



Scoping review of response shift methods: current reporting practices and recommendations

Tolulope T. Sajobi¹ · Ronak Brahmatt² · Lisa M. Lix³ · Bruno D. Zumbo⁴ · Richard Sawatzky^{2,5}

Accepted: 26 November 2017 / Published online: 5 December 2017
© Springer International Publishing AG, part of Springer Nature 2017

Abstract

Background Response shift (RS) has been defined as a change in the meaning of an individual's self-evaluation of his/her health status and quality of life. Several statistical model- and design-based methods have been developed to test for RS in longitudinal data. We reviewed the uptake of these methods in patient-reported outcomes (PRO) literature.

Methods CINHAHL, EMBASE, Medline, ProQuest, PsycINFO, and Web of Science were searched to identify English-language articles about RS published until 2016. Data on year and country of publication, PRO measure adopted, RS detection method, type of RS detected, and testing of underlying model assumptions were extracted from the included articles.

Results Of the 1032 articles identified, 101 (9.8%) articles were included in the study. While 54.5% of the articles reported on the *Then-test*, 30.7% of the articles reported on Oort's or Schmitt's structural equation modeling (SEM) procedure. Newer RS detection methods, such as relative importance analysis and random forest regression, have been used less frequently. Less than 25% reported on testing the assumptions underlying the adopted RS detection method(s).

Conclusions Despite rapid methodological advancements in RS research, this review highlights the need for further research about RS detection methods for complex longitudinal data and standardized reporting guidelines.

Keywords Response shift · Systematic review · Patient-reported outcomes

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11136-017-1751-x>) contains supplementary material, which is available to authorized users.

✉ Tolulope T. Sajobi
tolu.sajobi@ucalgary.ca

¹ Department of Community Health Sciences & O'Brien Institute for Public Health, Cumming School of Medicine, University of Calgary, 3280 Hospital Drive NW, Calgary, AB T2N 4Z6, Canada

² School of Nursing, Trinity Western University, Langley, Canada

³ Department of Community Health Sciences, University of Manitoba, Winnipeg, Canada

⁴ Department of Educational and Counselling Psychology, and Special Education, University of British Columbia, Vancouver, Canada

⁵ Centre for Health Evaluation and Outcome Sciences, Providence Health Care, Vancouver, Canada

Introduction

Patient-reported outcome (PRO) measures are increasingly used in clinical trials, population-based surveys, and cohort studies to understand the impact of disease burden and treatment on the disease course [1–3]. While the assessment of longitudinal change in PROs rests on the assumption that respondents' interpretations of a construct remain constant over time, this may not always be the case. Studies have shown that the meaning of patients' self-evaluations of their health status and quality of life may not be the same across different points in time, a phenomenon known as response shift (RS).

Although the concept of RS originated in management sciences [4, 5], Breetvelt and van Dam [6] are among the first researchers to discuss RS in health-related research. They attempted to explain the “paradox” they observed in health-related quality of life (HRQOL) assessment in cancer patients where studies have suggested no differences in HRQOL of cancer patients and healthy controls despite cancer patients' frequent reporting of high level of physical and/or psychological complaints. They concluded that

patient-reported HRQOL and psychological distress should be carefully interpreted. Sprangers and Schwartz [7] later proposed that the concept of RS could be used to explain some of the seemingly paradoxical or counterintuitive findings in the assessment of HRQOL in patients with terminal illnesses. Building on Golembiewski's [4] topology of longitudinal change in management science research, they defined RS in the context of health research as "a change in the meaning of one's evaluation of a construct as a result of a change in one's internal standards of measurement (recalibration), a change in one's values (reprioritization), or a change in one's definition of the construct (reconceptualization)" [7]. Sprangers and Schwartz further describe a theoretical model where RS occurs in response to a catalyst, which might be a significant health event such as a diagnosis, treatment, or surgical procedure.

