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

Detailed Report of published article


**Corresponding Author:**
Jeannie L. Haggerty
Associate Professor
Department of Family Medicine
McGill University

Postal Address:
Centre de recherche de St. Mary
Pavillon Hayes – Bureau 3734
3830, av. Lacombe
Montréal (Québec) H3T 1M5
Canada

Contact:
Tel: (514) 345-3511 ext 6332
Fax: (514) 734-2652
jeannie.haggerty@mcgill.ca
Differential Item Functioning in Primary Healthcare Evaluation Instruments: Detailed Report

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

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 that such differences are the result of item bias in how subgroups use evaluation scales and measures.

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

Method: A sample of 645 adult users balanced by French/English language (in Quebec and Nova Scotia, respectively), urban/rural residency, high/low education and overall healthcare experience responded to six validated instruments: the Primary Care Assessment Survey; the Primary Care Assessment Tool; the Components of Primary Care Index; the EUROPEP, the Interpersonal Processes of Care survey; and part of the Veterans Affairs National Outpatient Community Satisfaction Survey. 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 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 (Nova Scotia) than in the French (Quebec) respondents. However, we found evidence of differential item functioning in 73 (48%) of the 153 items, with the English version more discriminating generally than the French. Removing problem 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 problem 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 geography and educational status, but results suggest potential differences in the underlying construct in rural and low-education respondents.
Examining group differences in healthcare experience, whether across geographic locations or linguistic/ethnic groups, is essential to ensuring 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, it may adversely affect the conclusions of between-group comparisons by creating a false difference or failing to detect a true difference in the underlying construct.

We compared validated instruments that evaluate primary healthcare from the consumer perspective and that are thought to be most pertinent for the Canadian context. Among other objectives, we wanted to determine whether French and English versions of the instruments were equivalent and whether item or subscale performance differed systematically by urban/rural location or by high/low educational status.

Specific research questions
All the instruments used in our study were originally developed in English. In translation to French, some phrases proved problematic. For instance, for rating response options, one instrument – 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 questionnaire used the term “primary care provider” to refer to both person and place, for which there is no single French equivalent. Consequently, it was translated variously, depending on context, 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), and we hypothesized that measures of accessibility function differently by context. While we had no a priori concerns regarding educational achievement, we wanted to ensure that all instruments performed equally well in low-literacy groups, since we found considerable variation in readability between instruments.

Overview of differential functioning
The terms related to differential item functioning reflect the origins of this analysis 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 a question at different levels of student ability.

For example, if the probabilities of correctly answering a question 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 method of evaluating item performance has also been used to evaluate 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 $\geq 1$ (the “$a$” parameter) considered appropriate. 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 low-education respondents. Differences in item discrimination indicate that the question understood or interpreted differently by each subgroup. This would occur, for instance, when the French translation is not equivalent in meaning to the original English version.

Difficulty in attitudinal surveys refers to the probability of endorsing a specific response option for a given level of the construct being measured. For instance, in the PCAS Trust subscale that elicits agreement with statements using a five-point Likert scale of “1 = strongly disagree” to “5 = strongly agree,” if respondents are most likely to endorse the response “4 = agree” only when they have a very high overall score for trust, then the item is said to be “difficult.” If the probability of endorsing “4 = agree” is high even at low overall scores of trust, then the item is “easy.” When an item’s difficulty threshold varies by group membership, the item is said to exhibit differential difficulty functioning. Figure 2 illustrates the differential difficulty threshold between high- and low-education respondents, again for an item in the PCAS Trust subscale. 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., more difficulty) than will high-education respondents. Note that the difficulty differential is uniform across all levels of trust, whereas in Figure 1, where there is 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.

Comparisons between subgroups using an instrument that contains differentially functioning items are potentially biased. The potential impact of differential item functioning is assessed by removing problem items from the subscale or instrument,
recalculating the scores using only the purified scale (non-problem 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 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.

Method
Among instruments developed to measure the quality of primary healthcare from the consumer perspective, we identified six in the public domain that appeared of greatest relevance for Canada: the Primary Care Assessment Survey (Safran et al., 1998); the Primary Care Assessment Tool (PCAT) (Shi, Starfield, & Xu, 2001); the Components of Primary Care Index (CPCI) (Flocke, 1997); the EUROPEP (Wensing, Mainz, & Grol, 2000); the Interpersonal Processes of Care survey (IPC) (Stewart, Napoles-Springer, Gregorich, & Santoyo-Olsson, 2007); and the Veterans Affairs National Outpatient Community Satisfaction Survey (VANOCSS) (Borowsky et al., 2002). We only retained subscales that measured attributes in at least two instruments under study. We obtained permission to use the instruments from all the instrument developers.

