# Validation of a new concept: aptitudes of psychiatric nurses caring for depressed patients

Marc Haspeslagh, Kristof Eeckloo and Lucas B. Delesie

*Journal of Research in Nursing* 2012 17: 438 originally published online 16 November 2010
DOI: 10.1177/1744987110387489

*Journal of Research in Nursing*
17(5) 438–452

DOI: 10.1177/1744987110387489
jrn.sagepub.com

# Validation of a new concept: aptitudes of psychiatric nurses caring for depressed patients

**Marc Haspeslagh**
Psychiatric Nurse, Catholic University of Leuven, Belgium

**Kristof Eeckloo**
Research Fellow, Catholic University of Leuven, Belgium

**Lucas B. Delesie**
Professor, Catholic University of Leuven, Belgium

## Abstract

Validation is vital for operationalising a new concept into a measurement instrument. The measurement of human attributes is usually done with questionnaire items in ordered categories. Our objective was to validate a questionnaire capable of measuring, at the ordinal level, the aptitude of psychiatric nurses caring for depressed patients. We used expert panels, experimentation, categorical principal component analysis, and parametric and non-parametric item response theory to develop such a questionnaire and assess its validity. Expert panels delineated five aspects and 29 components of aptitude and formulated 32 items. Four consecutive exploratory experiments were performed to gauge and calibrate the items and their response categories into a semantic frame of reference and a socio-cultural and job context of nurses. This resulted in a questionnaire comprising three aspects of aptitude. Fourteen questionnaire items with a different number of response categories assessed aptitude. Appropriate techniques shed light onto how nurses understand and respond to items in the questionnaire. Before it can be reliably used in a different context, the questionnaire needs to be re-evaluated for validity. Moreover, validity needs to be re-established for translated versions. In conclusion, validation is a process. Understanding that the scope and limitations of a questionnaire develop as it is being used requires validity to be re-established at each step of development.

## Keywords

Aptitude, depressive disorder, multivariate analysis, psychiatric nursing, questionnaire design, validation studies as topic

**Corresponding author:**
Marc Haspeslagh, General Hospital Sint-Jan Brugge, Ruddershove 10, B-8000 Brugge, Belgium
Email: marc@haspeslagh.net

## Introduction

In psychiatric units head nurses face the daily challenge of assigning patients to nurses. Since nursing schedules are fixed weeks in advance, for any given day, head nurses have a limited pool of nurses to draw upon when assigning nurses to newly admitted patients. Thus, the 'fit' between a patient and a nurse is not always optimal, which can hinder the therapeutic relationship. The present study investigated one aspect of this allocation problem: competence and its sub-concept, aptitude.

It is commonly accepted that an individual's competence influences his or her performance. Competence consists of three sub-concepts: knowledge, skill, and aptitude (Nordhaug, 1993). Since all psychiatric nurses have a diploma and certificates, we assume that they have knowledge and skill. In all countries, possession of a diploma and certificates indicates that a nurse has met the minimum requirements for knowledge and skill. Thus, nursing diplomas and certificates are accepted as sufficient warrant competence assessments. Nordhaug (1993) stated that aptitude encompasses a person's natural talents, which can be applied to work and forms the basis for the development of knowledge and skill. He argued that aptitude is basic, since it underlies the development of knowledge and skill and cannot easily be developed. For psychiatric nurses, however, aptitude has not yet been operationalised. Nonetheless, the measurement of aptitude is crucial for investigating the problem of allocating psychiatric nurses to patients with depression. In several psychiatric units at general and psychiatric hospitals in Flanders, the Dutch-speaking region of Belgium, our study has focused on patients suffering from depression and their relationships to nurses that care for them. This paper focuses on measuring nurses' aptitude for caring for depressed patients. Specifically, we describe how to validate a questionnaire that measures aptitude at an ordinal level. We do so by using decades-old analysis techniques that are not commonly used today.

### *Literature*

While searching the literature on operationalising competence and aptitudes, we encountered several problems. Our search identified seven measurement options for operationalising aptitudes and guaranteeing validity: (1) take into account the job context (Milligan, 1998; Arnold, 2002); (2) take into account the semantic frame of reference and the socio-cultural context of the nurses (Watson et al., 2002); (3) combine asset and process approaches to management of aptitudes and generic and specific components of aptitude (Barnett, 1994); (4) consider variability of aptitudes over time (Barnett, 1994); (5) gather self/peer opinions of the nurses instead of the opinions of bosses or compliance to standards of good practice (Vuorinen et al., 2000; Meretoja and Leino-Kilpi, 2003); (6) during data processing, do not assume numeric levels of measurement but respect ordinal levels of measurement of aptitudes (Coombs, 1964; Young and Hamer, 1987); and (7) use an individual-level approach (idiographic) instead of a group-level approach (nomothetic) (Hand, 2004). These seven measurement options serve as the foundation for operationalising aptitudes. They guided the different choices we made as we developed and validated our aptitude questionnaire.

Validity and validation are subjects of continuing debate in the scientific literature. Classic reasoning divides validity into content-, construct-, and criterion-related validity (Carmines and Zeller, 1979). Streiner and Norman (2003) concluded that, unlike criterion validity, there is no one experimental design or statistic that is common to construct validational studies.

