under-reporting) severe than the actual situation. This might occur intentionally or unintentionally.

Rothmann (in press) investigated the use of a neural network to detect the random response threat. The neural network was trained with large random and non-random datasets and was then used to classify posterior cases as being random or non-random. The neural network classified 14,15% of individuals as random responders during cross-validation. The posterior classifications (random and non-random) were then subjected to factor analysis, internal consistency and structural equivalence tests. These tests showed that the neural network was accurate in its prediction as the non-random data proved to be valid, reliable and structurally equivalent to other non-random cases (on a group level), while the random data had problems with reliability, validity and structural equivalence. It is, however, still uncertain whether there were individual misclassified cases, because only group level statistical techniques were used.

Another statistical technique that can be useful for reliability and validity in individual cases is the Rasch model (Rasch, 1960), which is widely used for the calibration of measuring instruments in the human sciences (Bond & Fox, 2007). The Rasch model gains information about items measuring a construct based on the responses of a calibration sample of individuals. It then derives an estimate for the latent trait location of each individual and provides fit statistics for each item and each person. These fit statistics provide the researcher with more information on how well the items and individuals fit the Rasch model.

When an individual does not fit the Rasch model, it means that the individual did not respond to the items as would be expected, given the difficulties of the items and the individual's standing on the trait. Linacre (2002), Smith (1996) and Bond and Fox (2007) provide reasons for unacceptable person-fit statistics. These reasons are in line with the threats to protocol validity, building a case that fit statistics could be used as protocol validity indicators. If the responses to the items are not as would be expected, one could expect that one of the threats of protocol validity was evident. Although the fit statistics would not necessarily be able to determine the possible threat, they could provide evidence of a threat. Given that the Rasch

model provides information for individual cases, it would also be valuable to investigate the fit statistics for the Rothmann (in press) neural network to further investigate its prediction success.

The aim of this study was to evaluate the Rasch fit statistics for use as protocol validity indicators in the SAEHWS, and to compare these fit statistics to the neural network technique.

The Rasch model

In psychometrics, there are essentially two large bodies of theory to guide instrument development, namely Classical Test Theory (CTT) and Item Response Theory (IRT; Embretson & Reise, 2000). CTT was initially the leading theory for the development and analysis of standardised instruments. IRT has now replaced CTT to a large extent and has become the major theoretical framework in instrument development (Hambleton & Rogers, 1990).

CTT is based on the assumption that an individual has a true score and an observed score (Allen & Yen, 2002). The difference between the true score and the observed score is due to measurement error. CTT principles are evident in measurement methods ranging from reliability assessment and confirmatory factor analysis to scale development procedures. An advantage of CTT is that it relies on few assumptions; it is also relatively easy to interpret (Allen & Yen, 2002). The main criticism of CTT is that the observed score relies on the content of the test, meaning that individuals with similar trait levels may score differently depending on the item bias (Fan, 1998).

IRT was originally developed to overcome the problems associated with CTT (Hambleton & Rogers, 1990). The basic assumption of IRT is that the latent trait (the characteristic underlying the total scores) is independent of the content of the instrument. This implies that different items could be used to estimate the trait level for different individuals and the latent trait score would still be comparable. According to IRT, an individual with a high level of the trait being measured should have a high probability of endorsing the items (Fan, 1998). There are different IRT models with varying complexity. These models include the one-, two- and three-parameter IRT models.

The Rasch model is a unidimensional, latent trait model developed by the Danish mathematician, Georg Rasch (Rasch, 1960). It was originally developed, in its simplest form, for dichotomous data in the educational setting to calibrate tests in order to determine item difficulty and to derive person ability (Bond & Fox, 2007). Although some see the Rasch model as a special case of IRT, others (see Shaw, 1991) argue that the Rasch model is a practically and theoretically unique archetype. Nevertheless, the Rasch model is related to the paradigm of IRT (Andrich, 1989).

The Rasch model emphasises the basic criterion of invariance, which is a crucial feature of fundamental measurement (Bond & Fox, 2007). Invariance means that an instrument is required to work in the same way for all individuals. This implies invariant functioning across any group of respondents. For example, when comparing heights of men and women, it is assumed that the measuring tape works the same way for both genders.

The Rasch model states that the probability of a person to correctly answer an item is a logistic function of the person's ability minus the item difficulty (Bond & Fox, 2007). Each item is weighted according to how it is rated in the sample. Person ability is then derived from the answers to the items given the properties of the items. It is important to note that person ability refers to the level of the construct being measured (e.g. exhaustion). In the context of perception or attitude measurement, item difficulty refers to the intensity of the item rather than the difficulty of the item. Thus, certain items measure the variable (e.g. exhaustion) more intensely than others.

The model used in this article is a special case of the Rasch model, called the rating scale model (Andrich, 1978). The rating scale model allows for the analysis of polytomous or Likert scale data (Smith, 1996). The basic functioning of the Rasch model still applies to this model, except that the different categories on the scale are also weighted and taken into account (Bond & Fox, 2007). In this model, a person with the same ability (or level of the construct) will respond differently for items of different difficulties (intensities). These difficulties are weighted by extracting the thresholds between categories on the Likert scale. The threshold is concerned with the construct intensity required to rate a category higher on the scale (Shaw, Wright, & Linacre, 1992). The probability of a person to select a certain point on a scale is thus a logistic function of the person ability minus the item difficulty, plus

the difficulty of the threshold between the current scale category and the next category (Bond & Fox, 2007).

The required structure of response is a probabilistic Guttman pattern, which implies that for the same person ability, the probability to endorse an easy item has to be higher than the probability to endorse a more difficult item, and vice versa (Bond & Fox, 2007). When comparing persons with different abilities, a person with a higher ability is expected to endorse all items endorsed by a person with lower ability and additionally one or more difficult items. Thus, only certain response patterns are in accordance with the Guttman pattern. Since the responses are not necessarily required to be deterministic, but rather probabilistic, there is room for random variation. Bond and Fox (2007) explain that all response strings are possible; some are just less probable than others.

There are two popular chi-square-based fit statistics to determine how probable a person's responses are, namely infit and outfit. These statistics are reported as mean-squares in the form of chi-square statistics divided by their degrees of freedom (Linacre, 2002). The outfit statistic is an outlier-sensitive fit statistic that is more sensitive to responses where the item difficulty and person ability differs drastically (Linacre, 2002). Outfit is based on the conventional sum of squared standardised residuals. Meaning that, for each person, the standardised residuals (observed minus expected response) for the items are squared, summed and divided by the number of items (Bond & Fox, 2007). The infit statistic is an information weighted (inlier-sensitive) fit statistic that is more sensitive to responses where item difficulty and person ability are matched (Linacre, 2002). Infit is calculated like the outfit statistic, except that each squared standardised residual is first weighted by its variance before it is summed. The total is then divided by the sum of variances (Bond & Fox, 2007).

