RESEARCH PAPER

Differential Item Functioning in Primary Healthcare Evaluation Instruments by French/English Version, Educational Level and Urban/Rural Location

Le fonctionnement différentiel des items dans les instruments d'évaluation des soins de santé primaires en fonction des versions française ou anglaise, du niveau de scolarisation et du lieu de résidence urbain ou rural

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

FATIMA BOUHARAOUI, MSC
St. Mary's Research Centre, St. Mary's Hospital Center
Montreal, QC

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

Abstract

Evaluating the extent to which groups or subgroups of individuals differ with respect to primary healthcare experience depends on first ruling out the possibility of bias.

Objective: To determine whether item or subscale performance differs systematically between French/English, high/low education subgroups and urban/rural residency.

Method: A sample of 645 adult users balanced by French/English language (in Quebec and Nova Scotia, respectively), high/low education and urban/rural residency responded to six validated instruments: the Primary Care Assessment Survey (PCAS); the Primary Care Assessment Tool – Short Form (PCAT-S); the Components of Primary Care Index (CPCI); the first version of the EUROPEP (EUROPEP-I); the Interpersonal Processes of Care Survey, version II (IPC-II); and part of the Veterans Affairs National Outpatient Customer Satisfaction Survey (VANOCSS). We normalized subscale scores to a 0-to-10 scale and tested for between-group differences using ANOVA tests. We used a parametric item response model to test for differences between subgroups in item discriminability and item difficulty. We re-examined group differences after removing items with differential item functioning.
Results: Experience of care was assessed more positively in the English-speaking (Nova Scotia) than in the French-speaking (Quebec) respondents. We found differential English/French item functioning in 48% of the 153 items: discriminability in 20% and differential difficulty in 28%. English items were more discriminating generally than the French. Removing problematic items did not change the differences in French/English assessments. Differential item functioning by high/low education status affected 27% of items, with items being generally more discriminating in high-education groups. Between-group comparisons were unchanged. In contrast, only 9% of items showed differential item functioning by geography, affecting principally the accessibility attribute. Removing problematic items reversed a previously non-significant finding, revealing poorer first-contact access in rural than in urban areas.
Conclusion: Differential item functioning does not bias or invalidate French/English comparisons on subscales, but additional development is required to make French and English items equivalent. These instruments are relatively robust by educational status and geography, but results suggest potential differences in the underlying construct in low-education and rural respondents.

Résumé

Afin d'évaluer à quel point des groupes ou sous-groupes d'individus divergent quant à leur expérience en matière de soins de santé primaires, il faut d'abord éliminer les possibilités de biais.
Objectif : Déterminer si la performance d'un item ou d'une sous-échelle diffère systématiquement en fonction de la langue (français/anglais), des sous-groupes de scolarisation (élevée/faible) et du lieu de résidence (urbain/rural).
Méthode : Un échantillon de 645 adultes utilisateurs, équilibré en fonction de la langue (français : Québec et anglais : Nouvelle-Écosse), du niveau de scolarisation (élevé/faible) et du lieu de résidence (urbain/rural), a répondu aux six instruments validés suivants : Primary Care Assessment Survey (PCAS); Primary Care Assessment Tool – version courte (PCAT-S); Components of Primary Care Index (CPCI); la première version de l'EUROPEP (EUROPEP-I); Interpersonal Processes of Care Survey, version II (IPC-II); et une partie du Veterans Affairs National Outpatient Customer Satisfaction Survey (VANOCSS). Nous avons normalisé les scores des sous-échelles selon une échelle de 0 à 10 et nous avons vérifié les différences entre les sous-groupes au moyen de tests ANOVA. Nous avons utilisé un

modèle paramétrique de la théorie des réponses par items (IRT) pour tester les différences entre les sous-groupes selon le pouvoir discriminant des items et leur niveau de difficulté. Puis, nous avons réexaminé les différences entre les groupes après avoir retiré les items qui présentaient un fonctionnement différentiel (DIF).
Résultats : L'expérience de soins a été évaluée plus positivement au sein du groupe anglophone (Nouvelle-Écosse) par rapport au groupe francophone (Québec). Nous avons observé un fonctionnement différentiel d'item selon la langue anglais/français dans 48 % des 153 items : une discrimination différentielle dans 20 % des cas et une difficulté différentielle dans 28 % des cas. Les items anglais étaient généralement plus discriminants que les items français. Il n'y a pas eu de changement des différences français/anglais observées après le retrait des items problématiques. Le fonctionnement différentiel des items selon le niveau de scolarisation (élevé/faible) affectait 27 % des items, qui étaient généralement plus discriminants pour les groupes de scolarisation élevée. Les comparaisons entre les groupes n'ont pas montré de changement. Par contre, seulement 9 % des items montraient un fonctionnement différentiel en fonction du lieu géographique, affectant principalement l'accessibilité. Le retrait des items problématiques a provoqué le renversement d'un résultat préalablement non significatif, révélant un plus faible accès de premier contact dans les zones rurales par rapport aux zones urbaines.
Conclusion : Le fonctionnement différentiel des items ne cause pas de biais ou n'invalide pas les comparaisons français/anglais par sous-échelle, mais une adaptation supplémentaire est nécessaire pour développer des items équivalents en français et en anglais. Ces instruments sont relativement robustes en fonction du niveau de scolarisation et du lieu géographique, mais les résultats suggèrent des différences potentielles dans le construit sous-jacent, pour les répondants de niveau de scolarisation plus faible et des zones rurales.

