



Using multidimensional modeling to combine self-report symptoms with clinical judgment of schizotypy

Stéphanie M. van den Berg^{a,*}, Muirne C.S. Paap^a, Eske M. Derks^{b,1}, Genetic Risk and Outcome of Psychosis (GROUP) investigators²

^a University of Twente, Department of Research Methodology, Measurement, and Data-Analysis, Behavioral Sciences, De zui, P.O. Box 217, 7500 AE Enschede, The Netherlands

^b University Medical Center Utrecht, Rudolf Magnus Institute of Neuroscience, Department of Psychiatry, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands

ARTICLE INFO

Article history:

Received 27 April 2012

Received in revised form

27 August 2012

Accepted 5 September 2012

Keywords:

Item Response Theory (IRT)

Assessment

Measurement invariance

Liability

ABSTRACT

This study investigated psychometric properties of two widely used instruments to measure subclinical levels of psychosis, the Community Assessment of Psychic Experiences (CAPE) and the Structured Interview for Schizotypy-Revised (SIS-R), and aimed to enhance measurements through the use of multidimensional measurement models. Data were collected in 747 siblings of schizophrenia patients and 341 healthy controls. Multidimensional Item-Response Theory, Mokken Scale and ordinal factor analyses were performed. Both instruments showed good psychometric properties and were measurement invariant across siblings and controls. The latent traits measured by the instruments show a correlation of 0.62 in siblings and 0.47 in controls. Multidimensional modeling resulted in smaller standard errors for SIS-R scores. By exploiting correlations among related traits through multidimensional models, scores from one diagnostic instrument can be estimated more reliably by making use of information from instruments that measure related traits.

© 2012 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Subclinical psychotic experiences are prevalent in the general population (van Nierop et al., 2012; van Os et al., 2009). Even though they rarely transit into a clinical diagnosis of schizophrenia (prevalence 0.5–1%; McGrath et al., 2004), there is evidence for a

familial (genetic) continuity between subclinical psychotic experiences and clinical psychotic symptoms (Kendler and Walsh, 1995; Hanssen et al., 2003; van Nierop et al., 2012; Lataster et al., 2009). Subjects diagnosed with a psychotic disorder, such as schizophrenia, may be at the extreme high end of the liability distribution and score above the disease threshold, while subjects who score just below the disease threshold are not diagnosed with a psychotic disorder but may likely develop such a disorder in the future (Bak et al., 2003; Dominguez et al., 2010). The same may be true for other symptoms associated with schizophrenia, which can reveal themselves on the cognitive, interpersonal and emotional level (see e.g. Lenzenweger, 2010). Lenzenweger (2010) refers to the underlying liability for schizophrenia as “schizotypy”. It should be noted that some authors use the terms “schizotypy” and “subclinical psychosis” interchangeably. In this article, we use the term “schizotypy” as an overarching construct, including both “positive” and “negative” symptoms. Schizotypal symptoms can be measured in different ways. This study aims to show how information from two widely used screening instruments for schizotypy, one based on a psychiatric interview, the other based on a self-report questionnaire, can be combined using modern statistical techniques, resulting in increased measurement precision.

We focus on the Structured Interview for Schizotypy-Revised (SIS-R) and the Community Assessment of Psychic Experiences (CAPE). These instruments show good test–retest reliability and good inter-rater agreement (Kendler et al., 1989; Vollema and Ormel, 2000;

* Corresponding author. Tel.: +31 53 489 2422x3616.

E-mail addresses: stephanie.vandenberg@utwente.nl, m.c.s.paap@utwente.nl (S.M. van den Berg), e.m.derks@amc.uva.nl (E.M. Derks).

¹ Present address: Academic Medical Centre University of Amsterdam, Department of Psychiatry, PO Box 22660, 1100 DD Amsterdam, The Netherlands.

² GROUP investigators are: René S. Kahn, MD, PhD, Department of Psychiatry, Rudolf Magnus Institute of Neuroscience, University Medical Center Utrecht, Utrecht, the Netherlands; Don H. Linszen, MD, PhD, Department of Psychiatry, Academic Medical Centre, University of Amsterdam, Amsterdam, the Netherlands; Jim van Os, MD, PhD, South Limburg Mental Health Research and Teaching Network, EURON, Maastricht University Medical Centre, Maastricht, the Netherlands, and King's College London, King's Health Partners, Department of Psychosis Studies, Institute of Psychiatry, London, England; Durk Wiersma, PhD, Department of Psychiatry, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands; Richard Bruggeman, MD, PhD, Department of Psychiatry, University Medical Center Groningen, University of Groningen; Wiepke Cahn, MD, PhD, Department of Psychiatry, Rudolf Magnus Institute of Neuroscience, University Medical Center Utrecht; Lieuw de Haan, MD, PhD, Department of Psychiatry, Academic Medical Centre, University of Amsterdam; Lydia Krabbendam, PhD, South Limburg Mental Health Research and Teaching Network, EURON, Maastricht University Medical Centre; and Inez Myin-Germeys, PhD, South Limburg Mental Health Research and Teaching Network, EURON, Maastricht University Medical Centre.