It is obviously necessary to conduct validational studies for each new instrument developed. However, when the instrument is to measure a hypothetical construct, the task is an ongoing one (Streiner and Norman, 2003). Hand (2004) added some other elements to the reasoning about validity: Validity describes how well the measured variable represents the attribute being measured, or how well it captures the concept that is the target of measurement. From a statistical perspective, validity may be regarded as similar to bias in the sense that a biased measurement somehow misses the fundamental target, whereas reliability may be regarded as similar to variance in the sense that an unreliable measure varies between measurement occasions. One might even go so far as to say the measurement or measurement procedures themselves are not really the subject of the validation at all. Instead, it is the utility of the measurement that may or may not be valid.

Hand (2004) further stated that systematic errors could arise in many ways. The way individual respondents tend to respond to questions and the wording of questions are two main issues. There are two schools of thought about the question of subjectivity in the psychological and social measurement of individuals: (1) should one measure some property external to the mind of the subject? or (2) should one measure that property as perceived by the subject? Indeed, careful thought is needed about which measure to carry out. If the first school of thought were the aim, subjective bias would need to be removed. However, if the second school of thought were the aim, the subjective effects would have to be regarded as contributing to the measured value.

Hand (2004) concluded that it is not surprising that validating an instrument will cause the instrument to evolve, as understanding of its scope and limitations develop as it is used. Perhaps the best one can hope for is a carefully argued process of attempting to establish validity, so that other potential users can agree that what has been done is solid. The present study elaborates on validity in line with Hand's reasoning. We operationalised aptitude by using questionnaire items with corresponding response categories that are geared towards use by psychiatric nurses caring for depressed patients in psychiatric units of general and psychiatric hospitals in Flanders.

Meretoja and Leino-Kilpi (2003) found that, in five competence areas and one overall level of competence, managers assessed the level of competence significantly higher than did the nurses. Vuorinen et al. (2000) concluded that peer evaluation promotes nurses' professional development and on-the-job learning. Therefore, in this study, we chose to monitor the opinions of psychiatric nurses themselves and to exclude the opinions of managers.

To obtain an accurate measurement of aptitudes, a full experimental design is needed that will ensure that all nurses score all items of all aspects and components of aptitude for themselves as well as for their colleagues. It follows, then, that a 40-item questionnaire used for a team of 15 nurses, for example, would require every nurse to state 600 thoughtful and nuanced opinions. As scoring fatigue is unavoidable, the sincerity, accuracy, and reliability of all these opinions become questionable. Therefore, the aptitude questionnaire can only have a limited number of items.

## Methodology

Validity was established through a process that started with the delineation of aspects and components of aptitude and with the determination of items and response categories by expert panels of experienced psychiatric head nurses. This process continued through four

consecutive preliminary experiments (PE1, PE2, PE3, PE4), of which the data were analysed with appropriate statistical techniques that respect the ordinal measurement level of the aptitudes. The process ended with the main experiment and the analysis of these data for individual differences to determine the extent to which the aptitude measures of the construct were consistent with the 'best guesses' about the construct.

## *Step 1*

The first step consisted of specifying the domain of aspects and components. Since aptitude (Nordhaug, 1993) is a new concept in psychiatric nursing, a measurement instrument that specifically assesses the aptitude of psychiatric nurses is lacking. Different competence descriptions of psychiatric nurses are available (Hoot, 1995; American Nurses Association, 2000; DuPerron, 2001). However, none of them focuses on aptitudes, nor do they relate to specific patient populations, e.g. care for patients suffering from depression. To specify the domain of aspects and components, we used literature on psychiatric nursing (Peplau, 1952; Keltner et al., 2003; Videbeck, 2004), psychiatric nursing care plans (Schultz and Videbeck, 2002), and the literature and evidence on therapist variables (Beutler et al., 2004).

## *Step 2*

In the second step, two expert panels of experienced psychiatric head nurses selected the aspects and components of aptitude and formulated the items and response categories.

## *Step 3*

In the third step, we determined the extent to which the items tended to measure the same entity, several different entities, or many different entities by using empirical research and statistical analysis. The empirical research consisted of four consecutive preliminary experiments (PE1, PE2, PE3, PE4). During the four preliminary experiments, the wording of the items was adapted (gauging), and the wording of the items and the number and wording of the response categories were balanced (calibrating). Gauging and calibrating were necessary in order to connect the semantic frame of reference and the socio-cultural and job context of the psychiatric nurses under study. The statistical analysis was based on categorical principal component analysis (CatPCA), and parametric and non-parametric item response theory (IRT).

*Categorical principal component analysis.* We chose CatPCA (Gifi, 1990), because it allowed us to analyse the response data at nominal, ordinal, and ratio levels. By contrast, common PCA only permits analysis at the ratio level. CatPCA combines the stochastic vector model of PCA with the centroid model from multiple correspondence analysis (MCA) (Greenacre and Blasius, 1994). CatPCA focuses on the joint approximate representation of the items, the response categories, and the observations in a low-dimensional space, enabling the evaluation of the ordinal interdependence between items, response categories, and observations within the same frame of reference. The iterative algorithm transforms nonlinearly the ordinal response pattern of each item so that the overall fit of the solution improves. Overall fit means

maximum correlation between the transformed response categories and the dimensions in the low-dimensional space and maximum variance accounted for by the positions and distances between observations and all item response categories in the visualisation model. Part of the output of the analysis consists of transformation plots. These enable the evaluation of the balance between the wording of an item and the number and wording of response categories. Another part of the output of the analysis is eigenvalue structure and component loadings pattern. These enable the evaluation of whether or not the items measure one underlying structure and the determination of which items contribute to that structure.

