Journal of Alzheimer's Disease 86 (2022) 1485–1499
DOI 10.3233/JAD-200739
IOS Press

Review

Psychometric Properties of the Neuropsychiatric Inventory: A Review

Toni Saari[a,b,*], Anne Koivisto[a,c,d,e], Taina Hintsa[b], Tuomo Hänninen[c] and Ilona Hallikainen[a]
[a]*University of Eastern Finland, Neurology, Kuopio, Finland*
[b]*University of Eastern Finland, School of Educational Sciences and Psychology, Joensuu, Finland*
[c]*Kuopio University Hospital, Neurology, Kuopio, Finland*
[d]*University of Helsinki, Department of Neurosciences, Helsinki, Finland*
[e]*Helsinki University Hospital, Geriatrics, Department of Internal Medicine and Rehabilitation, Helsinki, Finland*

Accepted 28 July 2020
Pre-press 26 August 2020

Abstract. Neuropsychiatric symptoms cause a significant burden to individuals with neurocognitive disorders and their families. Insights into the clinical associations, neurobiology, and treatment of these symptoms depend on informant questionnaires, such as the commonly used Neuropsychiatric Inventory (NPI). As with any scale, the utility of the NPI relies on its psychometric properties, but the NPI faces unique challenges related to its skip-question and scoring formats. In this narrative review, we examined the psychometric properties of the NPI in a framework including properties pertinent to construct validation, and health-related outcome measurement in general. We found that aspects such as test-retest and inter-rater reliability are major strengths of the NPI in addition to its flexible and relatively quick administration. These properties are desired in clinical trials. However, the reported properties appear to cover only some of the generally examined psychometric properties, representing perhaps necessary but insufficient reliability and validity evidence for the NPI. The psychometric data seem to have significant gaps, in part because small sample sizes in the relevant studies have precluded more comprehensive analyses. Regarding construct validity, only one study has examined structural validity with the NPI subquestions. Measurement error was not assessed in the reviewed studies. For future validation, we recommend using data from all subquestions, collecting larger samples, paying specific attention to construct validity and formulating hypotheses *a priori*. Because the NPI is an outcome measure of interest in clinical trials, examining measurement error could be of practical importance.

Keywords: Alzheimer's disease, behavioral and psychological symptoms of dementia, dementia, measurement, Neuropsychiatric Inventory, neuropsychiatric symptoms, reliability, validity

INTRODUCTION

More than 40 million individuals worldwide are estimated to have Alzheimer's disease (AD) or other dementias [1], driving research efforts to understand these conditions. Neuropsychiatric symptoms (NPS), sometimes referred to as non-cognitive symptoms or behavioral and psychological symptoms of dementia [2], are disturbances in behavior, thought, and emotion related to neurocognitive disorders. These symptoms are common [3], but their measurement remains a challenge.

NPS are often assessed using informant questionnaires. The information obtained with them is not always easy to interpret, as the theories of NPS research either await formulation or their relationship to NPS measures is unclear [4]. Furthermore, NPS constructs overlap with one another [5, 6], and

*Correspondence to: Toni Saari, Yliopistonranta 1B, FIN–70210 Kuopio, Finland. Tel.: +358 50 325 9130; E-mail: toni.saari@uef.fi. Publication history: This manuscript was previously published in PsyArXiv, doi: 10.31234/osf.io/n8pv3.

ISSN 1387-2877 © 2022 – The authors. Published by IOS Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (CC BY-NC 4.0).

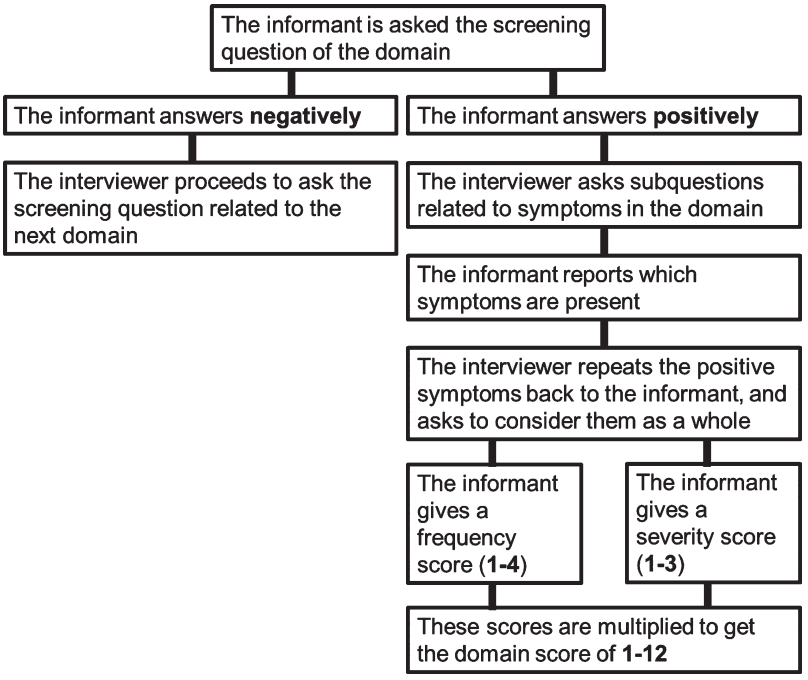

Fig. 1. The skip-question and scoring procedure of the Neuropsychiatric Inventory. This procedure is repeated until all domains are covered.

the associations between constructs, or the behavioral and psychological phenomena of interest, and measures (i.e., scores on an informant questionnaire) [7] remain incompletely described [8, 9]. The lack of clarity about measures arises from the relatively limited psychometric data available regarding most NPS scales.

The most commonly used scale in NPS research is the Neuropsychiatric Inventory (NPI) [10]. This scale is a flexible and a relatively brief informant questionnaire that screens for the most common NPS in AD and other neurocognitive disorders. It has been utilized in both research and the clinic, with applications in clinical trials, neuroimaging, descriptive psychopathology, and as a validity benchmark for developing subsequent NPS scales [11–13]. Modifications for more specific contexts also have been made, yielding the Neuropsychiatric Inventory-Questionnaire (NPI-Q) [14] for very brief screening, the Neuropsychiatric Inventory-Nursing Home for nursing home contexts [15], and the detailed Neuropsychiatric Inventory-Clinician version (NPI-C) [16], integrating information from the patient and informant. Variants of the NPI have been applied outside the context of dementia, including in stroke [17], delirium [18], and cognitively healthy elderly [19]. Some researchers predict that the NPI could be applied to central nervous system (CNS) disorder clinical trials in general, covering "anticonvulsants, Parkinson's disease therapies, hyperactivity/attention deficit disorder, pain therapies, CNS cancer therapies, and multiple sclerosis treatments" [11].

The NPI probes into 10 or 12 symptom domains depending on the version used. The 10 original domains are: delusions, hallucinations, agitation, depression, anxiety, euphoria, apathy, disinhibition, irritability, and aberrant motor behavior [10]. Sleep disturbances and appetite disturbances were added later [12]. Two important features distinguish the NPI from several conventional scales in psychiatry and neurology that rate the severity of individual symptoms on a Likert scale, e.g., from 1 to 5, and sum these values for the total score.

