RESEARCH PAPER

An Overview of Confirmatory Factor Analysis and Item Response Analysis Applied to Instruments to Evaluate Primary Healthcare

Aperçu de l'analyse factorielle confirmatoire et de l'analyse de réponse par item appliquées aux instruments d'évaluation des soins primaires

DARCY A. SANTOR, PHD
School of Psychology, University of Ottawa
Ottawa, ON

JEANNIE L. HAGGERTY, PHD
Department of Family Medicine, McGill University
Montreal, QC

JEAN-FRÉDÉRIC LÉVESQUE, MD, PHD
Centre de recherche du Centre hospitalier de l'Université de Montréal
Montréal, QC

FREDERICK BURGE, MD, MSC
Department of Family Medicine, Dalhousie University
Halifax, NS

MARIE-DOMINIQUE BEAULIEU, MD, MSC
Chaire Dr Sadok Besrour en médecine familiale
Centre de recherche du Centre hospitalier de l'Université de Montréal,
Montréal, QC

DAVID GASS, MD
Department of Family Medicine, Dalhousie University
Halifax, NS

RAYNALD PINEAULT, MD, PHD
Centre de recherche du Centre hospitalier de l'Université de Montréal,
Montréal, QC

Abstract

This paper presents an overview of the analytic approaches that we used to assess the performance and structure of measures that evaluate primary healthcare; six instruments were administered concurrently to the same set of patients. The purpose is (a) to provide clinicians, researchers and policy makers with an overview of the psychometric methods used in this series of papers to assess instrument performance and (b) to articulate briefly the rationale, the criteria used and the ways in which results can be interpreted. For illustration, we use the case of instrument subscales evaluating accessibility. We discuss (1) distribution of items, including treatment of missing values, (2) exploratory and confirmatory factor analysis to identify how items from different subscales relate to a single underlying construct or sub-dimension and (3) item response theory analysis to examine whether items can discriminate differences between individuals with high and low scores, and whether the response options work well. Any conclusion about the relative performance of instruments or items will depend on the type of analytic technique used. Our study design and analytic methods allow us to compare instrument subscales, discern common constructs and identify potentially problematic items.

Résumé

Cet article présente un aperçu des approches analytiques que nous avons utilisées pour évaluer le rendement et la structure des mesures qui servent à évaluer les soins de santé primaires : six instruments ont été appliqués simultanément au même groupe de patients. L'objectif est (a) de fournir aux cliniciens, aux chercheurs et aux responsables de politiques, un aperçu des méthodes psychométriques utilisées dans cette série pour évaluer le rendement de l'instrument et (b) d'articuler brièvement l'analyse raisonnée, les critères employés et les façons dont peuvent être interprétés les résultats. À titre d'exemple, nous avons utilisé le cas des sous-échelles qui servent à évaluer l'accessibilité. Nous discutons (1) la distribution des items, y compris le traitement des valeurs manquantes, (2) les analyses factorielles exploratoires et confirmatoires afin de voir comment les items de différentes sous-échelles sont liés à un seul construit (ou sous-dimension) sous-jacent et (3) l'analyse de réponse par item pour voir si les items permettent de discriminer les différences entre les unités qui présentent des scores élevés et faibles, et pour voir si les choix de réponses fonctionnent bien. Toute conclusion sur le rendement relatif des instruments ou des items dépend du type de technique analytique employé. La conception et les méthodes analytiques de cette étude permettent de comparer les sous-échelles des instruments, de discerner les construit communs et de repérer les items potentiellement problématiques.

Psychometric scales and instruments have been used to assess virtually every component of healthcare, whether to identify gaps in service, to assess special needs of patients, or to evaluate the performance and efficiencies of programs, organizations or entire healthcare systems. In this special issue of *Healthcare Policy*, we examine the performance of several instruments that assess different attributes of primary healthcare from the patient's

perspective. Any conclusion about the relative performance of instruments and items from those instruments will depend on the type of analytic technique used to assess performance.

The purpose of this paper is to describe the analytic approach and psychometric methods that we used to assess and compare the performance of the instrument subscales. We articulate the rationale of different approaches, the criteria we used and how results can be interpreted. For illustration, we use the case of instrument subscales evaluating accessibility, described in detail elsewhere in this special issue (Haggerty, Lévesque et al. 2011).

Data Sources

All the instruments used in this study have previously been validated and meet standard criteria for validity and reliability. The goal of the study was to extend this process of validation and compare the performance of the six instruments for measuring core attributes of primary healthcare for the Canadian context. Our intention is not to recommend one instrument over another, but to provide insight into how different subscales measure various primary healthcare attributes, such as accessibility and care.