Examining group differences in healthcare experience, whether across geographic locations or linguistic/ethnic groups, is essential to ensuring that healthcare is delivered as equitably and effectively as possible. However, observed differences between two groups do not necessarily imply true differences unless it can be demonstrated that the evaluation scales and measures function similarly in both groups. To interpret group differences, we must first rule out any bias in how individuals answer questions.

Differential item functioning (sometimes called item bias) occurs when, at the same level of the underlying construct, responses differ significantly by group membership. If several items in a subscale demonstrate differential item functioning, this may adversely affect the conclusions of between-group comparisons by creating a false difference or failing to detect a true difference.

Among instruments developed to measure the quality of primary healthcare from the patient's perspective, we identified six in the public domain that appeared of greatest relevance for Canada: the Primary Care Assessment Survey (PCAS) (Safran et al. 1998); the adult Primary Care Assessment Tool – Short form (PCAT-S) (Shi et al. 2001); the Components

of Primary Care Index (CPCI) (Flocke 1997); the first version of the EUROPEP (EUROPEP-I) (Wensing et al. 2000); the Interpersonal Processes of Care Survey, version II (IPC-II) (Stewart et al. 2007); and the Veterans Affairs National Outpatient Customer Satisfaction Survey (VANOCSS) (Borowsky et al. 2002).

We wanted to determine whether French- and English-language versions of the instruments were equivalent and whether item or subscale performance differed systematically by high/low educational status or by urban/rural location.

Specific research questions

All the instruments used in our study were originally developed in English, and thus equivalence with French versions was a major concern. In translation to French, some phrases proved problematic. For example, for rating response options, the European French version of the EUROPEP translated "poor" as "médiocre" (second-rate), whereas Quebec translators rendered it as "mauvais" (bad). The English question "How often…," followed by frequency response options "always," "usually" and "sometimes," was translated in French as "Combien de fois…" (How many times), which naturally elicits a count rather than a frequency response. Finally, one instrument used the term "primary care provider" to refer to both person and place, for which there is no single French equivalent. Consequently, the French term varied by the context of the question specifying as "source habituelle de soins" (usual source of care), "clinique" (clinic) or "médecin" (physician).

Our concern about differential functioning by geographic area arose from previous studies in which rural residents reported better accessibility than did residents of metropolitan areas (Haggerty et al. 2007). We hypothesized that measures of accessibility may function differently by context. We had no a priori concerns regarding educational achievement, but we wanted to ensure that all instruments performed equally well in low-literacy groups, because we found considerable variation in readability among instruments.

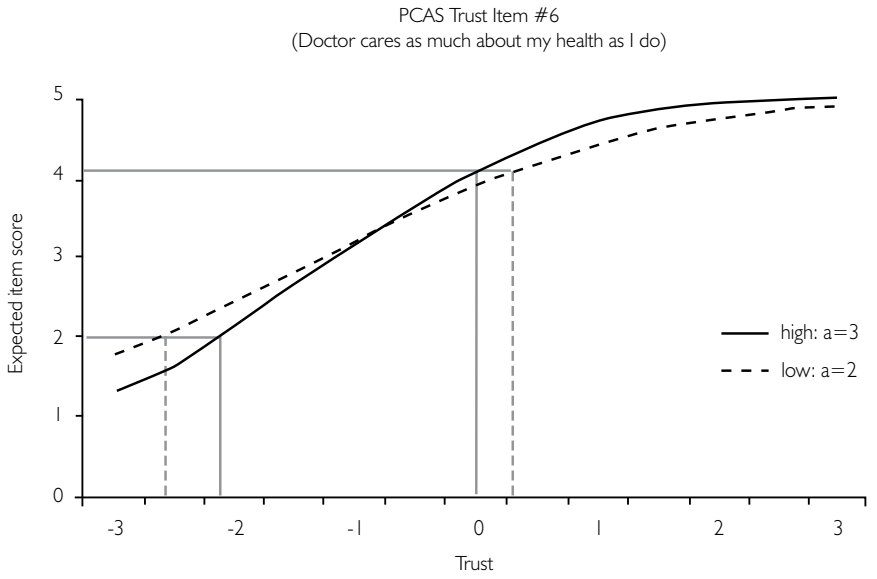
Overview of differential functioning

The language of differential item functioning analysis reflects its origins in educational assessment. The method, developed to assess the performance of questions that estimate a student's understanding of a topic, evaluates performance in two ways – first in terms of *discriminability*, i.e., how well the item can differentiate between individuals with different levels of ability, and then in terms of *difficulty*, i.e., how hard it is to answer correctly a question at different levels of student ability.

For example, if the probability of answering correctly changes depending on the student's level of ability and can detect even a small difference in ability between two individuals, then the question has good discriminability. If a student has a 50% probability of responding correctly only in the high range of ability, then the question or item is considered difficult; if 50% probability is achieved in the low range of ability, then it is considered easy. A good instrument includes questions with difficulty thresholds across the entire ability range, each with good discriminability.

This approach has also been used to evaluate item performance of attitudinal surveys. Discriminability is an item's sensitivity to differences between individuals on the construct being measured (e.g., trust in the provider) and is represented with a slope in item response models. The steeper the slope, the more discriminating the item, with slopes ≥1 (the "a" parameter) considered appropriate; i.e., each unit increase in the item predicts a unit increase in the underlying construct. Ideally, the item's slope should not differ among subgroups; if it does, the item has differential discriminability. Figure 1 illustrates differential discriminability by educational level for an item in the Trust subscale from the Primary Care Assessment Survey (PCAS), showing higher discrimination in high- than in low-education respondents. Differences in item discrimination indicate that the question is understood or interpreted differently by each subgroup. This would occur, for example, when the French translation is not equivalent in meaning to the original English version.