There have been significant theoretical and methodological advances since the initial theoretical model of RS in HRQOL research [8–11]. Of note, is the advancement in our theoretical understanding of adaptation and appraisal processes associated with RS. Rapkin and Schwartz [11] showed that RS involves a change in processes by which individuals interpret and respond to questions about their health status or quality of life. This has further broadened our knowledge of RS processes and its consequences on PRO measurement [10–13]. A commonly reported consequence of RS is over-estimation or under-estimation of longitudinal change in PRO scores, which conventional statistical analyses may not reveal in longitudinal PRO studies. RS can also be viewed as an operationalization of psychological adaptation to chronic diseases over time [11–13]. In other words, RS can be regarded as a desirable change in values and experiences, triggered by a significant health event that consequently influences the appraisal processes that drive an individual's response and interpretation of questions about their health and well-being. For example, in a study of 53 patients treated for full-thickness cartilage defects [14], patients thought they felt worse before surgery in retrospect than they did at the time (recalibration RS). Similarly, Sajobi, Fiest, and Wiebe [15] investigated reprioritization RS in patient-reported HRQOL in a randomized controlled trial investigating the effects of surgical treatment in patients with temporal epilepsy. They found that patients who received surgical treatment were likely to de-emphasize being free of seizures while placing increased emphasis on social functioning aspects of their HRQOL 1 year after surgery, which is a clinically desired treatment goal for persons with epilepsy.

In addition to theoretical advancements, there have been developments in statistical methods to detect and adjust for RS in longitudinal data [13, 16]. Early research about RS detection primarily relied on design-based approaches, such as the *Then-test* method [17]. However, this method

is sensitive to recall bias and cannot be used in secondary analyses of longitudinal data. Statistical approaches that involve exclusive reliance on modeling methods have been developed to test a priori RS hypotheses in secondary analyses of longitudinal data. These include Schmidt's structural equation model (SEM) method [18], Oort's four-step (SEM) method [19, 20], item response theory (IRT) models/methods [21, 22], mixed-effects regression models [23], latent trajectory analysis of residuals method [24], relative importance analysis method [25], classification and regression tree (CART) models [26], and random forest regression models [27] (see Table 1 for a summary of these methods). Schmitt's technique [18], which uses covariance structure analysis with SEM, operationalizes RS as change in SEM parameters between two measurement occasions using Golembiewski [4]'s typology of change. Thus, this methodology can only detect reconceptualization and uniform recalibration RS. Oort [19] later proposed a 4-step approach for sequentially testing different forms of RS by comparing models in which the different measurement model parameters are constrained to be equal over time. Although Oort's SEM based on maximum likelihood estimation is commonly applied to detect RS using domain-level data, this methodology has been extended to detect RS in ordinal item-level data using robust maximum likelihood [28] and diagonally weighted least squares estimation approaches [29, 30]. One advantage of the Oort's procedure is its versatility in detecting all three forms of RS. More recently, IRT models, which also use changes in measurement parameters as an operationalization of RS, have been proposed to detect recalibration and reprioritization RS at the item level. Anota [21] proposed a Linear Logistic with Relaxed Assumptions (LLRA) model to detect recalibration RS in retrospective assessment data. More recently, RespOnse Shift ALgorithm in Item response theory (ROSALI), which uses longitudinal partial credit and graded partial credit IRT as measurement models, was proposed to test for recalibration and reprioritization RS [22]. ROSALI relies on an operationalization of non-uniform and uniform recalibration, and reprioritization as change(s) in the values of discrimination parameters in polytomous IRT models. However, ROSALI is currently limited to uni-dimensional IRT models and can therefore not be used to detect reconceptualization RS. Although the SEM and IRT approaches for identifying RS are different, both methods operationalize RS based on tests of changes in latent variable measurement model parameters and can also be used to detect RS at the item level.

In addition to the above latent variable models for RS detection, trajectory analysis approaches based on mixed-effects regression models have been developed to test for RS by modeling of longitudinal change in PROs while accounting for within-subject variation via random effects [23, 24]. However, these regression-based methods are sensitive to