Linacre (2002) explains that outfit statistics are more likely to indicate lucky guesses and careless mistakes, while infit statistics report smaller differences in the comprehension of the items. Infit statistics are usually used during scale construction to identify problems with the measurement items. It is for this reason that infit was used to assess the fit of the Exhaustion items. Outfit, on the other hand, is used to identify problematic cases to discard during scale construction, and will thus be used as an indicator of person-fit for the purposes of protocol validity.

Table 1 contains recommendations from Wright and Linacre (1994) for the interpretation of parameter-level mean-square fit statistics. These recommendations follow informal simulation studies and experience in analysing a large number of datasets.

Table 1

Interpretation of Parameter-level Mean-square Fit Statistics

Mean square	**Implication for Measurement**
> 2,00	Distorts or degrades the measurement system.
1,50 – 2,00	Unproductive for construction of measurement, but not degrading.
0,50 – 1,50	Productive for measurement.
< 0,50	Less productive for measurement, but not degrading. May produce misleadingly high reliability coefficients.

Mean-square values greater than 2,00 indicate distortion in the measurement, possibly posing problems for protocol validity. Smith (1996) provides reasons for these high mean-square values. Firstly, respondents could have avoided engaging in the rating scale, creating a problematic relationship between the responses and the construct. Secondly, the items might have been misunderstood. Finally, respondents might have responded randomly to the items. These reasons help build the case as to why fit statistics could be used as protocol validity indicators.

Values between 1,50 and 2,00 are unproductive for measurement, but do not imply a corrupt case. This category defines a grey area in the fit statistics. Some of the cases in this category might still be valid, while others, depending on the measurement structure, might be due to problematic responses (see Smith, 1996). Mean-square values between 0,50 and 1,50 imply that the case fits the Rasch model optimally, and would thus also indicate good protocol validity.

Values below 0,50 indicate overfitting, meaning that the responses are not corrupted, but too predictable. Although they are too predictable, producing misleadingly high reliability coefficients, they still agree with the model (Smith, 1996). This category is more applicable to the creation of the measurement system, as it provides information that the categories of

the rating scale are not being used with enough variance. For the purposes of this article, the overfitting cases were regarded as valid.

METHOD

Research design

A survey design was used to reach the specific research objectives (Huysamen, 2001). In this type of research, data is collected by posing questions and recording people's responses. The data will be gathered from a survey data archive (see Whitley, 2002). The survey archive contains responses to survey questions in the well-being audits and demographic data concerning the respondents. This data is kept on computer databases. Survey archives are useful because they have been collected for research purposes; consequently, great care was taken to ensure the reliability and validity of the data.

Participants

To make the results in this article comparable with the results from Rothmann (in press), the same study population was used. The study population consisted of 11 097 participants from several industries in South Africa, including financial, engineering, mining, human resources and manufacturing. In Rothmann (in press), the large sample (n=11097) was divided into two smaller samples by means of random sampling. Sample 1 (n=5549) would be used to build models, and sample 2 (n=5548) for cross-validation. In the present study, the cross-validation sample 2 was used to calculate the fit statistics and compare them to the neural network developed by Rothmann (in press). Table 2 shows the characteristics of the samples. The mean age of the participants was 40,81 (SD=10,02). There were 3 394 (61,18%) male respondents. Looking at racial groups, 25,34% of the respondents were black and 39,26% white. In terms of qualification, the sample included 2 729 (49,19%) respondents with grade 12 or lower .

Table 2
Characteristics of the Participants

Item	Category	Sample ($n = 5548$)	
		Frequency	*Percentage*
Gender	Male	3394	61,18
	Female	2154	38,83
Race	Black	1406	25,34
	White	2178	39,26
	Coloured	230	4,15
	Indian	117	2,11
	Other	6	0,11
	Missing	1611	29,00
Qualification	Up to grade 12	2729	49,19
	Certificate	695	12,53
	Diploma or degree	838	15,10
	Postgraduate qualification	622	11,21
	Missing	664	11,97

In addition, Rothmann (in press) generated 5548 (equal to sample 2) random cases from a uniform distribution to compare results with. This sample was referred to as the random sample, while the study population was referred to as the non-random sample.

Measuring instrument

One subscale of the South African Employee Health and Wellness Survey (SAEHWS), namely Exhaustion, was used for the purposes of this study. The SAEHWS is a self-report instrument based on the dual-process model of work-related well-being (Rothmann & Rothmann, 2006). The SAEHWS is based on the assumption that employees' perceptions and experiences represent important information regarding the wellness climate in the organisation. The SAEHWS instrument measures an employee's health and wellness status, relates the data to the organisational climate and compares the results to the South African norm (Rothmann & Rothmann, 2006). The SAEHWS measures organisational climate, wellness, health and lifestyle, organisational commitment, and personal variables. Exhaustion was measured with 5 items (like "I feel tired before I arrive at work") on a 7-point scale

varying from 0 (*never*) to 6 (*always*). Exploratory and confirmatory analyses were used to assess the factor structures of all the components of the measurement model of the SAEHWS (Rothmann & Rothmann, 2006). Furthermore, it was found that the factor structures of the measuring instruments are equivalent for different ethnic groups and organisations. The internal consistencies of all the subscales of the SAEHWS are acceptable ($\alpha > 0{,}70$).

Statistical analyses

Statistical analyses were conducted with SPSS 16.0 (SPSS, 2008) and Winsteps 3.68 (Linacre, 2009). Descriptive statistics (e.g. means and standard deviations) were used. Pearson's product-moment correlation (Tabachnick & Fidell, 2001) was used to investigate the relationship between variables. Exploratory factor analyses, specifically principal component analyses (Kline, 1994), were conducted to determine the validity of the construct that was measured in this study. Coefficient alpha (Cronbach, 1951) was used to assess reliability, as it contains important information regarding the proportion of variance of the items of a scale in terms of the total variance explained by the particular scale.

In addition, the following statistics were used in conjunction with the Rasch model (Bond & Fox, 2007). First, Rasch reliability was used to provide an estimate of the reproducibility of the measures. Rasch reliability is a more conservative estimate for the ratio of real person variance than Cronbach's α (Linacre, 2002). Second, item measures (indicated by δ) were used to assess the severity of the item's measurement of the latent construct. Last, infit statistics were used to assess how accurately or predictably the items fit the Rasch model and outfit statistics were used to assess person-fit for the purposes of protocol validity.

