Structural validity and age-based differential item functioning of the French Nottingham Health Profile in a sample of surgery patients

Jacques Juhel and Anne-Claire Gaillot

Abstract—The Nottingham Health Profile (NHP) is a popular and extensively used generic measure of health related quality of life (HRQoL) in patients suffering from various clinical diseases. It is made up of 38 dichotomous items divided into 6 health areas: loss of energy (En), pain experience (P), limitations in physical mobility (Pm), negative emotional reactions (Em), sleeping problems (Sl) and social isolation (So). The aim of this study was to further investigate factorial structure, scaling and psychometric properties of a reduced French version of the NHP in a sample of adult surgery patients (N=560). Item factor analysis (IFA) for dichotomous items was used to identify the factorial structure of the reduced French NHP. Differential item functioning (DIF) was examined with respect to age. The method used for DIF detection was a logistic regression technique extended by using Item Response Theory (IRT) based trait estimates. Results suggested the existence of three physical health factors (En, P and Pm) that could eventually be combined on a higher level, and one psychological factor measuring sleeping problems and anxiety level of patients. The fact that two items of the psychological scale showed DIF with respect to age suggested caution in the use of these items with elderly patients. Overall, the factor structure of the reduced version of the French NHP was found to be conceptually well-founded and demonstrated satisfactory measurement invariance among younger and older surgery patients.

Index Terms—Differential Item Functioning, Factor analysis for dichotomous items, Health Related Quality of Life, Nottingham Health Profile.

I. INTRODUCTION

Health-related quality of life (HRQoL) is a multidimensional construct referring to patients’ perceptions of the impact of disease and treatment on their quality of life. It is now thought to encompass several fundamental domains related to physical, psychological and social functioning [1]. Published literature has shown that it was interesting to consider HRQoL to describe, beyond clinical observations, the negative functional impact of disease and its medical and surgical treatment on physical well-being, perceived emotional state and social isolation of patients. Thus, HRQoL measures are now used as benchmarks and integrated into a larger process of assessment of therapeutic success.

Assessment of HRQoL generally rests on the use of self-administered questionnaires exploring a limited number of areas in the studied population [2]. These instruments have been developed in two main directions [3]. Disease-specific measures provide information specific to a target population of patients with a given disease, such as cancer [4], immunology [5] or in orthopedics and traumatology [6]. Generic measures [7] are used in any disease population or to provide comparisons between different pathological groups. Each of these two kinds of instruments has its own advantages and disadvantages. Generic questionnaires cannot be as accurate as specific instruments that target a given disease but they can be applied to a broader population, for example to patients experiencing pain and functional limitations after surgery.

Le Nottingham Health Profile (NHP) [8-10] is a widely used generic HRQoL questionnaire developed in the early 1970s and intended for primary health care, to provide a brief indication of a patient’s subjective distress and activity limitations. It comprises 38 binary items that assess perceived emotional, physical and social health problems. These items are grouped into six unidimensional subscales: energy loss (e.g. I'm tired all the time), pain experience (e.g. I have unbearable pain), negative emotional reactions (e.g. I'm feeling on edge), sleeping problems (e.g. I take tablets to help me sleep), social isolation (e.g. I feel I am a burden to people), and limitations of physical mobility (e.g. I'm unable to walk at all). Each item is assigned a weighted value; the sum of all weighted values in a given domain adds up to 100 with higher scores indicating a greater degree of subjective distress. The NHP is most suitable for measuring HRQoL in people with chronic disabling disorders. For example, using it has been recommended for patients suffering from lower limb atherosclerosis disease [11], knee osteoarthritis [12], rheumatoid arthritis [13], hereditary neuromuscular diseases [14] and chronic lower limb ischaemia [15]. The NHP is also
widely used in people expected to experience relatively high levels of distress after surgery [16-18] as well as in assessment of quality of life in older people [19-21]. The use of NHP, however, is more appropriate in patients with pronounced disability than in the general population [22].