We administered the six instruments to 645 health service users in Nova Scotia and Quebec: the Primary Care Assessment Survey (PCAS, Safran et al. 1998); the adult version of the Primary Care Assessment Tool – Short (PCAT-S, Shi et al. 2001); the Components of Primary Care Index (CPCI, Flocke 1997); the first version of the EUROPEP (EUROPEP-I, Grol et al. 2000); the Interpersonal Processes of Care – 18-item version (IPC-II, Stewart et al. 2007); and the Veterans Affairs National Outpatient Customer Satisfaction Survey (VANOCSS, Borowsky et al. 2002). The sample was balanced by overall rating of primary healthcare, high and low level of education, rural and urban location, and English and French language (Haggerty, Burge et al. 2011).

Distribution of Responses

The first step was to examine the distribution of the responses, flagging as problematic items where a high proportion of respondents select the most negative (floor effect) or the most positive (ceiling effect) response options, or have missing values. Missing values, where respondents failed to respond or wrote in other answers, may indicate questions that are not clear or are difficult to understand. We expected low rates of true missing values because truly problematic items would be eliminated during initial validation by the instrument developers, but remained sensitive to items that are problematic in the Canadian context.

However, some instruments offer response options such as "not applicable" (EUROPEP-I) or "not sure" (PCAT-S), which count as missing values in analysis because they cannot be interpreted as part of the ordinal response scale. They represent a loss of information as they cannot be interpreted.

Missing information, for whatever reason, is problematic for our study given that missing information on any item means that data for the entire participant is excluded from factor analysis (listwise missing), compromising statistical power and potentially introducing bias.

In the case of accessibility subscales, 340 of the 645 respondents (53%) were excluded from factor analysis because of missing values, of which 267 (79%) were for selecting "not sure" or "not applicable" options. We examined the potential for bias by testing for differences between included and excluded respondents on all relevant demographic and healthcare use variables. We also imputed values for most of the missing values using maximum likelihood imputation (Jöreskog and Sörbom 1996), which uses the subject's responses to other items and characteristics to impute a likely value. Then, we repeated all the factor analyses to ensure that our conclusions and interpretations remained unchanged, and reducing the possibility of bias. Nonetheless, the high proportion of missing values in some instances is an important limitation of our study, and needs to be considered in the selection of instruments.

Subscale Scores

Next, we examined the performance by subscale. Subscale scores were mostly calculated as the mean of item values if over 50% of the items were complete. This score was not affected by the number of items, and it reflects the response options. For example, a subscale score of 3.9 in Organizational Access on the PCAS corresponds approximately to the "4=good" option on the response scale of 1 (very poor) to 6 (excellent). But it is difficult to know how this compares to the score of 3.6 on the EUROPEP-I Organization of Care for a similar dimension of accessibility from 1 (poor) to 5 (excellent). To compare the subscales between different response scales, we normalized scores to a common 0-to-10 metric using the following formula:

New score = {(raw score – minimum possible) / (raw maximum – raw minimum)} * 10

So, the normalized mean for PCAS Organizational Access, 5.9, is seen to be considerably lower than 6.5 on the EUROPEP-I Organization of Care, and the PCAS variance is lower than the EUROPEP-I (1.8 vs. 2.4). Thus, if accessibility were measured in one population using the PCAS and in another using the EUROPEP-I, the scores of the EUROPEP-I would be expected to be higher than the PCAS, even if there were no difference in accessibility between the two populations.

Reliabilities

The reliability of each subscale was evaluated using Cronbach's coefficient α, which estimates how much each item functions as a parallel, though correlated, test of the underlying construct. Cronbach's α ranges from 0 (items completely uncorrelated, all variance is random) to 1 (each item yields identical information), with the convention of .70 indicating a minimally reliable subscale. The subscales included in our study all reported adequate to good internal consistency. The coefficient α is sensitive to sample variation as well as the number and quality of items. Given that our study sample was selected to overrepresent the extremes of poor and excellent experience with primary healthcare relative to a randomly selected sample, we expected our alpha estimates to meet or exceed the reported values.

We then calculated Pearson correlations among the subscale scores, controlling for educational level, geography and language (partial correlation coefficients) to account for slight deviations from our original balanced sampling design. We expect high correlations between subscales mapped to the same attribute (convergent validity) and lower correlations with subscales from other attributes (some degree of divergent validity), indicating that the items in the subscales are indeed specific to that attribute and that respondents appropriately distinguish between attributes. Pearson correlation coefficients indicate expected relationships observed in factor analysis. If we observe high correlations within an attribute, then we would expect all the items from those subscales to "load" on a common factor. So, for example, after the correlation analysis for accessibility, we observed that the PCAT-S First-Contact Utilization subscale correlated less strongly with the other accessibility subscales than it did with relational continuity. We had high expectations that its items would not only form a separate factor, but also that it would relate poorly to accessibility as a whole.