The Multilayer Perceptron (MLP) neural network was used for predictive modelling. The MLP is a feed-forward neural network that can be trained to store knowledge, based on the relationship between the dependent and independent variables, and to predict values for posterior cases. Neural networks were used for the following reasons:

- Neural networks can approximate a linear or non-linear relationship, depending on the relationship in the data (Haykin, 1998).
- A model does not have to be hypothesised in advance (Haykin, 1998).
- Minimal demands are made on assumptions (SPSS, 2008).

Cross-validation (Tabachnick & Fidell, 2001) was used to ensure repeatability by testing the model against an unknown sample. If the model is based on one sample and tested against an unknown sample of cases, the efficiency of the model can be determined for classifying posterior unknown cases.

Tucker's coefficient of congruence phi (ϕ) was used to compute structural equivalence between factors for different samples (Tucker, 1951). Structural equivalence can be used to prove differences in factor structures for non-random and random predicted cases, confirming the validity of the neural network prediction. Tucker's ϕ is defined by the following formula:

$$\phi = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$

In this formula, x_i and y_i represent the respective component loadings. Tucker's ϕ ranges from -1,00 via 0 to +1,00 (perfect similarity). Values above 0,95 can be taken to indicate factorial similarities, while values below 0,85 show unavoidable incongruencies (Van de Vijver & Leung, 1997).

The better-than-chance effect size index I was used to determine the success of the neural network (Huberty & Lowman, 2000). This index adjusts the observed hit rate of a category for incidental correct classification of cases. In other words, it indicates if the classification was correct by chance or not. The better-than-chance index is calculated by the following formula:

$$I = \frac{H_o - H_e}{1 - H_e}$$

In this formula, H_o represents the observed hit rate (correct classifications divided by total cases), while H_e represents the chance rate, which is the proportional prior probabilities of classification. Huberty and Lowman (2000) provides guidelines for the interpretation of I. Values below 0,10 could be seen as a small effect, while values above 0,35 represent a large effect.

Research procedure

First, the Exhaustion items on both datasets were tested against the Rasch model with the Winsteps program. The item and person measures and the fit statistics were then evaluated to investigate whether there were problems with the Exhaustion dimension. If problems with the fit statistics were evident for the Exhaustion items, the person-fit indicators would be less reliable (Bond & Fox, 2007), which would be problematic for the aim of this study. The person measures and fit statistics were then exported from Winsteps for further statistical analyses.

Subsequently, the cross-validation data in the sample was categorised according to the person outfit mean-square statistics. The categorisation was based on the criteria for fit described in Wright and Linacre (1994). Outfit mean-square statistics below 0,50 defined the overfitting category, 0,50 to 1,50 the good fit category, 1,50 to 2,00 the slightly underfitting category, and greater than 2,00 the underfitting category. Descriptive statistics, reliability and validity were then assessed for each of the categories. Structural equivalence was computed for each of the categories with every other category.

The random sample was also categorised according to the person outfit mean-square statistics for comparison with the non-random data. The same categorisation of outfit statistics was followed. Descriptive statistics, reliability and validity were assessed for each of the categories and structural equivalence tests were conducted on every category with its non-random counterpart.

Lastly, the outfit statistics were compared with the outcomes of the Rothmann (in press) neural network. The comparison was done by assessing outfit descriptive statistics for the neural network predicted categories and a cross tabulation of the outfit categories versus the neural network categories. Furthermore, internal consistency, a principal component analysis and structural equivalence were computed for each of the outfit categories within each neural network prediction to get a clearer picture of the nature of the cases the neural network misclassified. The better-than-chance index I was also calculated to determine the neural network's prediction success in spite of misclassified outfit statistics.

RESULTS

Calculating Rasch model fit statistics

The Rasch model fit statistics were computed on the 11 097 cases with the Winsteps program (Linacre, 2009). The Rasch item reliability was 1,00, providing evidence that the items would measure similarly if measured again. Table 3 shows the item measure, infit and outfit statistics for each item.

Table 3

Rasch Model Item Fit Statistics for Exhaustion

Item	δ	Infit Mean Square	Outfit Mean Square
Item 1	-0,24	1,00	0,99
Item 2	-0,16	0,94	0,93
Item 3	0,79	1,26	1,22
Item 4	-0,45	0,99	0,98
Item 5	0,06	0,82	0,81

It is evident that Item 3 is the most severe measurement of Exhaustion (δ = 0,79) and Item 4 the least severe (δ = -0,45). The criterion for acceptable item fit was chosen as an infit value of below 1,30 and above 0,75 (Bond & Fox, 2007). All items fulfilled the infit criteria, indicating that the items contribute to a single underlying construct. Even the oufit values, which tend to be more volatile due to their sensitivity for outliers, match these criteria.

In terms of person statistics, the Rasch person reliability is 0,84. The mean person measure (δ) for exhaustion levels amounts to -0,31 (*SD* = 1,14). The root mean square error (RMSE) of the model is 0,45. The mean of both the infit and outfit person statistics are 0,99 (*SD*=1,12). This provides sufficient evidence about the fit and applicability of the Rasch model for the Exhaustion construct.

Fit statistics, traditional validity and reliability on non-random data

Table 4 contains descriptive statistics, reliability and validity measures for different outfit mean-square categories for the cross-validation non-random data. The statistics for each of the categories are used to investigate the validity and reliability for different levels of fit according to the Rasch outfit statistics.

Table 4

Descriptive statistics, Cronbach α and Factor Analysis for Different Outfit Categories on Non-random data

Outfit mean-square	*N*	%	Mean	*SD*	α	Components extracted	Component % variance
< 0,50	2436	43,90	0,26	0,13	0,94	1	81,80
0,50 - 1,50	2032	36,63	0,88	0,28	0,86	1	64,08
1,50 - 2,00	380	6,86	1,70	0,15	0,73	1	48,43
> 2,00	700	12,61	3,41	1,40	0,30	2	30,47

The largest number of cases (43,90%) has an outfit mean-square below 0,50, with a mean of 0,26 (*SD* = 0,13). When introduced in a principal component analysis, one component is extracted (Eigenvalues > 1), explaining 81,80% of the variance. These cases produce a Cronbach α of 0,94. A total of 2 032 cases (36,63%) has an outfit mean-square between 0,50 and 1,50, with a mean of 0,88 (*SD* = 0,28) and a Cronbach α of 0,86, also extracting one component explaining 64,08% of the variance. The slightly underfitting category (outfit mean-square between 1,50 and 2,00) consists of only 6,86% of the cases with an outfit mean of 1,70 (*SD* = 0,15). Interestingly, these cases still yield one component that explains 48,43% of the variance, and contribute to an alpha of 0,73. 12,61% of the cases had an outfit mean-square greater than 2,00, with a mean of 3,41 (*SD*=1,40). These cases yield two components, with the first component only explaining 30,47% of the variance. These cases show low internal consistency ($\alpha = 0{,}30$).