Table 1 A overview of response shift detection methods

Response shift method	Description	Advantages	Limitations
<i>Then-test</i> method	This method relies on a pre-post study design to test for response shift (RS). <i>Then-test</i> involves including one or more retrospective pre-test measures to detect recalibration response shift	<ul style="list-style-type: none"> • Can be used to estimate recalibration RS effect size • Can be used to test multiple RS hypotheses while controlling for overall Type I error 	<ul style="list-style-type: none"> • <i>Then-test</i> is sensitive to recall bias • It cannot be used to detect reprioritization or reconceptualization • It cannot be used to test RS in secondary analyses of longitudinal data without any retrospective pre-test design
Statistical methods			
Structural equation model (SEM)	<p>SEM is a multivariate procedure that combines confirmatory factor analysis and techniques for assessing equivalence (i.e., invariance) of latent variable measurement model parameters to test RS hypotheses. There are two dominant SEM approaches for detecting RS:</p> <ul style="list-style-type: none"> • Schmitt's SEM approach uses covariance structure analysis with SEM to identify both reconceptualization and recalibration RS • Oort's SEM approach involves 4-steps for sequentially testing different forms of RS by comparing models in which the different measurement model parameters are constrained to be equal over time 	<ul style="list-style-type: none"> • Well-developed theory and method that is easily implementable in most statistical software packages • Oort's SEM approach can be used to test hypotheses about the three forms of RS • SEM can be used to estimate RS effect sizes • SEM can be used in conjunction with multiple testing methods to control overall familywise Type I error associated with test of multiple RS hypotheses • SEM can be used to detect RS using item-level data based on robust maximum likelihood estimators and diagonally weighted least squares estimators 	<ul style="list-style-type: none"> • SEM approaches tests RS hypotheses at the group level • SEM may require large sample sizes with increasing number of domains or items • Schmidt SEM can only detect uniform recalibration and reconceptualization RS but not reprioritization RS
Item response theory (IRT)	<p>The use of IRT to examine RS involves testing changes in the relationships between discrete items and corresponding latent factors in terms of a logistic link function. The two main IRT approaches to RS detection are as follows:</p> <ul style="list-style-type: none"> • Anota's IRT model is based on linear logistic model with relaxed assumption (LLRA), a IRT model for measuring change across two or more measurement occasions and can detect recalibration RS • The Response Shift Algorithm in IRT (ROSALI) method uses longitudinal partial credit and graded partial credit IRT models as measurement models to test for recalibration and reprioritization RS. As in the SEM approach, tests of specific measurement model-parameter constraints can be conducted to detect corresponding forms of response shift 	<ul style="list-style-type: none"> • IRT can be used to detect RS at the item level • IRT does not rely on the assumption of normality because it uses a measurement model adapted for categorical items • It can be used to estimate RS effect sizes • The LLRA requires neither items' unidimensionality nor multidimensionality to estimate calibration RS in the data • IRT can be used in conjunction with multiple testing methods to control overall familywise Type I error associated with tests of multiple RS hypotheses 	<ul style="list-style-type: none"> • IRT requires large sample sizes as the number of items increases in order to aid accurate estimation of item-level parameters • The existing IRT approaches for RS detection cannot yet detect reconceptualization RS but would require a multidimensional IRT to do so • Anota LLRA model was developed using retrospective test-pretest data but has not been investigate in secondary longitudinal data

Table 1 (continued)

Response shift method	Description	Advantages	Limitations
Group-based trajectory modeling of mixed-effects regression residuals	This method uses subject-specific residuals, the difference between predicted and observed HRQOL scores over time, obtained from mixed-effects regression used to classify patients into discrete RS groups	<ul style="list-style-type: none"> It identifies subsamples with look-alike patterns of RS This methodology can accommodate incomplete data under the assumption of missing at random 	<ul style="list-style-type: none"> This method cannot be used to test hypotheses about re-prioritization RS It is sensitive to model misspecification This methodology relies on the assumptions of multivariate normality
Mixed-effects regression models with interactions	This methodology models longitudinal global HRQOL scores as a function of HRQOL domain scores. Significant tests of interaction effects are used as evidence of RS	<ul style="list-style-type: none"> Validity has been evaluated using simulation techniques Subject-specific variation in the interaction can be modeled Mixed-effects regression can accommodate incomplete longitudinal data 	<ul style="list-style-type: none"> This method can only be applied to data on an instrument with both global and domain-specific HRQOL scores This methodology relies on the assumption of multivariate normal distribution
Relative importance measures	This is a group-based method that examines changes in relative importance of HRQOL domains over time. Measures of relative importance derived from logistic regression or discriminant analysis are used to rank order the HRQOL domains according to their importance in discriminating between groups. Hypothesis about re-prioritization RS is tested using either of the following approaches: <ul style="list-style-type: none"> changes in relative importance weights associated with each domain obtained from discriminant function analysis or logistic regression over time changes in rank ordering of the HRQOL domains over two measurement occasions which finds the linear combination Changes in relative importance weights or ranks are used as evidence of RS 	<ul style="list-style-type: none"> This methodology is a group-based approach to detect re-prioritization response shift This methodology can accommodate the use of multiple testing procedures to control for overall Type I error Relative importance analysis based on logistic regression does not rely on the assumption of normality 	<ul style="list-style-type: none"> It requires two or more independent groups It is sensitive to strong domain correlations and susceptible to collinearity Relative importance analysis based on discriminant function analysis is sensitive to non-normality Cannot be used to estimate RS effect size
Classification and regression tree (CART) method	CART is a non-parametric regression method based on recursive partition of the data into increasingly homogeneous subgroups	<ul style="list-style-type: none"> CART is a non-parametric method and does not rely on any distributional assumption about the data It can be used to detect the three forms of RS 	<ul style="list-style-type: none"> CART is not useful for inferential testing of RS hypotheses Cannot be used to estimate RS effect size