The first distinguishing feature of the NPI is the skip-question format (Fig. 1). In the NPI, informants are first asked screening questions related to the 10 or 12 symptom domains. Following a positive answer, the informant is asked six to nine subquestions per domain. Following a negative answer, the interviewer proceeds to the next domain without asking the subquestions in the screen-negative domain. This procedure saves time but is not without psychometric problems because of the assumption that the subquestions are only relevant if there is a positive answer on the screening question [20].

The second unique property is the scoring system. A screen-positive domain is given a domain score based on the product of the frequency (1 = occasionally, 2 = often, 3 = frequently, 4 = very frequently) and severity (1 = mild, 2 = moderate, 3 = severe) of the symptoms in the behavioral domain. The range is then 1 to 12 per subdomain, and scores of 5, 7, 10, and 11 are not possible. A total score for the NPI can be calculated as a measure of general level of psychopathology (maximum of 120 and 144 for 10 and 12-domain versions, respectively).

This scoring method was chosen with clinical trials in mind: it was considered that the effect of a treatment could manifest separately in reduced symptom frequency or severity [12]. This rationale has been criticized [8], as has the ability of informants to judge the severity of symptoms on behalf of patients [21]. In the NPI-Q, the frequency score was omitted as redundant, because the severity score correlates highly with caregiver distress and the frequency scores in the original NPI [14]. Acknowledging the variability in scoring symptoms of differing frequency and severity under a domain, Connor et al. [22] recommend that the interviewer repeat the positive subquestion symptoms to the informant, asking the informant to consider these symptoms as a whole in determining the severity and frequency scores. This approach addresses the challenge of weighing symptoms of differing frequency and severity by relying on the holistic judgment of the informant [22]. In sum, the subquestions are considered to reflect the same unitary construct as the screening question, the informant condenses the subquestion symptoms into single frequency and severity scores, and these scores are then combined to arrive at the multiplication score of 1 to 12.

In the development of the NPI, a host of psychometric properties were examined [10, 12]. Although these qualities are important, they are only a subset of the features that are usually desired in scale validation [23, 24]. Previous reviews of the NPI [8, 11, 13, 25, 26] have summarized these data, taking them at face value, and solidified the notion that the psychometric data are sufficient and indicative of good reliability and validity. However, drawing inferences based only on the reported metrics may have led to biased conclusions for at least two reasons: the conditions for stricter psychometric analyses have not been met, or the useful properties come at the expense of unexamined properties through overoptimization [24]. Indeed, recent reports suggest that common scales in behavioral sciences either report insufficient psychometric data to evaluate their utility [27], or when examined more closely, have suboptimal validity evidence [24]. All scales face these challenges, but the unconventional structure and scoring method of the NPI may cause additional difficulties in validation.

Despite wide use of this scale, several knowledge gaps persist regarding the NPI. For instance, aspects of construct validity, such as correlations with other tests, are rarely mentioned. Similarly, differences between translated versions and scoring procedures should be considered. Given these cautionary observations and the aspiration that the NPI could be used as an outcome measure in CNS disorders more broadly [11], a thorough review of its psychometric basis is timely and needed. The aim of this narrative review is to examine the psychometric properties of the NPI. Instead of focusing only on what has been reported, we highlight which psychometric properties are seldom included in the translation and validation studies of the NPI. The identified gaps are subsequently discussed as opportunities for refining the NPI.

Review procedures

As noted, existing reviews of the NPI [8, 13, 11, 25, 26] have involved only its reported psychometric properties, without addressing whether these data are sufficient in light of general psychometric recommendations. To address this gap, we used a comprehensive framework for the evaluation of psychometric properties in development and validation studies of the NPI. For the psychometric properties we use in this review, we drew on recent psychometric recommendations for health and behavioral sciences [27, 28] which overlap in major aspects. For example, in addition to the often-reported properties of reliability and internal consistency, these two papers emphasize the importance of structural validity as a part of construct validity.

The Consensus-based Standards for the Selection of Health Measurement Instruments (COSMIN) guidelines for systematic reviews of patient-reported outcomes [28], form the first component of our framework. The guidelines were specifically developed for outcomes that are directly assessed by the patient, so for inclusion in this review, we used only those criteria for selecting good measurement properties (Table 1).

The second component of our review framework is a study of and recommendations for construct validation in social and personality psychology [27].

Table 1
COSMIN guidelines for items to be assessed in systematic reviews of PROMs [28]

Psychometric property
Content validity
Structural validity
Internal consistency
Reliability
Measurement error
Hypotheses testing for construct validity
Cross-cultural validity/measurement invariance
Criterion validity
Responsiveness

Table 2
Flake et al. [27] phases for construct validation

Phase	Validity evidence
Substantive	Literature review and construct conceptualization
	Item development and scaling selection
	Content relevance and representativeness
Structural	Item analysis
	Factor analysis
	Reliability
	Measurement invariance
External	Convergent and discriminant
	Predictive or criterion
	Known groups differences

In that study, the authors interpreted construct validation broadly as "the process of integrating evidence to support the meaning of a number which is assumed to represent a psychological construct" [27]. The authors synthesized decades of psychometric research and suggest that the process of construct validation can be divided into three consecutive phases: substantive, structural, and external (Table 2).

For study selection, TS screened and reviewed full English-language texts of validation or translation studies regarding the 10- or 12-domain NPI published between 1994 (the first NPI publication) and April 2020. The studies were selected from the references in the most recent review of the NPI [11], and complemented with Medline, PsycINFO, Scopus, and Cochrane database searches with the string "neuropsychiatric inventory" AND ("translation" OR "validity" OR "reliability"). Because the NPI was designed to assess psychopathology in dementia, we included papers where the sample comprised mostly patients with dementia. We did not review articles pertaining to later iterations of the NPI (e.g., NPI-Q), and we assessed only English texts [28]. Thus, the Japanese [29], Dutch [30], and Spanish [31] translations were not included because only the related abstracts were available in English. We did retain a study that used an alternative scoring method for the NPI because the authors assessed psychometric properties that the other studies had not addressed [21]. Ultimately, we included 14 studies in this review.

VALIDITY AND RELIABILITY OF THE NPI

Characteristics of included studies

Table 3 shows the reported psychometric properties of the 14 included studies. The values are the product of frequency and severity (frequency × severity), and the coefficients are correlations, percentages, or alpha values unless otherwise indicated. "N/R" indicates that the property was not reported in the reviewed study.

Sample sizes in the included studies were rather modest (M = 89.9, SD = 51.3, range 29–219) although the more recent studies indicated a trend to increasing sample sizes. Samples comprised mostly patients with AD, although in three studies, the proportion of participants with vascular dementia was equal or near equal to that of patients with AD. For inter-rater and test-retest analyses, a subset of the total sample was commonly used. Of the 14 studies, 4 (29%) included unaffected controls. The most frequent study setting was tertiary care, e.g. university hospital neurology or psychiatry outpatient clinics. Overall, the validation studies followed the procedures set forth in the original NPI study, although with some variation.

Psychometric properties

We examined the reviewed psychometric properties in the order in which they are listed in Table 3. The content validity of the original NPI was established through a Delphi panel review of domains arising from clinical experience and existing NPS measures [10, 11]. The translated versions of the NPI showed variability in describing the translation procedures, although in some of the studies, the authors went to great lengths and used pilot studies to ascertain the cultural sensitivity of the measure.