All the instruments were originally developed in English, but two had been subsequently validated in French: the EUROPEP and the PCAT. For our study, we translated remaining instruments French, then back-translated them into English. Discrepancies between the two English versions guided modifications to the French translation. The reading level of the back-translated English version was higher than the original, so we adapted the French version for a low-literacy audience. To achieve equivalency, the bilingual researchers made final adjustments that were validated in group discussions with bilingual volunteers.

Study population
The target population for this study were adult, Canadian, primary healthcare users, undifferentiated by age or health condition. Eligible subjects were aged ≥ 18 years with a regular source of healthcare that they had consulted in the previous 12 months. To maximize the statistical efficiency to examine our main contrasts of interest, we used a sampling design that balanced the study population by French/English language, urban/rural location and educational level. 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 questionnaires were administered in Nova Scotia and French questionnaires in Quebec.

Urban location was defined as census metropolitan areas and rural, as areas more than one hour’s travel from a metropolitan area, including remote areas in Quebec, more than four hours’ travel from the nearest census metropolitan area. We used an age-sensitive
Differential Item Functioning in Primary Healthcare Evaluation Instruments: Detailed Report

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: completed only high school and were under 45 years old; completed 10 years of school and were 45 to 55 years old; or completed less than eight years and were over 55 years old (Smith & Haggerty, 2003).

Analysis
We examined the distribution of missing values by language, geography and educational achievement. 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. We compared subscale scores by language, geography and education using regression modelling controlling for the other design variables as well as for overall experience, using $\alpha = 0.05$, 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, geography and education.

There are different ways of assessing differential item functioning, but they all 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 & Ramsay, 1998; Kristjansson, Aylesworth, McDowell, & Zumbo, 2005; Reeve, 2006; Teresi & 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 problem items from the latent variable and repeated the process until we found no differentially functioning items. We used a critical value of $\alpha = 0.01$ to indicate statistical significance because lower values detected trivial differences.

It is important to note that the parametric item response model is sensitive to small differences (i.e. conservative) because item behavior may not necessarily correspond to the parametric assumptions, and the likelihood of finding a significant result will increase with large samples such as ours. However, this method did allow us to identify a subset of items entirely free of differential item functioning or item bias, which we needed to rule out before interpreting group 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 overall experience of care, French/English language, urban/rural location and high/low education, the 645 respondents were not equally distributed in
Differential Item Functioning in Primary Healthcare Evaluation Instruments: Detailed Report

these categories. Only 149 (23%) indicated their overall experience as poor, compared to 232 (36%) and 264 (41%) as average and excellent, respectively. The French sample (Quebec, n = 302) was more balanced by geography and education than was the English sample (Nova Scotia, n = 343). The English group was more urban (59% vs. 49%, \( \chi^2 = 6.7, p < 0.096 \)), more likely to have a high school reading level (75% vs. 55%, \( \chi^2 = 27.5, p < 0.001 \)) and also more likely to perceive their health as good or excellent, to be affiliated to a family doctor rather than a clinic and to have longer affiliations (Table 1). Wait times for appointments were better among English, urban and high-education respondents.

Descriptive results

The number of missing values was not systematically higher by language or educational achievement. Missing frequencies differed by geographic location for only five items; in every case, they occurred more in rural than in urban populations, but they were not consistently in one instrument or attribute. English respondents showed a higher tendency than the French to select the “don’t know / not applicable” option offered by the PCAT and EUROPEP instruments.

The normalized subscale scores, grouped by primary care attribute, are compared by language, education and geography in Table 2. 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. These differences remained when we controlled for other design variables and health status. Only one subscale differs by education, the PCAT First-contact Access, with less positive assessments in low- than in high-education groups. We found urban/rural differences for two subscales, with rural respondents indicating more positive assessments than urban respondents in PCAT First-contact Utilization and less positive assessments in PCAS Trust.