The NHP and its various translations are quite well-documented with regard to face validity, content validity and criterion validity [11-25]. Studies of NHP’s internal construct validity, both important in clinical practice and research, are fewer. Though the a priori six factor model is consistent with the traditional measurement model of NHP, it has been found that his model did not fit well the data [26-27]. Alternative factor models should be formally tested with appropriate multivariate methods such as confirmatory factor analysis (CFA) for dichotomous items or multidimensional item response theory (MIRT). For example, the hypothesis that two second-order factoring, physical health on the one hand, emotional problems on the other hand, are sufficient to account for the organization of the questionnaire, seems to be also considered [1]. Furthermore, item behavior vis-à-vis the foreseen dimensions of the underlying factor structure of the NHP is insufficiently known, in some cases it may have difficulty discriminating with different patient groups experiencing lower or higher levels of disability (skewed distributions of scores, ceiling and floor effects) [15, 24-25]. This question is compounded by the loss of reliability among several other types of patients (e.g. primary care, musculoskeletal diseases, chronic renal failure and cardiovascular disease) of the social isolation [28-29] and energy loss [30] subscales. Another important aspect of psychometric evaluation is related with formal testing for measurement invariance and differential item functioning (DIF) in such characteristics as gender, age and other group variables. Though measurement equivalence is a necessary condition for valid comparison among different subgroups [31, 32], assessment of DIF in the NHP is not common and has only recently benefited from advances in scaling techniques [15, 25, 33-34]. Consequently, NHP’s validity as an argument [35] may not be evaluated as positively as it should be for an instrument meant to be used in a variety of both clinical and research purposes.

The main goal of the present study was to further investigate scaling and psychometric properties of the French version of the NHP [36] in a sample of adult patients from surgery and to increase the acceptability of the questionnaire by reducing the number of its items. Item factor analysis (IFA) for dichotomous items [37] was used to identify the factorial structure of the French NHP. Since a significant age-related decline of subjective health has been shown in older people [38], a logistic regression DIF technique extended by using Item Response Theory (IRT) based trait estimates [39, 40] was also implemented to examine possible DIF with respect to age.

II. Method

A. Participants

Patients enrolled in the study were 560 (56.96% females) with mean age of 51.6 years (standard deviation – SD=19.34, range=18-92 years). The number of adults aged 65 years and over was 178. They were suffering from various diseases (digestive, gynecological, orthopedic, ortho-rhino- laryngeal, proctology, urology and vascular), and had been hospitalized five days on average (SD=4.85, range=1-42 days) after surgery.

B. Procedure and measures

The study protocol was performed in accordance with the standards of the local ethical committee. After patients gave written informed consent, the French version of the NHP was self-administered and completed by patients the morning of their release from the hospital. The French NHP was used to assess the severity of patients’ energy loss (3 items), pain experience (8 items), negative emotional reactions (9 items), sleeping problems (5 items), social isolation (5 items), and limitations of physical mobility (8 items). For each scale, a score was computed by adding together the number of affirmative answers, using weighted scoring of the French NHP [36], and expressing the number as a percentage, ranging from 0 (best health status) to 100 (worth health status). No patients expressed any difficulties completing the questionnaire.

C. Statistical analysis

Item factor analysis

Limited- versus full-information methods may be envisaged for factor analyzing items rated on a dichotomous response format. Although a full information approach to binary factor analysis has good optimality properties, the method is computationally demanding with a large number of items and several latent variables resulting in numerous dimensions of numerical integration. Thus, we conducted IFA models based on the tetrachoric correlation matrix, that is the matrix of the estimates of the correlations between the latent continuous variables \( y^* \) underlying the observed binary variables \( y \). Following recent recommendations [41], we decided to estimate factor models using the robust weighted least squares with adjustments for the mean and variance (WLSMV) estimator available in Mplus 6.1 [42]. We first carried out a preliminary exploratory factor analysis in a structural equation modeling (EFA-SEM) approach [43] in order to determine the minimum number of factors that must be assumed. Then we specified and tested several models using confirmatory factor analysis (CFA) to approach a solution with as few cross-loadings as possible. Assessment of model fit [44] was based upon: 1) the robust goodness of fit statistic (\( \chi^2 \)) that can be considered analogous to the Satorra-Bentler-scaled chi-square statistic for continuous data; 2) the comparative fit index (CFI) and the Tucker-Lewis Index (TLI) with values >.95 considered representative of a well-fitting model; 3) the root mean square
error of approximation (RMSEA) with value less than .05 indicating that the model fits well in the population and the 90% confidence interval around the RMSEA value (a narrow confidence interval represents a good degree of precision); 4) the weighted root mean square residual (WRMR) with values <.95 regarded as indicative of good model fit with continuous as well binary data. Difference in fit between non nested models (EFA-SEM) was assessed by changes in CFI and RMSEA with ΔCFI<.01 and ΔRMSEA<.015 indicating no substantial modification in model fit [46]. Difference in fit between nested models was evaluated during model re-specification search by chi-square difference tests (Mplus DIFFTEST function [45]). Since Bayesian estimation can be simply seen as a computational tool for getting estimates that are analogous to ML estimation had it been possible [47], Bayesian estimates of the parameters of the final measurement model were also computed.