With few exceptions, the studies did not include item analysis data containing, e.g., response distributions and item-total correlations. Baiyewu et al. [36] reported the distributions of the domain frequency scores, and Davidsdottir et al. [42] reported the item-total correlations between domain scores and total NPI scores. In the latter case, the lowest and also statistically non-significant correlations were

Table 3
Psychometric properties reported in the development, validation and translation studies of the Neuropsychiatric Inventory

Authors, year and version	Sample	Setting	Content validity	Structural validity: item analysis	Structural validity: factor analysis or IRT	Internal consistency (Cronbach's alpha)	Test-retest reliability	Interrater reliability	Measurement error	Convergent validity	Discriminant validity	Criterion validity	Cross-cultural validity and measurement invariance	Known groups differences
Cummings 1994; 10-domain [10]	40 (20 AD, 9 VaD, 11 other dementia) for convergent validity; 45 for interrater (42 AD, 1 VaD, 2 other dementia), of which 20 participated in test-retest; 20 caregivers for responsiveness	University or Veterans Affairs dementia clinic or clinical trial participants	Established via Delphi panel	N/R	N/R	Whole scale 0.88, severity 0.88, frequency 0.87	0.51–1	0.89–1	N/R	0.33–0.76 (BEHAVE-AD, HAM-D)	22% domains correlated; MMSE correlated -0.31 to -0.39 with De, Di, An and AMB; age correlated 0.38 with Ap	N/R	N/R	4.5% FP rate; differences across MMSE strata
Binetti et al., 1998; 10-domain [32]	50 Italian (AD); 50 American (AD) for cross-cultural validity analyses	N/R	Back-translation	N/R	N/R	Whole scale 0.76, individual domains 0.68–0.74	Total score 0.78	0.84–1	N/R	N/R	N/R	N/R	Stratified by MMSE, Italian patients had higher total and Ap and AMB scores than American patients	Differences across MMSE strata
Choi et al., 2000; 12-domain [33]	92 Korean (43 AD, 32 VaD, 11 FTLD, 3 PDD, 1 PSP, 1 NPH, 1 TBI); 49 controls; 29 for test-retest reliability analysis, 7 of which were in the control group	Tertiary care	Back-translation, reviewed by an expert group, piloted in 10 patients and further modified.	N/R	N/R	Whole scale 0.85, severity 0.82, frequency 0.81	0.43–0.78	N/R	N/R	N/R	N/R	N/R	N/R	0–22.4% FP rate; differences across MMSE strata
Leung et al., 2001; 10-domain [34]	62 Chinese (41 AD, 16 VaD, 5 other dementia), 29 of which were in the inter-rater reliability analysis	Tertiary care	Back-translation, appraisal by psychiatric experts	N/R	N/R	Whole scale 0.84, severity 0.86, frequency 0.79	N/R	0.92–1	N/R	0.48–0.77 (BEHAVE-AD, HAM-D)	N/R	N/R	N/R	0–11.7% FN and 2.1% FP rate; differences across MMSE strata

Table 3
(Continued)

Authors, year and version	Sample	Setting	Content validity	Structural validity: item analysis	Structural validity: factor analysis or IRT	Internal consistency (Cronbach's alpha)	Test-retest reliability	Interrater reliability	Measurement error	Convergent validity	Discriminant validity	Criterion validity	Cross-cultural validity and measurement invariance	Known groups differences
Fuh et al., 2001; 12-domain [35]	95 Taiwanese (AD); 86 of which were in the test-retest analysis	Tertiary care	Back-translation, reviewed by expert panel	N/R	PCA using domain frequency scores*	Whole scale 0.78, severity 0.78, frequency 0.74	10-domain scale frequency 0.88, severity 0.84, 12-domain scale frequency 0.85, severity 0.82, domains 0.37–0.76 for frequency and 0.34–0.79 for severity	N/R	N/R	N/R	Ha, Di and SD correlated significantly with CDR; MMSE correlated –0.25 with AMB	N/R	N/R	N/R
Baiyewu et al., 2003; 12-domain [36]	40 Nigerian (39 AD, 1 nonspecific dementia), 10 of which were in test-retest and 15 in inter-rater reliability studies	Community	Back-translation, harmonization using established procedures	Distribution for frequency scores	N/R	Whole scale 0.90, severity 0.73, frequency 0.73	Total score 0.81	Whole scale 0.99	N/R	N/R	MMSE correlated –0.32 to –0.47 with De, Ha, and Ag, and –0.33 with total score; ADL 0.33 to 0.5 with Dep, An, SD, and 0.32 with total score	N/R		No differences in NPI total scores stratified by CDR
Politis et al., 2004; 12-domain [37]	29 Greek (AD)	Tertiary care	Back-translation	N/R	N/R	Whole scale 0.76, domains 0.69–0.76	N/R	N/R	N/R	0.48–0.68 (BPRS)	N/R	De, Ag, Apa, AMB, Ir, SD and total score distinguished patients referred for 'behaviors causing fear' vs 'behaviors causing embarassment'	N/R	N/R

Table 3
(Continued)

Authors, year and version	Sample	Setting	Content validity	Structural validity: item analysis	Structural validity: factor analysis or IRT	Internal consistency (Cronbach's alpha)	Test-retest reliability	Interrater reliability	Measurement error	Convergent validity	Discriminant validity	Criterion validity	Cross-cultural validity and measurement invariance	Known groups differences
Kørner et al., 2008; 12-domain [38]	72 Icelandic (59 AD, 8 VaD, 5 other dementia); 29 controls; 84 of the combined sample participated in the test-retest analysis and 17 in the inter-rater reliability analysis	Tertiary care	Back-translation	N/R	Loevinger coefficients 0.25 for whole scale severity, frequency and total scores*	N/R	Total score 0.88, no statistically significant differences between domains' change scores	Whole scale 0.94	N/R	N/R	N/R	N/R	Rasch analyses indicate measurement invariance for sex	Differences in total score across dementia severity
Camozzato et al., 2008; 12-domain [39]	36 Brazilian (AD), all of which participated in test-retest and inter-rater reliability studies	Tertiary care	Back-translation, adaptation to ensure cultural and educational comprehension	Distribution for frequency scores	N/R	Severity 0.7	Total score 0.82, 0.86 severity, 0.82 frequency; domains 0.4–0.97	Total score 0.98, severity 0.96; domains 0.12–0.91	N/R	N/R	N/R	N/R	N/R	N/R
Gallo et al., 2009; 12-domain NPI-A [21]	124 (62 AD, 43 VaD, 19 mixed dementia)	Outpatient memory assessment program	N/R	N/R	PCA for all items under the 12 domains, resulting in 3-component structure	All items 0.96, domains 0.57–0.91	N/R	N/R	N/R	N/R	N/R	N/R	N/R	N/R
Wang et al., 2012; 12-domain [40]	219 Mainland Chinese (AD)	Tertiary care	Back-translation, appraisal by psychiatric experts	N/R	PCA using domain scores*	Whole scale 0.69	Total score 0.96, domains 0.66–0.95	N/R	N/R	N/R	N/R	N/R	N/R	N/R

Table 3
(Continued)