FIGURE 1. Differential discriminability between high- and low-education respondents for item in the PCAS Trust subscale



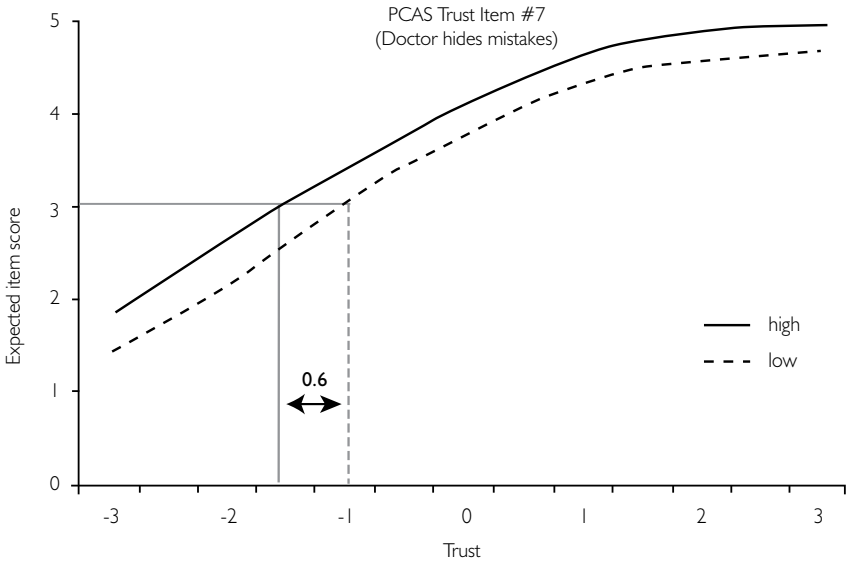
Note: Difficulty threshold is lower (easy) in low- than in high-education respondents at low levels of trust, but higher (more difficult) at high levels of trust.

Difficulty in attitudinal surveys refers to the probability of endorsing a specific response option for a given level of the construct being measured. When an item's difficulty threshold varies by group membership, the item is said to exhibit differential difficulty functioning, e.g., in the PCAS Trust subscale that elicits agreement with statements using a five-point Likert scale of "1 = strongly disagree" to "5 = strongly agree." Figure 2 illustrates the differential difficulty threshold between high- and low-education respondents. Note that for the same expected item score of 3, low-education respondents will have 0.6 higher level of trust on the

standardized trust score (i.e., the item is more difficult) than will high-education respondents. The difficulty differential is uniform across all levels of trust, whereas in Figure 1, showing differential discriminability, the difficulty differential is not uniform across levels of trust. Consistent differences in difficulty thresholds in a scale's items may point to differences in how response options are interpreted.

The potential impact of differential item functioning is assessed by removing problematic items from the subscale or instrument, recalculating the scores using only the purified scale (non-problematic or "anchor" items) and comparing the group values again. If the between-group comparison using the purified scale reaches a different conclusion, differential item functioning is said to have an impact, and using the original scale could give biased measures. If the comparison remains essentially unchanged (typically, when differences are minor or in different directions), differential item functioning is said to have no impact.

FIGURE 2. Differential difficulty between high- and low-education respondents for item in the PCAS Trust subscale, showing equal discriminability (a=1.64) and uniform difficulty threshold (b=0.6) across all levels of trust



Method

Study population

The target population for this study was adult, Canadian, primary healthcare users, undifferentiated by age or health condition. The sample was selected to be balanced by French/English language, high/low educational level and urban/rural location. We also stratified by excellent, average and poor primary care experience based on a single screening question: "Overall, has your experience of care from your regular family doctor or medical clinic been excellent, poor or average?"

Participants responded to all six instruments and provided socio-demographic and utilization information. Data were collected between February and July 2005. English-language questionnaires were administered in Nova Scotia and French-language questionnaires in Quebec.

Urban location was defined as census metropolitan areas; rural, as more than one hour's travel from a metropolitan area; remote (Quebec only), more than four hours' travel. We used an age-sensitive cut-off to denote educational achievement as a proxy for reading level. Subjects were considered to have a high-school reading level or lower if they had (a) completed only high school and were under 45 years old, (b) completed 10 years of school and were 45 to 55 years old or (c) completed less than eight years and were over 55 years old (Smith and Haggerty 2003).

Analysis

We examined the distribution of missing values by language, educational achievement and geography. The score for each subscale, calculated as the mean of items, was normalized to a 0-to-10 scale to permit comparisons on a common metric (formula, Table 2). We compared normalized subscale scores by language, education and geography using regression modelling controlling for the other design variables as well as for overall experience, using $\alpha=.05$ to denote statistical significance, despite multiple testing, to maintain a high sensitivity to potential differences. We conducted exploratory factor analysis to examine whether factor resolution for the subscales was the same by language, education and geography.

All methods of assessing differential item functioning consist of examining the distribution of responses in the subgroups of interest when they are conditioned on the same level of the underlying construct or latent variable (Santor and Ramsay 1998; Kristjansson et al. 2005; Reeve 2006; Teresi and Fleishman 2007). In this study, we used parametric item response analysis using Multilog software (Du Toit 2003) to test for differential discrimination and difficulty across all options within an item using a chi-square test. The latent variable was the total score of the subscale. We first assessed significant between-group differences in the discrimination parameter. If none was found, we fixed the discrimination parameter to be equal between groups and tested for uniform and non-uniform differences in the difficulty threshold across response options. We retested the discrimination parameter after removing problematic items from the latent variable and repeated the process until we found no differentially functioning items. We used a critical value of $\alpha=.01$ to indicate statistical significance because lower values detected trivial differences.