Evaluation of scale reliability and detection of DIF

We used the two-parameter logistic (2PL) model to ensure unidimensionality and to estimate the composite reliability-related point and interval of each scale of the French NHP. Since coefficient alpha cannot be considered a dependable estimator of scale reliability with binary items, reliability analyses were carried out within the latent variable modeling methodology following the procedure proposed by Raykov and colleagues [48].

DIF analysis was concerned with identifying items that give different probabilities of being endorsed by two groups differing in age (less than 65 years vs. 65 years or more), after controlling for an estimate of the matching variable (observed scale score or trait level). Two types of DIF can be distinguished from one another. Uniform DIF occurs when the difference in probabilities of endorsing an item - between subjects from different groups but with the same observed scale score or latent trait level - is independent of the matching criterion level. When non-uniform DIF effect is present, the magnitude of the item-group interaction varies according to the matching criterion. Furthermore, DIF items can bias the matching criterion’s level and affect the validity of the measures of DIF. Purification, which is an iterative process that successively removes items flagged as DIF, is one way for controlling this bias.

A large number of detection methods are available to identify items that manifest DIF [31, 32, 39, 47]. One widely used approach is logistic regression [40, 49] which has been recently extended into a hybrid framework [50] by using IRT based trait estimates, as increasingly recommended for being applied to HRQoL outcomes. Following the logistic regression approach, the probability of endorsing the item is related to the latent trait level θ, group membership g, and the interaction θ*g between the grouping variable and the matching criterion. For each item, three nested models can be written respectively as:

Model 1 : \[ \ln \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 \theta, \]
Model 2 : \[ \ln \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 \theta + \beta_2 g, \]
Model 3 : \[ \ln \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 \theta + \beta_2 g + \beta_3 \theta g. \]

Detection of DIF was performed with the lordif package available from the Comprehensive R Archive Network [50]. Uniform DIF was evaluated by comparing twice the difference in loglikelihood values for model 1 and model 2 to a chi-square distribution with one degree of freedom (number of groups minus one). The presence of non-uniform DIF was assessed in the same way by testing the interaction (model 2 vs. 3). Finally, DIF magnitude was statistically tested against the empirical threshold values for the statistics and McFadden pseudo R2 statistics [51] obtained by generating multiple datasets under the no-DIF hypothesis.

III. RESULTS

One should first note the high acceptability of the questionnaire as evidenced by the very low rate of missing data (0% for 12 items; 0.18% for 10 items; 0.36% for 4 items; 0.54% for 9 items; 0.72% for 3 items). Table 1 lists the endorsement rates (p) for the 38 items of the French NHP and figure 1 shows box plots for each scale score (see [36] for item weights of the French NHP).

TABLE 1
ENDORSEMENT RATE FOR THE 38 ITEMS OF THE FRENCH NHP (SURGERY PATIENTS, N=560).

Overall, half of the patients reported sleeping problems (0.48) when about one third of them reported experiencing a loss of energy (.35) and pain (.31). Patients also pointed out
impaired physical mobility (19) and negative emotional reactions (11). Finally, a feeling of social isolation emerged from very few patients (03). Note that several items of the emotional reactions and social isolation scales items provide information only with severe problems on these health dimensions, which reduces obviously their interest on this sample. On the whole, the general health status explored with these surgery patients was a little worse than that in the general population (23.65 vs. 14.38, [22]).