Konings et al., 2006). Konings et al. (2006) showed that the positive and negative dimensions correspond closely to the positive and negative dimensions of the SIS-R. Details on the factorial structure of CAPE and studies on its reliability and validity can be found elsewhere (Stefanis et al., 2002; Brenner et al., 2007; Konings et al., 2006 and citations therein). Previous studies suggest a multifactorial structure for symptoms associated with schizotypy (e.g., Vollema and Hoijtink, 2000; Kendler et al., 1991). However, very few studies involving the CAPE or SIS-R have thus far employed modern test theory (Item Response Theory, IRT), whereas its advantages are being increasingly recognized. Also within the field of psychiatry, the popularity of IRT has been on the rise, both for analyzing the psychometric properties of questionnaires (e.g., Egberink and Meijer, 2011; Paap et al., 2011b), as well as scrutinizing formal diagnoses (Langenbucher et al., 2004; Paap et al., 2011a). IRT provides a conceptual and statistical framework for studying the internal structure of a scale, possible violations of measurement invariance across subpopulations, and measurement precision across trait level (Reise and Waller, 2009). Moreover, it allows the assessment of correlated traits using multidimensional measurement models.

Our main aim is to enhance the estimation of SIS-R scores by using information contained in the correlation between SIS-R and CAPE scores, through the use of multidimensional IRT (MIRT) models. Briefly, MIRT models are IRT models where several latent traits are related to a fairly large number of items, where these latent traits are allowed to be correlated (Reckase, 2009). As psychopathological items are usually endorsed by relatively few healthy individuals, it is difficult if not impossible to distinguish among individuals with medium or low trait levels. This is reflected in the large number of healthy subjects with minimum scores on the SIS-R, among whom no further distinction can be made. Since the CAPE was specifically designed to assess symptoms in low-scoring individuals, it would be an important advantage for both research and clinical work if the information contained in CAPE items could be somehow used to improve the precision of the estimation of subclinical psychotic symptoms based on the SIS-R. Here we will use CAPE items to enhance measurement precision of the SIS-R scores by modeling two correlated latent traits, one for CAPE items and one for SIS-R items, through a MIRT model.

Before we combine the information from the CAPE and SIS-R, we will investigate the dimensionality of the instruments separately using three complementary methods: Mokken Scale Analysis (MSA), multidimensional Item Response Theory models (MIRT), and ordinal factor analysis (FA). In addition, we will test whether the assessment of schizotypy is influenced by individual characteristics, such as being a sibling of a schizophrenia patient. It is not unlikely that siblings interpret items differently compared to community controls, as they have been in close personal contact with a psychotic family member: they probably have better knowledge of what might be involved regarding certain symptom descriptions. As a consequence, the item score of a given person may depend not only on the latent dimensions of interest but will also depend on individual characteristics (Mellenbergh, 1989; Meredith, 1993). Such a violation of measurement invariance complicates a fair comparison of liability scores across groups.

2. Methods

2.1. Subjects

The data were collected as part of the Genetic Risk and Outcome of Psychosis (GROUP) project (www.group-project.nl), a longitudinal observational study focusing on the factors that make people vulnerable to develop psychosis (GROUP, 2011). Eligible siblings of schizophrenia patients had to fulfill the criteria of (1) age between 18 and 50 (extremes included), (2) fluent in Dutch, and (3) able and willing to give written informed consent. Eligible healthy controls had to

fulfill the criteria of (1) age between 18 and 50 (extremes included), (2) no lifetime psychotic disorder, (3) no first-degree family member with a lifetime psychotic disorder, (4) fluent in Dutch, and (5) able and willing to give written informed consent. In the present study we included a sample of 1088 subjects (639 siblings of schizophrenia patients and 327 controls with CAPE data; 746 siblings and 339 controls with SIS data) who had been assessed at the research center in Utrecht, Groningen, or Amsterdam. The mean age of controls was 31 years (S.D.=10.5; 41.5% male) and the mean age of the siblings was 27 years (S.D.=8.0; 46.3% male).

2.2. Measures

The Dutch versions of the Community Assessment of Psychic Experiences (CAPE) and The Revised Structure Interview for Schizotypy (SIS-R) were assessed. The CAPE is a self-report tool measuring lifetime subthreshold psychotic experiences. It consists of 42 items assessing the frequency (rated on a 4-point Likert scale) of subclinical psychotic experiences in the following three domains: positive symptoms (20 items), negative symptoms (14 items) and depression symptoms (8 items).

The SIS-R (Kendler et al., 1989; Vollema and Ormel, 2000) is an interview instrument that measures a broad range of schizotypal symptoms and signs by applying standardized rating and scoring procedures (four response categories). The shortened version of the SIS-R used in this study describes schizotypy in two dimensions: positive schizotypy (7 items) and negative schizotypy (8 items). It should be noted that we consider both the CAPE and SIS-R to be indicators of schizotypy, even though the CAPE refers to the measured construct as “subclinical psychosis”; both measures include subscales tapping into both positive and negative symptoms.