Finally, we re-examined group differences with a series of standard ANOVA tests using subscale scores based on the subset of items found to be free of differential item functioning.

Results

The six instruments contained 153 validated items. Despite attempts to balance the sample equally by French/English language, high/low education and urban/rural location, the 645 respondents were not equally distributed. The English-language group was more urban (59% vs. 49%, $\chi^2=6.7$, $p<.096$), more likely to have a high-school reading level (75% vs. 55%, $\chi^2=27.5$,

$p<.001$) and also more likely to perceive their health as good or excellent, to be affiliated with a family doctor rather than a clinic and to have longer affiliations (Table 1). Wait times for appointments were better among English-speaking, high-education and urban respondents.

TABLE 1. Characteristics of the study sample compared by language, geography and education (only statistically significant differences are shown*)

Characteristic	Total (n = 645) % (n)	Language % (n)		Geography % (n)		Education % (n)	
		English (n=343)	French (n=302)	Urban (n=351)	Rural (n=294)	High (n=424)	Low (n=221)
Overall experience of care							
Poor	23.1 (149)	–		–		–	
Average	36.0 (232)						
Excellent	40.9 (264)						
Mean age in years (SD)	48.0 (14.9)	–		46.4 (15.0)	49.8 (14.4)**	49.5 (13.9)	45.1 (13.2)**
Per cent female	64.7 (414)	–		–		–	
Mean years of education (SD)	13.0 (3.4)	13.5 (3.0)	12.4 (3.6) ***	13.5 (3.4)	12.4 (3.3) ***	14.4 (2.8)	10.3 (2.6) ***
Regular provider							
Physician	94.1 (607)	96.8 (332)	91.1 (275) **	–		96.2 (408)	90.1 (199) **
Clinic only	5.9 (38)	3.2 (11)	8.9 (27)			3.8 (16)	9.9 (22)
Per cent indicating health status as good or excellent	38.0 (241)	45.6 (155)	29.2 (86) ***	–		42.5 (179)	29.0 (62) ***
Per cent with chronic health problem	61.6 (392)	–		–		–	
Mean usual wait time for appointment			***		***		*
Less than 2 days	35.2 (220)	53.2 (177)	14.7 (43)	43.7 (149)	25.0 (71)	38.5 (158)	28.8 (62)
2 to 7 days	32.6 (204)	36.3 (121)	28.4 (83)	33.4 114)	31.7 (90)	32.9 (135)	32.1 (69)
7 days to 2 weeks	11.8 (74)	8.7 (29)	15.4 (45)	11.1 (38)	12.7 (36)	12.0 (49)	11.6 (25)
2 weeks to 4 weeks	9.3 (58)	1.2 (4)	18.5 (54)	5.6 (19)	13.7 (39)	7.6 (31)	12.6 (27)
More than 4 weeks	11.0 (69)	0.6 (2)	23.0 (67)	6.2 (21)	16.9 (48)	9.0 (37)	14.9 (32)
Mean number of visits to other providers (SD)							
Specialist	4.2 (5.1)	–		–		–	
Other providers	9.4 (19.2)	–		–		–	

* p≤.05
** p≤.01
*** p≤.001

Descriptive results

The number of missing values was not systematically higher by language or educational achievement, but was higher in rural than in urban respondents for five items, but they were not

consistently in one instrument or attribute. English-speaking respondents showed a higher tendency than the French speakers to select the "don't know/not applicable" option when offered.

The normalized subscale scores, grouped by primary care attribute, are compared by language, education and geography in Table 2. Strikingly, subscale scores are systematically higher (more positive assessment) in the English than in the French subgroup, with the exception of the CPCI Coordination of Care subscale. The study design does not allow us to determine whether the difference is due to true differences between Nova Scotia and Quebec or differential functioning between English- and French-language versions.

TABLE 2. Comparison of mean normalized subscale scores by language, geography, education and total sample (only statistically significant differences are shown***)

Questionnaire	Developer's Subscale Name	# of Items in Subscale	Language English (SD) French (SD)	Education Low (SD) High (SD)	Geography Rural (SD) Urban (SD)	Overall Mean (SD)
Accessibility						
PCAS	Organizational Access	6	6.31 (1.73) 5.52 (1.85)***	–	–	5.94 (1.83)
PCAT	First-Contact Utilization	3	9.26 (1.36) 8.89 (1.80)**	–	9.25 (1.35) 8.95 (1.76)*	9.10 (1.60)
PCAT	First-Contact Access	4	6.36 (2.43) 4.71 (2.54)***	5.31 (2.66) 5.83 (2.56)*	–	5.60 (2.60)
EUROPEP	Organization of Care	7	7.05 (2.09) 5.92 (2.50)***	–	–	6.51 (2.36)
Comprehensiveness of Services						
PCAT	Comprehensiveness (services available)	4	8.54 (1.73) 6.71 (2.84)***	–	–	7.72 (2.47)
CPCI	Comprehensive Care	6	8.16 (1.85) 7.23 (2.24)***	–	–	7.72 (2.09)
Relational Continuity						
PCAS	Visit-Based Continuity	2	8.78 (1.60) 7.86 (2.48)***	–	–	8.35 (2.11)
PCAS	Contextual Knowledge	5	6.09 (2.24) 5.73 (2.31)*	–	–	5.92 (2.28)
PCAT	Ongoing Care	4	7.38 (2.13) 6.89 (2.54)**	–	–	7.15 (2.34)
CPCI	Accumulated Knowledge	8	7.21 (2.42) 6.75 (2.54)*	–	–	6.99 (2.49)
CPCI	Patient Preference for Regular Physician	5	8.27 (1.72) 7.02 (2.12)***	–	–	7.68 (2.01)