Based on usual recommendations [52], items with endorsement rates between .15 and .85 were considered acceptable. As a result, we did not keep any items from the social isolation subscale and retained only two items (Em3 and Em4) from the negative emotional reactions subscale. In addition, two items (PM3 and PM4) were removed from the physical mobility subscale. This prior reduction process led to exclude 14 unsuitable items. Thus, 24 items were kept for further structural analysis.

A. Structural validity of the 24-item French NHP

The design of the 24-item French NHP suggests that five factors related to loss of energy (En: 3 items), pain experience (P: 8 items), physical mobility limitations (Pm: 6 items), negative emotional reactions (Em: 2 items) and sleeping problems (Sl: 5 items) can be expected. This initial five-factor structure was first tested with CFA. The values for the goodness of fit indicators were χ²=762 with 221 degrees of freedom, CFI=.913, TLI=.901, RMSEA=.066 and WRMR=1.707 which suggests a poor fit to the data. A bifactor model was then fit with each item loading on a common factor of poor HRQoL and on a specific domain (En, P, Pm, Em and Sl). The 6 factors were constrained to be mutually uncorrelated so that all covariance was partitioned either into loadings on the common factor or onto the specific factors. This CFA bifactor model also showed an insufficient degree of fit (χ²=584 with 229 degrees of freedom, CFI=.943, TLI=.931, RMSEA=.053 and WRMR=1.332).

We tested then several models with a decreasing number of factors using exploratory factor analysis within a SEM framework (EFA-SEM). The five-factor model resulted in χ²=218 with 148 degrees of freedom, CFI=.989, TLI=.981, RMSEA=.029 and WRMR=0.601. The values for the goodness of fit indicators for the four-factor model were χ²=288 with 167 degrees of freedom, CFI=.981, TLI=.970, RMSEA=.036 and WRMR=.740. Finally, we examined the three-factor model (χ²=467 with 187 degrees of freedom, CFI=.955, TLI=.939, RMSEA=.052 and WRMR=1.050) which demonstrated a lower quality data-model fit. Taken together, the goodness of fit indicators for the five-factor model and the four-factor model show that these models both fit the data well in this study. However, since the values of the difference between the two models in the CFI fit measure (ΔCFI=.008) and in the RMSEA measure (ΔRMSEA=.007) were not indicative of a significant loss of fit, we retained the more parsimonious four-factor hypothesis.

In the next CFA step, we attempted to closely reproduce the relationships between the 24 items of the French-NHP keeping number of cross-loadings as low as possible. Two four-factor models without cross-loadings were first compared. In the first model, items from the En and Em scales were hypothesized to measure the same factor, other factors being P, Pm and Sl (χ²=716, df=246, CFI=.924, TLI=.915, RMSEA=.058 and WRMR=1.546). In the second model, we hypothesized that Em items and Sl items were indicators of the same factor with En, P and Pm as the three other factors (χ²=608, df=246, CFI=.942, TLI=.935, RMSEA=.051 and WRMR=1.432). Since the first model showed a somewhat poorer fit than the second one (ΔCFI>.01), we retained the second hypothesis for further specification search. The strategy was then to inspect modification indices (MIs) which reflect the expected improvement in model fit if a fixed parameter were freely estimated. Cross-loadings with a large MI were released until the model was considered to present a satisfactory fit. As a first step, the loadings of P7 on Pm (Δχ²=41.8, df=1, p<.001), Pm on En (Δχ²=41.69, df=1, p<.001) and P4 on Pm (Δχ²=19.44, df=1, p<.01) were successively freed. In a second time, we tested a model in which the loading of P7 on Pm was fixed to 0. This hypothesis led to an acceptable model fit (χ²=480, df=243, CFI=.962, TLI=.957, RMSEA=.042 and WRMR=1.235). We decided to retain this model despite the presence of two bi-dimensional items (Pm7 as a measure of Pm and En; P4 as a measure of Pm and P). Finally, we used a multiple indicators multiple causes (MIMIC) model for estimating the effects of age and gender on the four underlying factors of the 24-item French NHP. Resulting goodness of fit indexes showed that the model was reasonably consistent with

![Box-and-whisker diagram of weighted sum scores (N=560 surgery patients).](image-url)
the data ($\chi^2=523$, df=284, CFI=.962, TLI=.957, RMSEA=.039 and WRMR=1.175). Table 2 shows Bayesian parameter estimates of this latter model with a posterior predictive p-value equals .053, indicating that the observed-data statistic was not significantly larger than what would have been generated by the model.