2.3. Statistical analyses

2.3.1. Assessing dimensionality of CAPE and SIS-R

Three complementary techniques were used to investigate the dimensionality of the CAPE and SIS-R: Mokken Scale Analysis, parametric IRT analysis, and ordinal factor analysis. Mokken Scale Analysis (MSA; Mokken, 1971; Sijtsma et al., 2011) was applied using the software package Mokken Scale Analysis for Polytomous items (MSP5.0; Molenaar and Sijtsma, 2000). MSA is a non-parametric type of IRT analysis. MSA can be used to uncover the dimensionality (factorial structure) of the data, and at the same time identifies scales that allow an ordering of individuals on an underlying one-dimensional scale using the unweighted sum of item scores. In order to determine which items belong together and form a scale, scalability coefficients are calculated. Similar to the item-rest correlation, the scalability coefficient expresses the degree to which an item is related to other items in the scale. The scalability coefficient can be seen as a ‘corrected’ correlation: the correlation between items is divided by the maximum expected correlation given the items’ marginal score-frequency distributions. Dimensionality was investigated using MSP5.0’s automated item selection procedure (AISP) that aims to find one-dimensional clusters of items. These clusters were identified by running the AISP several times in a row, each time increasing the lower bound scalability coefficient (also known as the user-specified constant, *c*). Following (Sijtsma and Molenaar, 2002; see also Meijer et al., 2011), we ran the AISP repeatedly for increasing values of *c*. The resulting sequence of outcomes indicates whether the data set is one-dimensional or multidimensional. Sijtsma and Molenaar (2002) provide the following guidelines. In case of one unidimensional scale for all items, the typical sequence is (1) most or all items are in one scale, (2) one smaller scale is found, and (3) one or a few small scales are found and several items are excluded. In multidimensional datasets the typical sequence is (1) most or all items are in one scale, (2) two or more scales are formed, and (3) two or more smaller scales are formed and several items are excluded. For a recent empirical application of this procedure see Wismeijer (2012).

Parametric IRT models have the same basic assumptions as Mokken models: one-dimensionality, monotonicity and local independence (Reise and Waller, 2009). The difference is that where the Mokken scale merely assumes a non-decreasing relation between the probability of a positive response as a function of trait level, IRT models assume a parametric form for this relationship, either a logistic function or a normal probability distribution, so that IRT models are more restrictive than Mokken models, but allow for the possibility that some items are better indicators for a trait than others. The specific model used here was the Generalized Partial Credit Model (GPCM, Muraki, 1992) for polytomous items. Moreover, we applied multidimensional extensions of the GPCM, where we assumed that individuals have two or more latent trait levels, which might be correlated. Each latent trait is coupled to a fixed set of items, for instance the positive or the negative symptom items on the SIS-R, so that each latent trait can be interpreted through the items associated with it (Béguin and Glas, 2001). Marginal Maximum Likelihood estimation was used. Model fit was ascertained by computing absolute differences between expected and observed item scores for high, average and low scoring individuals. An absolute difference smaller than 0.10 was interpreted as sufficient item fit (cf. Van den Berg et al., 2010). The parametric IRT analyses were applied using the package MIRT (Glas, 2010).

Factor analysis (FA) for ordinal data is similar to the IRT analysis (Takane and de Leeuw, 1987), save for the assumptions about the location parameters in the case of polytomous items. Exploratory factor analyses of the SIS-R data compared the fit of factor models with one to three factors, while the fit of one to four factor models was compared for the CAPE data. In addition, confirmatory models were fitted using the dimensions of positive, negative and depressive symptoms. Factor analyses were carried out using Mplus (Muthén and Muthén, 1998–2000). Robust weighted maximum likelihood estimation (WLSMV) was used and the ordinal nature of the data was taken into account. The fit of the FA models was evaluated using the Root Mean Square Error of Approximation (RMSEA) and the Comparative Fit Index (CFI; Hu and Bentler, 1995). RMSEA smaller than 0.05 indicates good fit, ranging from 0.05 to 0.08 reasonable fit, 0.08 to 0.10 medium fit, and larger than 0.10 poor fit (e.g., Byrne, 2001). A CFI above 0.95 usually indicates good model fit, while values between 0.90 and 0.95 indicate acceptable fit (Hu and Bentler, 1995).

2.3.2. Assessing measurement invariance

In the IRT framework, the general term for violations of the assumption of measurement invariance (Meredith, 1993) is Differential Item Functioning (DIF). DIF is indicated when the model parameters for any item are different across groups, while correcting for any mean difference in liability. DIF was investigated for both CAPE and SIS-R using the MIRT software, comparing siblings and controls. MIRT computes absolute differences between expected and observed average item scores per group, under the assumption of measurement invariance, and tests whether these differences are statistically significant using Lagrange multiplier tests.