*Parametric item response theory.* We chose parametric IRT (Embretson, 2003) because it enabled us to investigate item difficulty and item discrimination. Parametric IRT also places the subjects on a latent variable (aptitude) based on their item scores on the questionnaire. Within IRT we used the option of the generalised partial credit model (GPCM), which is appropriate for analysing personality scale responses where subjects rate their beliefs, or respond to statements on a multi-point scale (Embretson and Reise, 2000). The GPCM analysis produced cumulative item response function (IRF) curves, which enabled us to evaluate visually the item difficulty and item discrimination.

*Non-parametric item response theory.* We chose non-parametric IRT (Sijtsma and Molenaar, 2002) because it facilitated interpretation of test performance and facilitated our search for subscales in questionnaires with several items of ordinal measurement level. The evaluation of test performance was based on the double monotonicity model, which is less restrictive than parametric IRT models because it assumes only monotone increasing and nonintersecting IRF curves (Sijtsma and Molenaar, 2002). The monotone increasing property corresponds to the ordinal level of measurement of the items of a questionnaire, whereas the nonintersecting property corresponds to the invariant item ordering. This means that if there is a fixed sequence of developmental stages, and the items tap into sub-abilities corresponding to the stages, we would expect an ordering of difficulty of the items that corresponds to the developmental stages and that is the same for all subjects. The search for subscales depends on fixing the lower bound parameter. A higher lower bound translates to a stronger scale in the sense of a more accurate ordering of subjects on the latent variable (i.e. aptitude) by means of their total score based on all selected items. Hence, a scale is considered weak when the Loevinger H coefficient $< 0.4$; moderate when $0.4 \leq H < 0.5$; and strong when $H \geq 0.5$. Sijtsma and Molenaar (2002) warn that the definition of a scale is mathematical: 'It does not guarantee an operational relation of the measured variable to the intended hypothetical construct'.

## Step 4

In the fourth step, we performed subsequent individual differences studies to determine the extent to which supposed measures of the construct are consistent with 'best guesses' about the construct. Here, the response data of the main experiment involving 14 psychiatric units and 119 nurses were aggregated into aptitude construct scores, and the individual differences of the psychiatric nurses were investigated along three ordered categories of professional functioning: novice, proficient, and master. Finally, associations with the nurses' demographic variables were investigated.

## Results

Two expert panels comprising two and three experienced head nurses of a psychiatric unit delineated five aspects and 29 components of aptitude (Table 1). These aspects and components were selected from a list of aptitude aspects and components from pertinent literature. They agreed on the wording of 32 items, for which they delineated four response categories—(1) seldom, (2) now and then, (3) regularly, and (4) most of the time—resulting in the initial version of the questionnaire.

We anticipated that, to achieve a valid questionnaire, three to four preliminary experiments were needed to gauge and calibrate the questionnaire to the semantic frame of reference and socio-cultural and job context of Flemish psychiatric nurses. The reduction of the number of items and the gauging and calibrating of items and their corresponding response categories were based on the use of the statistical techniques described in the methods section. Next, we present examples of how these techniques were used.

### *Categorical principal component analysis*

CatPCA analysis of PE2 identified several items having rather equivocal measurement. Item two (v02), 'I can keep my thoughts on the patient', has five response categories. The transformation plot of Figure 1 shows the result of the analysis.

The x-axis of the graph represents the different response categories. The y-axis represents the quantification value that is used to transform the response category in order to optimise the global solution of the CatPCA. The analysis level is ordinal and corresponds to the measurement level of the items. The first response category ('.') was not used in this preliminary experiment. The three highest response categories (3, 4, and 5) were quantified as equal, indicating that the nurses did not differentiate between the wording of these three response categories. Thus, according to this result, the five ordinal categories should be reduced to only two response categories: a dichotomous measurement at a nominal level instead of a measurement at an ordinal level.

To capture the nuanced opinions of the nurses, we sought to obtain an ordinal measurement. Hence, we reworded item two, 'I can keep my thoughts on the patient', to 'the patient in front of me gets my full attention'. We also altered the wording of the five response categories. Figure 2 shows the transformation plot resulting from the analysis of data from PE3. After PE2 some items were dropped and the remaining items were re-worded and re-ordered. As a result, the re-worded item two was labelled v09.

Analysis of the data from PE3 revealed that the nurses did not select the lowest response category. Apparently, they had difficulty expressing negative opinions. The other response categories were quantified differently, resulting in a monotonic, increasing quasi-linear transformation. This indicated that the nurses clearly differentiated the wording of the different response categories in relation to the wording of the item. Thus, the intended ordinal measurement was realised for all response categories, except for the lowest category. To realise usage of all the response categories and to capture a nuanced opinion, we adjusted the wording of the response categories again to ensure that all the response categories for all the items will be used. The modified questionnaire was assessed in PE4.