### Table 2

<table>
<thead>
<tr>
<th>Item</th>
<th>En</th>
<th>P</th>
<th>Pm</th>
<th>SI/Em</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>En1</td>
<td>.693 (.045)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.225 (.014)</td>
</tr>
<tr>
<td>En2</td>
<td>.693 (.049)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.705 (.110)</td>
</tr>
<tr>
<td>En3</td>
<td>.897 (.035)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.155 (.127)</td>
</tr>
<tr>
<td>P1</td>
<td>-</td>
<td>539 (.052)</td>
<td>-</td>
<td>-</td>
<td>-.184 (.091)</td>
</tr>
<tr>
<td>P2</td>
<td>-</td>
<td>428 (.065)</td>
<td>-</td>
<td>-</td>
<td>.921 (.077)</td>
</tr>
<tr>
<td>P3</td>
<td>-</td>
<td>645 (.047)</td>
<td>-</td>
<td>-</td>
<td>.180 (.106)</td>
</tr>
<tr>
<td>P4</td>
<td>-</td>
<td>511 (.081)</td>
<td>-</td>
<td>-</td>
<td>.423 (.111)</td>
</tr>
<tr>
<td>P5</td>
<td>-</td>
<td>911 (.034)</td>
<td>-</td>
<td>-</td>
<td>.514 (.138)</td>
</tr>
<tr>
<td>P6</td>
<td>-</td>
<td>-</td>
<td>994 (.019)</td>
<td>-</td>
<td>.806 (.106)</td>
</tr>
<tr>
<td>P7</td>
<td>-</td>
<td>-</td>
<td>534 (.069)</td>
<td>-</td>
<td>.835 (.129)</td>
</tr>
<tr>
<td>Sl</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>918 (.019)</td>
<td>.965 (.069)</td>
</tr>
<tr>
<td>Em</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.367 (.062)</td>
</tr>
</tbody>
</table>

Note. En: energy loss; P: pain experience; Pm: limitations of physical mobility; Em: negative emotional reactions; Sl: sleeping problems; SI: social isolation. MIMIC=multiple indicators multiple causes; LR: Likelihood Ratio; CI: confidence interval.

The first three identified latent factors belong to the domain of physical health: loss of energy (e.g. “I soon run out of energy”), pain experience (e.g. “I’m in pain when I’m standing”) and limitations in physical mobility (e.g. “I have trouble getting up and down stairs”). The magnitude of residual correlations between these factors (with age and gender statistically controlled) makes it possible to group together, as a large body of research notably suggests, all these items in one more general physical health scale. Indeed, some items of the pain scale have loadings consistent with this grouping (see the loading of P2 on En, the loading of P7 on Pm and the loadings of P4 on En and Pm). The fourth latent variable (e.g. “I sleep badly at night”) that essentially measures sleeping problems and, to a lesser degree, patients anxiety level, corresponds to the mental or psychological component of HRQoL. As was to be expected, the second-order physical health factor, as measured by En, P and Pm, and the (first-order) psychological factor (Sl/Em) were positively correlated (residual correlation r=.321, SE=.053, p<.001).

Table 2 also shows a significant positive effect of age (younger vs. older) on loss of energy ($\beta$=.093, SE=.049, p<.05), limitations in physical mobility ($\beta$=.097, SE=.050, p<.05) and sleeping problems / negative emotional reactions ($\beta$=.121, p<.01). A similar trend was observed for pain experience, though the effect did not reach statistical significance ($\beta$=.081, SE=.050, p=.53). More broadly, the beta estimate in the regression from the second-order physical health latent variable on age was .119 (SE=.050, p<.05). All these results are entirely consistent with increased chronic health limitations in older people. Turning now to the effects of gender on subjective health status, one can see in table 2 that compared to men, women reported a small but significant increased loss of energy ($\beta$=.079, SE=.043, p<.05). This result might suggest that women might not cope as well physically as men. However, the non significant global effect of gender on physical health ($\beta$=-.016, SE=.039, p>.05) weakens this interpretation (results of this analysis not reported here but available upon request from the first author).