TABLE 2. Continued

Questionnaire	Developer's Subscale Name	# of Items in Subscale	Language English (SD) French (SD)	Education Low (SD) High (SD)	Geography Rural (SD) Urban (SD)	Overall Mean (SD)
Management Continuity						
PCAS	Integration	6	7.28 (1.94) 6.41 (2.14)***	–	–	6.90 (2.07)
PCAT	Coordination	4	7.92 (2.36) 7.13 (2.86)***	–	–	7.57 (2.62)
CPCI	Coordination of Care	8	6.37 (1.33) 6.86 (2.59)**	–	–	6.60 (2.03)
VANOCSS§	Coordination of Care (overall), number of problems	6	2.33 (1.81) 2.80 (1.91)**	–	–	2.51 (1.88)
VANOCSS§	Specialty Provider Access (number of problems)	4	0.42 (0.78) 0.78 (0.97)**	0.79 (0.94) 0.52 (0.88)*	–	0.62 (0.91)
Interpersonal Communication						
PCAS	Communication	6	7.71 (1.92) 6.90 (2.22)***	–	–	7.33 (2.11)
PCAS	Trust	8	7.86 (1.66) 7.17 (1.83)***	–	7.36 (1.74) 7.68 (1.80)*	7.53 (1.78)
CPCI	Interpersonal Communication	6	7.60 (2.28) 6.72 (2.27)***	–	–	7.19 (2.32)
EUROPEP	Clinical Behaviour	16	8.14 (1.84) 7.54 (2.26)***	–	–	7.85 (2.07)
IPC-II	Communication (elicited concerns, responded)	3	8.40 (1.79) 7.15 (2.35)***	–	–	7.81 (2.16)
IPC-II	Communication (explained results, medications)	4	7.66 (2.37) 7.12 (2.59)**	–	–	7.40 (2.49)
IPC-II	Decision-Making (patient-centred decision-making)	4	5.82 (3.06) 4.97 (3.19)***	–	–	5.41 (3.15)
Respectfulness						
PCAS	Interpersonal Treatment	5	7.83 (2.01) 7.00 (2.26)***	–	–	7.44 (2.17)
IPC-II	Hurried Communication	5	8.44 (1.54) 7.52 (1.88)***	–	–	8.01 (1.77)
IPC-II	Interpersonal Style (compassionate, respectful)	5	8.35 (2.16) 7.66 (2.46)***	–	–	8.02 (2.33)
IPC-II	Interpersonal Style (respectful office staff)	4	9.05 (1.72) 8.47 (1.91)***	–	–	8.78 (1.83)

TABLE 2. Continued

Questionnaire	Developer's Subscale Name	# of Items in Subscale	Language English (SD) French (SD)	Education Low (SD) High (SD)	Geography Rural (SD) Urban (SD)	Overall Mean (SD)
Whole-Person Care						
PCAT	Community Orientation	3	5.31 (2.75) 4.44 (2.93)***	–	–	4.88 (2.87)
CPCI	Community Context	2	7.28 (2.71) 5.55 (3.32)***	–	–	6.47 (3.13)

* $p \leq .05$

** $p \leq .01$

*** $p \leq .001$

[§] The VANOCSS scores are not normalized; the score represents the number of problems reported.

Only one subscale differs by education, the PCAT-S First-Contact Access, with fewer positive assessments in low- than in high-education groups. Rural respondents indicate more positive assessments than urban respondents in PCAT-S First-Contact Utilization and fewer positive assessments in PCAS Trust.

Most subscales had similar factor resolution by subgroup. Three subscales found two factors (eigenvalue >1) in one group and the expected single factor in the other: the CPCI Coordination of Care (management continuity) subscale had two factors in English; the CPCI Preference for Regular Physician (relational continuity) had two in rural; and the PCAS Trust (interpersonal communication) subscale had two in French-speaking and in low-education respondents.

Differential functioning

Because of space constraints, we report only summary results at a subscale level; item-specific results are available upon request. The discriminability of individual items is reported in the attribute-specific papers elsewhere in this special issue of the journal. Table 3 shows the number of items within each subscale that are free of differential functioning and would be considered pure or anchor items for making valid comparisons between subgroups.

The French/English comparison exhibited the most differential item functioning and urban/rural, the least. Of the 153 items, only 80 (52%) were free of French/English differential functioning, compared to 111 (73%) in high/low education and 139 (91%) in urban/rural location.

Of the items with differential French/English functioning, one-third (24/73) were important differences in discriminability or difficulty. Overall, 41% (30/73) of items showed differences in discriminatory capacity, but only 13 of these had discriminability differentials greater than 1 (Figure 1 demonstrates a discriminability differential of 1). English items tended to be more discriminating, but only four items discriminated adequately in English and poorly in French, all from the CPCI. For example, agreement with the statement, "If I am sick I would always contact this doctor first" (CPCI Preference for Regular Physician) had a discrimination value of 1.63 in English and 0.87 in French.