Authors, year and version	Sample	Setting	Content validity	Structural validity: item analysis	Structural validity: factor analysis or IRT	Internal consistency (Cronbach's alpha)	Test-retest reliability	Interrater reliability	Measurement error	Convergent validity	Discriminant validity	Criterion validity	Cross-cultural validity and measurement invariance	Known groups differences
Malakouti et al., 2012; 12-domain [41]	100 Iranian (diagnosis of dementia), 50 of which participated in inter-rater reliability analyses, of these, 30 were randomly selected for test-retest; the other 50 participated in convergent validity analyses; 49 controls	Convenience sample	Back-translation, appraisal by researchers, pilot study in four caregivers	N/R	N/R	Whole scale 0.8, domains 0.73–0.82	0.51–0.95	0.59–0.98	N/R	0.3–0.9 (PANSS, GDS-15)	MMSE correlated –0.34 to –0.56 with Ap, SD, Ag, AMB, and –0.49 with NPI total score	N/R	N/R	Ag, An, Ir and Eu elevated in controls; differences across MMSE strata
Davidsdottir et al., 2012; 12-domain [42]	38 Icelandic (19 AD, 19 VaD)	Tertiary care	Back-translation by a translator blinded to the original NPI, pilot study	Item-total correlations 0.25–0.69	N/R	Whole scale 0.81, frequency 0.76, severity 0.78	Total score 0.86, domains 0.38–0.96	N/R	N/R	0.18–0.9 (BEHAVE-AD, GDS-30)	N/R	N/R	N/R	Ap scores associated with disease severity
Ferreira et al., 2015; 12-domain [43]	166 European Portuguese ("cognitive deficits" 60%)	Nursing home	Translated, details unavailable in English	N/R	N/R	Whole scale 0.76, domains 0.71–0.77	Total score 0.91, domains 0.3–0.98	N/R	N/R	0.17 Depression domain with GDS	MMSE correlated –0.17 to –0.18 with De, Di and AMB	N/R	N/R	N/R

Combined items from Prinsen et al. [28] as well as Flake, Pek, and Hehman [27]. Values are correlations, percentages or coefficient alpha for frequency × severity scores, unless otherwise indicated. *This is not a method of structural validation in the traditional sense, as it aims to find higher-order structures, not that the existing structures perform as intended. Abbreviations for NPI domains: De, delusions; Ha, hallucinations; Ag, agitation; Dep, depression; Eu, euphoria; An, anxiety; Ap, apathy; Di, disinhibition; Ir, irritability; AMB, aberrant motor behavior; SD, sleep disturbances; AED, appetite and eating disturbances. Other abbreviations: AD, Alzheimer's disease; BEHAVE-AD, Behavioral Pathology in Alzheimer's Disease Rating Scale; BPRS, Brief Psychiatric Rating Scale; CDR, Clinical Dementia Rating scale; FN, false negative; FTLD, fronto-temporal lobar degeneration; FP, false positive; GDS, Geriatric Depression Scale; HAM-D, Hamilton Depression Rating Scale; IRT, item response theory; MMSE, Mini-Mental State Examination; NPI-A, Neuropsychiatric Inventory-Alternative; NPH, normal pressure hydrocephalus; PANSS, Positive and Negative Symptoms Scale; PSP, progressive supranuclear palsy; VaD, vascular dementia

hallucinations ($r = 0.25$) and euphoria ($r = 0.30$), whereas the highest was depression ($r = 0.69$, $p < 0.001$). None of the studies reported data on the subquestion response distributions.

One group relied on factor analysis (FA) or item response theory (IRT) methods to analyze whether the subquestions reflect a single construct implied by the screening question. This group, Gallo et al. [21], used an alternative scoring method, with the frequencies of all NPI subquestions rated by 124 caregivers of patients with either AD or vascular or mixed dementia. This approach allowed the researchers to conduct a principal component analysis (PCA), with results suggesting that the subquestions in the depression, anxiety, apathy, irritability, and disinhibition domains form unified components, as envisioned in the development of the NPI. However, the rest of the subquestions from the remaining seven domains loaded onto two or three components, suggesting that these domain scores address more than one psychopathological construct. Davidsdottir et al. [42] used IRT, and found that the NPI total score did not reflect a unidimensional construct. Wang et al. [40] used PCA to observe that the domain scores could be grouped into syndromes, respectively.

Internal consistency, or coefficient alpha, was estimated in almost all studies. The studies did vary somewhat in whether alpha was estimated for 1) the domain scores of frequency, severity, and frequency × severity, or 2) the total score combining all 10 or 12 domain scores, either frequency, severity, or frequency × severity. Generally, the values for alpha can be interpreted at least as being adequate. It is also not entirely evident how the domain-specific alphas were calculated in each instance.

Of the 14 studies, 11 involved assessment of test-retest reliability. This aspect of reliability was a strength in these studies, particularly for the total score (range, 0.78–0.96, available in seven studies). Individual domains with the lowest test-retest reliability varied among studies (and among cultures): appetite disturbances [33], disinhibition [44], irritability [10, 42], and agitation [41]. Inter-rater reliability was evaluated in half of the studies and ranged from 0.12 to 1.0. Measurement error was not assessed in the reviewed studies.

Regarding construct validity properties, convergent validity was estimated with the Behavioral Pathology in Alzheimer's Disease Rating Scale (BEHAVE-AD) [45] as the scale of comparison in three of the six studies reporting these analyses. The BEHAVE-AD includes subscales for delusions, hallucinations, anxiety and phobias, affective disturbance, activity disturbance, and sleep as well as a total score. The Hamilton Rating Scale for Depression [46] and variants of the Geriatric Depression Scale [47] were used for correlations with depression domain score. With few exceptions [41], the convergent validity of euphoria, disinhibition, irritability, apathy, and appetite disturbances was not assessed in these studies.

Discriminant validity, or establishing that two theoretically distinct constructs are not correlated, was examined mostly between Mini-Mental State Examination (MMSE) and NPI domain and total scores. With the exception of the original NPI validation study [10], the independence of theoretically distinct NPS constructs was not examined. From the original NPI study [10], we can learn that 22% domains (frequency and severity scores separately) were correlated, with some examples provided. Politis et al. examined criterion validity as an index of external validity, in which some NPI domain scores and the total score distinguished different causes for referral [37].

In one study, the authors directly examined cross-cultural validity, in which Italian patients matched for MMSE with their U.S. counterparts had higher scores on apathy and aberrant motor behavior and higher total scores [32]. In an Icelandic study, Rasch analyses were used to suggest measurement invariance for gender [42]. Known groups differences were examined between controls and patients, or by stratifying the patient group with MMSE scores or disease severity. (For a review of studies examining differences in NPI scores between neurological disorders, see reference [11]). Among the studies, the highest false-positive rate for a screening question in controls was 22% for depression [33], whereas the false-negative frequency was more than 11% for depression, sleep, and appetite disturbances in the study that examined this outcome [34].