Exploratory and Confirmatory Factor Analysis

Subscales from different instruments that were designed to assess the same primary healthcare attribute should relate to a single underlying factor or construct. We used both exploratory and confirmatory factor analysis to examine this premise, as well as to determine whether items across subscales related to sub-dimensions within the attribute.

Exploratory analyses

Exploratory factor analysis is a descriptive technique that can detect an overarching structure that explains the relationships between items in a parsimonious way. We used the common factor analysis procedure computed in SAS v. 9.1 (SAS 2003). This procedure identifies how much items can be represented by a smaller group of variables (i.e., common factors) that account for as much of the variability in the data as possible. The procedure assigns an eigenvalue to each factor that corresponds to the total variance in item responses that can be explained by the factor. Typically, factors with eigenvalues greater than 1.0 are retained.

This procedure also computes how strongly each individual item maps on to each factor. "Factor loadings" range from –1.0 to 1.0 and can be interpreted much like a correlation coefficient. These indicate (a) the extent to which all items relate to one or more distinctive factors, (b) how strongly each item is related to each factor (and whether the item should be retained or eliminated within a factor) and (c) how much variation in responses to items can be accounted for by each factor or subgroup. We considered items with factor loadings $\geq |.4|$ as strongly related to the underlying factor. It is important to note that common factor analysis assumes a normal distribution; items with highly skewed distributions will affect both the loadings and the extent to which factors can be easily interpreted (Gorsuch 1983).

Results of an exploratory factor analysis for items in subscales from three different instruments assessing accessibility are presented in Table 1. Factor loadings are presented for each item showing how each item is related to three distinct factors. The first factor has a large

eigenvalue (7.84) and accounts for approximately 41% of variance in the responses given to items, compared to just 6% for the second factor (eigenvalue=1.19) and less than 1% for the third. As a result, only two factors would be considered worth interpreting. This confirms our expectation based on the correlation analysis that the PCAT-S First-Contact Utilization subscale might not fit with other accessibility subscales. The two important underlying factors could be characterized as timeliness and accommodation (Haggerty, Lévesque et al. 2011).

TABLE 1. Factor loadings from an oblique exploratory principal component analysis for accessibility items drawn from four measures of accessibility

		Factor 1 Eigen=7.84	Factor 2 Eigen=1.19	Factor 3 Eigen=0.77
PCAS: Accessibility (Organizational Access)				
PS_oa1	How would you rate the convenience of your regular doctor's office **location**?	—	—	—
PS_oa2	How would you rate the **hours** that your doctor's office is open for medical appointments?	—	—	—
PS_oa3	How would you rate the usual **wait** for an appointment when you are sick and call the doctor's office asking to be seen?	0.41	0.49	—
PS_oa4	How would you rate the amount of time you **wait** at your doctor's office for your appointment to **start**?	—	0.91	—
PS_oa5	Thinking about the times you have needed to **see or talk to** your doctor, how would you rate the following: ability to **get through to** the doctor's office by phone?	—	0.51	—
PS_oa6	Thinking about the times you have needed to **see or talk to** your doctor, how would you rate the following: ability to **speak to** your doctor by phone when you have a question or need medical advice?	—	0.55	—
PCAT-S: Accessibility (First-Contact Utilization)				
PT_fcu1	When you need a regular general checkup, do you go to your Primary Care Provider before going somewhere else?	—	—	0.78
PT_fcu2	When you have a new health problem, do you go to your Primary Care Provider before going somewhere else?	—	—	0.74
PT_fcu3	When you have to see a specialist, does your Primary Care Provider have to approve or give you a referral?	—	—	—
PCAT-S: Accessibility (First-Contact Accessibility)				
PT_fca1	When your Primary Care Provider is **open** and you get sick, would someone from there see you the same day?	0.61	—	—
PT_fca2	When your Primary Care Provider is **open**, can you get advice quickly over the phone if you need it?	0.64	—	—
PT_fca3	When your Primary Care Provider is **closed**, is there a phone number you can call when you get sick?	0.73	—	—
PT_fca4	When your Primary Care Provider is **closed** and you get sick **during the night**, would someone from there see you that night?	0.65	—	—
EUROPEP-I: Accessibility (Organization of Care)				
EU_oc1	Preparing you for what to expect from specialist or hospital care	0.51	—	—
EU_oc2	The helpfulness of staff (other than the doctor)	0.41	—	—
EU_oc3	Getting an appointment to suit you	0.51	—	—
EU_oc4	Getting through to the practice on the phone	0.42	—	—

TABLE 1. Continued

		Factor 1 Eigen=7.84	Factor 2 Eigen=1.19	Factor 3 Eigen=0.77
EU_oc5	Being able to speak to the general practitioner on the telephone	0.68	—	—
EU_oc6	Waiting time in the waiting room	—	0.87	—
EU_oc7	Providing quick services for urgent health problems	0.64	—	—

Note: Factor loadings smaller than 0.40 have not been presented.