TABLE 3. Number of items free from differential item functioning (discrimination or difficulty) within each validated subscale by language, geography and education

Developer's Subscale Name	Number of Items without Differential Item Functioning		
	Language (Province) English/French	Education Low/High	Geography Urban/Rural
Accessibility			
PCAS Organizational Access	4/6 (67%)	6/6 (100%)	5/6 (83%)
PCAT First-Contact Utilization	2/3 (67%)	3/3 (100%)	3/3 (100%)
PCAT First-Contact Accessibility	1/4 (25%)	3/4 (75%)	2/4 (50%)
EUROPEP Organization of Care	3/7 (43%)	7/7 (100%)	5/7 (71%)
Subtotal	*10/20 (50%)*	*19/20 (95%)*	*15/20 (75%)*
Comprehensiveness of Services			
PCAT Comprehensiveness (services available)	2/4 (50%)	4/4 (100%)	4/4 (100%)
CPCI Comprehensive Care	3/6 (50%)	4/6 (67%)	6/6 (100%)
Subtotal	*5/10 (50%)*	*8/10 (80%)*	*10/10 (100%)*
Relational Continuity			
PCAS Visit-Based Continuity	0/2 (0%)	0/2 (0%)	2/2 (100%)
PCAS Contextual Knowledge	3/5 (60%)	5/5 (100%)	4/5 (80%)
PCAT Ongoing Care	1/4 (25%)	4/4 (100%)	3/4 (75%)
CPCI Accumulated Knowledge	1/8 (13%)	4/8 (50%)	7/8 (88%)
CPCI Patient Preference for Regular Physician	2/5 (40%)	4/5 (80%)	4/5 (80%)
Subtotal	*7/24 (29%)*	*17/24 (71%)*	*20/24 (83%)*
Management Continuity			
PCAS Integration	3/6 (50%)	5/6 (83%)	6/6 (100%)
PCAT Coordination	2/4 (50%)	4/4 (100%)	3/4 (75%)
CPCI Coordination of Care	3/8 (38%)	5/8 (63%)	6/8 (75%)
VANOCSS Coordination of Care (overall), number of problems	6/6 (100%)	6/6 (100%)	6/6 (100%)
VANOCSS Specialty Provider Access, number of problems	4/4 (100%)	4/4 (100%)	4/4 (100%)
Subtotal	*18/28 (64%)*	*24/28 (86%)*	*25/28 (89%)*
Interpersonal Communication			
PCAS Communication	5/6 (83%)	5/6 (83%)	6/6 (100%)
PCAS Trust	1/8 (13%)	1/8 (13%)	8/8 (100%)
CPCI Interpersonal Communication	1/6 (17%)	2/6 (33%)	6/6 (100%)
EUROPEP Clinical Behaviour	10/16 (63%)	15/16 (94%)	16/16 (100%)

TABLE 3. Continued

Developer's Subscale Name	Number of Items without Differential Item Functioning		
	Language (Province) English/French	Education Low/High	Geography Urban/Rural
IPC-II Communication (elicited concerns, responded)	3/3 (100%)	0/3 (0%)	2/3 (67%)
IPC-II Communication (explained results, medications)	4/4 (100%)	2/4 (50%)	4/4 (100%)
IPC-II Decision-Making (patient-centred decision-making)	4/4 (100%)	3/4 (75%)	4/4 (100%)
Subtotal	*28/47 (60%)*	*28/47 (60%)*	*46/47 (98%)*
Respectfulness			
PCAS Interpersonal Treatment	4/5 (80%)	0/5 (0%)	5/5 (100%)
IPC-II Hurried Communication	4/5 (80%)	5/5 (100%)	5/5 (100%)
IPC-II Interpersonal Style (compassionate, respectful)	1/5 (20%)	5/5 (100%)	5/5 (100%)
IPC-II Interpersonal Style (respectful office staff)	2/4 (50%)	2/4 (50%)	3/4 (75%)
Subtotal	*11/19 (58%)*	*12/19 (63%)*	*18/19 (95%)*
Whole-Person Care – Community Orientation			
PCAT Community Orientation	1/3 (33%)	3/3 (100%)	3/3 (100%)
CPCI Community Context	0/2 (0%)	0/2 (0%)	2/2 (100%)
Subtotal	*1/5 (20%)*	*3/5 (60%)*	*5/5 (100%)*
Number of subscales with no differential item functioning	5/29 (17%)	12/29 (41%)	18/29 (62%)
Number of subscales where ≥50% of items exhibit differential functioning	12/29 (41%)	6/29 (21%)	0/29 (0%)

Of the 43 items with differential French/English difficulty, only 11 had differentials over 0.5, which is approximately the magnitude illustrated in Figure 2. The pattern of differences does not support a systematic difference between English and French when "poor" is translated as "médiocre" versus "mauvais," and it appears that frequency response scales were understood equivalently in both French and English. However, the difficulty threshold for the "fortement en désaccord" option is consistently more positive than for "strongly disagree" across several subscales and two instruments (CPCI and PCAS Trust). The response option "strongly disagree" seems to be more negative than "fortement en désaccord." We found no systematic direction of difficulty differences for "strongly agree."

By education, 43% (18/42) of differentially functioning items were due to differential discriminability, with seven being differentials >1. Items tended to have higher discrimination values in high-education respondents, although the reverse was seen for respectfulness. Only 12 of the remaining 24 items had difficulty differentials >0.5. The items tend to be more discriminating and difficult in the high-education than in the low-education groups; specifically,

low-education respondents have a higher probability of responding positively at lower levels of the construct of interest (communication, respectfulness). One of the largest observed difficulty differentials was in the PCAS Interpersonal Treatment (respectfulness) subscale, where all items had differential functioning, with an average difficulty threshold being 0.8 higher for high-education than low-education respondents.

By geography, there were only 14 differentially functioning items, with four out of seven discriminability differentials being >0.5. All were in accessibility and relational continuity. All items were more discriminating in urban than in rural groups.