FA of the NPI

Only one of the reviewed studies included an analysis of the structural validity of the subquestions. Here, it is useful to follow the terminology that these authors used, in which they distinguished between an "item score" and a "domain score" [21]. The researchers modified administration of the NPI, asking all subquestions without the screening questions, and the item score referred to the individual frequency rating of a subquestion. It is important to note that here, the

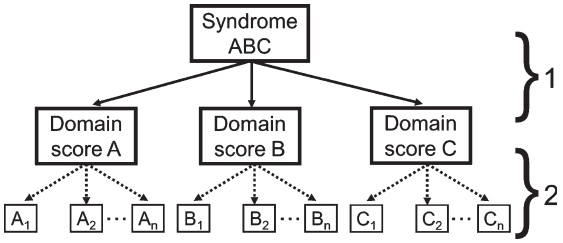

Fig. 2. Two levels of factor analyses of the Neuropsychiatric Inventory. 1) The syndrome level is where the majority of the NPI factor analytic studies has taken place. The aim of this research is to explore whether latent variables (syndrome ABC), often called a syndrome, would underlie the correlations between different domain scores (A, B, C) of the NPI. These studies implicitly assume that the domain scores can be used as useful indicators for a latent variable. This is the level of factor analytic research reviewed by Canevelli et al. [48]. 2) The subquestion level is critical for establishing structural validity, but it has not been extensively studied, indicated by the dashed lines. The aim of this research is to show that the subquestions (e.g.., A_1, A_2, ... A_n) address a unidimensional construct (the one suggested by the screening question), justifying their use in scoring the domain. Structural validity studies can reveal, for example, that the relationships between subquestions and the latent construct are not strong enough, that the subquestions under a single domain address more than one construct like in the study by Gallo et al., or that subquestions from a domain could reflect some other construct instead, or in addition to, the one it is intended to. To explore potential cross-loadings (e.g., C_1 to domain C and A), asking all subquestions from the informant without screening questions is required.

score refers to a single item, not to the domain. As mentioned earlier, the domain score is typically used in NPI research and is based on the screening procedure, subquestion probing, informant judgment of the holistic frequency and severity of the symptom domain, and ultimately the product of the frequency and severity scores. This painstaking description is crucial for unraveling FA studies of the NPI, in which domain scores are used similarly to simpler scores, such as a Likert rating on a single symptom.

The NPI domain scores have been subjected to numerous FAs, in which the objective has been to derive syndromes of related symptom domains. For example, a psychotic syndrome could include delusion, hallucination, and aberrant motor behavior domain scores of the NPI. This *post hoc* line of research was not specified in the development of the NPI. It also has a different aim from FA, using item scores acquired via alternative scoring methods, as in the Gallo et al. study, to demonstrate whether the subquestions refer to a single, independent construct in the first instance (Fig. 2). Structural validity based on subquestion data would substantiate the inferences concerning any domain score (including zero) and further combinations of domain scores in syndrome models.

In a review of the FA literature related to the NPI, Canevelli et al. [48] found that none of the studies produced a unidimensional factor structure (as sometimes approximated in the validation studies reviewed in the present paper), and no two studies identified the same factor structure. The authors also reported that an exhaustive factor model was not established in almost 30% of NPI FAs (i.e., some domains often failed to associate with any factor, commonly aberrant motor behavior, appetite disturbances, and sleep disturbances). The NPI domain syndromes are also temporally unstable [49]. These findings are important for two reasons. First, they call into question the utility of estimating the internal consistency of the entire NPI, as was done in some studies reviewed here. It is theoretically unlikely and not even desired that the combination of all domains reflects a unidimensional structure. Second, in addition to other factors, such as sample size issues and study population variability, the heterogeneity of the syndrome structures may stem from domain scores being suboptimal indicators for a latent variable, given the assumption made with little empirical support that they reflect only one construct.

DISCUSSION

We examined the psychometric properties of the NPI in a framework including those properties pertinent for construct validation and health-related outcome measurement in general. In agreement with previous reviews, we found that aspects of reliability are major strengths of the NPI, perhaps explaining the measure's "gold standard" status [26]. The inter-rater and test-retest reliabilities are particularly key in clinical trial settings. Although flexibility and relatively quick administration were not properties explicitly included in the review, they also are major strengths of the NPI. Thus, it is not surprising that the NPI has been applied in several settings where a rating scale of similar scope has been lacking.

The main issue, however, is that the reported properties cover only some of the generally used psychometric properties, representing what might be necessary but insufficient reliability and validity evidence for the NPI. The psychometric data appear to have significant gaps, at least in part because the small sample sizes in the studies precluded more comprehensive analyses. Regarding construct validity, we

find it concerning that only one study [21] examined structural validity with the NPI subquestions. Those authors suggested that more than half of the subquestions currently considered to reflect a single domain could represent two or three constructs. These findings should be considered preliminary, however, because this group used exploratory methods and a small sample size relative to the number of items. If some domains are indeed multidimensional, this factor could limit their utility as validity benchmarks for other scales (and vice versa), given the lack of clarity about the contributions these different constructs make to a correlation [50, 51]. Discriminant validity, or the extent to which theoretically distinct constructs do not correlate with one another, was mainly investigated regarding global cognition. This assessment was important for demonstrating that the NPI cannot be reduced to a measure of dementia severity. However, it would also be pertinent to show that theoretically distinct neuropsychiatric symptom constructs do not correlate in undesired ways. The existing FA literature on NPI domain scores addresses a different aim altogether, namely one of exploring possible higher-order syndromes, and should not be confused with construct validation in the traditional sense.

Coefficient alpha was used in the studies as an estimate of internal consistency of either the entire NPI or domains of it. The interpretability of coefficient alpha hinges on several assumptions, the most pertinent being that the items form a unidimensional latent variable model [23, 52, 53]. Most studies did not include examination of this property, which would have required larger sample sizes and alternative scoring methods [20, 23]. As the authors of one of the reviewed studies noted [42], using coefficient alpha for the entire NPI seems counterintuitive because the domains should not form an internally consistent aggregate; rather, the aim is to assess 10 to 12 distinct symptom domains.

Data for a subset of patients enrolled in the studies were analyzed in the test-retest and inter-rater reliability analyses. Although the reporting was transparent regarding how many patients were included in these analyses, the number of patients who were screen-positive for the assessed domains in these subsets was less obvious. For example, the original NPI study [10] had no patients with euphoria in the test-retest analysis, and a correlation of 1.0 between hallucination severity scores at the first and second evaluations suggests that few patients were assessed in this analysis.

None of the reviewed studies included assessment of measurement error. In our experience, this finding is typical in the scale validation literature, and its significance should be interpreted in the context of the intended use of the scale. Because the NPI scoring method was chosen with clinical trials in mind [12], information about measurement error could aid in interpreting trial results where the clinical significance of changes in outcome measures is of interest. Sample sizes in trials of NPS with dementia are increasingly large—some authors even suggest that they are too large [54]—and thus, sufficiently powered to detect even minute changes in the primary outcome, often some variant of the NPI. Thus, it could be of practical importance to assess the lower limit of changes that the NPI can detect and the magnitude of change that can be considered clinically important [55]. Additional challenges may arise from the non-parametric distribution of domain scores.

Taken together, the reviewed studies suggest that the validation and translation of the NPI seem to have followed homogenous procedures in the last two decades. Although this consistency has the benefit of allowing estimation of stability of some properties across studies and cultures, it has led to systematic neglect of some aspects of validity. Below, we outline some recommendations for future research.

Regardless of which variant of the NPI is going to be further refined, the sample sizes need to be larger for FA methods in structural validation. Some authors have recommended rules of thumb regarding minimum sample sizes in FA, such as 200 to 300 [23]. Others have refrained from such suggestions, however, because properties such as the communality of the items can lead to reliable FAs in smaller groups [56]. Nonetheless, this aspect can also be considered a question of research priorities. To advance the study of AD in large collaborative projects, neuroimaging can be performed on thousands of unaffected controls, at-risk individuals, and those with a diagnosis of a neurocognitive disorder, whereas collecting item-level data on NPS scales is possible for a fraction of the cost. Gathering data for further validation of the NPI can be achieved alongside many such pioneering research projects.