Confirmatory analyses

Confirmatory factor analysis differs from exploratory factor analysis by allowing the investigator to impose a structure or model on the data and test how well that model "fits." The "model" is a hypothesis about (a) the number of factors, (b) whether the factors are correlated or uncorrelated and (c) how items are associated with the factor. Models with different configurations are compared using structural equation modelling. Statistical software packages produce various "goodness-of-fit" statistics that capture how well the implied variance–covariance matrix of the proposed model corresponds to the observed variance–covariance matrix (i.e., how items from the instrument actually correlate). Confirmatory factor analysis attempts to account for the covariation among items (ignoring error variance), whereas common factor analysis accounts for the "common variance" shared among items, differentiating variance attributable to an underlying factor and error variance. Although similar in spirit, factor loadings are computed in fundamentally different ways from common factor analysis and should be interpreted differently.

Our premise was that different measures of an attribute can still be viewed as indicators (i.e., items) assessing the same underlying construct despite being drawn from different instruments employing different phrasing and response scales. Testing this hypothesis allows researchers and policy makers to view similarly the results obtained from different measures of, say, accessibility.

Figure 1 presents the results of a confirmatory factor analysis for accessibility subscales. Figure 1A depicts a standard unidimensional model, where every item is linked to the same, single underlying construct, namely accessibility. Constructs (designated with ellipses) are linked to (designated with arrows with loading coefficients) individual items (designated with rectangular boxes). The model shows that most items load strongly (i.e., factor loading greater than .90) on the latent construct called Access, but that some items do not (e.g., loadings of .71 and .78) or they have high residual error, shown to the right of the item.

The performance of the model is evaluated by examining the ensemble of "goodness-of-fit" statistics, such as the comparative fit index (CFI), the normed fit index (NFI), consistent Akaike's information criterion (CAIC) or the root mean squared error of approximation (RMSEA). They all assess in different ways the discrepancy between the pattern of variances and covariances implied by the model and the actual pattern of variances and covariances observed in the data (see Kline 1998 for an in-depth review of basic issues in structural equation modelling). If the implied pattern is close to what is observed in the data, then the model is said to fit – it accurately accounts for the manner in which items are interrelated.

Fit statistics for the model in Figure 1A were all good. Unlike the usual interpretation of significance, lower chi-squared (χ^2) values suggest better fit. The χ^2 value was 649 with 152 degrees of freedom and was significant, which might indicate the model does not fit well (though χ^2 is sensitive to large samples such as ours). However, other fit statistics, such as NFI, CAIC and GFI (results not shown), which take into consideration both the sample size and model complexity, were all well above the conventional criterion of .90 of "good fit." The RMSEA of .104 is higher than the .05 criterion indicating good fit, but is still reasonable. Altogether, these results suggest that although items were drawn from distinct subscales, response to questions can be accounted for by a single underlying construct, namely, accessibility.

However, we might hypothesize that because items were drawn from subscales with different numbers of response options and formats, the pattern of responses would be even better explained by a model that explicitly locates individual items with the subscale from their parent instrument (first-order factor) and then links these first-order factors within the general construct of accessibility (a second-order factor). Figure 1B depicts this second-order, multidimensional model. Fit statistics for this model were also extremely good. Again, NFI, CAIC and GFI were all .98. The χ^2 value was 514 (with 148 degrees of freedom).

Comparing models

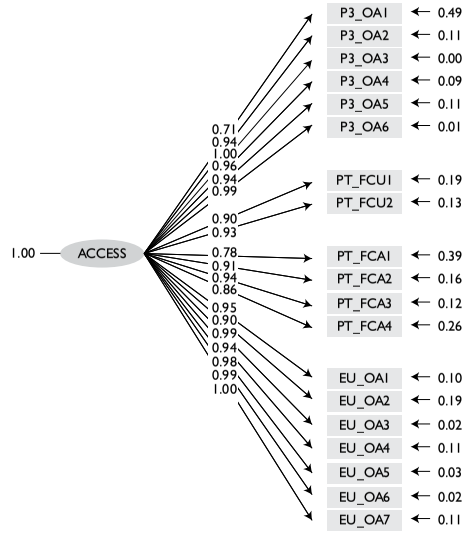
One of the strengths of a confirmatory factor analysis is the ability to compare "nested" models, where one model is a simpler version of a more complex model. Because these models differ only in the number of paths that are being estimated, χ^2 values for one model can be subtracted from the other and the significance of the difference evaluated. The χ^2 difference between the simple model in Figure 1A and the more complex model in Figure 1B (649 – 514 = 135, 4 *df*) is statistically significant, so we can infer that the complex model is more valid. Some of the variability in how individuals respond to questions is not just determined by the underlying construct of accessibility, but also by the specific measure from which the question is drawn.