Table 4 compares the subscale scores by language, education and geography after we removed items with differential functioning. Of the 29 subscales, only five (17%) are free from French/English differential functioning, compared to 12 (41%) in education and 18 (62%) in geography. Valid comparison by language was impossible for subscales with no remaining non-problematic items: PCAS Visit-Based Continuity, PCAT-S Community Orientation and CPCI Community Context. Comparisons based on less than 50% of the original items must be interpreted cautiously; this affects 12 (41%) subscales by language, six (21%) by education and none by geography. However, the results show that, for language, the conclusions are essentially unchanged from those of Table 2: assessments for all attributes remain more positive in English (Nova Scotia) than in French (Quebec). The previous more positive French scores on CPCI Coordination of Care disappear in the purified subscale.

TABLE 4. Subscale comparisons by language, geography and education using purified subscale scores (free of items with differential item functioning)

Developer's Subscale Name	Language (Province) French/English	Education High/Low	Geography Urban/Rural
Accessibility			
PCAS Organizational Access	+	NS	NS
PCAT First-Contact Utilization	+	NS	+
PCAT First-Contact Access	+	+ / NS	NS / +
EUROPEP Organization of Care	+	NS	NS
Comprehensiveness of Services			
PCAT Comprehensiveness (services available)	+	NS	NS
CPCI Comprehensive Care	+	NS	NS
Relational Continuity			
PCAS Visit-Based Continuity	+ / 0	NS / 0	NS
PCAS Contextual Knowledge	+	NS	NS
PCAT Ongoing Care	+ / NS	NS	NS
CPCI Accumulated Knowledge	+ / NS	NS	NS

TABLE 4. Continued

Developer's Subscale Name	Language (Province) French/English	Education High/Low	Geography Urban/Rural
CPCI Patient Preference for Regular Physician	+	NS	NS
Management Continuity			
PCAS Integration	+	NS	NS
PCAT Coordination	+	NS	NS
CPCI Coordination of Care	+ / –	NS	NS
VANOCSS Coordination of Care (overall), number of problems	+	NS	NS
VANOCSS Specialty Provider Access, number of problems	+	+	NS
Interpersonal Communication			
PCAS Communication	+	NS	NS
PCAS Trust	+ / NS	NS	+
CPCI Interpersonal Communication	+	NS	NS
EUROPEP Clinical Behaviour	+	NS	NS
IPC-II Communication (elicited concerns, responded)	+	NS / 0	NS
IPC-II Communication (explained results, medications)	+	NS	NS
IPC-II Decision-Making (patient-centred decision-making)	+	NS	NS
Respectfulness			
PCAS Interpersonal Treatment	+	NS / 0	NS
IPC-II Hurried Communication	+	NS	NS
IPC-II Interpersonal Style (compassionate, respectful)	+	NS	NS
IPC-II Interpersonal Style (respectful office staff)	+	NS	NS
Whole-Person Care			
PCAT Community Orientation	+	NS	NS
CPCI Community Context	+ / 0	NS / 0	NS

The symbol "+" indicates that previous positive differences between categories remain positive; "NS" indicates that previously non-significant differences remain non-significant. Symbols separated by "/" indicate a change between original and purified subscale results; "–" indicates negative and "0" indicates that no items remained on which to test the purified result.

For the high/low education comparison, the previous difference on PCAT-S First-Contact Access disappears, and no other scores are statistically different. However, it is difficult to conclude that non-significant differences by education are valid on the IPC Communication and PCAS Interpersonal Treatment (respectfulness) subscales, because no items were free from differential functioning. The difficulty differential may be such that non-significant difference in Table 2 may be masking an actual difference for these subscales.

When the urban and rural groups are compared using the purified subscales, rural scores become significantly lower than urban scores for PCAT-S First-Contact Access (likelihood of obtaining same-day needed care from regular provider), but the higher rural score persists in PCAT-S First-Contact Utilization (tendency to contact the regular provider first). The previous difference with respect to PCAS Trust disappears.

Discussion and Conclusion

We found that assessments of primary healthcare attributes were systematically more positive by English- than French-speaking respondents despite an a priori expectation of equivalency. Without analyzing differential functioning, it is difficult to determine whether this difference is due to differences in the Quebec and Nova Scotia healthcare systems or to problems with measurement equivalency of the French and English versions. The answer seems to be both. We found substantial differential item functioning between English and French versions. However, the systematically more positive assessments in Nova Scotia persist even after removing problematic items. The differences in wait times and proportion having a regular physician also support the existence of a real difference.

These results suggest that continued refinement is needed to ensure that French-language versions are equivalent to the original English-language instruments, but that most of the differential functioning is minor and has minimal impact on comparisons at the subscale level. The parametric item response models detected differences as small as 0.4 in discriminability and small differences in difficulty; rarely was discriminability compromised in French and adequate in English. Rather, the differences meant that an item showed good discrimination in one group and slightly better in the other, so that overall, the functioning of the items and scales was acceptable despite differential functioning.

In some cases, our results helped us detect slight shifts in meaning in French translations. For example, the English word "ability" in the PCAS Organizational Accessibility subscale was translated as "facilité," "to get through to the practice by telephone" and as "possibilité," "to talk to the doctor by telephone." The former resulted in differential discriminatory capacity, but not the latter, suggesting that "possibilité" is a more equivalent translation for "ability" in this context than is "facilité." Likewise, the varied translation of "primary care provider" in the PCAT-S instrument may have introduced differential functioning by creating specific, limited terms in the French-language versions while retaining a broad and flexible term in English. In other cases, we could not identify the source of non-equivalence, suggesting differences in cultural interpretation or in interacting with the healthcare system.