2.3.3. Combining CAPE and SIS-R data

Using the MIRT software, all CAPE and SIS-R items were combined in one scale. Next, it was tested whether a two-dimensional IRT model would fit the CAPE and SIS-R item data better, where one dimension related to only the CAPE items, and the other dimension only related to the SIS-R items, while allowing for a correlation between the two dimensions. Subsequently it was determined to what extent the application of such a two-dimensional model could improve the estimation of individual trait levels, compared to estimates based on one test only. One of the advantages of multidimensional IRT models is that when estimating latent trait levels for a given trait, say formally assessed schizotypy through the SIS-R, the information concerning the level on the second trait, say, self-reported schizotypy as measured by the CAPE, is taken into consideration. The higher the correlation between the two traits, the more influence the information on trait CAPE has on the estimation of trait SIS-R, and vice versa. Therefore, even if the two traits are not the same, the assessment of a particular trait can be improved by using information on the related trait.

This idea is illustrated in Fig. 1, where Model 1 may refer to a one-dimensional measurement model for items belonging to the SIS-R. Latent variable θ then represents the schizotypy construct measured by the SIS-R instrument. Model 2 on the other hand refers to a multidimensional measurement model, where two constructs are measured: latent variable θ may again represent the SIS-R related construct, whereas latent variable ζ may represent the construct that is assessed using the CAPE self-report items. The model allows for correlation r between these two constructs.

When we interpret Model 1 in Fig. 1 as a representation of a 2-parameter IRT measurement model, we can quantify the standard error of measurement of θ using the test information function (Lord, 1980). This information function only takes into account the information coming from the θ items. When modeling is extended to include items that measure a correlated construct ζ , such as in a multidimensional model in Model 2, this results in extra information on latent

variable θ , because of correlation r (if $r > 0$). Such statistical borrowing of information results in higher test information content, and because measurement error variance is inversely related to test information, therefore smaller the standard errors of measurement. Thus, we compared the standard errors of measurement for θ under Model 1 with the standard errors of measurement for θ under Model 2, and expected these errors to be smallest in Model 2.

3. Results

3.1. Item data

A large proportion of the items showed no or very few observations in the higher answer categories. For SIS-R, the item responses for six items were therefore dichotomized (absent vs. mild/moderate/severe). For the remaining nine SIS-R items, only the two highest response categories of SIS-R items were collapsed. For CAPE, only the *voodoo* item was left unchanged. Seventeen items were dichotomized. For the remaining CAPE items only the two highest response categories were collapsed.

3.2. Dimensionality of CAPE and SIS-R

The series of Mokken Scale analyses run for the CAPE data resulted in a pattern typical for unidimensionality: (1) most or all items in one scale, (2) one smaller scale was found, and (3) one or a few small scales were found and several items were excluded. A similar picture emerged when analyzing the SIS-R items; only the item referring to *blunted affect* was flagged during the analysis, because it displayed a negative association with at least one other item in the scale.

For the CAPE, a three-dimensional parametric IRT model, with dimensions related to positive, negative and depression symptoms, fitted significantly better than a one-dimensional model ($\chi^2=569.97$, d.f.=12, $P<0.05$), but when inspecting item fit (being good generally, with absolute differences ≤ 0.10), this multidimensional model gave no better item fit. Because of fairly large sample size, the minimal increase in model fit was statistically significant, but not large enough to show its added value at the level of item fit. The estimated correlations among the three traits in the multidimensional model for both siblings and controls were between 0.63 and 0.82, indicating that the traits related to positive, negative and depression symptoms show considerable overlap in the CAPE data.

Similar IRT results were obtained for the SIS-R items related to positive and negative symptoms: the two-dimensional model fitted significantly better than the one-dimensional model ($\chi^2=74.85$, d.f.=5, $P<0.05$), but no improvement in item fit

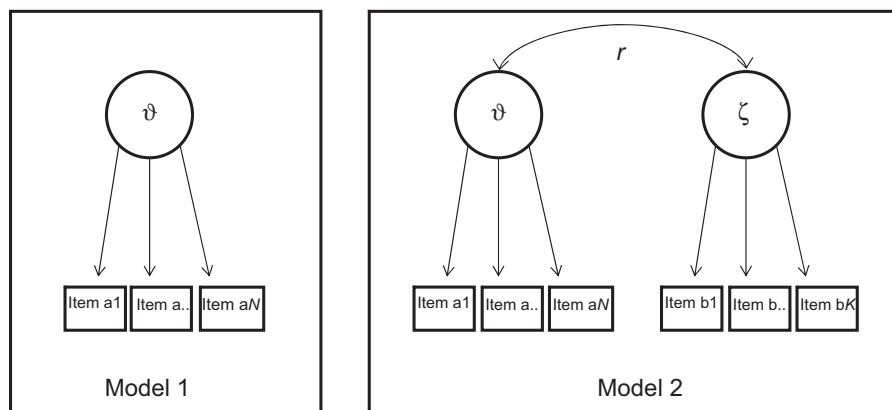


Fig. 1. Enhancement of the precision of scores: methodology. In Model 1, trait θ is measured using a number of test items. In Model 2, trait θ is also defined as being measured by the same items, but the model is extended to include a trait ζ that is associated with θ through correlation r . The extra information on correlated trait ζ adds to the information regarding θ , which results in smaller standard errors for estimated θ scores.

was observed, being generally good. Correlations for the two traits in the multidimensional model were 0.69 and 0.66, for siblings and controls, respectively.