Table 2 presents the eigenvalue structure of the data of PE4. All three dimensions of the analysis had an eigenvalue greater than one. According to the common rule, these three dimensions were retained. The first dimension had an explained variance of 37.26%, which

**Table 1.** Aspects and components of aptitude selected by expert panels

| Reference | Aspect | Component |
|---|---|---|
| Luborsky (1985) | Therapist's adjustment, skill, and interest in helping people | |
| Videbeck (2004) | | Genuine interest |
| Crits-Christoph (1991) | | Use of treatment manual |
| Beutler (2004) | | Flexibility of applying manual |
| Authors' experience | | Follow pace of the patient |
| Videbeck (2004) | The quality of the therapist–patient relationship | |
| Lafferty (1989) | | Empathic understanding |
| Authors' experience | | Understanding context |
| Peplau (1952) | Therapeutic relationship | |
| Lambert (1994) | | Trust: nurses congruence |
| | | Trust: nurses reliability |
| Videbeck (2004) | | Acceptance: not become upset |
| | | Acceptance: avoids being judgemental |
| | | Positive regard: appreciates patient as unique and worthwhile human being |
| | | Positive regard: unconditional non-judgemental attitude |
| Schultz (2002) | | Self awareness: knows own values |
| | | Self awareness: knows own beliefs |
| Videbeck (2004) | | Therapeutic use of self: uses own coping skills |
| | | Therapeutic use of self: uses own perceptions |
| Beutler (2004) | Procedural aspects | |
| Gunderson (1978) | | Distribution of responsibilities |
| | | Distribution of decision making |
| | | High levels of interaction between patient and staff: no down time |
| | | Clarity of the therapeutic programme for patient |
| | | Clarity in the leadership of the therapeutic programme |
| Schultz (2002) | Depression | |
| | | Ineffective coping: engage in reality-based interactions |
| | | Ineffective coping: expresses feelings directly with congruent verbal and non-verbal messages |
| | | Impaired social interactions: can guide patient into social contact with others |
| | | Impaired social interactions: initiates interaction with others |
| | | Self-care: stimulates patient to take up self care |
| | | Self-care: builds progressive steps in approach |
| | | Chronic low self-esteem: verbalises increased feelings of self-worth |
| | | Chronic low self-esteem: makes plans for the future consistent with personal strengths |

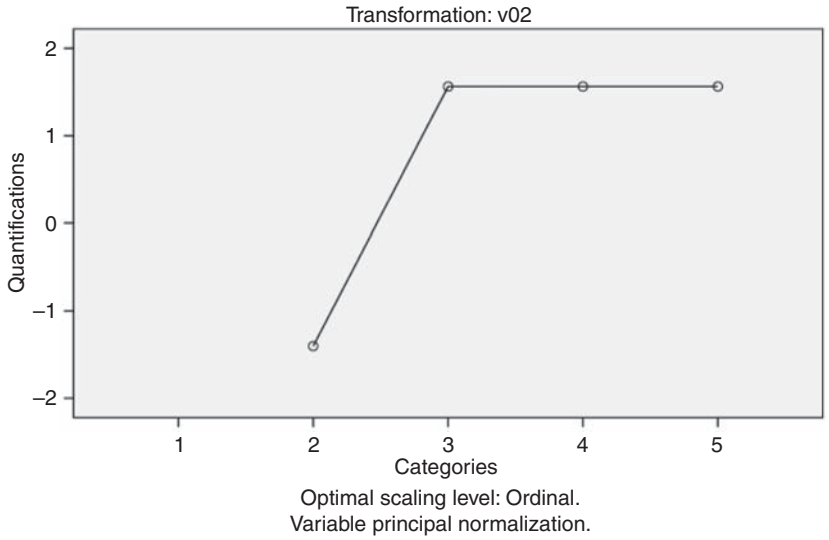


**Figure 1.** Transformation plot of item 2 of data analysis from PE2

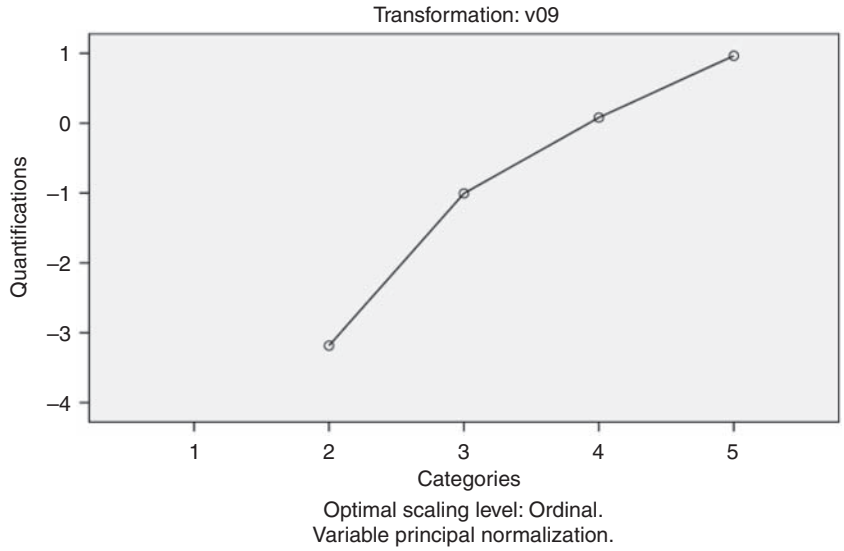


**Figure 2.** Transformation plot of item 9 of data analysis from PE3

was more than twice the explained variance of the second dimension. From this eigenvalue structure we concluded that the questionnaire measures one overall dimension with two supplementary dimensions that nuance the overall dimension.