The vast majority of subscales had similar factor resolution by subgroup. Three subscales found two factors with 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) subscale had two in rural; and the PCAS Trust (interpersonal communication) subscale had two in French and two in low-education respondents.

Differential functioning

Because of space restrictions, we report only summary results at a subscale level; item-specific results are available upon request. 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 29 subscales, only five (17%) are free from French/English differential functioning, compared to 12 (41%) in education and 18 (62%) in geography.
Differential Item Functioning in Primary Healthcare Evaluation Instruments: Detailed Report

Differential item functioning affected more than 50% of items in 12 (41%) subscales by language, six (21%) by education and none by geography. Of the items with differential French/English functioning, 41% (30/73) showed differences in discriminatory capacity; of which 13 out of 30 had discriminability differentials greater than 1 (Figure 1 demonstrates a discriminability differential of 1). English items had significantly higher discriminability values than their French counterparts, except for management continuity attributes, where French items were more discriminating. Only four items discriminated adequately in English and poorly in French; all were in the CPCI. For instance, 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.

Of the 43 items with differential difficulty, 11 had differentials over 0.5, which is approximately the magnitude illustrated in Figure 2. The pattern of differences do not support a systematic difference between English and French when “poor” is translated as “médiocre” vs. “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 higher than for “strongly disagree” across several subscales and two instruments (CPCI and PCAS Trust). It takes a higher level of the construct, for example trust, for French respondents to endorse the next most positive category after “strongly disagree”; we found no systematic direction of difficulty differences for “strongly agree.”

By education, 42% (18/42) of differentially functioning items were due to differential discriminability, with seven being differentials > 1. Differentially discriminating items tended to have higher discrimination values in high-education respondents, although the reverse was seen for respectfulness. Only 12 of the 24 items had difficulty differentials > 0.5. Items tended to have higher difficulty thresholds in the high-education than in the low-education groups, such that high-education respondents only endorsed more positive response options at higher levels of the construct (e.g. communication). One of the largest observed difficulty differentials was in the PCAT Interpersonal Treatment (respectfulness) subscale, where all items had differential functioning, with an average difficulty differential of 0.8 for the whole scale and with low-education respondents endorsing all options at a lower level of respectfulness.

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 design variable after we removed items with differential functioning. Valid comparison by language was impossible for subscales with no remaining non-problem items: PCAS Visit-based Continuity, PCAT Community Orientation and CPCI Community Context. Comparisons based on less than 50% of the original items must be interpreted cautiously. However, the results show that, for language, the conclusions are essentially unchanged from those of Table 2: assessments
Differential Item Functioning in Primary Healthcare Evaluation Instruments: Detailed Report

for all attributes are more positive in English than in French. The previous more positive French scores on CPCI Coordination of Care disappear in the purified subscale.

For the high/low education comparison, the previous difference on PCAT First-contact Access is no longer evident and no other scores are statistically different between these groups, despite differential functioning on interpersonal communication and respectfulness. On the IPC Communication and PCAS Interpersonal Treatment (respectfulness) subscales, no items were free from differential functioning, so it is difficult to conclude that non-significant differences by education are valid. The items are more discriminating and difficult in the high-education than in the low-education groups; specifically, high-education respondents only endorsed the more positive response options at higher levels of the construct (communication, respectfulness). Thus low-education respondents would have a higher probability of responding positively at lower levels of the construct of interest, meaning the 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, the previous difference with respect to PCAS Trust disappears, but the higher rural score persists in PCAT First-contact Utilization (tendency to contact first the regular provider), and rural scores become significantly lower than urban scores for PCAT First-contact Access (likelihood of obtaining same-day needed care from regular provider).

Discussion
We found that assessments of primary healthcare attributes were systematically more positive by English than French respondents despite an a priori expectation of equivalency. Without analyzing differential functioning, it is difficult to determine whether this 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 problem items.

The general similarity of conclusions when using the original vs. the purified subscales by language suggests that the French versions of the instruments are largely equivalent to the original English versions when used as complete subscales. The minimal impact of differential item functioning at the subscale level is likely due to the sensitivity of our testing. Parametric item response models detected differences as small as 0.4 in discriminability (a parameter or slope) and small differences in difficulty. In only five items was discriminability compromised in one group while remaining adequate in another. 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.