Ordinal factor analyses for the CAPE showed highest CFI and lowest RMSEA for an exploratory four-factor solution (CFI=0.96, RMSEA=0.03). A confirmatory three-factor model for positive, negative and depression symptoms showed good fit index values (CFI=0.91, RMSEA=0.05); these were better than for a one-factor model (CFI=0.83, RMSEA=0.07). Estimated correlations among the three factors ranged from 0.3 to 0.5. Similar results were obtained for the SIS: the best fit index values were found for the exploratory 3-factor solution (CFI=0.95, RMSEA=0.06). The confirmatory 2-factor structure for positive and negative items showed better model fit (CFI=0.87, RMSEA=0.08), than the one-dimensional structure (CFI=0.83, RMSEA=0.09). The estimate for the correlation between the two factors in the confirmatory analysis was 0.4.

3.3. Measurement invariance

The one-dimensional IRT-based SIS-R scale showed significant DIF for one item (psychotic phenomena), tested at a Type I error rate of 0.01 because of multiple testing. However, the difference between observed and expected item scores was less than 0.03 (cf. Van den Berg et al., 2010). The one-dimensional IRT model for the 42 CAPE items showed significant DIF for six items, with the largest absolute effect for the item related to lack of energy: siblings scored higher on this item than expected, after correction for their generally higher scores. This deviation was 0.08 and in a scale of 42 items this effect can be regarded negligible for all practical purposes.

3.4. Combining CAPE and SIS-R data

Given that the MSA and IRT analyses favored the one-dimensional solutions, we chose to use one-dimensional models for the CAPE and SIS-R data in this analysis. The two-dimensional model, with one dimension related to all CAPE items and one dimension related to all SIS-R items (cf Fig. 1, Model 2), fitted significantly better than a one-dimensional model for all CAPE and SIS-R items ($\chi^2=421.33$, d.f.=5, $P<0.05$). The estimated correlation between the two traits was 0.62 in siblings, and 0.47 in controls. These correlations are lower than the ones observed among the CAPE and SIS-R subscales (reported above). This can be interpreted as CAPE and SIS-R measuring distinguishable but related traits. Regarding item fit, most items showed absolute difference much smaller than 0.10. Only one item showed an absolute difference of 0.11 (*telepathy*) but only in the group of healthy controls.

Next, individual scores on the latent traits were estimated, for both the two-dimensional and the two one-dimensional IRT models. In Fig. 2, individuals are plotted with on the horizontal axis their estimated score on the basis of their SIS data alone in a one-dimensional model (ie Model 1 in Fig. 1), and on the vertical axis their estimated score on again the SIS dimension, but now based on the two-dimensional model with the 42 CAPE items on the second dimension (ie Model 2 in Fig. 1). Correlation between the two estimates is 0.94. Both estimated scores are related to the same trait, defined by the SIS items, but in the second case, information on the second correlated trait (CAPE) is statistically borrowed to fine-tune the estimated SIS score. Fig. 2 shows that this approach particularly increases the precision of the measurement at the lower end of the scale. Ignoring CAPE data leaves quite a few individuals with an estimated score of -1.1 on the SIS trait. These 168 individuals have zero scores on all SIS items and are therefore psychometrically indistinguishable from each other.

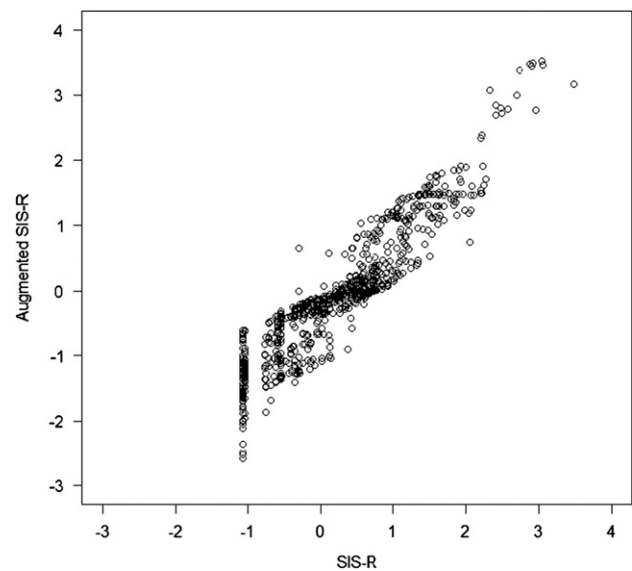


Fig. 2. Enhancement of the precision of scores: results. Individual estimated scores for SIS-R, once under a simple one-dimensional model for SIS-R items, ignoring CAPE data (horizontal axis), and once using a multidimensional model including CAPE data on the second, correlated dimension (vertical axis).

But since these individuals do differ in their CAPE item scores, the range of their estimated theta on the SIS dimension broadens to $[-2.5, -0.5]$, based on the two-dimensional model. The average standard error of the θ estimate also dropped from 0.59 to 0.54. Focusing only on the lower half of the scale, the drop was from 0.70 to 0.61.

Estimated scores for the CAPE trait in a one-dimensional model correlated 0.99 with estimated scores for the CAPE dimension in a two-dimensional model where the CAPE scores were fine-tuned using information from the SIS-R scores. The SIS-R items did not add much information for the estimation of CAPE scores, as the average standard error of the score estimates dropped from 0.32 to 0.31.