Table 3 presents the component loadings pattern of the data analysis in PE4. On the first dimension, all the items indicated by v01 to v14 had positive values, indicating that the scale

**Table 2.** Eigenvalue structure of data from PE4

| | | Variance accounted for | |
|---|---|---|---|
| Dimension | Cronbach's alpha | Total (eigenvalue) | Per cent of variance |
| 1 | .870 | 5.216 | 37.260 |
| 2 | .648 | 2.511 | 17.934 |
| 3 | .427 | 1.657 | 11.832 |
| Total | .962[a] | 9.384 | 67.025 |

[a]Total Cronbach's alpha is based on the total eigenvalue.

**Table 3.** Component loadings of data from PE4

| | Dimension | | |
|---|---|---|---|
| Item | 1 | 2 | 3 |
| v01 | 0.296 | 0.552 | −0.519 |
| v02 | 0.266 | 0.567 | −0.509 |
| v03 | 0.007 | 0.296 | 0.705 |
| v04 | 0.860 | −0.222 | −0.065 |
| v05 | 0.803 | −0.513 | −0.057 |
| v06 | 0.635 | −0.088 | 0.030 |
| v07 | 0.664 | 0.131 | 0.106 |
| v08 | 0.801 | −0.529 | −0.053 |
| v09 | 0.563 | 0.589 | 0.096 |
| v10 | 0.569 | 0.496 | 0.095 |
| v11 | 0.519 | 0.490 | 0.135 |
| v12 | 0.741 | 0.269 | 0.008 |
| v13 | 0.811 | −0.513 | −0.056 |
| v14 | 0.265 | 0.109 | 0.754 |

was a unipolar scale, with a dominance relationship between the scale values. By contrast, the second and third dimensions comprised positive and negative values. This means that these dimensions were bipolar scales that measured the proximity to the items with the greatest positive and greatest negative component loadings. For example, for the second dimension, the loadings for items v09 and v02 are positive, whereas the loadings for items v08, v013, and v05 are negative. From this pattern of component loadings we conclude that we have one unipolar scale and two bipolar scales. All items contributed positively to the unipolar scale. This pattern indicated an invariant item ordering and measured developmental stage of aptitude; i.e. novice, proficient, and master. This means that the subjects can be ordered along their developmental stage. The pattern of component loadings of the two bipolar scales indicated a balance between either the items on the positive side or the items on the negative side. This means that the subjects tended to distribute themselves to one of the sides of the two scales. Hence, they can be typified along these two scales.

### *Parametric item response theory*

Parametric IRT analysis of PE1 and PE2 revealed that the nurses could not distinguish two items: 'I find that it is clear for the patient what the therapeutic programme is' and 'I find that it is clear for the patient who is in charge of their therapeutic programme'. Hence, these two items were combined into one item. This resulted in 31 items for the questionnaire in PE3. CatPCA analysis of the data from PE3 showed seven items having rather equivocal measurement. These were dropped from further IRT analysis. Figure 3 shows the IRF curves of the 24 remaining items.

The x-axis of the graph represents the levels of the latent variable (aptitude). The y-axis represents the cumulative probability of the item categories after the IRT model was applied. The six IRF curves that course approximately through the diagonal of the graph (labelled with blue '+', red 'x', and cyan '*' symbols) deviate from the other IRF curves. These IRF curves represent the lowest three response categories of the two remaining items of the procedural aspect of the questionnaire. At 0.5 cumulative probability, the two deviating IRF curves of the lowest response categories (blue '+') are situated at about −2.25 on the latent variable. The IRF curves of the lowest response categories of the remaining items are situated between −4.0 and −3.5 on the latent variable. This higher location on the latent variable illustrates the item difficulty in that response category for these two items. The slope of the six deviating IRF curves is lower than the slope of the other IRF curves in the lowest three response categories. This illustrates less discriminatory power of these two items in the lowest three response categories. On the basis of this IRT analysis, we concluded that the nurses did not interpret the procedural aspect as being part of the aptitude concept. Hence,

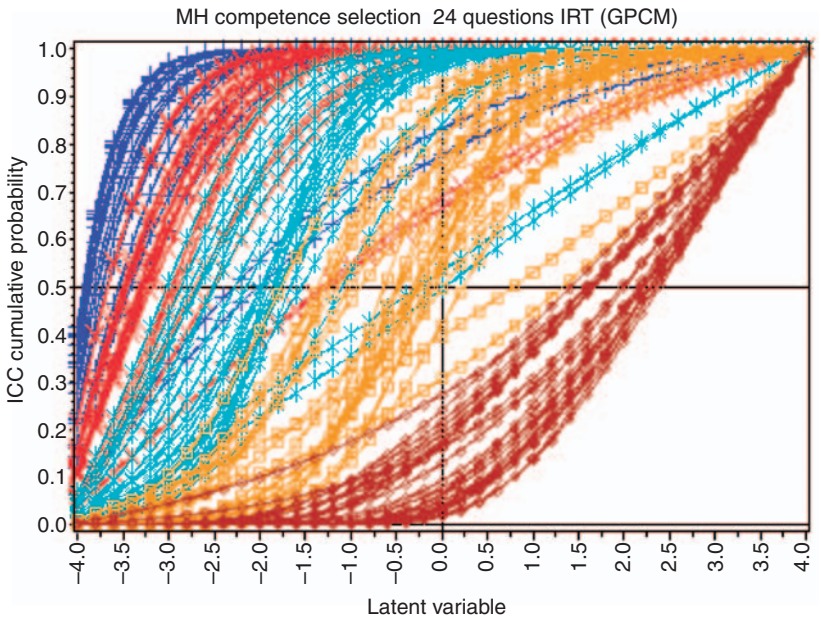


**Figure 3.** Item response function (IRF) curves of items of data analysis from PE3

the items of the procedural aspect were dropped in the questionnaire that was assessed in the next preliminary experiment.