Nonetheless, our analysis suggests further work is needed to achieve French/English equivalence for many items. The tendency toward lower discriminability of French translations of English items means that item statements need to be examined carefully.
In some cases, our results helped us detect slight shifts in meaning in French translations. For instance, the English word “ability” in the PCAS Organizational Accessibility subscale was translated as “facilité” with “to get through to the practice by telephone” and as “possibilité” with “to talk to the doctor by telephone.” The former resulted in differential discriminatory capacity, but not the latter, suggesting that “possibilité” is more equivalent to “ability” in this context than is “facilité.” Likewise, the varied translation of “primary care provider” in the PCAT instrument may have introduced differential functioning by creating specific, limited terms in the French versions while retaining a broad and flexible term in English. In other cases, we could not identify the source of non-equivalence, suggesting that the differences may be in cultural interpretation or in the way of interacting with the healthcare system. In sum, while evaluators can be relatively secure about the absence of bias in French/English comparisons of subscale measures, they need to exercise caution in selecting individual items as indicators or in making item banks for computer adaptive testing.

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 “strongly disagree” is a more negative notion than “fortement en disaccord.” ‘Disagree’ may not be equivalent in sense and meaning to “désaccord.” Whereas in English, “disagree” is the negative expression of “agree,” in French “désaccord” may be a slightly different concept from “accord.” The more appropriate equivalent may be “pas du tout d’accord” (not at all in agreement) or even “complètement en désaccord” (completely in disagreement). Other studies recommend caution in establishing extremes of response scales. For instance, some authors have found that “dissatisfied” is not the same as “not satisfied” (Eriksen, 1995; Coyle, 1999). Systematic differences in how agreement response options are interpreted may explain the high level of differential functioning by language in the CPCI, which uses a disagree/agree response format for all items.

In this study, we assumed that the original English version is the gold standard and that French versions must be modified to achieve equivalence. However, the qualitative component of this study also suggests some original English statements should be modified to be more valid or precise. For instance, both English and French respondents preferred other rating or reporting scales to the disagree/agree response options, and English respondents expressed confusion about the meaning of “primary care provider” (Haggerty et al., 2009).

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 under-detect 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 was less present by urban/rural residence but specifically affected the attribute of accessibility. Rural residents are more likely than urban residents to consult the regular provider for all types of care (PCAT First-contact Utilization), but when they are sick they are less likely to obtain same-day care from their provider (PCAT First-contact Access), a difference that 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. In another study, we demonstrated that care-seeking patterns differ by geographic areas (Haggerty, Robarge, Lévesque, Geoffrion, & Loignon, 2009). Thus, the studied instruments may not be addressing barriers that are specific to rural areas. It is important to ensure that instruments used to profile accessibility in rural areas be free from differential item functioning by geographic context.

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 the assumptions of the model. The sample’s balance by language, geography and education increased our statistical efficiency for detecting differences in these groups. However, some differences we found may by spurious, due to multiple testing, and our small sample size did not provide sufficient power for us to split the sample and validate the results in the split halves. Furthermore, the analytic software we used was highly sensitive to even small differences in difficulty threshold, so some of the statistically significant differences may not be meaningful. The latent variables against which item performance was evaluated sometimes consisted of very few items, so the representation of the construct may not be very robust. Pooling of items across instruments into common constructs could provide a more reliable result of item performance. Finally, although using parametric item response models permits statistical testing of differences, it also imposes assumptions (e.g. sigmoid curves) that may mask nuances in item or response scale behaviour that provide the key to understanding the source of the differential functioning. Ideally, all differences should also be modelled non-parametrically for a more fine-grained analysis.

**Conclusion**

We found evidence of substantial differential item functioning between French and English versions of instruments that evaluate primary healthcare from the consumer perspective. However, at a subscale level, this differential functioning did not introduce bias because the conclusions were largely unchanged when problem items were removed. Overall, in this non-representative sample, we found that assessments of care were more positive in Nova Scotia (using English instruments) than they were in Quebec (using French versions). For geographic area, we found evidence to suggest measures of first-contact accessibility may be biased. We found little evidence of bias by educational status, but we observed that items tended to be more discriminating and difficult in high-education groups, suggesting that instruments may not be addressing some dimensions of experience that are important to low-education patients.
Differential Item Functioning in Primary Healthcare Evaluation Instruments: Detailed Report

Reference List


Differential Item Functioning in Primary Healthcare Evaluation Instruments: Detailed Report