We also test a model that groups items within sub-dimensions of accessibility. Figure 1C depicts a second-order, multidimensional model in which items are grouped within two first-order factors, namely, the timeliness of service and the extent to which patients' access barriers are accommodated, which are themselves part of a broader, second-order factor: accessibility. This model says there are two components of the more general construct of accessibility, and that these transcend specific instrument subscales.

It is important to note that not all models can be compared directly. The model in Figure 1C does not include the items from the PCAT First-Contact Utilization subscale; it differs from the model in Figure 1A by more than the number of paths. To test the validity of this model, we compared its own restricted version rather than the model depicted in Figure 1A and found that grouping items within sub-factors of timeliness and accommodation is superior to the one-dimensional model (χ^2 426 – 364 = 52, 3 degrees of freedom).

FIGURE 1. Results of a confirmatory factor analysis for accessibility subscales

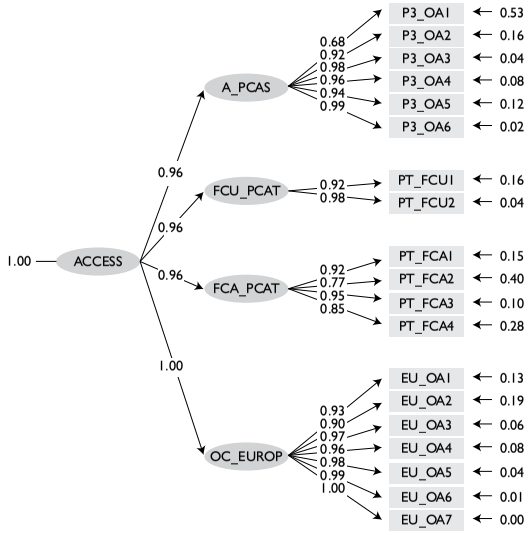
1A. Standard unidimensional model



Chi-square = 649.37, *df* = 152, *p*-value = 0.00000, RMSEA = 0.104

Note: This type of model says that variability in response to individual items can be explained by a single underlying factor, namely accessibility, irrespective of the scale from which items are drawn.

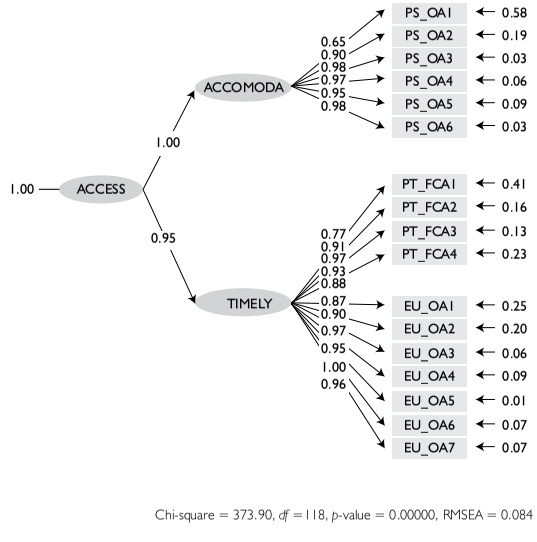
1B. Second-order, multidimensional model



Chi-square = 514.03, *df* = 148, *p*-value = 0.00000, RMSEA = 0.090

Note: Results show that the standard, second-order multidimensional model, in which individual items are located first within the scale from which they come (first-order factor) and are then linked to the general construct of accessibility (a second-order factor), also fit the data well, but not better than the unidimensional model.

1C. Second-order, multidimensional model

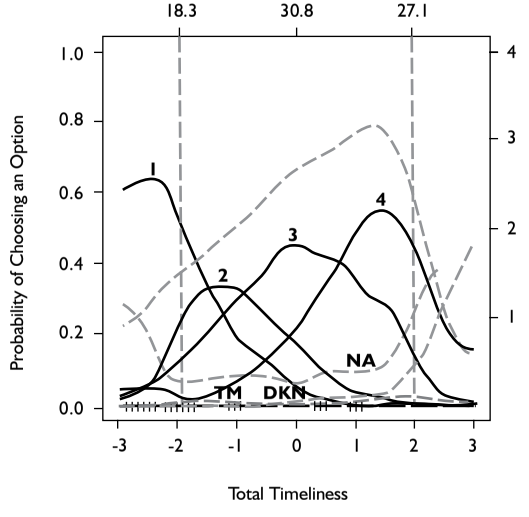


Chi-square = 373.90, *df* = 118, *p*-value = 0.00000, RMSEA = 0.084

Note: Results show that a revised, second-order multidimensional model, in which items are first located within two different facets of accessibility, namely the timeliness of service and the extent to which patient's needs are accommodated (first-order factors) and then linked to the general construct of accessibility (a second-order factor), also fit the data well, but not better than the unidimensional model.