### *Non-parametric item response theory*

Non-parametric IRT analysis of the data from PE4 revealed different solutions along the fixing of the lower bound. Table 4 shows the results when the lower bound was set to 0.5 (strong scale). The scale analysis retained three scales of three items each. The three scales correspond to the three components with an eigenvalue of >1 of the CatPCA. The items that contributed to the scales differed, however, with regard to the component loadings pattern of the CatPCA.

On the basis of the results of the analyses of the data in PE4, we modified the wording of the items and corresponding response categories. This was the last time the items were modified. We decided not to elaborate further on the validity of the questionnaire in a fifth preliminary experiment for the following reasons: The CatPCA transformation plots of all the items already showed monotonically increasing transformations, the IRT analysis showed no deviating items, and the non-parametric IRT confirmed the CatPCA structure with three scales.

This version of the questionnaire was assessed in the main experiment, which was conducted in 14 psychiatric units. One hundred and nineteen nurses participated, resulting in 1,283 opinions about nurses' aptitude. The aggregation of these data into aptitude construct scores requires an explanation of the different aggregation procedures, which is beyond the scope of this paper. It is sufficient to state that the aggregation was founded on the CatPCA and led to three aptitudes construct scores: (1) aptitude for caring for depressed patients, (2) aptitude for using boundaries, and (3) aptitude for empowering patients. The aptitude for caring for depressed patients' scores ordered the nurses according to having less to more aptitude for caring for depressed patients which enabled us to divide the nurses into three ordered categories of professional stage: (1) novice, (2) proficient, and (3) master. The two remaining aptitudes provide an indication of the therapeutic style of a nurse.

The study of individual differences consisted of the correlation of these scales with demographic and other identification variables of the nurses. The main construct score, 'aptitude for caring for depressed patients', was not correlated with age or number of years since graduation of the nurses. This confirms the conceptualisation of aptitude by Nordhaug (1993): Aptitude encompasses one person's natural talents and cannot easily be developed. This is another indication of the validity of the measurement with this questionnaire in the given context.

### *Synthesis of the development of the questionnaire*

IRT analysis of data from PE1 and PE2 revealed that the nurses could not distinguish two items. Hence, we combined these two items into one item. CatPCA of all the items assessed in PE2 revealed that the nurses found 21-item descriptions to be unclear. These items were reworded in as-neutral-Dutch language as possible to capture a professionally accepted difference in personal preference. CatPCA of data from PE3 revealed seven items with rather equivocal measurement. These items were dropped. Parametric IRT analysis of the data from PE3 showed that the IRF curves of the items of the procedural aspect deviated distinctly. Thus, these items were dropped. The same IRT analysis showed that the IRF

**Table 4.** Results of non-parametric item response theory (IRT) analysis of data from PE4

| Scale analysis results | | | | |
|---|---|---|---|---|
| ****ANALYSIS – 1, **SCALE 1** **** | | | | |
| Final scale 1 | | Number of items: 3 | | Number of steps: 2 |
| Lowerbound: 0.50 | | Adjusted alpha: 0.00044 | | Critical Z: 3.33 |
| Scalability coefficients, | Loevinger's H weighted | | | |
| Scale coefficient | H = 0.57 | Scale Z = 8.83 | | |
| Item coefficients | | | | |
| Item | Label | Mean | H wgt | Z |
| V07 | ik stuur het verpleegplan i | 2.71 | 0.57 | 7.24 |
| V05 | ik hou in mijn acherhoofd | 3.30 | 0.58 | 7.36 |
| V04 | ik maak de waarden van de p | 3.61 | 0.55 | 7.06 |
| ****ANALYSIS – 1, **SCALE 2** **** | | | | |
| Final scale 2 | | Number of items: 3 | | Number of steps: 2 |
| Lowerbound: 0.50 | | Adjusted alpha: 0.00069 | | Critical Z: 3.20 |
| Scalability coefficients, | Loevinger's H weighted | | | |
| Scale coefficient | H = 0.57 | Scale Z = 8.57 | | |
| Item coefficients | | | | |
| Item | Label | Mean | H wgt | Z |
| V10 | ik pas me aan, aan het temp | 2.85 | 0.56 | 6.80 |
| V09 | mijn volle aandacht gaat | 3.01 | 0.60 | 7.42 |
| V12 | ik laat de patient zijn | 4.31 | 0.56 | 6.82 |
| ****ANALYSIS – 1, **SCALE 3** **** | | | | |
| Final scale 3 | | Number of items: 3 | | Number of steps: 2 |
| Lowerbound: 0.50 | | Adjusted alpha: 0.0013 | | Critical Z: 3.02 |
| Scalability coefiicients, | Loevinger's H weighted | | | |
| Scale coefficient | H = 0.53 | Z = 8.30 | | |
| Item coefficients | | | | |
| Item | Label | Mean | H wgt | Z |
| V08 | ik hou rekening met mijn st | 2.57 | 0.50 | 6.28 |
| V13 | ik stimuleer de patient tot | 4.01 | 0.53 | 6.83 |
| V14 | ik begeleid de patient naar | 4.18 | 0.56 | 7.19 |