4. Discussion

The three approaches to assess dimensionality led to seemingly divergent results. The Mokken Scale analyses and the IRT analyses seemed to favor one-dimensional models for both CAPE and SIS-R, whereas the ordinal factor analyses seemed to favor multidimensional models for correlated traits (although the estimated correlations among the dimensions were moderate). Divergence of conclusions across methods can be ascribed to differences in model assumptions and different procedures to study dimensionality (i.e., statistical tests, item fit, comparative fit indices, exploratory MSA using increasing threshold values for the scalability coefficient, allowing for correlations among traits). For instance, Mokken analysis focuses on selecting items that discriminate well between persons (large scalability coefficients), whereas factor analysis and IRT do not (they allow for low factor loadings/ discrimination parameters, respectively, see also *Sijtsma and Meijer, 2007*). In addition, we note that other factors than the statistical technique used can affect the dimensionality pattern, such as the particular sample being analyzed here (healthy individuals). This could explain the seeming discrepancy between our findings and those of *Vollema and Hoijtink (2000)* who found clear support for a multidimensional pattern in a sample of psychiatric patients. Taken together, however, our results suggest the presence of a broadly defined latent trait

(schizotypy) that assesses a person's tendency to experience psychotic phenomena, be it that some additional clustering can be identified. Based on our results, we recommend the test-user (researcher, clinician) to either use the one-dimensional scale score or to take the correlation between the subscales into account when interpreting subscale scores (i.e., that they are not independent).

One important assumption when comparing subpopulations is that the observed scores are not influenced by group membership except for differences in the underlying trait. A higher mean score on schizotypy in one population should represent *generally* higher scores on all symptoms, not just one or two symptoms in particular. Our analyses showed that the CAPE and SIS-R schizotypy scales are measurement invariant across siblings and community controls. We therefore conclude that these scales allow for an unbiased comparison across siblings of schizophrenia patients and other members of the general population.

Results showed that the CAPE is more precise in what it purports to measure than the (shortened) SIS-R, as can be predicted from the number of items but can also be gauged from the average standard error of the latent trait estimates in the IRT models. Moreover, the CAPE gives more information across the entire range of trait values, whereas the shortened SIS-R has poor resolution particularly at the low end of the scale, at least in the population of healthy individuals and siblings of schizophrenia patients. But note that the complete SIS-R should be more precise; here only the shortened version was used in the data collection.

Both instruments are useful, one providing self-reports, the other a clinician's report, but they do not necessarily concur. Without a gold standard that everybody agrees on, they should ideally complement each other. Generally, the clinician's judgment is deemed more reliable and valid than a self-report, although self-reported experiences can nevertheless be used to augment the clinician's report by adding extra information. This fine-tuning of the clinician's judgment can be formalized through the application of a two-dimensional IRT model and estimating the trait value related to the SIS-R items. As we have shown here, SIS-R scores will then be more reliable (i.e., smaller standard error of measurement) and will show more variation, particularly at the lower end of the scale. This makes it possible to distinguish among subjects of average and low levels of SIS-R defined schizotypy, and therefore to detect subclinical levels of schizotypy.

Here we showed that it was not possible to combine all CAPE and all SIS-R items in one simple measurement model with one latent variable, as the correlation of the two latent traits was significantly lower than 1. As an alternative to multidimensional modeling, a strategy could consist of finding a subset of CAPE items that directly map onto the 'SIS-R' trait, increasing the measurement precision in a different way (cf. test linking, Kolen and Brennan, 2004). For each individual study, which approach is best should be empirically determined. The approach proposed here of making scores more precise through the application of multidimensional measurement models can be applied in many other instances where multiple measurements exist of a psychiatric disorder using different diagnostic instruments, particularly in those cases where one suspects that the two (or more) instruments do not show complete overlap in the constructs being measured.

Limitations of this study include the observation that the SIS-R scale used here was the shortened version. Increase in measurement precision would probably have been less dramatic if the full version had been used in the GROUP study. Another limitation is that the SIS-R and CAPE measured are not unequivocally one-dimensional, as shown here. The method proposed here of combining two measures through a MIRT model would work best

with two clearly one-dimensional constructs. Alternatively, the multidimensionality modeling might be extended to more than two dimensions that includes a multidimensional structure for trait 1 and another multidimensional structure for trait 2, with some higher-order correlational structure for correlations between multidimensional traits 1 and 2. However, this would require larger samples sizes than we have here. Nevertheless, we feel the method proposed here is very helpful in clinical studies with multiple related but independent measures, where it is not always that obvious how to combine these into one sensible index. Future work should look at how helpful this method is in clinical studies, for example, whether such a newly constructed measure shows higher correlations with covariates.