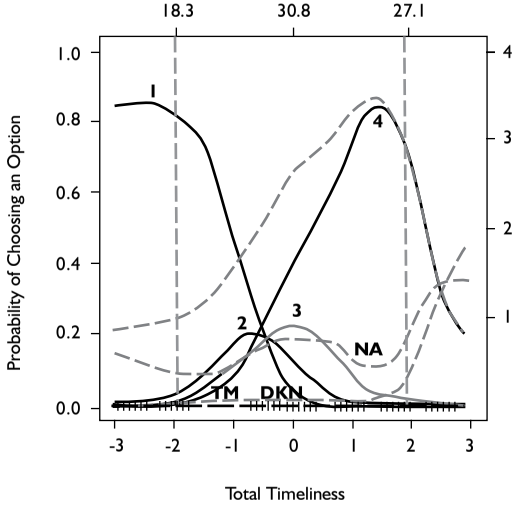
FIGURE 2. Response graphs for three items drawn from the PCAT-S and EUROPEP-I subscales assessing construct of timeliness within accessibility

2A.



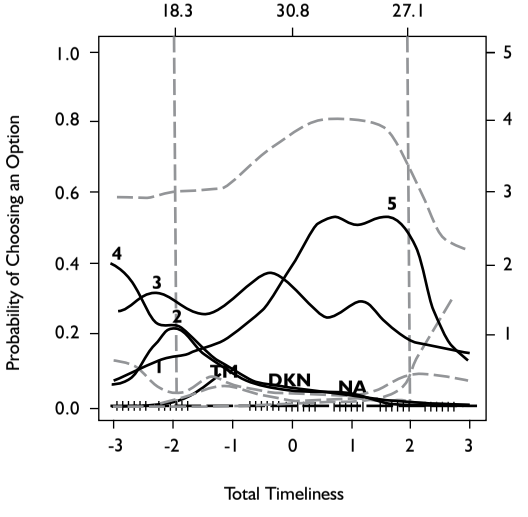
Note: Option characteristic curves (solid lines) and expected total scores (broken lines) for item 2, "When your Primary Care Provider is open, can you get advice quickly over the phone if you need it?" from the PCAT-S are modelled as a function of total scores on these measures (top axis) and standard normal quantiles (bottom axis). Results show that this item performs well relative to other items. First, option characteristic curves cover the entire range of scores. Second, the region in which curves are most likely to be endorsed is in a distinct region that corresponds to the weight assigned to each option. The probability of true missing (TM), not applicable (NA) and don't know (DK) responses did not vary substantially as a function of total scores for the majority of participants, falling between –2 and +2 standard normal quantiles.

2B.



Note: Option characteristic curves (solid lines) and expected total scores (broken lines) for item 3, "When your Primary Care Provider is closed, is there a phone number you can call when you get sick?" from the PCAT-S are modelled as a function of total scores on these measures (top axis) and standard normal quantiles (bottom axis). Results show difficulties with some options from this item. First, only two options, namely options 1 and 4, are being endorsed with any frequency, suggesting that this item is essentially dichotomous. The probability of endorsing options 2 or 3 is relatively small compared to options 1 and 4. The probability of true missing (TM), not applicable (NA) and don't know (DK) responses did not vary substantially as a function of total scores for the majority of participants, falling between –2 and +2 standard normal quantiles.

2C.



Note: Option characteristic curves for item 2 from the EUROPEP-I, "The helpfulness of staff (other than the doctor)." Results show a number of difficulties with this item. First, the probability of endorsing option 3 is greater than probability of endorsing option 2 at all levels of total scores. Second, most option characteristic curves do not change rapidly with changes in total scores, which is captured in an essentially flat expected item score. The probability of true missing (TM), not applicable (NA) and don't know (DK) responses did not vary substantially as a function of total scores for the majority of participants.

Item Response Models

Item response analysis evaluates how well questions and individual response options perform at different levels of the underlying construct being evaluated. They provide a fine-grained level of analysis that can be used to evaluate how well individual items and options discriminate among individuals at both high and low levels of the construct and can identify items and options that should ideally be revised or, if necessary, discarded.

Defining a shared or common underlying continuum against which items can be compared is crucial, because any result will be contingent on the appropriateness of this common underlying dimension. Each item's performance was modelled two ways, as a function of (1) items drawn from the single original subscale and (2) all items from any subscale that appear to measure a common construct, for example, timeliness within accessibility. Items are likely to perform better when modelled as a function of just the instrument from which they were drawn. However, because our goal is to compare the relative performance of all items (which may come from different instruments) that are believed to assess a similar construct (i.e., accessibility), we report on items modelled on the shared or common underlying dimension (e.g., timeliness within accessibility).