*Final Scale 1 items:* v07, I direct the care plan as a function of the patient; v05, I am aware that I view the patient through my own eyes; v04, I discuss the patient's values.
*Final Scale 2 items:* v10, I accommodate to the pace of the patient; v09, the patient in front of me gets my full attention; v12, I let the patient express his feelings of self-worth.
*Final Scale 3 items:* v08, I take into account my strengths and weaknesses; v13, I stimulate the patient to understand on his own; v14, I accompany the patient to encourage more social interactions.

curves of the items of the aspects 'quality of the therapeutic relationship' and 'therapeutic relationship' coincide, leading us to combine these items. CatPCA of the data in PE4 indicated that the nurses had difficulty with the verb 'can'. Consequently, all item descriptions were reworded in the present tense without using 'can'. Parametric IRT analysis of the data from PE4 showed that the IRF curves of the eight items of the aspect 'depression' correlated two-by-two. Thus, four items were dropped.

Frequency distribution analysis of the data from PE1 yielded left-skewed distributions over the four response categories. Thus, we inserted a fifth response category in between the original third and fourth response categories. CatPCA of the data from PE3 and frequency distribution analysis showed that the distributions differed remarkably between the items of the different aptitude aspects. Therefore, we used a different number of response categories for the items of the different aptitude aspects: five response categories for the four items of 'therapist adjustment, skill, and interest in helping patients'; six response categories for the six items of 'therapeutic relationship'; and seven response categories for the four items of 'caring for depressed patients'.

The length of the questionnaire was shortened step-by-step during the four preliminary experiments. The initial version had 32 items. Combining two items after PE2 yielded a questionnaire with 31 items. Analysis of the data of PE3 revealed seven items with equivocal measurement and four items of the procedural aspect that were not part of aptitude. Thus, we deleted eleven items. Combining the aspects 'quality of the therapeutic relationship' and 'therapeutic relationship' resulted in two items being dropped. We dropped four of the eight items of the aspect 'caring for depressed patients', since their IRF curves correlated two-by-two. This pruning resulted in a questionnaire with three aspects and fourteen components with fourteen corresponding items and a different number of response categories for the items along the aptitude aspect.

## Discussion

Validation of a questionnaire through use of classic PCA does not generate insight into how nurses understand and respond to items and response categories, since this method does not generate transformation plots. However, CatPCA provides these plots, and with them, insight into nurses' understanding of a questionnaire. Hence, the wording of the items and corresponding response categories, and the balance between the wording of an item and the number of response categories, can be adapted according to that insight. Moreover, CatPCA enables analysis on nominal, ordinal, and ratio levels—whatever the measurement level of the data is—enabling the explorative analysis of the data and improving insight into the scope and limitations of the questionnaire. CatPCA has existed for decades but is not commonly used in the development of questionnaires for nursing research.

Usually, once the validity of a questionnaire is established, the questionnaire is used in different contexts without re-evaluating its validity. However, validity is bound to how the questionnaire is used and in what context. Usage of a given questionnaire in contexts in which it was not made to be used compels the re-evaluation of its validity. As a consequence, usage of this questionnaire in another context may require adaptation of the questionnaire to that context.

The results of the present study do not support the suggestion by Meretoja et al. (2004) of using the same assessment questionnaire across national and/or language borders. Translation of a questionnaire does not guarantee that it is still gauged and calibrated to

the culture, semantic frame of reference, and job context of that particular target population. Streiner and Norman (2003) argued that there are five types of equivalencies that must be met before a translation of a questionnaire can be judged as valid: (1) conceptual equivalence, (2) item equivalence, (3) semantic equivalence, (4) operational equivalence, and (5) measurement equivalence. Hence, validity results from one version of a questionnaire are not transposable onto its translated versions. Validity of the translated version must be re-established prior to its use.

## Conclusion

Validation is a process. Understanding the scope and limitations of a questionnaire develops over time, as the questionnaire is used. Thus, validity must be re-established according to its ever-developing scope and limitations. The questionnaire measuring the aptitudes of psychiatric nurses caring for depressed patients is valid for this use and context.

### Key points

- Establishing validity is a process.
- The use of appropriate statistical techniques facilitates the gauging and calibrating of the questionnaire, hence establishing validity.
- Analyses using the appropriate statistical techniques generate insight into how nurses respond to and interpret items and response categories of a questionnaire.
- The validation process is to be re-iterated for each use of a questionnaire in another context.

### Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

### References

American Nurses Association (2000) *Scope and Standards of Psychiatric Mental Health Nursing Practice*. Washington DC: American Nurses Publishing.

Arnold L (2002) Assessing professional behavior: yesterday, today, and tomorrow. Academic medicine. *Journal of the Association of American Medical Colleges* 77(6): 502–515.

Barnett R (1994) *The Limits of Competence: Knowledge, Higher Education and Society*. Buckingham: The Society for Research into Higher Education & Open University Press.