### B. Scales composite reliability and age-related DIF

Evaluation of the reliability-related point ($\rho$) and interval (CI) estimates was carried out as a byproduct of fitting the 2PL IRT model to each scale using Mplus. Likelihood Ratio (LR) chi-square for the robust maximum likelihood run showed acceptable model fit ($\chi^2=50.110$, df=21, p=.0004) with $\rho=.800$, SE=.014 and 95%-CI: (.772, .828); experience of pain scale ($\chi^2=79.201$, df=51, p=.0069) with $\rho=.721$, SE=.019 and 95%-CI: (.684, .757); limitations in physical mobility scale ($\chi^2=278.883$, df=238, p=.0354) with $\rho=.864$, SE=.008 and 95%-CI: (.849, .879); sleeping problems and anxiety scale (LR $\chi^2=191.682$, df=113, p=.0354) with $\rho=.736$, SE=.021 and 95%-CI: (.694, .777). These results showed that the reliability of the four scales of the 24-item French NHP could be considered satisfactory.

As discussed above, age was shown to be a significant predictor of En, Pm and Sl/Em. Therefore, the presence of DIF (both uniform and non-uniform) was tested under the logistic regression framework for each of these three scales. The reference and focal groups were defined as younger (< 65 years) and older (65 years and more). Flag DIF items were
based on the LR $\chi^2$ test as the detection criterion and a corresponding flagging criterion of .05. A Monte Carlo simulation procedure ($n=1000$ and $\alpha=0.05$) was invoked to obtain empirical thresholds for McFadden’s pseudo $R^2$ as the magnitude measure. Using these settings, lordif did not flag any item of the En and Pm scales as displaying age-related DIF. However, two items of the SL/Em scale (Em3: “I’m feeling on edge” and Sl1: “I take tablets to help me sleep”) were flagged. The trait distributions for the younger and older groups as well as diagnostic plots for each flagged item are presented in figure 2.

The existence of a uniform DIF was supported by result of the LR $\chi^2$ 1-df test comparing model 1 and model 2 [Pr($\chi^2$)$<.006$]. Though significant, magnitude of these effects was weak with a non-uniform pseudo $R^2$ of .008 for Em3 and a uniform pseudo $R^2$ of .011 for Sl1. The very low impact of these DIF items on total expected score at any sleeping problems and anxiety level for younger or older patients can be visualized by comparing the test characteristic curves (TCC) presented in the top right of figure 1. Differential test functioning seems rather uniform with older patients having a slightly higher probability of endorsing sleeping problems and anxiety items at low level of the trait measured by these items.

Finally, we evaluated the gain in reliability resulting from deleting Em3 (“I’m feeling on edge”) and Sl1 (“I take tablets to help me sleep”) of the scale. Analysis revealed that dropping Em3 from the set of items measuring sleeping problems and anxiety was associated with a significant gain in reliability ($\Delta \rho=0.013$, SE=.007, $p<.05$) yielding a slightly higher estimate ($\rho=0.749$) of reliability for the revised scale. Contrariwise, dropping Sl1 from the set of items did not produce a significant gain in reliability ($\Delta \rho=.009$, SE=.006, $p>.05$).

**IV. DISCUSSION**

The objective of the present study was to examine in a pooled sample of surgery patients the structural validity of the French version of the Nottingham Health Profile, a generic measure of perceived health status. The present findings first demonstrated the low sensitivity of fourteen of thirty eight items of the French NHP. Of the fourteen items removed, twelve items measured the psychological component of the French NHP (negative emotional reactions, sleeping problems and social isolation) when only two measured its physical component (loss of energy, pain experience, limitations in physical mobility). Evaluation in surgery patients of the physical component of French NHP thus seems to be more relevant than that of its psychological component. It is particularly interesting to note that all the items in the social isolation sub-scale were eliminated, as already observed in item reduction of the Spanish NHP [33, 34]. In this respect, one may wonder whether it might be advisable to assess perceived health status in surgery patients with items measuring anxiety level or coping strategies to manage pain and stress, rather than with items such as “I feel that life is not worth living” and “I’m finding it hard to get on with people”.