Acknowledgments

This work was supported by the Geestkracht programme of the Dutch Health Research Council (ZON-MW, Grant no. 10-000-1002); the EU Seventh Framework Programme (consortium name: EU-GEI) and matching funds from the Participating universities and mental health care organizations (Site Amsterdam: Academic Psychiatric Centre AMC, Ingeest, Arkin, Dijk en Duin, Rivierduinen, ErasmusMC, GGZ Noord Holland Noord; Site Utrecht: University Medical Centre Utrecht, Altrecht, Symfora, Meerkanten, Riagg Amersfoort, Delta; Site Groningen: University Medical Center Groningen, Lentis, GGZ Friesland, GGZ Drenthe, Adhesie, Mediant, GGZ De Grote Rivieren and Parnassia psycho-medical centre; Site Maastricht: Maastricht University Medical Center, GGZ Eindhoven, GGZ Midden-Brabant, GGZ Oost-Brabant, GGZ Noord-Midden Limburg, MondriaanZorggroep, Prins Claus-centrum Sittard, RIAGG Roermond, Universitair Centrum Sint-Jozef Kortenberg, CAPRI University of Antwerp, PC Ziekeren Sint-Truiden, PZ Sancta Maria Sint-Truiden, GGZ Overpelt, OPZ Rekem). Eske Derks is financially supported by the Netherlands Scientific Organization (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, gebied Maatschappij-en Gedragwetenschappen: NWO/MaGW, Grant no. VENI-451-080-010). We are grateful for the generosity of time and effort by the patients and their families, healthy subjects, and all researchers who make this GROUP project possible.

References

- Bak, M., Delespaul, P., Hanssen, M., de Graaf, R., Vollebergh, W., van Os, J., 2003. How false are false positive psychotic symptoms? *Schizophrenia Research* 62, 187–189.
- Béguin, A.A., Glas, C.A.W., 2001. MCMC estimation and some fit analysis of multidimensional IRT models. *Psychometrika* 66, 471–488.
- Brenner, K., Schmitz, N., Pawliuk, N., Fathalli, F., Jooper, R., Ciampi, A., King, S., 2007. Validation of the English and French versions of the Community Assessment of Psychic Experiences (CAPE) with a Montreal community sample. *Schizophrenia Research* 95, 86–95.
- Byrne, B.M., 2001. Structural equation modeling with AMOS, EQS, and LISREL: comparative approaches to testing for the factorial validity of a measuring instrument. *International Journal of Testing* 1, 55–86.
- Dominguez, M.D., Saka, M.C., Lieb, R., Wittchen, H.U., van Os, J., 2010. Early expression of negative/disorganized symptoms predicting psychotic experiences and subsequent clinical psychosis: a 10-year study. *American Journal of Psychiatry* 167, 1075–1082.
- Egberink, I.J.L., Meijer, R.R., 2011. An IRT analysis of Harter's Self-Perception Profile for Children (SPPC) or why strong clinical scales should be distrusted. *Assessment* 18, 201–212.
- Genetic risk and outcome in psychosis (GROUP) investigators, 2011. Evidence that familial liability for psychosis is expressed as differential sensitivity to cannabis: an analysis of patient-sibling and sibling-control pairs. *Archives of General Psychiatry* 68, 138–147.
- Glas, C.A.W., 2010. Preliminary Manual of the Software Program Multidimensional Item Response Theory (MIRT). Department of Research Methodology, Measurement and Data-Analysis, University of Twente, Enschede, The Netherlands.
- Hanssen, M., Peeters, F., Krabbendam, L., Radstake, S., Verdoux, H., van Os, J., 2003. How psychotic are individuals with non-psychotic disorders? *Social Psychiatry and Psychiatric Epidemiology* 38, 149–154.