Beutler LE, Malik M, Alimohamed S, Harwood TM, Talebi H, Noble S and Wong E (2004) Therapist Variables. In: Lambert MJ (ed.) *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change*. New York: John Wiley & Sons, 00–00.

Carmines EG and Zeller RA (1979) *Reliability and Validity Assessment*. Beverly Hills, California: Sage Publications.

Coombs CH (1964) *A Theory of Data*. New York: John Wiley & Sons.

Crits-Christoph P, Baranackie K, Kurcias JS, Beck AT, Carroll K, Perry K, et al. (1991) Meta-analysis of therapist effects in psychotherapy outcome studies. *Psychotherapy Research* 1: 81–92.

DuPerron B (2001) *Registered Psychiatric Nurses: Competency Profile for the Profession in Canada*. Edmonton: Alberta Health and Wellness in partnership with Registered Psychiatric Nurses of Canada.

Embretson SE and Reise SP (2000) *Item Response Theory for Psychologists*. Mahwah, NJ: L. Erlbaum Associates.

Embretson SE (2003) *The Second Century of Ability Testing: Some Predictions and Speculations*. Princeton, NJ: Educational Testing Service.

Gifi A (1990) *Nonlinear Multivariate Analysis*. New York: John Wiley & Sons.

Greenacre M and Blasius J (1994) *Correspondence Analysis in the Social Sciences*. London: Academic Press.

Gunderson JG (1978) Defining the therapeutic processes in psychiatric milieus. *Psychiatry* 41(4): 327–335.

Hand DJ (2004) *Measurement Theory and Practice; The World through Quantification*. London: Arnold.

Hoot S (1995) *Standards of Practice for Mental Health Nursing in Australia*. Greenacres South Australia: Australian and New Zealand College of Mental Health Nursing.

Keltner NL, Schwecke LH and Bostrom CE (2003) *Psychiatric Nursing*. St. Louis: Mosby.

Lafferty P, Beutler LE and Crago M (1989) Differences between more and less effective psychotherapists: a study of select therapist variables. *Journal of Consulting and Clinical Psychology* 57(1): 76–80.

Lambert MJ, Bergin AE and Garfield SL (1994) The effectiveness of psychotherapy. In: Bergin AE and Garfield SL (eds) *Handbook of Psychotherapy and Behavior Change*. New York: John Wiley & Sons, 143–189.

Luborsky L, McLellan AT, Woody GE, O'Brien CP and Auerbach A (1985) Therapist success and its determinants. *Archives of General Psychiatry* 42(6): 602–611.

Meretoja R and Leino-Kilpi H (2003) Comparison of competence assessments made by nurse managers and practising nurses. *Journal of Nursing Management* 11(6): 404–409.

Meretoja R, Isoaho H and Leino-Kilpi H (2004) Nurse competence scale: development and psychometric testing. *Journal of Advanced Nursing* 47(2): 124–133.

Milligan F (1998) Defining and assessing competence: the distraction of outcomes and the importance of educational process. *Nurse Education Today* 18(4): 273–280.

Nordhaug O (1993) *Human Capital in Organizations: Competence, Training, and Learning*. Oslo: Scandinavian University Press.

Peplau H (1952) *Interpersonal Relations in Nursing*. New York: G.P. Putnam.

Schultz JM and Videbeck SL (2002) *Lippincott's Manual of Psychiatric Nursing Care Plans*. Philadelphia: Lippincot.

Sijtsma K and Molenaar IW (2002) *Introduction to Nonparametric Item Response Theory Measurement Methods for the Social Sciences*. Thousand Oaks: Sage Publications.

Streiner DL and Norman GR (2003) *Health Measurement Scales: a Practical Guide to their Development and Use*. New York: Oxford University Press.

Videbeck SL (2004) *Psychiatric Mental Health Nursing*. Philadelphia: Lippincott Williams & Wilkins.

Vuorinen R, Tarkka MT and Meretoja R (2000) Peer evaluation in nurses' professional development: a pilot study to investigate the issues. *Journal of Clinical Nursing* 9(2): 273–281.

Watson R, Stimpson A, Topping A and Porock D (2002) Clinical competence assessment in nursing: a systematic review of the literature. *Journal of Advanced Nursing* 39(5): 421–431.

Young F and Hamer R (1987) *Multidimensional Scaling; History, Theory, and Applications*. Hillsdale: Lawrence Erlbaum Associates.

**Marc Haspeslagh** (RPN, MNS, MS) is a registered psychiatric nurse. He received masters' degrees in nursing science and in statistics from the Catholic University Leuven. He is a PhD student at the Catholic University Leuven. His research field is management of psychiatric nurses. During his hospital career, he has had different positions in the nursing department and is now senior advisor of the board. His specialty area is internal control, audit, and strategic planning and management.

**Kristof Eeckloo** (LLM, PhD) is a research fellow at the School of Public Health of the Catholic University of Leuven. His research fields are hospital governance, hospital–physician relationships, public reporting, and healthcare management models. He holds a master's degree in law and a PhD in medical sciences. He is a visiting scientist at Harvard University.

**Lucas Delesie** (DI, MS, PhD) received his DI from the Catholic University of Leuven and MS and PhD in operations research from the University of Pennsylvania. His field is exact, in contrast to assumption-based or simplified, information management, data mining, and visualization. He has worked in a hospital federation, hospital group, and university hospital. He dealt with health policy and health financing systems development. He is professor emeritus at the Catholic University of Leuven.