The skip-question format of the NPI poses specific problems for assessing construct validity at the level of individual items. In the end, the domain scores are a result of the informant ascribing some numbers to the whole of subquestions in the domain, so that analyzing these subquestions should be the basis for any future NPI validation projects. Ensuring this

aim, however, requires an alternative scoring method in which informants are asked all subquestions, not only those tied to positive screening answers. The issue could be further explored in a test-retest setting assessing whether informants answer similarly to the standard NPI screening and related subquestions and the alternative version with subquestions only. This line of inquiry could delineate the limitations of having statistical contingencies between subquestions and screening questions. A perhaps more feasible approach would be further development of the NPI-C in which all items are queried. However, the extensive number of items on the NPI-C poses statistical and practical challenges.

Perhaps because of the relatively limited theoretical literature available, researchers often have not explicitly stated their hypotheses regarding convergent and discriminant validity. Statistical significance between two scales should not be a deciding factor *per se,* and the magnitude of correlations is what is important. The COSMIN guidelines suggest some generic hypotheses regarding systematic reviews of the convergent and discriminant validity of patient-reported outcomes: for example, scales measuring the same construct should correlate at 0.50 or higher, whereas those measuring distinct constructs should correlate under 0.30 [28]. With two decades of validating, developing, and translating the NPI, the literature is sufficient to allow formulation of *a priori* hypotheses regarding these properties.

This review has focused mostly on the traditional latent variable modeling approach to psychometrics. Rather recently, however, it has been suggested that modeling scale items as networks of conditional relationships could resonate with the nature of the constructs more accurately [57, 58]. The core idea of network models is that variables are connected to one another through conditional relationships, so that only relationships that have taken into account the shared variance of all symptoms are included. The result is a sparse network, depicting unique relationships among items. Networks correspond with the notion that symptoms of psychiatric disorders interact so that the presence of one symptom is likely to increase the possibility of another symptom. Regarding NPS, network methods have already been used in a few studies [59, 60]

Even more recent developments concern the so-called hybrid models, which combine both latent variable and network properties, such as using the residuals of latent variables to estimate networks [61]. A clinical interpretation of these models could be as follows: the latent variables represent a change in brain structure or metabolism associated with the disease, whereas the networks represent the remaining psychological and behavioral phenomena unaccounted for by the latent variable and their relationships with one another. This interpretation is compatible with etiological accounts of NPS [62].

The NPI literature can be confusing because the domain scores are often called "symptoms" or "items." Given the lack of clarity about which and how many symptoms are referenced in each situation and that the scores result from a multiphasic process, we suggest using the term "domain scores." The subquestions of the NPI in turn can be considered to reflect symptoms through item scores. These distinctions preserve the separation between constructs and measures [7], and attach the NPI to psychiatric and neurological scales more generally, in which symptoms often are more narrowly defined.

Finally, we have sought to include in this review some actionable suggestions for using and developing future NPS scales. Developers would benefit from proceeding according to the substantive, structural, and external phases outlined in Table 2. These phases build on one another, so that having data on a latter property is difficult to interpret in the absence of previous properties. Although this compilation of information is relatively recent [27], the literature itself is not. Indeed, the classical methods presented in the construct validity literature have been used successfully in NPS scale development [6].

CONCLUSIONS

The NPI is considered the gold standard for NPS scales and has been used to deepen understanding of NPS in both research and clinical contexts in the past two decades. Based on our review, the reliability of the NPI is its greatest psychometric strength. However, because the screening and scoring rely on information related to individual subquestions, we were concerned to find that only one study examined the structural validity of the subquestions in the NPI, whereas other investigations of construct validity were limited in scope. These findings point to uncertainty regarding whether the current NPI formulation reflects the intended constructs. Future research should target addressing these gaps and aiming for conceptual clarity.

ACKNOWLEDGMENTS

TS was supported by grants from the Finnish Brain Foundation and the Finnish Cultural Foundation. We thank UEF Neurology and the Brain Research Unit for providing facilities to carry out this research. We also thank administrative assistant Mari Tikkanen for invaluable help.

Authors' disclosures available online (https://www.j-alz.com/manuscript-disclosures/20-0739r1).

REFERENCES

[1] Nichols E, Szoeke CEI, Vollset SE, Abbasi N, Abd-Allah F, Abdela J, Aichour MTE, Akinyemi RO, Alahdab F, Asgedom SW, Awasthi A, Barker-Collo SL, Baune BT, Béjot Y, Belachew AB, Bennett DA, Biadgo B, Bijani A, Bin Sayeed MS, Brayne C, Carpenter DO, Carvalho F, Catalá-López F, Cerin E, Choi JYJ, Dang AK, Degefa MG, Djalalinia S, Dubey M, Duken EE, Edvardsson D, Endres M, Eskandarieh S, Faro A, Farzadfar F, Fereshtehnejad S-M, Fernandes E, Filip I, Fischer F, Gebre AK, Geremew D, Ghasemi-Kasman M, Gnedovskaya EV, Gupta R, Hachinski V, Hagos TB, Hamidi S, Hankey GJ, Haro JM, Hay SI, Irvani SSN, Jha RP, Jonas JB, Kalani R, Karch A, Kasaeian A, Khader YS, Khalil IA, Khan EA, Khanna T, Khoja TAM, Khubchandani J, Kisa A, Kissimova-Skarbek K, Kivimäki M, Koyanagi A, Krohn KJ, Logroscino G, Lorkowski S, Majdan M, Malekzadeh R, März W, Massano J, Mengistu G, Meretoja A, Mohammadi M, Mohammadi-Khanaposhtani M, Mokdad AH, Mondello S, Moradi G, Nagel G, Naghavi M, Naik G, Nguyen LH, Nguyen TH, Nirayo YL, Nixon MR, Ofori-Asenso R, Ogbo FA, Olagunju AT, Owolabi MO, Panda-Jonas S, Passos VM de A, Pereira DM, Pinilla-Monsalve GD, Piradov MA, Pond CD, Poustchi H, Qorbani M, Radfar A, Reiner RC, Robinson SR, Roshandel G, Rostami A, Russ TC, Sachdev PS, Safari H, Safiri S, Sahathevan R, Salimi Y, Satpathy M, Sawhney M, Saylan M, Sepanlou SG, Shafieesabet A, Shaikh MA, Sahraian MA, Shigematsu M, Shiri R, Shiue I, Silva JP, Smith M, Sobhani S, Stein DJ, Tabarés-Seisdedos R, Tovani-Palone MR, Tran BX, Tran TT, Tsegay AT, Ullah I, Venketasubramanian N, Vlassov V, Wang YP, Weiss J, Westerman R, Wijeratne T, Wyper GMA, Yano Y, Yimer EM, Yonemoto N, Yousefifard M, Zaidi Z, Zare Z, Vos T, Feigin VL, Murray CJL (2019) Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol* **18**, 88-106.

[2] Finkel SI, Costa e Silva J, Cohen G, Miller S, Sartorius N (1996) Behavioral and psychological signs and symptoms of dementia: A consensus statement on current knowledge and implications for research and treatment. *Int Psychogeriatr* **8**, 497-500.

[3] Zhao QF, Tan L, Wang HF, Jiang T, Tan MS, Tan L, Xu W, Li JQ, Wang J, Lai TJ, Yu JT (2016) The prevalence of neuropsychiatric symptoms in Alzheimer's disease: Systematic review and meta-analysis. *J Affect Disord* **190**, 264-271.