We used a non-parametric item response model to examine item performance on the common underlying factor (Ramsay 2000). This is an exploratory approach (Santor and Ramsay 1998), and these techniques have been used previously to examine the psychometric properties of self-report items and to evaluate item bias (Santor et al. 1994; Santor et al. 1995; Santor and Coyne 2001). A detailed description of the algorithm used to estimate response curves has been published elsewhere (Ramsay 1991).

We supplemented our non-parametric models with parametric item response modelling to estimate the discriminatory capacity of each item within its own original subscale using Multilog (Du Toit 2003). Discriminability (the "a" parameter) indicates an item's sensitivity to detect differences among individuals ranked on the construct being measured (e.g., accessibility). It can be viewed as a slope, with a value of 1 considered the lower limit for acceptable discriminability, i.e., each unit increase in the item predicts a unit increase in the underlying construct. Items with lower discriminability in the parent construct invariably performed poorly on the common underlying dimension with non-parametric item response modelling.

Examining item performance

To illustrate how item response models can be used to evaluate item and response option performance, Figures 2A, 2B and 2C show item response graphs for three items drawn from the PCAT-S and EUROPEP-I subscales assessing construct of timeliness within accessibility. Figure 2A presents a relatively well-performing item from the PCAT-S; Figures 2B and 2C illustrate some difficulties in the other two items.

In the Figure 2 graphs, the total expected score for timeliness is presented at the top of the plot; below the horizontal axis on the bottom, it is represented as standard normal scores. Expressing scores as standard normal scores is useful because it is informative about the propor-

tion of a population above or below integer values of standard deviations from the mean score. So in the graphs we can see that −2 SD corresponds to a total timeliness score of 18.3, the mean is 30.8, and +2 SD is 27.1. Extreme values on curves need to be interpreted with caution because, by definition, sample sizes are small in the tails of the overall distribution of scores.

The overall performance of the item is captured in the steepness of the slope of the characteristic curve (the topmost dashed lines in Figure 2A–C), which expresses item discriminability, the relationship between the cumulated item score and the total score in the construct (e.g., total timeliness). Given that we calculated items from different instruments as a function of a common continuum, slopes can be compared directly to assess performance across different subscales.

Several important features of item performance are illustrated in Figure 2A for an item from the PCAT-S. First, each of the option characteristic curves (a solid line probability curve for each response option) increases rapidly with small increases in timeliness. For example, the probability of option 1 being endorsed increases rapidly from 0.0 to 0.6 over a narrow region of timeliness, −3.0 to −1.5. Second, each option tends to be endorsed most frequently in a specific range of timeliness. For example, option 2 is more likely to be endorsed than any other option within the timeliness range of −1.0 to 0.0. Third, the regions over which each option is most likely to be endorsed are ordered, left to right, in the same way as the option scores (weights, 1 to 4). That is, the region in which option 2 is most likely to be endorsed falls between the regions in which option 1 and option 3 are most likely to be endorsed. Finally, together, the options for an item span the full continuum of accessibility, from −3 to +3. Most positive options are endorsed only at high levels of timeliness (e.g., option 5), whereas most negative options are endorsed only at low levels of timeliness (e.g., option 1).

In contrast, Figure 2B shows an item from the PCAT-S with four response options, but only options 1 and 4 are endorsed frequently. Options 2 and 3 do not provide any meaningful additional information, and the response scale functions essentially as a binary option. However, the responses cover specific and distinct areas of timeliness, making the item very discriminating, as illustrated by the steep slope of the item characteristic curve.

Figure 2C illustrates a problematic item. The response option curves do not peak rapidly nor in specific areas of timeliness, and the response options do not seem to be ordinal. The item characteristic curve is almost flat, showing little capacity for discrimination. It does not perform well to measure timeliness, which may not be surprising given that it asks about helpfulness of staff.

Conclusion

Each of the techniques described above offers a different method of examining item and subscale performance; applied together, they offer a comprehensive assessment of how the selected instruments measure performance of core primary healthcare attributes. The attribute-specific results are presented in individual papers elsewhere in this special issue.

The strength of this study was our analysis across instruments, which allowed us to identify sub-dimensions within an attribute. Sometimes a sub-dimension is unique to one subscale;

sometimes, more than one is represented. This approach will help program evaluators select the measures appropriate for their needs. Another consideration will be the missing values, and evaluators may choose not to offer "not sure" or "not applicable" options to minimize information loss. As with any study, results are sample-dependent, and items that do not function well in the present sample may still function well in a different sample of individuals or a different healthcare setting. However, the results of our study show that most of these measures can be used with confidence in the Canadian context. Ideally, any difficulties identified should be viewed as opportunities for improvement, potentially by rewriting, rewording or clarifying questions.