Table 5 contains the structural equivalence (Tucker's ϕ) between the outfit categories. It is evident that the component loadings are all highly equivalent ($\phi \geq 0{,}95$), except for the outfit mean-square statistics above 2,00. This category shows unavoidable incongruencies with the

other categories (ϕ <0,85). Because this category yields a two-factor model, Tucker's ϕ was calculated on the first factor.

Table 5

Tucker's ϕ of the Outfit Mean-square Categories

Infit mean-square	1	2	3
1. < 0,50	1,00	-	-
2. 0,50 - 1,50	1,00	1,00	-
3. 1,50 - 2,00	0,99	1,00	1,00
4. > 2,00	0,72	0,73	0,75

Fit statistics, traditional validity and reliability on random data

Table 6 contains descriptive statistics, reliability and validity measures for different outfit mean-square categories for the cross-validation random data.

Table 6

Descriptive Statistics, Cronbach Alpha and Factor Analysis for Different Outfit Categories on Random Data

Outfit mean-square	*n*	%	Mean	*SD*	α	Components extracted	Component % variance
< 0,50	116	2,09	0,34	0,13	0,95	1	82,14
0,50 - 1,50	919	16,56	1,05	0,28	0,78	1	53,69
1,50 - 2,00	599	10,80	1,76	0,15	0,44	1	31,23
> 2,00	3914	70,55	3,82	1,36	-0,57	4	22,33

Most of the random cases (70,55%) have an outfit value greater than 2,00, with a mean of 3,82 (*SD* = 1,36). There is indeed low internal consistency (α < 0,70) and, as would be expected from random data, four components are extracted from a principal component analysis. The first component explains 22,33% of the variance. A total of 599 (10,80%) cases have outfit values between 1,50 and 2,00, with a mean of 1,76 (*SD* = 0,15). In contrast with the non-random data, these cases also lack internal consistency (α < 0,70), but still extract one component that explains 31,23% of the variance. A total of 919 (16,56%) of the cases fit

satisfactorily with an outfit of between 0,50 and 1,50 (mean = 1,05; *SD* = 0,28). These cases produce internally consistent responses and yield one component explaining 53,69% of the variance in a principal component analysis. Only 116 (2,09%) cases have an outfit mean square below 0,50, with a mean of 0,34 (*SD* = 0,13). These cases have high internal consistency ($\alpha = 0{,}95$) and yield one component that explains 82,14% of the variance.

Table 7 shows the structural equivalence between the principal component analyses for the non-random (Table 4) and the random data (Table 6). It is evident that the only category that is not equivalent for both the non-random and random data is where outfit is greater than 2,00 ($\phi < 0{,}95$).

Table 7

Structural Equivalence between Non-random and Random Outfit Categories

Outfit mean-square	Tucker's ϕ
< 0,50	1,00
0,50 - 1,50	1,00
1,50 - 2,00	0,97
> 2,00	-0,15

Rasch model fit statistics and neural network comparison

Table 8 contains the descriptive statistics of the outfit mean-square statistics for the different classifications of the Rothmann (in press) neural network. Outfit mean-squares range from 0,02 to 7,61, with a mean of 0,71 (*SD* = 0,74) for the non-random classified cases. The random classified cases have a mean outfit mean-square of 2,64 (*SD* = 1,59) ranging from 0,36 to 9,90.

Table 8

Outfit Descriptive Statistics of the Different Neural Network Classifications

Neural network classification	*n*	Outfit Minimum	Outfit Maximum	Outfit Mean	Outfit *SD*
Non-random	4763	0,02	7,61	0,71	0,74
Random	785	0,36	9,90	2,64	1,59

Table 9 shows the cross tabulation of the neural network predicted categories against the outfit categories. Only 5,31% of the non-random predicted cases have an outfit mean-square statistic above 2,00, while 55,92% of the random predicted cases have an outfit mean-square statistic above 2,00. Surprisingly, 23,31% of the random cases indicate good fit.

Table 9

Cross Tabulation of the Neural Network Prediction versus the Outfit Categories

Neural network prediction	**Rasch model Outfit statistics**							
	< 0,50		*0,50 - 1,50*		*1,50 - 2,00*		*> 2,00*	
	Freq	%	Freq	%	Freq	%	Freq	%
Non-random	2440	51,23	1852	38,88	218	4,58	253	5,31
Random	3	0,38	183	23,31	160	20,38	439	55,92

To determine the success of the neural network in spite of the misclassifications (in Table 9), the better-than-chance index was calculated. Outfit values greater than 2,00 were seen as misclassification in the non-random category, and values smaller than 2,00 defined misclassification in the random category. If the prior probabilities are set to the probability of an outfit value being greater or smaller than 2,00 in the total non-random sample (see Table 4), the total chance-hit rate is 0,78. The observed hit rate is 0,89. This provides us with a better-than-chance index of 0,51, which still shows a large effect ($I \geq 0,35$).

Table 10 contains the internal consistencies, principal component analyses and structural equivalence of the outfit categories for the neural network predictions. The non-random predicted cases show acceptable internal consistencies and component structures for the outfit categories below 1,50. The random predicted cases show problematic component structures and low internal consistency for the cases with outfit values greater than 2,00.

The non-random predicted cases with higher outfit values and the random cases with lower outfit values are of interest. The cases with outfit values between 1,50 and 2,00 prove to be internally consistent for the non-random predicted cases (α=0,76), but not for the random predicted cases (α=0,54). Their component structures, however, are still equivalent to cases with good fit (ϕ=0,99) for both predictions although they extract a second component in a principal component analysis.

It is further evident that the non-random predicted cases with outfit values greater that 2,00 have problems with internal consistency (α=0,60) and structural equivalence (ϕ = 0,79). The random predicted cases that present good outfit statistics between 0,50 and 1,50, show acceptable internal consistency (α=0,78). Interestingly, the structural equivalence is slightly lower (ϕ=0,97) than would be expected, given that the equivalence was tested for other cases within this same category, but is still acceptable.