[4] Cummings JL (1997) Theories behind existing scales for rating behavior in dementia. *Int Psychogeriatr* **8**, 293-300.

[5] Cummings J, Mintzer J, Brodaty H, Sano M, Banerjee S, Devanand DP, Gauthier S, Howard R, Lanctôt K, Lyketsos CG, Peskind E, Porsteinsson AP, Reich E, Sampaio C, Steffens D, Wortmann M, Zhong K (2015) Agitation in cognitive disorders: International Psychogeriatric Association provisional consensus clinical and research definition. *Int Psychogeriatr* **27**, 7-17.

[6] Marin RS, Firinciogullari S, Biedrzycki RC (1993) The sources of convergence between measures of apathy and depression. *J Affect Disord* **28**, 117-124.

[7] Edwards JR, Bagozzi RP (2000) On the nature and direction of relationships between constructs and measures. *Psychol Methods* **5**, 155-174.

[8] Perrault A, Oremus M, Demers L, Vida S, Wolfson C (2000) Review of outcome measurement instruments in Alzheimer's disease drug trials: Psychometric properties of behavior and mood scales. *J Geriatr Psychiatry Neurol* **13**, 181-196.

[9] Seignourel PJ, Kunik ME, Snow L, Wilson N, Stanley M (2008) Anxiety in dementia: A critical review. *Clin Psychol Rev* **28**, 1071-1082.

[10] Cummings JL, Mega M, Gray K, Rosenberg-Thompson S, Carusi DA, Gornbein J (1994) The Neuropsychiatric Inventory: Comprehensive assessment of psychopathology in dementia. *Neurology* **44**, 2308-2308.

[11] Cummings J (2020) The Neuropsychiatric Inventory: Development and applications. *J Geriatr Psychiatry Neurol* **33**, 73-84.

[12] Cummings JL (1997) The Neuropsychiatric Inventory: Assessing psychopathology in dementia patients. *Neurology* **48**, S10-S16.

[13] Lai CK (2014) The merits and problems of Neuropsychiatric Inventory as an assessment tool in people with dementia and other neurological disorders. *Clin Interv Aging* **9**, 1051-1061.

[14] Kaufer DI, Cummings JL, Ketchel P, Smith V, MacMillan A, Shelley T, Lopez OL, DeKosky ST (2000) Validation of the NPI-Q, a brief clinical form of the Neuropsychiatric Inventory. *J Neuropsychiatry Clin Neurosci* **12**, 233-239.

[15] Wood S, Cummings JL, Hsu MA, Barclay T, Wheatley MV, Yarema KT, Schnelle JF (2000) The use of the Neuropsychiatric Inventory in nursing home residents: Characterization and measurement. *Am J Geriatr Psychiatry* **8**, 75-83.

[16] de Medeiros K, Robert P, Gauthier S, Stella F, Politis A, Leoutsakos J, Taragano F, Kremer J, Brugnolo A, Porsteinsson AP, Geda YE, Brodaty H, Gazdag G, Cummings J, Lyketsos C (2010) The Neuropsychiatric Inventory-Clinician rating scale (NPI-C): Reliability and validity of a revised assessment of neuropsychiatric symptoms in dementia. *Int Psychogeriatr* **22**, 984-994.

[17] Wong A, Cheng ST, Lo ESK, Kwan PWL, Law LSN, Chan AYY, Wong LK-S, Mok V (2014) Validity and reliability of the Neuropsychiatric Inventory questionnaire version in patients with stroke or transient ischemic attack having cognitive impairment. *J Geriatr Psychiatry Neurol* **27**, 247-252.

[18] Leonard M, McInerney S, McFarland J, Condon C, Awan F, O'Connor M, Reynolds P, Meaney AM, Adamis D, Dunne C, Cullen W, Trzepacz PT, Meagher DJ (2016) Comparison of cognitive and neuropsychiatric profiles in hospitalised elderly medical patients with delirium, dementia and comorbid delirium-dementia. *BMJ Open* **6**, e009212.

[19] Yoro-Zohoun I, Nubukpo P, Houinato D, Mbelesso P, Ndamba-Bandzouzi B, Clément J, Dartigues J, Preux P, Guerchet M, for the EPIDEMCA Group (2019)

Neuropsychiatric symptoms among older adults living in two countries in Central Africa (EPIDEMCA study). *Int J Geriatr Psychiatry* **34**, 169-178.

[20] Borsboom D, Fried EI, Epskamp S, Waldorp LJ, van Borkulo CD, van der Maas HLJ, Cramer AOJ (2017) False alarm? A comprehensive reanalysis of "Evidence that psychopathology symptom networks have limited replicability" by Forbes, Wright, Markon, and Krueger (2017). *J Abnorm Psychol* **126**, 989-999.

[21] Gallo JL, Schmidt KS, Libon DJ (2009) An itemized approach to assessing behavioral and psychological symptoms in dementia. *Am J Alzheimers Dis Other Demen* **24**, 163-168.

[22] Connor DJ, Sabbagh MN, Cummings JL (2008) Comment on administration and scoring of the Neuropsychiatric Inventory in clinical trials. *Alzheimers Dement* **4**, 390-394.

[23] Clark LA, Watson D (1995) Constructing validity basic issues in objective scale development. *Psychol Assess* **7**, 309-319.

[24] Hussey I, Hughes S (2020) Hidden invalidity among 15 commonly used measures in social and personality psychology. *Adv Methods Pract Psychol Sci* **3**, 166-184.

[25] Woods D, Buckwalter K (2018) Taking another look: Thoughts on behavioral symptoms in dementia and their measurement. *Healthcare* **6**, 1-14.

[26] Jeon YH, Sansoni J, Low LF, Chenoweth L, Zapart S, Sansoni E, Marosszeky N (2011) Recommended measures for the assessment of behavioral disturbances associated with dementia. *Am J Geriatr Psychiatry* **19**, 403-415.

[27] Flake JK, Pek J, Hehman E (2017) Construct validation in social and personality research: Current practice and recommendations. *Soc Psychol Personal Sci* **8**, 370-378.

[28] Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, Terwee CB (2018) COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res* **27**, 1147-1157.

[29] Hirono N, Mori E, Ikejiri Y, Shimomura T, Hashimoto M, Yamashita H, Ikeda M (1997) Japanese version of the Neuropsychiatric Inventory–a scoring system for neuropsychiatric disturbance in dementia patients. *No To Shinkei* **49**, 266-271.

[30] Kat M, de Jonghe JFM, Aalten P, Kalisvaart K, Dröes R-M, Verhey F (2002) Neuropsychiatric symptoms of dementia: Psychometric aspects of the Dutch Neuropsychiatric Inventory (NPI). *Tijdschr Gerontol Geriatr* **33**, 150-155.

[31] Vilalta-Franch J, Lozano-Gallego M, Hernández-Ferrándiz M, Llinàs-Reglà J, López-Pousa S, López O (1999) The Neuropsychiatric Inventory. Psychometric properties of its adaptation into Spanish. *Rev Neurol* **29**, 15-19.

[32] Binetti G, Mega MS, Magni E, Padovani A, Rozzini L, Bianchetti A, Trabucchi M, Cummings JL (1998) Behavioral disorders in Alzheimer disease: A transcultural perspective. *Arch Neurol* **55**, 539-544.