Table 1 (continued)

Response shift method	Description	Advantages	Limitations
Random forest regression	This is an ensemble learning method derived by repeatedly sampling from the original data with replacement to generate a multitude of CART. The importance of a domain is quantified using the average variable importance (AVI) which is estimated as a mean relative decrease in the trees' performance when the observations of this variable are randomly permuted. The presence of reprioritization response shift among the domains can be examined graphically by plotting the domain-specific AVIs against time. Reprioritization response shift is considered to be present among the domains if the AVI curves for two or more domains intersect	<ul style="list-style-type: none"> • This non-parametric method does not rely on the any distributional assumption • It can be used to assess reprioritization RS across two or more measurement occasions 	<ul style="list-style-type: none"> • RF is prone to over-fitting • Cannot be used to estimate RS effect size

AVI average variable importance, HRQOL health-related quality of life, IRT item response theory, RF random forest, SEM structural equation modeling

model misspecification, especially when data are not collected on important explanatory variables. Another approach involves the use of relative importance analysis methods [25], including discriminant analysis and/or logistic regression, to estimate the relative importance of HRQOL domains with respect to their ability to discriminate between group memberships (e.g., active versus inactive disease) at each measurement occasion. Reprioritization RS is operationalized as change in the relative importance of the HRQOL domains over time. CART relies on a non-parametric recursive partitioning of the data into increasing homogenous subgroups using relevant explanatory variables [26]. The occurrence of each form of RS can be hypothesized by detecting differences in the way the explanatory variables account for differences in HRQOL scores within subgroups. Random forest regression, a variant of the CART method, was developed to test for reprioritization RS over multiple occasions [27]. Other qualitative methods for RS detection have been proposed in the literature [31–33], but their discussion is beyond the scope of this article.

Statistical RS detection methods may not always result in the same conclusions about the presence of RS in a set of data, and may not be equally sensitive to detect different types of RS effects. This may be due, in part, to the different operational definitions of RS that underlie the methods and the fact that they are not equally sensitive to detect [different types of] RS effects under different data analytic conditions. In addition, the different methods do not all rely on the same derivational assumptions. For example, latent variable methods focus on the detection of RS in terms of a change in measurement model parameters, whereas several of the regression-based and relative importance methods focus on manifestations of RS in terms of group differences. The non-parametric methods focus on describing RS without testing the statistical significance of observed changes. While a number of studies have attempted to demonstrate the convergent validity of two or more RS detection methods using empirical datasets, the findings from these studies have been conflicting [30, 34–37]. For example, Ahmed et al. [34] compared the SEM and *Then-test* methods for testing for RS in stroke patients and concluded that these methods result in different conclusions about the presence of RS. In contrast, Visser, Oort, and Sprangers [35] confirmed the convergent validity of then *Then-test* and SEM in detecting RS in cancer patients. Lix and colleagues [25, 36, 37] also investigated the presence of RS in health-related quality of life of a population-based cohort of individuals with inflammatory bowel disease using relative importance analysis, SEM, and random effects regression. While all the methods confirmed the presence of RS, the type and magnitude of RS effects detected varied depending on which method was used.

Despite the availability of a range of statistical-based methods for RS detection, there has been limited

investigation regarding the uptake of these methods [38]. Schwartz, Bode, Repucci, Becker, Sprangers, and Fayers [12] conduct a meta-analysis of 26 articles on response shift studies focusing specifically on the magnitude and clinical significance of RS and showed that RS effects are typically small. This review focused on studies that adopted SEM and/or *Then-test* methods and did not report on other statistical methods. We addressed this limitation by conducting a scoping review to describe the uptake of statistical methods for RS in longitudinal studies of PROs.