Table 10

Internal Consistency, Factor Analysis and Structural Equivalence for the Neural network Prediction and Outfit Categories Cross Tabulation

Prediction	**Outfit**	α	**Components***	**First Component Variance %**	ϕ^+
Non-random	< 0,50	0,94	1	81,80	1,00
	0,50 - 1,50	0,87	1	65,07	1,00
	1,50 - 2,00	0,76	2	51,60	0,99
	> 2,00	0,60	2	44,12	0,79
Random	< 0,50	*N/A*	*N/A*	*N/A*	*N/A*
	0,50 - 1,50	0,78	1	53,51	0,97
	1,50 - 2,00	0,54	2	36,07	0,99
	> 2,00	-0,05	3	25,88	0,08

* Number of components with Eigenvalues > 1

$^+$ Tucker's ϕ calculated on the first component against the 0,50 – 1,50 category component for the entire sample.

DISCUSSION

The aim of this study was to evaluate the Rasch fit statistics for use as protocol validity indicators in the SAEHWS and to compare these fit statistics to the Rothmann (in press) neural network technique. It was evident from the literature that the reasons provided for misfitting outfit mean square statistics are related to the threats to protocol validity (Smith, 1996; Linacre, 2002). Furthermore, 87,39% of the non-random data showed outfit values under 2,00 while 70,55% of the random data showed outfit values above 2,00. It also seemed

that outfit values above 2,00 had problems with internal consistency and provided dissimilar structures in principal component analyses.

In comparison to the Rothmann (in press) neural network, 94,69% of the predicted non-random cases proved outfit values less than 2,00. The random prediction showed less success, with only 55,92% of the cases having outfit values larger than 2,00 and 23,31% showing good outfit statistics between 0,50 and 1,50.

When looking at the cases with outfit statistics below 0,50 in both datasets, extremely high internal consistency is evident with α=0,94 for non-random cases and α=0,95 for random cases. Outfit statistics below 0,50 could be seen as overfitting cases, meaning that they are too predictable according to the Rash model (Linacre, 2002). The high internal consistencies are confirmed by Bond and Fox (2007) where they state that the statistical implications of overfitting are deflated standard errors and inflated reliabilities. However, Bond and Fox (2007) point out that there are no practical implications of overfitting. Smith (1996) describes that overfitting cases could be due to the underuse of categories on the scale. This has a practical implication for protocol validity, because a symptom of fixed responding is the underuse of categories on the scale. Fortunately, fixed responding could easily be detected by an algorithm that finds patterns in the data. For these reasons, it could only be assumed that the overfitting cases are valid when controlling for the fixed response threat.

The cases with outfit statistics between 0,50 and 1,50 are described as good fitting cases (Linacre, 2002). Only 36,63% of the non-random cases matched these criteria, providing some evidence that the Exhaustion scale could be optimised, although the item fit statistics are sufficient. On the downside, 16,56% of the random cases also matched these criteria. This provides evidence that the random data could have good fit statistics due to chance. It is important to take note when using the outfit statistics as protocol validity indicators that good fit does not completely eliminate threats to protocol validity. Although certain threats are evident, it is possible that the response pattern could still show good fit. A potential problem with the outfit statistics in this category is fixed responding at the extreme points of the scale – in other words, cases where all the items have been answered with either zeros or sixes. These cases are seen as minimum or maximum item measures by the Rasch model algorithm, and are awarded an outfit mean square of 1,00. This problem could be minimised by wording

items in the opposite direction (semantic opposites; see Goldberg & Kilkowski, 1985), or again by controlling for fixed response with a pattern detection algorithm.

Cases with outfit values between 1,50 and 2,00 were described as unproductive but not degrading (Linacre, 2002). This contrast is confirmed by the results. The non-random sample for these outfit statistics proved internal consistency (α=0,73) while the random sample showed problems (α=0,44). Although the internal consistencies differed between groups, the components extracted in a principal component analysis were equivalent for the non-random and random data ($\phi \geq$0,95). From these findings, it is evident that outfit values between 1,50 and 2,00 should be handled with caution. Moderate threats should be expected although the case could still be valid. Protocol validity threats to this outfit category might include moderate random responding, mildly inconsistent over- or under-reporting or any other threat that causes moderate inconsistency in response.

Analyses of the cases with outfit values above 2,00 provide a good indication of problematic responses. Linacre (2002) describes these cases as distortion to the measurement system. In both the non-random and random data samples, unstable factor structures and low internal consistencies were evident. In addition, there was no structural equivalence of components extracted in the principal component analyses. When cases show outfit values greater than 2,00, threats to protocol validity should be expected. Bond and Fox (2007) describe these cases as too unpredictable, and state that conclusions cannot be drawn about the level of measurement in these specific cases. There is, however, no definite indication of the specific threats. Threats might include severe random responding, over- or under-reporting or any other threat that causes severe inconsistency in response.

One of the limitations of the Rothmann (in press) study was the uncertainty whether certain classified non-random cases were in fact valid and reliable. The results in this article show that only 5,31% of the non-random classified cases were problematic according to the outfit statistics. This serves as evidence that the non-random classification of the neural network performed well, given that it was blindly trained with two datasets, one with responses from individuals and the other with computer-generated random data.

The random prediction was somewhat less accurate, with only 55,92% of the cases that had outfit values greater than 2,00. In fact, 23,31% of the random cases had good fit statistics

between 0,50 and 1,50. The validity and reliability tests on these misclassification categories provided interesting results. It was evident that although the random predicted cases with outfit values between 0,50 and 1,50 were internally consistent, they showed slight (but still acceptable) invariance with non-random cases from that category. The non-random and random predicted cases with outfit values between 1,50 and 2,00 extracted two components, of which the first component was equivalent to cases with outfit values of 0,50 to 1,50. The internal consistency was higher for the non-random predicted cases than for the random predicted cases. This evidence shows that even though the outfit values were similar, there were still some minor differences between the non-random and random data.

Rothmann (in press) calculated the better-than-chance index of the neural network after training. The index was 0,72, which provided evidence that the neural network performed well in classification ($I \geq 0,35$). The better-than-chance index was calculated again in this article to investigate the performance of the neural network given the misclassifications. Although the index was somewhat lower ($I=0,51$) given the misclassifications, the neural network still proved a large effect size ($I \geq 0,35$), providing evidence that the neural network is a good predictor of cases with misfitting outfit statistics.