Our approach to internal validity of the 24-item French NHP was based on item factor analysis of dichotomous items that is on a probit multidimensional IRT model. This approach produced a structure describing perceived health problems on 4 dimensions: loss of energy, limitations in physical mobility, pain and sleeping problems/emotional reactions. Note that we were led, as in some previous studies [34], to aggregate sleep and emotional reactions items within the same mental scale. Additionally, the high correlations (.71 to .84) between the first 3 dimensions suggest that a second-order factor model
could provide a satisfactory fit to the observations from the physical domain. Turning now to the dichotomous conceptualization underlying interpretation of the 24-item French NHP, the one-dimensional view of health of the NHP implies the existence of a strong correlation between the physical and psychological scales. The observation of an average correlation (>.31) between these two scales makes it preferable using two scales (physical, sleep/emotional) rather than a summary scale. These results thus confirm those from previous published studies that distinguished scores for each health component [1, 27]. Finally, reliability as referring to how well a factor in an IRT model is captured by a sum of binary items was good (> .80) for loss of energy and limitations in physical mobility scales, and satisfactory (> .70) for pain scale and sleeping problems and anxiety scale.

CFAs analyses were also extended to MIMIC modeling with the objective of assessing the effects of age on subjective health as measured by the 24-item French NHP. In accordance with expectations, we found a significant age-related decline of subjective health in the physical domain above the age of 65 years and older. In agreement with the well-documented subjective sleep disorders among elderly patients [53], and unlike amazing findings of a previous study [28], the results also showed the increase of age-related sleeping problems associated with negative emotional reactions. Additionally, it should be noticed that older surgery patients are more at risk of experiencing symptoms of anxiety (such as feeling tense, on edge) than a “low” mood. In comparison to younger surgery patients, symptoms of anxiety in older patients might be an appropriate situational response to physical distress being more compromised at baseline. Therefore, the finding of an effect of age on subjective health supports the use of age specific reference data of the French NHP.

Finally, we considered the question of whether or not DIF may occur across age groups. We found no evidence of DIF for items measuring physical distress (loss of energy, pain, limitations of physical mobility). However, examination of the sleeping problems and anxiety item set for DIF based on age demonstrated two items with DIF, indicating that compared with younger patients, older patients had higher probability of endorsing the item, especially at low level of trait. Thus, caution should be exercised in the use of these items though neither of them displayed high magnitude DIF, and their impact on total scale score was minimal for low levels of the trait assessed. One option for treating items with DIF could be to discard them from the scale. Item “I’m feeling on edge” could thus be dropped inasmuch as excluding this item significantly increased the reliability of the sleeping problems and anxiety scale. On the other hand, since DIF did not really impact total score of the sleeping problems and anxiety scale, which has a small number of items, another option could be to tag DIF items (“I’m feeling on edge” and “I take tablets to help me sleep”) that should not be administered to elderly surgery patients.

Two limitations of this study should be kept in mind when interpreting the findings. The analyses were based on a heterogeneous group of patients suffering diverse illnesses and having undergone various surgical procedures. In addition, patients in this study were hospitalized for varying periods before they were administered the French NHP. Therefore, whether the structural validity, reliability and measurement invariance of the 24-item French NHP can be replicated across specific groups of patients having undergone the same surgical procedure and having been hospitalized an equal number of days, remains an open question.

In conclusion, we reduced the original 38-item French NHP in a shorter version comprising 24 items and displaying much less floor effects, thus increasing the possibility to differentiate between patients and facilitating detection and assessment of change in subjective health over time. This reduced version offers some valuable benefits. From a practical point of view, the 24-item NHP is obviously shorter to administer than the 38-item NHP. From a methodological point of view, the psychometric properties of items were subjected to an explicit process of evaluation. But the main advantage comes from the illumination provided in terms of internal validity of the questionnaire. Found to be adequate and demonstrating satisfactory measurement invariance among younger and older surgery patients, the factor structure of the French NHP is clearly stated in this sample of patients. However, additional studies with more targeted pathologies are needed to generalize these conclusions.

REFERENCES