Methods

Data sources and search strategy

With the help of a health information scientist/librarian, we searched seven library databases (Biomed, CINAHL, EMBASE, Medline, ProQuest, PsycINFO, and Web of Science) for the use of the term “*response shift*,” “*response shift effects*,” “*longitudinal measurement invariance*,” or “*retrospective bias*” as keywords in the title, abstract, or meta-data of the article. The searches were limited to documents written in English and published or accepted for publication on or before December 31, 2016.

Article selection

Relevant articles were those that (1) reported on original research focused on testing for RS in a health-related study, (2) used the *Then-test* or a statistical method to test for RS in empirical data, and (3) used a PRO measure. We excluded studies that had no clear definition of RS, non-systematic or narrative reviews, commentaries, editorials, letters, case-control studies, case reports, and simulation studies. These inclusion and exclusion criteria were applied by first screening the titles and abstract and subsequently reviewing the full text of potentially relevant articles. The citations were randomly divided into 3 groups, which were each independently reviewed by 2 of the 3 investigators (TTS, RB, RS) to decide whether the full-text article should be reviewed. A weighted kappa statistic was calculated to assess agreement between the reviewers. Full-text articles were similarly divided to determine the final selection of articles. In cases where there were disagreements, the disagreements were resolved through discussion and consensus. See the Online Appendix for the list of all articles included in this review.

Synthesis of review results and quality assessment

We extracted the following information from each study and subsequently summarized the information across studies using descriptive techniques (i.e., frequencies and

percentages). First, in order to understand differences in uptake of RS detection methods across regions, patient populations, and PRO instruments, we collected information on the country of affiliation for the first author, patient population, type of PRO measure, and disease type. Second, we evaluated current reporting practices by extracting information about the type of RS detection method(s), type of RS tested and/or detected, and reporting RS effect sizes detected. Third, we extracted information about the type of statistical assumptions evaluated for the *Then-test* and other statistical methods. Multiple tests of hypotheses are prone to arise when testing RS across multiple scale domains or instruments. We therefore specifically examined the use and reporting of methods to control Type I error rates in articles that adopted the *Then-test*, SEM, IRT, or relative importance analysis methods. Strong between-domain correlations in PRO instruments are likely to lead to variable multicollinearity, which can result in inflated standard errors of regression coefficients. We therefore also examined the adoption and reporting on procedures to address multicollinearity. For methods that assume either a univariate or multivariate normal distribution, such as conventional SEM based on maximum likelihood estimators, and relative importance analysis based on discriminant analysis, we examined the articles for evidence of descriptive or inferential analyses of distributional assumptions or robust inference methods to address departures from normality. We also examined the adoption and reporting of information about missing data (both type and patterns) and methods to accommodate missing data.

Results

Descriptive characteristics of included articles

A total of 1032 articles were initially identified, while 101 (9.8%) articles satisfied our inclusion criteria and were included in this review (Fig. 1). The three reviewers had a high rate of agreement in the screening of articles; the estimated weighted kappa was 96.0%. Figure 2 describes the frequency distribution of the number of articles that adopted design-based and statistical methods by year of publication. An overall increasing trend in the number of publications over the years was observed for statistical methods for RS detection, with 61 (60.4%) articles published between 2005 and 2016. More importantly, between 2012 and 2016, we see that statistical-based RS detection methods were increasingly being adopted. But the uptake of *Then-test* methods appears to be stable over time. Figure 3 describes the most frequent characteristics of the included articles. When we summarized the articles by country of the lead author, 22 (21.8%) were published by researchers in the Netherlands, 21 (20.8%) by researchers in Canada, and 21 (20.8%) articles