One reason for the misclassified random cases might be that the neural network was over-trained (see SPSS, 2008), implying that too much noise was inferred during training. This problem could have been solved by increasing the number of cases in the testing sample (see Rothmann, in press). Another reason might be that the random sample used to train the neural network was either insufficient or contradicting. The neural network might have performed better if trained with a sample containing cases with outfit values larger than 2. Of course, this defeats the purpose of using a neural network as a protocol validity indicator if outfit statistics can be calculated. Given the scientific background of the Rasch model (see Bond & Fox, 2007), the literature behind the fit statistics (Smith, 1996; Linacre, 2002) and the evidence in this study, it can be concluded that the fit statistics provide a more accurate indication of general protocol invalidity than the neural network in Rothmann (in press).

The neural network, however, did prove to be reasonably effective in discriminating between datasets. The outfit statistics only provide information about how the case fits the Rasch model. Neural networks can derive a custom model for specific samples, given that it was

trained appropriately. Deriving custom models from samples creates the possibility of using neural networks to detect specific threats within cases.

It is evident that outfit statistics are more powerful indicators of protocol invalidity than validity. Large outfit statistics imply problematic responses, but smaller outfit statistics do not guarantee protocol validity. Certain threats might still be evident, even if the outfit statistics show good fit. Specific protocol validity threats, especially fixed response, should be controlled for in addition to the outfit statistics. Neural networks have the ability to discriminate between datasets, and it might be worth investigating whether neural networks can be used to detect specific threats in conjunction with the outfit statistics, especially when outfit statistics are acceptable.

There were several limitations to this study. Firstly, outfit statistics were only evaluated against random data with a uniform distribution. Different threats were not tested against these outfit statistics. Secondly, the Rasch model was only implemented in the rating scale context and other facets of measurement were not taken into account. Thirdly, methods for calculating outfit statistics on posterior cases without re-running the entire Rasch model algorithm were not investigated. Finally, participants were not followed up to investigate whether they had in fact provided invalid responses.

RECOMMENDATIONS

Rasch model fit statistics have proved to be effective protocol invalidity indicators. It is recommended that these fit statistics be implemented in conjunction with other measures in the SAEHWS. When a profile has high outfit statistics ($\geq$ 2,00), the facilitator can be sure that there was a problem with the validity of the protocol. However, if the outfit statistics are smaller than 2,00, no guarantees can be made about the validity of the protocol. Other measures – such as a pattern recognition algorithm that detects fixed response – should also be evaluated by the facilitator. When any problems are evident, the facilitator should conduct an interview with the participant to determine whether there were problems with the completion of the instrument. The interview can be structured within Ben-Porath's (2003) framework for protocol validity threats by asking questions such as "Did you read all the questions before providing an answer?", "Did you understand all the questions?", and "Do you feel you were totally honest in answering all the questions?".

It is recommended that further research be conducted to investigate the usefulness of fit statistics as indicators of protocol validity. It would be beneficial to gather data where specific threats were evident. Participants could be asked to complete an instrument accurately, randomly, time-constrained and from sketched scenarios where they would over- and under-report. Fit statistics should be investigated for these different threats. It would then also be beneficial to train similar neural networks with data from specific threats to determine whether it can be used to test a protocol for the existence of a specific protocol validity threat. These neural networks can then be used in accordance with the outfit statistics.

In addition, less computationally intensive methods should be investigated for calculating outfit statistics on posterior cases. If certain values are pre-calculated on a large sample and stored, it should be possible to determine the outfit of a posterior case with fewer calculations. It might also be possible to train a neural network to predict outfit statistics from the answers on the items.

Further research should be done into the multi-faceted Rasch model (Bond & Fox, 2007) and its fit statistics. This model is an extension of the unidimensional Rasch model that takes other aspects of measurement into account (Bond & Fox, 2007). The SAEHWS is supported by a causal model (Rothmann & Rothmann, 2006), making it possible to include the causes of certain dimensions as facets to be controlled for in the multi-faceted Rasch model. When including more dimensions, the fit statistics will have a more holistic picture of the context of certain responses and might serve as more powerful protocol validity indicators.

REFERENCES

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory.* Long Grove, IL: Waveland Press.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika,* 43, 561-73.

Andrich, D. (1989), Distinctions between assumptions and requirements in measurement in the Social sciences. In J. A. Keats, R. Taft, R. A. Heath & S. Lovibond (Eds.), *Mathematical and theoretical systems* (pp. 7-16). North Holland, Amsterdam: Elsevier Science Publishers.

Ben-Porath, Y. S. (2003). Self-report inventories: Assessing personality and psycho-pathology. In J. R. Graham & J. Naglieri (Eds.) Vol. X: *Handbook of assessment psychology* (pp. 554-575). New York: Wiley.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah NJ: Lawrence Erlbaum Associates.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58,* 357-381.

Goldberg, L. R., & Kilkowski, J. M. (1985). The prediction of semantic consistency in self-descriptions: Characteristics of persons and of terms that affect the consistency of responses to synonym and antonym pairs. *Journal of Personality and Social Psychology, 48,* 82-98.

Hambleton, R. K., & Rogers, J. H. (1990). Using item response models in educational assessments. In W. Schreiber & K. Ingenkamp (Eds.), *International developments in large-scale assessment* (pp. 155-184). England: NFER-Nelson.

Haykin, S. (1998). *Neural networks: A comprehensive foundation* (2nd ed.). New York: Macmillan College Publishing.

Huberty, C.J. & Lowman, L.L. (2000). Group overlap as a basis for effect size. *Educational and Psychological Measurement,* 60, 543-563.

Huysamen, G. K. (2001). *Methodology for the social and behavioural sciences.* Cape Town: Oxford University Press.

Kline, P. (1994). *An easy guide to factor analysis.* London: Routledge.

Linacre, J. M. (2002). Cronbach alpha or Rasch reliability: Which tells the "truth"? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2009). *WINSTEPS®: Rasch measurement computer program.* Beaverton, OR: Winsteps.com

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmarks Paedagogiske Institut.

Rothmann, S., & Cooper, C. L. (2008). *Organizational and work psychology.* London: Hodder Education.

Rothmann, S. (2008, April). *Psychological fitness: Concept and measurement.* Paper presented at the SASOM Conference, Pretoria.

Rothmann, S. (in press). The use of neural networks to establish protocol validity.

Rothmann, J. C., & Rothmann, S. (2006). *The South African Employee Health and Wellness Survey: User manual.* Potchefstroom: Afriforte (Pty) Ltd.