[33] Choi SH, Na DL, Kwon HM, Yoon SJ, Jeong JH, Ha CK (2000) The Korean version of the Neuropsychiatric Inventory: A scoring tool for neuropsychiatric disturbance in dementia patients. *J Korean Med Sci* **15**, 609-615.

[34] Leung VPY, Lam LCW, Chiu HFK, Cummings JL, Chen QL (2001) Validation study of the Chinese version of the neuropsychiatric inventory (CNPI). *Int J Geriatr Psychiatry* **16**, 789-793.

[35] Fuh JL, Liu CK, Mega MS, Wang SJ, Cummings JL (2001) Behavioral disorders and caregivers' reaction in Taiwanese patients with Alzheimer's disease. *Int Psychogeriatr* **13**, 121-128.

[36] Baiyewu O, Smith-Gamble V, Akinbiyi A, Lane KA, Hall KS, Ogunniyi A, Gureje O, Hendrie HC (2003) Behavioral and caregiver reaction of dementia as measured by the neuropsychiatric inventory in Nigerian community residents. *Int Psychogeriatr* **15**, 399-409.

[37] Politis AM, Mayer LS, Passa M, Maillis A, Lyketsos CG (2004) Validity and reliability of the newly translated Hellenic Neuropsychiatric Inventory (H-NPI) applied to Greek outpatients with Alzheimer's disease: A study of disturbing behaviors among referrals to a memory clinic. *Int J Geriatr Psychiatry* **19**, 203-208.

[38] Kørner A, Lauritzen L, Lolk A, Abelskov K, Christensen P, Nilsson FM (2008) The Neuropsychiatric Inventory—NPI. Validation of the Danish version. *Nord J Psychiatry* **62**, 481-485.

[39] Camozzato AL, Kochhann R, Simeoni C, Konrath CA, Pedro Franz A, Carvalho A, Chaves ML (2008) Reliability of the Brazilian Portuguese version of the Neuropsychiatric Inventory (NPI) for patients with Alzheimer's disease and their caregivers. *Int Psychogeriatr* **20**, 383-393.

[40] Wang T, Xiao S, Li X, Wang H, Liu Y, Su N, Fang Y (2012) Reliability and validity of the Chinese version of the neuropsychiatric inventory in mainland China: Reliability and validity of the Chinese version of the NPI. *Int J Geriatr Psychiatry* **27**, 539-544.

[41] Malakouti SK, Panaghi L, Foroughan M, Salehi M, Zandi T (2012) Farsi version of the Neuropsychiatric Inventory: Validity and reliability study among Iranian elderly with dementia. *Int Psychogeriatr* **24**, 223-230.

[42] Davidsdottir SR, Snaedal J, Karlsdottir G, Atladottir I, Hannesdottir K (2012) Validation of the Icelandic version of the Neuropsychiatric Inventory with caregiver distress (NPI-D). *Nord J Psychiatry* **66**, 26-32.

[43] Ferreira AR, Martins S, Ribeiro O, Fernandes L (2015) Validity and reliability of the European Portuguese version of Neuropsychiatric Inventory in an institutionalized sample. *J Clin Med Res* **7**, 21-28.

[44] Ferreira D, Cavallin L, Granberg T, Aguilar C, Vellas B, Tsolaki M, Kłoszewska I, Soininen H, Lovestone S, Simmons A, Wahlund LO, Westman E; AddNeuroMed consortium and for the Alzheimer's Disease Neuroimaging Initiative (2016) Quantitative validation of a visual rating scale for frontal atrophy: Associations with clinical status, APOE e4, CSF biomarkers and cognition. *Eur Radiol* **26**, 2597-2610.

[45] Reisberg B, Auer SR, Monteiro IM (1997) Behavioral pathology in Alzheimers disease (BEHAVE-AD) rating scale. *Int Psychogeriatr* **8**, 301-308.

[46] Hamilton M (1967) Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol* **6**, 278-296.

[47] Yesavage JA, Sheikh JI (1986) Geriatric Depression Scale (GDS): Recent evidence and development of a shorter version. *Clin Gerontol* **5**, 165-173.

[48] Canevelli M, Adali N, Voisin T, Soto ME, Bruno G, Cesari M, Vellas B (2013) Behavioral and psychological subsyndromes in Alzheimer's disease using the Neuropsychiatric Inventory: Behavioral subsyndromes in Alzheimer's disease. *Int J Geriatr Psychiatry* **28**, 795-803.

[49] Connors MH, Seeher KM, Crawford J, Ames D, Woodward M, Brodaty H (2018) The stability of neuropsychiatric subsyndromes in Alzheimer's disease. *Alzheimers Dement* **14**, 880-888.

[50] Smith GT, McCarthy DM, Zapolski TCB (2009) On the value of homogeneous constructs for construct validation,

theory testing, and the description of psychopathology. *Psychol Assess* **21**, 272-284.
[51] Strauss ME, Smith GT (2009) Construct validity: Advances in theory and methodology. *Annu Rev Clin Psychol* **5**, 1-25.
[52] Cortina JM (1993) What is coefficient alpha? An examination of theory and applications. *J Appl Psychol* **78**, 98-104.
[53] Tavakol M, Dennick R (2011) Making sense of Cronbach's alpha. *Int J Med Educ* **2**, 53-55.
[54] Hulshof TA, Zuidema SU, Janus SIM, Luijendijk HJ (2020) Large sample size fallacy in trials about antipsychotics for neuropsychiatric symptoms in dementia. *Front Pharmacol* **10**, 1701.
[55] de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM (2006) Minimal changes in health status questionnaires: Distinction between minimally detectable change and minimally important change. *Health Qual Life Outcomes* **4**, 54.
[56] MacCallum RC, Widaman KF, Zhang S, Hong S (1999) Sample size in factor analysis. *Psychol Methods* **4**, 84-99.
[57] Borsboom D, Cramer AOJ (2013) Network analysis: An integrative approach to the structure of psychopathology. *Annu Rev Clin Psychol* **9**, 91-121.
[58] Fried EI, Cramer AOJ (2017) Moving forward: Challenges and directions for psychopathological network theory and methodology. *Perspect Psychol Sci* **12**, 999-1020.
[59] Saari TT, Hallikainen I, Hintsa T, Koivisto A (2020) Network structures and temporal stability of self- and informant-rated affective symptoms in Alzheimer's disease. *J Affect Disord* **276**, 1084-1092.
[60] van Wanrooij LL, Borsboom D, Moll van Charante EP, Richard E, van Gool WA (2019) A network approach on the relation between apathy and depression symptoms with dementia and functional disability. *Int Psychogeriatr* **31**, 1655-1663.
[61] Epskamp S, Rhemtulla M, Borsboom D (2017) Generalized network psychometrics: Combining network and latent variable models. *Psychometrika* **82**, 904-927.
[62] Geda YE, Schneider LS, Gitlin LN, Miller DS, Smith GS, Bell J, Evans J, Lee M, Porsteinsson A, Lanctôt KL, Rosenberg PB, Sultzer DL, Francis PT, Brodaty H, Padala PP, Onyike CU, Ortiz LA, Ancoli-Israel S, Bliwise DL, Martin JL, Vitiello MV, Yaffe K, Zee PC, Herrmann N, Sweet RA, Ballard C, Khin NA, Alfaro C, Murray PS, Schultz S, Lyketsos CG (2013) Neuropsychiatric symptoms in Alzheimer's disease: Past progress and anticipation of the future. *Alzheimers Dement* **9**, 602-608.