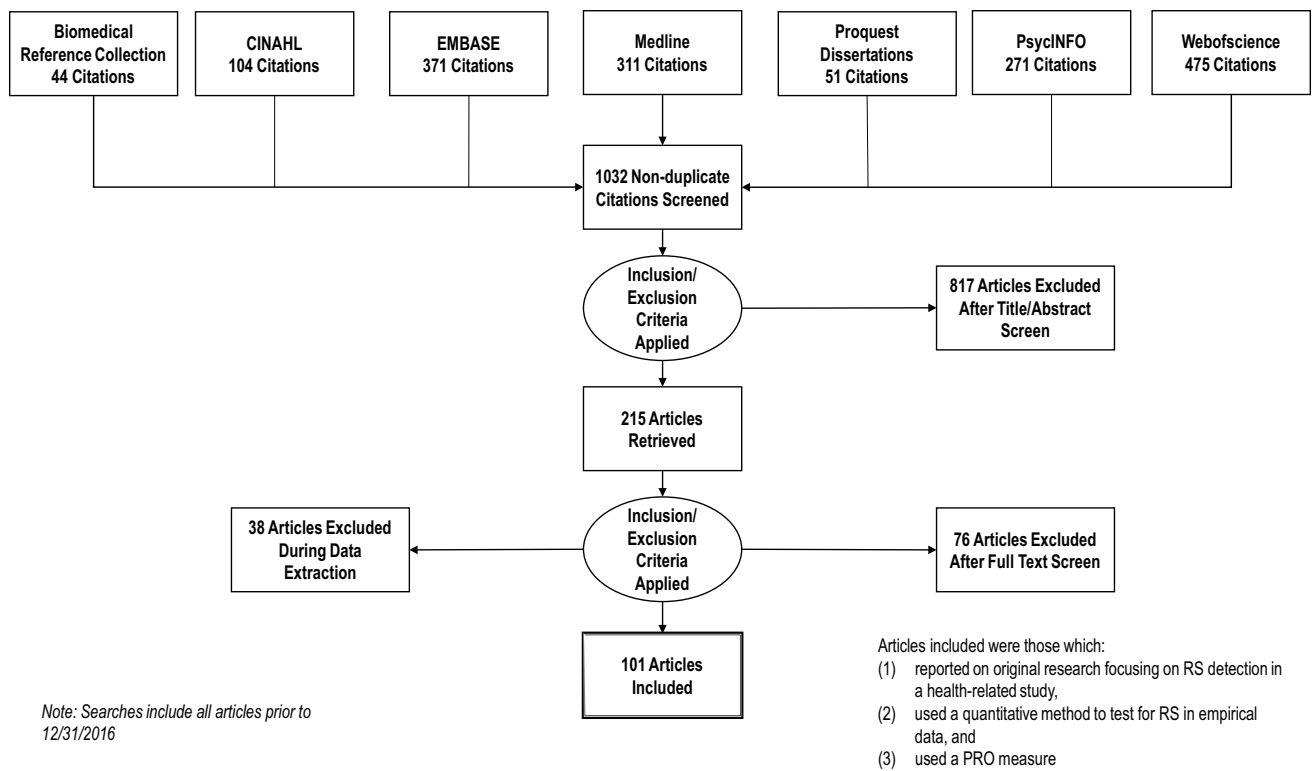
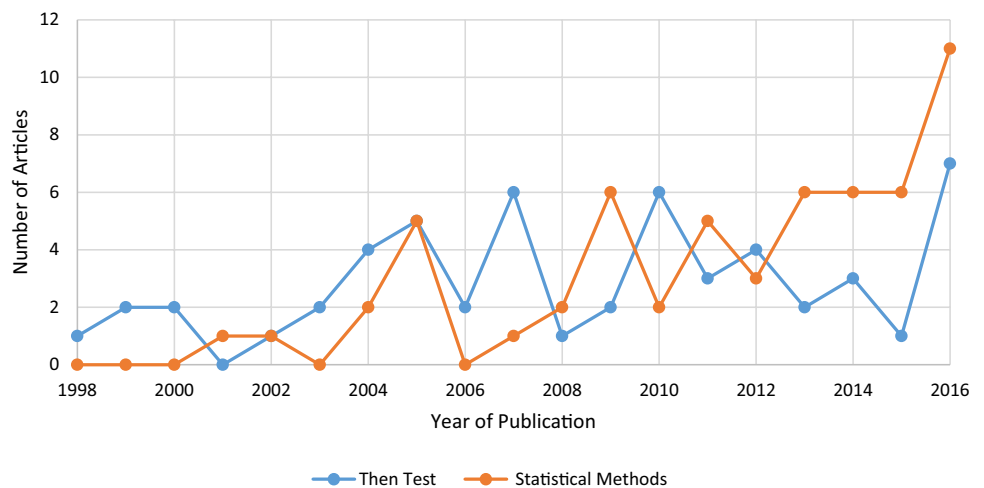