Scherbaum, C. A., Finlinson, S., Barden, K., & Tamanini, K. (2006). Applications of item response theory to measurement issues in leadership research. *The Leadership Quarterly, 17*, 366-386.

Shaw, F. (1991). Descriptive IRT versus prescriptive Rasch. *Rasch Measurement Transactions, 5*(1), 131.

Shaw, F., Wright, B., & Linacre, J. M. (1992). Disordered steps? *Rasch Measurement Transactions, 6*(2), 225.

Smith, R. M. (1996). Polytomous mean-square fit statistics. *Rasch Measurement Transactions, 10*(3), 516-517.

SPSS Inc. (2008). *SPSS 16.0 for Windows*. Chicago, IL: SPSS Inc.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Needham Heights, MA: Allyn & Bacon.

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies.* Personnel Research Section Report No. 984. Washington, DC: Department of the Army.

Van de Vijver, F. J. R., & Leung, K. (1997). *Method and data analysis for cross-cultural research*. Beverly Hills, CA: Sage.

Whitley, B. E. (2002). *Principles of research in behavioral science* (2nd ed.). Boston, MA: McGraw-Hill.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*, 370.

CHAPTER 4

CONCLUSIONS, LIMITATIONS AND RECOMMENDATIONS

In this chapter, conclusions are made concerning the literature findings and the results of the empirical study. Furthermore, the limitations of the present study are discussed and recommendations are presented to the organisations and with a view to future research.

4.1 CONCLUSIONS

The general aim of this study was to establish an efficient, real-time method/indicator for determining protocol validity in web-based instruments. More specifically, the objectives were to 1) study a framework for the threats to protocol validity; 2) develop and evaluate a protocol validity indicator that detects random responses in electronic well-being surveys; 3) evaluate the IRT fit statistics for use as protocol validity indicators 4) compare the IRT fit statistics with the developed protocol validity indicator; and 5) discuss the practical implications of implementing the protocol validity indicators in an online wellness instrument.

The following conclusions can be made regarding the specific objectives:

A framework for the threats of protocol validity

Knowing the possible threats to protocol validity is fundamental in the application of a measuring instrument. If threats are to be anticipated and detected, one should have a framework for these threats. Different threats also have different causes and consequences for instrument scoring. The threats are mainly divided into categories that reflect the role of item content in the response behaviour. Ben-Porath (2003) classifies the threats to protocol validity into two broad categories, namely non-content-based invalid responding and content-based invalid responding. These categories reflect the role of the instrument item content in invalid responding.

Non-content based responding is where the item meaning and content plays no role in the response provided by the participant and includes mainly three different behaviours, namely

non-responding, fixed responding and random responding. There might be several reasons for these behaviours. Typically, an unwilling participant might show fixed and random response behaviours with varying intensities throughout the instrument. A participant that does not understand the items due to language or comprehension deficiencies might also respond randomly (even if it is not intentionally), or fail to respond at all. Instrument presentation also plays a role because response recording errors might present themselves as random responses.

Content-based invalid responding refers to participants portraying themselves different to the real situation. Participants with a certain motive may intentionally try to distort their responses in a positive or negative way. The distortion might be total fabrication or mere exaggeration. Unfortunately, intentional distortion is not the only problem. Self-report instruments rely on individuals' subjective perceptions of the instrument content. There are several factors that can influence that perception. Negative emotionality is a personality trait that causes individuals to perceive their environment more threatening than it really is. Other self-misperceptions or social desirability may also cause participants to minimise psychological difficulties or negative characteristics.

Minimising these threats is not just about detecting them. It should be ensured that instrument content are valid, unbiased, non-threatening and easily interpretable. The presentation of the instrument should also minimise the opportunity for error. Participants should also understand why they are completing the instrument and who the applicable parties are that have access to results, in order to gain cooperation and honesty.

The development and evaluation of a protocol validity indicator

A protocol validity indicator was developed by utilising neural networks. Neural networking is a technology that is mainly used in the field of data mining to infer models from a large dataset and predict posterior values based on that model. In this study, neural networks was used to identify random response in individual cases by classifying it as belonging to either a random or non-random sample of cases. The random and non-random samples differed in properties like response distribution. Neural networks showed that they were effective in discriminating between datasets with different properties. This is mainly due to the non-linear inference capability of neural networks.

It is important to understand the paradigm of using neural networks for discrimination. Where other approaches might assume a model and fit the data to the model, neural networks infer models from the data. This means that every model will be unique, depending on the data used. This stresses the critical role the datasets have in ensuring the success of the neural network. In this study, computer-generated random data was used to train the neural network. Although the neural network did an acceptable job in discriminating between random and non-random data, it still made some errors due to the overlap in dataset properties.

The overlap between datasets (in this study) is a function of the odds of a random response string being potentially non-random. Using a larger amount of items will decrease the overlap between datasets. Items that one would strongly expect to relate in a certain way to other items would also reduce overlap. One advantage of neural networks is that a testing sample can be included during training that will reduce the noise inferred in the model, thereby also minimising the effect of overlapping properties between datasets.

It would be safe to conclude that a neural network is an effective tool for determining potential random responses given certain considerations. Firstly, enough data should be available. Secondly, a rejection dataset should be identified where the threat was evident. As in this study, a random dataset could always be generated to train the neural network to discriminate, but it is still unclear whether random responses by participants would be similar to generated random data. Thirdly, the rejection dataset should not have a large amount of overlap with the valid dataset. When these considerations are met, neural networks can show optimum performance in discriminating between datasets for the determination of potential random responses.

The evaluation of IRT fit statistics for use as protocol validity indicators

In contrast to the neural network paradigm of inferring a model from data, the fit statistics calculated in IRT illustrate to what degree an individual's responses fit the expected, predefined IRT model. This model is customised by item parameters such as the intensity of the construct measured by the item (termed item difficulty in IRT). In this study the Rasch model fit statistics were used. The Rasch model relates to the one-parameter (1-PL) IRT model, because it only takes item difficulty into account (in conjunction with person ability).

Fit statistics can either overfit or underfit the predefined model. Overfit refers to overly predictable responses while underfit refers to unpredictable responses.