ACKNOWLEDGEMENTS

This research was funded by the Canadian Institutes of Health Research (CIHR).

Correspondence may be directed to: Darcy A. Santor, PhD, School of Psychology, Faculty of Social Sciences, University of Ottawa, 136 Jean-Jacques Lussier, Ottawa, ON K1N 6N5; tel.: 613-562-5799; fax: 613-562-5147; e-mail: dsantor@uottawa.ca.

REFERENCES

Borowsky, S.J., D.B. Nelson, J.C. Fortney, A.N. Hedeen, J.L. Bradley and M.K. Chapko. 2002. "VA Community-Based Outpatient Clinics: Performance Measures Based on Patient Perceptions of Care." *Medical Care* 40(7): 578-86.

Du Toit, M. 2003. *IRT from SSI: Bilog-mg, Multilog, Parscale, Testfact.* Lincolnwood, IL: Scientific Software International.

Flocke, S. 1997. "Measuring Attributes of Primary Care: Development of a New Instrument." *Journal of Family Practice* 45(1): 64-74.

Gorsuch, R.L. 1983. *Factor Analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Grol, R., M. Wensing and Task Force on Patient Evaluations of General Practice. 2000. "Patients Evaluate General/Family Practice: The EUROPEP Instrument." Nijmegen, Netherlands: Centre for Quality of Care Research, Raboud University.

Haggerty, J.L., F. Burge, M.-D. Beaulieu, R. Pineault, C. Beaulieu, J.-F. Lévesque et al. 2011. "Validation of Instruments to Evaluate Primary Healthcare from the Patient Perspective: Overview of the Method." *Healthcare Policy* 7 (Special Issue): 31–46.

Haggerty, J.L., J.-F. Lévesque, D.A. Santor, F. Burge, C. Beaulieu, F. Bouharaoui et al. 2011. "Accessibility from the Patient Perspective: Comparison of Primary Healthcare Evaluation Instruments." *Healthcare Policy* 7 (Special Issue): 94–107.

Jöreskog, K.G. and D. Sörbom. 1996. *LISREL 8: User's Reference Guide.* Chicago: Scientific Software International.

Kline, R.B. 1998. *Principles and Practice of Structural Equation Modeling.* New York: Guilford Press.

Ramsay, J.O. 1991. "Kernel Smoothing Approaches to Nonparametric Item Characteristic Curve Estimation." *Psychometrika* 56: 611–30.

Ramsay, J.O. 2000. *TESTGRAF: A Program for the Graphical Analysis of Multiple-Choice Test and Questionnaire Data (Computer Program and Manual).* Montreal: McGill University, Department of Psychology. Retrieved June 11, 2011. <ftp://ego.psych.mcgill.ca/pub/ramsay/testgraf/TestGrafDOS.dir/testgraf1.ps>.

Safran, D.G., J. Kosinski, A.R. Tarlov, W.H. Rogers, D.A. Taira, N. Lieberman and J.E. Ware. 1998. "The Primary Care Assessment Survey: Tests of Data Quality and Measurement Performance." *Medical Care* 36(5): 728–39.

Santor, D.A. and J. C. Coyne. 2001. "Examining Symptom Expression as a Function of Symptom Severity: Item

Performance on the Hamilton Rating Scale for Depression." *Psychological Assessment* 13: 127–39.

Santor, D.A. and J.O. Ramsay. 1998. "Progress in the Technology of Measurement: Applications of Item Response Models." *Psychological Assessment* 10: 345–59.

Santor, D.A., J.O. Ramsay and D.C. Zuroff. 1994. "Nonparametric Item Analyses of the Beck Depression Inventory. Examining Gender Item Bias and Response Option Weights in Clinical and Nonclinical Samples." *Psychological Assessment* 6: 255–70.

Santor, D.A., D.C. Zuroff, J.O. Ramsay, P. Cervantes and J. Palacios. 1995. "Examining Scale Discriminability in the BDI and CES-D as a Function of Depressive Severity." P*sychological Assessment* 7: 131–39.

SAS Institute. 2003. *SAS User's Guide: Statistics* (Version 9.1). Cary, NC: SAS Institute.

Shi, L., B. Starfield and J. Xu. 2001. "Validating the Adult Primary Care Assessment Tool." *Journal of Family Practice* 50(2): n161w–n171w.

Stewart, A.L., A.M. Nápoles-Springer, S.E. Gregorich and J. Santoyo-Olsson. 2007. "Interpersonal Processes of Care Survey: Patient-Reported Measures for Diverse Groups. *Health Services Research* 42(3 Pt. 1): 1235–56.