- Hu, L.-T., Bentler, P., 1995. Evaluating model fit. In: Hoyle, R.H. (Ed.), *Structural Equation Modeling: Concepts, Issues, and Applications*. Sage, London, pp. 76–99.
- Kendler, K.S., Lieberman, J.A., Walsh, D., 1989. The Structured Interview for Schizotypy (SIS): a preliminary report. *Schizophrenia Bulletin* 15, 559–571.
- Kendler, K.S., Ochs, A.L., Gorman, A.M., Hewitt, J.K., Ross, D.E., Mirsky, A.F., 1991. A pilot multitrait twin study. The structure of schizotypy. *Psychiatry Research* 36, 19–36.
- Kendler, K.S., Walsh, D., 1995. Schizotypal personality disorder in parents and the risk for schizophrenia in siblings. *Schizophrenia Bulletin* 21, 47–52.
- Kolen, M.J., Brennan, R.L., 2004. *Test Equating, Scaling, and Linking: Methods and Practices*, 2nd ed. Springer, New York.
- Konings, M., Bak, M., Hanssen, M., van Os, J., Krabbendam, L., 2006. Validity and reliability of the CAPE: a self-report instrument for the measurement of psychotic experiences in the general population. *Acta Psychiatrica Scandinavica* 114, 55–61.
- Langenbucher, J.W., Labouvie, E., Martin, C.S., Sanjuan, P.M., Bavy, L., Kirisci, L., Chung, T., 2004. An application of item response theory analysis to alcohol, cannabis, and cocaine criteria in DSM-IV. *Journal of Abnormal Psychology* 113, 72–80.
- Lataster, T., Myin-Germeyns, I., Derom, C., Thiery, E., van Os, J., 2009. Evidence that self-reported psychotic experiences represent the transitory developmental expression of genetic liability to psychosis in the general population. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 150, 1078–1084.
- Lenzenweger, M.F., 2010. *Schizotypy and Schizophrenia: The View From Experimental Psychology*. Guilford Press, New York.
- Lord, F.M., 1980. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- McGrath, J., Saha, S., Welham, J., el Saadi, O., MacCauley, C., Chant, D., 2004. A systematic review of the incidence of schizophrenia: the distribution of rates and the influence of sex, urbanicity, migrant status and methodology. *BMC Medicine* 2, 13.
- Meijer, R.R., de Vries, R.M., van Bruggen, V., 2011. An evaluation of the Brief Symptom Inventory-18 using Item Response Theory: which items are most strongly related to psychological distress? *Psychological Assessment* 23, 193–202.
- Mellenbergh, G., 1989. Item bias and item response. *International Journal of Educational Research* 13, 127–143.
- Meredith, W., 1993. *Measurement Invariance, Factor Analysis and Factorial Invariance*. Psychometrika 58, 525–543.
- Mokken, R.J., 1971. *A Theory and Procedure of Scale Analysis*. Mouton, The Hague, The Netherlands.
- Molenaar, I.W., Sijtsma, K., 2000. *MSP5 for Windows*. ProGAMMA, Groningen, The Netherlands.
- Muraki, E., 1992. A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement* 16, 159–176.
- Muthén, L., Muthén, B., 1998–2000. *Mplus User's Guide*. Muthén & Muthén, Los Angeles, CA.
- Paap, M.C.S., Kreukels, B.P.C., Cohen-Kettenis, P.T., Richter-Appelt, H., de Cuypere, G., Haraldsen, I.R., 2011a. Assessing the utility of diagnostic criteria: a multi-site study on gender identity disorder. *Journal of Sexual Medicine* 8, 180–190.
- Paap, M.C.S., Meijer, R.R., Van Bebbber, J., Pedersen, G., Karterud, S., Hellem, F.M., Haraldsen, I.R., 2011b. A study of the dimensionality and measurement precision of the SCL-90-R using item response theory. *International Journal of Methods in Psychiatric Research* 20, e39–e55, <http://dx.doi.org/10.1002/mpr.347>.
- Reckase, M.D., 2009. *Multidimensional Item Response Theory*. Springer, New York.
- Reise, S.P., Waller, N.G., 2009. Item response theory and clinical measurement. *Annual Review of Clinical Psychology* 5, 27–48.
- Sijtsma, K., Meijer, R.R., 2007. Nonparametric item response theory and related topics. In: Rao, C.R., Sinharay, S. (Eds.), *Handbook of Statistics 26: Psychometrics*. Elsevier, Amsterdam, The Netherlands, pp. 719–746.
- Sijtsma, K., Meijer, R.R., van der Ark, L.A., 2011. Mokken scale analysis as time goes by: an update for scaling practitioners. *Personality and Individual Differences* 50, 31–37.
- Sijtsma, K., Molenaar, I.W., 2002. *Introduction to Nonparametric Item Response Theory*. Sage, Thousand Oaks, CA.
- Stefanis, N.C., Hanssen, M., Smirnis, N.K., Avramopoulos, D.A., Evdokimidis, I.K., Stefanis, C.N., Verdoux, H., van Os, J., 2002. Evidence that three dimensions of psychosis have a distribution in the general population. *Psychological Medicine* 32, 347–358.
- Takane, Y., de Leeuw, J., 1987. On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika* 52, 393–408.
- Van den Berg, S.M., Heuven, H.C.M., Van den Berg, L., Duffy, D.L., Serpell, J.A., 2010. Evaluation of the C-BARQ as a measure of stranger-directed aggression in three common dog breeds. *Applied Animal Behaviour Science* 124, 141–146.
- van Nierop, M., van Os, J., Gunther, N., Myin-Germeyns, I., de Graaf, R., ten Have, M., van Dorsselaer, S., Bak, M., van Winkel, R., 2012. Phenotypically continuous with clinical psychosis, discontinuous in need for care: evidence for an extended psychosis phenotype. *Schizophrenia Bulletin* 38, 231–238.
- van Os, J., Linscott, R.J., Myin-Germeyns, I., Delespaul, P., Krabbendam, L., 2009. A systematic review and meta-analysis of the psychosis continuum: evidence for a psychosis proneness-persistence-impairment model of psychotic disorder. *Psychological Medicine* 39, 179–195.
- Vollema, M.G., Hoijtink, H., 2000. The multidimensionality of self-report schizotypy in a psychiatric population: an analysis using multidimensional Rasch models. *Schizophrenia Bulletin* 26, 565–575.
- Vollema, M.G., Ormel, J., 2000. The reliability of the structured interview for schizotypy-revised. *Schizophrenia Bulletin* 26, 619–629.
- Wismeijer, A.A.J., 2012. Dimensionality analysis of the Thought Suppression Inventory: combining EFA, MSA, and CFA. *Journal of Psychopathology and Behavioral Assessment* 34, 116–125.