In this study it was investigated whether the fit statistics provide an adequate indication of protocol validity. The result was that when responses underfit, protocol validity issues may be expected. These underfitting responses are unpredictable by nature and one would have expected the respondent to answer differently. Overfitting responses also have an implication because it shows the possible underuse of categories on a scale. Fixed response would typically result in overfitting responses. The problem is that not all overfitting responses are invalid, especially if a Likert scale is used. In fact, when using Cronbach's alpha, one typically wants responses to different items of the same dimension to not vary too much. Thus, in certain cases, overfitting responses could also be valid. Adequate fitting responses seemed to be less problematic, but this does not guarantee protocol validity. Participants that get a maximum or minimum score on the dimension are awarded perfect fit. The Rasch model cannot compute fit statistics for these scores and participants are given the benefit of the doubt. The problem is that fixed response at the extreme end of the scale will also result in a minimum or maximum score, and the responses will be considered a perfect fit.

Given these considerations it can be concluded that fit statistics are better indicators of protocol invalidity than protocol validity because validity could not be guaranteed by adequate or overfitting fit statistics. In addition, these fit statistics would only be good indicators of protocol invalidity when used in conjunction with an algorithm that detects fixed responses. Fixed response is not specific to a dimension, meaning that it occurs on a page regardless of the dimension measured. This makes it possible to design a pattern recognition algorithm that can detect fixed response in electronic instruments by looking at all the responses throughout the instrument.

Comparing IRT fit statistics with the neural network

In comparison to the neural network technique, it may be argued that fit statistics are the superior indicators of protocol invalidity. This is mainly because the fit statistics have a more structured, scientific way of determining invalid responses. Certain items measure the construct more intense than others, and one would expect that posterior responses should also fit this pattern. The neural network creates a model from the data, which makes it difficult to

be sure that certain response patterns won't be misclassified, especially because it is difficult to provide for all the considerations of training datasets. While neural networks have potential for other applications in protocol validity, the conclusion is thus that fit statistics provide a better indication of general protocol invalidity. It can, however, not determine what specific threat was evident, but just provide an indication of whether a certain threat (or a collection thereof) has corrupted a case so much that it fails to fit the predefined IRT model.

Practical implications of implementing the protocol validity indicators

In terms of implementing these methods electronically, both the neural network and the IRT fit statistics are computationally intensive methods that would not be possible to calculate by hand. Although computational power is available in modern computers, it is important to note that in the context of internet based surveys computational power is easily limited because of a large number of users utilising the system simultaneously. Algorithms should be optimised to minimise the calculations.

Neural networks utilise extensive computational power to infer a model. The advantage of neural networks is that the model is only inferred once, after which the model is saved in synaptic weights. Posterior classification then only involves a few straightforward algebraic calculations. It is for this reason that neural networks show large potential for implementation in online surveys.

IRT fit statistics has a larger implication for online surveys. The only aspect that can be pre-processed would be the item parameters. After a survey has been completed, the person parameters need to be estimated. The fit statistics can be calculated only after the person parameters have been estimated. Calculating the fit statistics involve the implementation of an algorithm like Maximum Likelihood Estimation (MLE). MLE iterates to find the parameters until it converges to a certain criterion. Although modern computers can implement MLE gracefully, the necessary computational power can increase drastically as the number of concurrent users increase, especially if MLE needs to estimate based on a large number of polytomous items. Methods should be investigated that guarantees instant convergence of the algorithm used. If effective methods are implemented, the fit statistics also have great potential for implementation in online surveys.

4.2 LIMITATIONS OF THE PRESENT STUDY

Apart from the limitations mentioned in articles 1 and 2 there are limitations to the study in general. The main limitation is that only statistical methods were investigated for use as potential protocol validity indicators. Therefore there was not a specific focus on content-based indicators such as social desirability scales. Another limitation is that the data used from the survey archive did not have any indication on which specific threats were evident. Participants that were identified with invalid protocols were not followed up to determine if they had in fact responded invalidly. Finally, this study focused on only one dimension of a large instrument. It is unsure how other dimensions would behave with the methods investigated in this study.

4.3 RECOMMENDATIONS

Recommendations pertaining to the organisation as well as recommendations for future research are made in this section.

4.3.1 Recommendations for the organisation

It is recommended that an algorithm be implemented that calculates the fit statistics online directly after completion of both the SAPFI and the SAEHWS. The indicator should present itself on the report for interpretation by facilitators. Facilitators should evaluate the indicator before making decisions based on the results of the survey. Because the indicator can't provide for the specific threats evident, the facilitator should conduct an interview with the participant to investigate possible threats to the validity of the protocol. This interview can be structured within the framework for protocol validity threats (Ben-Porath, 2003) by asking questions, such as "Did you read all the questions before providing an answer?", "Did you understand all the questions?", and "Do you feel you were totally honest in answering all the questions?".

4.3.2 Recommendations for future research

The following recommendations can be made for future research:

- The effect of predicting validity and reliability with more than five items as inputs in the neural network should be investigated. It would be beneficial to include semantically opposite items in this research, as more items would minimise the probability of random data being valid and reliable by coincidence. The relationship between the size of the training sample, the number of independent variables and the prediction success of the neural network should be investigated. This should include an attempt to determine the optimal number of cases that would prevent overtraining, yet provide the neural network with enough information to make accurate predictions.
- It should be investigated whether it is more effective to include items for several constructs into one neural network for use as a validity and reliability score, or to use multiple neural networks for each construct for creating an aggregated validity and reliability score.
- It is proposed that the relationship between the validity, reliability and pseudo-probabilities be investigated. If a linear relationship is found, the pseudo-probabilities could potentially be used to indicate a severity score for protocol validity. Norms, benchmarks and cut-off points should also be considered for the pseudo-probabilities. These norms, benchmarks and cut-off points can then be used by facilitators to decide about the validity and reliability of cases in different situations.
- Research that focuses on the existence of specific threats needs to be conducted. For example, participants could be asked to complete an instrument accurately, randomly, time-constrained and from sketched scenarios where they would over-and under-report. Fit statistics should be investigated for these different threats. Neural networks should also be trained with data from specific threats to determine whether it can be used to test a protocol for the existence of a specific protocol validity threat. These neural networks can then be used in accordance with the fit statistics.
- Less computationally intensive methods should be investigated for calculating outfit statistics on posterior cases.
- It should be investigated whether the multi-faceted Rasch model provides a better indication of protocol validity. Other related dimensions from the same instrument

can be included as facets in the model. When including more dimensions, the fit statistics will have a more holistic picture of the context of certain responses and might serve as more powerful protocol validity indicators.

REFERENCES

Ben-Porath, Y. S. (2003). Self-report inventories: Assessing personality and psychopathology. In J. R. Graham & J. Naglieri (Eds.) Vol. X: *Handbook of assessment psychology* (pp. 554-575). New York: Wiley.