We did not detect systematic patterns in difficulty differences that would suggest difference in how response options or scales function in these groups, with the exception of the agree/disagree response scale by French/English. The observed difference suggests that "disagree" may not be equivalent in sense and meaning to "désaccord." In French, "désaccord" seems to be a slightly different concept from, rather than the opposite of, "accord" (agree). It may be

analogous to the finding that "dissatisfied" is not the same construct as "not satisfied" (Eriksen 1995; Coyle 1999). This difference explains the high level of differential functioning in the CPCI, which uses a disagree/agree response format. We recommend that French-language versions of "disagree" response options be reformulated as extremes of agreement such that "pas du tout d'accord" (not at all in agreement) is the equivalent of "strongly disagree."

In this study, we assumed that the original English-language version is the gold standard and that French-language versions must be modified to achieve equivalence. However, results from discussion groups also suggest that some original English statements should be modified to be more valid or precise. For example, English-speaking respondents expressed confusion about the meaning of "primary care provider" (Haggerty et al. 2011), and we believe that the specificity that was required for the French translation resulted in a more precise measure.

It is a tribute to the instrument developers that the instruments and subscales mostly perform equivalently across high- and low-education groups. However, differences in difficulty thresholds, especially in attributes such as respectfulness and interpersonal communication, suggest some measures may systematically underdetect true differences in experience between high- and low-education patients. Higher difficulty thresholds in high-education patients would be consistent with higher expectations among these respondents, a finding that has been repeatedly observed in studies of satisfaction (Crow et al. 2002).

Differential item functioning by urban/rural residence specifically affected the attribute of accessibility. The finding that rural residents are less likely than urban residents to obtain same-day care from their provider when they are sick (PCAT-S First-Contact Access) becomes evident only when differentially functioning items are removed. This is a concern because urban/rural comparisons of accessibility have important implications for health planners' decisions on health services location to optimize equity of access.

The strength of this study is that the same questionnaires were administered to each subject, so that the underlying construct can be directly compared across groups rather than relying on model assumptions of equivalence. However, some differences we found may be spurious owing to multiple testing, and the analytic software we used was highly sensitive to even small differences in difficulty threshold, so that some of the statistically significant differences may not be meaningful. Furthermore, removing problematic items from subscales may compromise construct representation so that comparing subscale scores before and after removal of problematic items is no longer meaningful.

We feel comfortable recommending the use of the French-language versions, while continuing to refine them. At a subscale level, the differential functioning did not introduce bias because the conclusions were largely unchanged when problematic items were removed. We recommend that the original versions be reviewed where translation has posed a problem. We recommend caution in interpreting rural and urban comparisons of access, and urge the development of unbiased measures. We found little evidence of bias by educational status and are confident in recommending these instruments for a broad educational spectrum of patients.

ACKNOWLEDGEMENTS

This research was funded by the Canadian Institute for Health Research. During this study, Jeannie L. Haggerty held a Canada Research Chair in Population Impacts of Healthcare at the Université de Sherbrooke. The authors wish to thank Beverley Lawson for conducting the survey in Nova Scotia and Christine Beaulieu in Quebec, and Donna Riley for support in preparation and editing of the manuscript.

Correspondence may be directed to: Jeannie L. Haggerty, Associate Professor, Department of Family Medicine, McGill University, St. Mary's Research Centre, Hayes Pavilion – Suite 3734, 3830 Lacombe Ave., Montreal QC H3T 1M5; tel.: 514-345-3511 ext. 6332; fax: 514-734-2652; e-mail: jeannie.haggerty@mcgill.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.

Coyle, J. 1999. "Exploring the Meaning of 'Dissatisfaction' with Health Care: The Importance of 'Personal Identity Threat'." *Sociology of Health and Illness* 21: 95–124.

Crow, R., H. Gage, S. Hampson, J. Hart, A. Kimber, L. Storey and H. Thomas. 2002. "The Measurement of Satisfaction with Healthcare: Implications for Practice from a Systematic Review of the Literature." *Health Technology Assessment* 6: 1–244.

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

Eriksen, L.R. 1995. "Patient Satisfaction with Nursing Care: Concept Clarification." *Journal of Nursing Measurement* 3: 59–76.

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

Haggerty, J.L., R. Pineault, M.-D. Beaulieu, Y. Brunelle, J. Gauthier, F. Goulet and J. Rodrigue. 2007. "Room for Improvement: Patients' Experiences of Primary Care in Quebec Before Major Reforms." *Canadian Family Physician* 53: 1056–57.

Haggerty, J.L., C. Beaulieu, B. Lawson, D.A. Santor, M. Fournier and F. Burge. 2011. "What Patients Tell Us about Primary Healthcare Evaluation Instruments: Response Formats, Bad Questions and Missing Pieces." *Healthcare Policy* 7 (Special Issue): 66–78.

Kristjansson, E., R. Aylesworth, I. McDowell and B.D. Zumbo. 2005. "A Comparison of Four Methods for Detecting Differential Item Functioning in Ordered Response Items." *Educational and Psychological Measurement* 65: 935–53.

Reeve, B. 2006. *An Introduction to Modern Measurement Theory.* Rockville, MD: National Cancer Institute. Retrieved May 12, 2011. <http://appliedresearch.cancer.gov/areas/cognitive/immt.pdf>.

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.O. Ramsay. 1998. "Progress in the Technology of Measurement: Applications of Item Response Models." *Psychological Assessment* 10: 345–59.

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

Smith, J.L. and J. Haggerty. 2003. "Literacy in Primary Care Populations: Is It a Problem?" *Canadian Journal of Public Health* 94: 408–12.

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.

Teresi, J.A. and J.A. Fleishman. 2007. "Differential Item Functioning and Health Assessment." *Quality of Life Research* 16(Suppl. 1): 33–42.

Wensing, M., J. Mainz and R. Grol. 2000. "A Standardised Instrument for Patient Evaluations of General Practice Care in Europe." *European Journal of General Practice* 6: 82–87.