Fig. 1 Flow diagram illustrating selection strategy for Included Articles

Fig. 2 Uptake of design-based and statistics-based response shift methods by year of publication



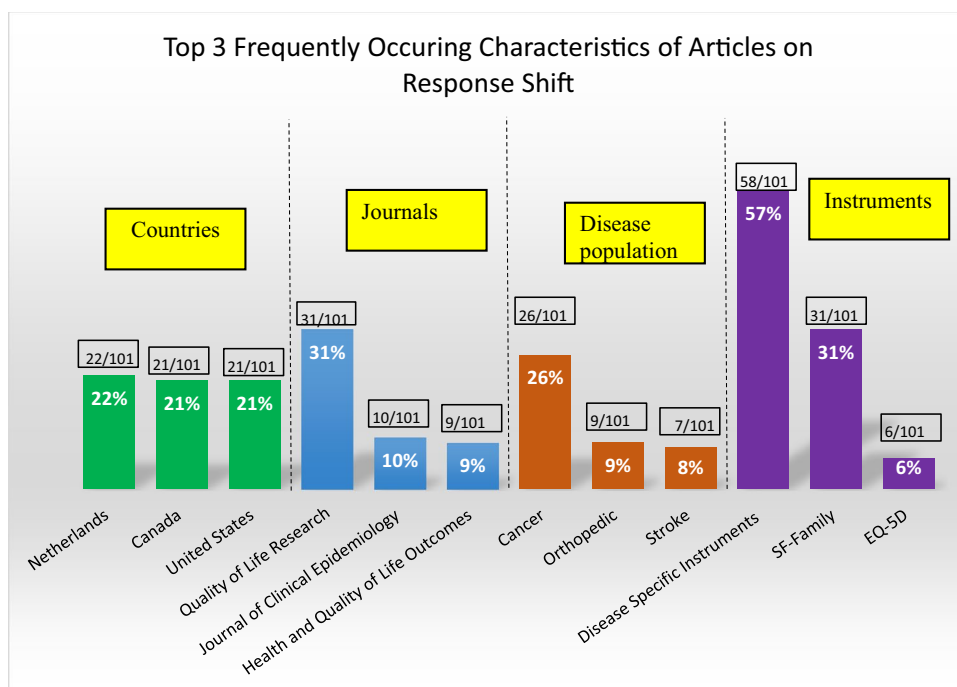
by researchers in the United States. Specifically, 31 articles (30.7%) were published in *Quality of Life Research*, while an additional 19 articles (18.8%) were published in either *Journal of Clinical Epidemiology* or *Health and Quality of Life Outcomes*. 26 articles (25.7%) investigated the detection of RS in cancer populations, 9 (8.9%) articles focused on RS in relation to orthopedic conditions, while 7 (8.0%) articles tested RS in stroke populations. In terms of the different types of PRO measures utilized, 58 (57.4%) of the articles

reported on at least one disease-specific measure, 31 (27.1%) reported on different versions of the SF-36 and the SF-12 [39], while 6 (6.3%) articles reported on the EQ-5D [40].

Uptake of RS detection methods

Tables 2 and 3 and Fig. 4 describe the RS detection methods, types of RS detected, and reporting practices for the *Then-test* and other statistical methods. Of the 101 articles

Fig. 3 Most frequently occurring characteristics (Country, Journal, Disease, and Instruments adopted) of Included Articles. *SF-Family* Short-form family of surveys (i.e., SF-6, SF-12, SF-36), *EQ-5D* 5-dimensional EuroQol Questionnaire



included in this review, 55 (54.5%) adopted the *Then-test* method, 26 (25.7%) articles adopted Oort's SEM approach, while 5 (5.0%) articles adopted Schmitt's SEM approach method. RS detection methods developed more recently were, as expected, less frequently used. These include relative importance analysis (4.0%), IRT (4.0%), and CART methods (3.0%). In addition, 6 (5.0%) of the articles used linear regression analysis of change scores to test for RS. When we focused on articles published between 2012 and 2016 in order to understand the uptake of newly developed statistical methods, only 9 (20.0%) of the 45 published articles within this period, adopted any of the more recently developed RS detection methods.

96 articles (95.0%) reported detecting RS, while only 5 articles (5.0%) reported the absence of RS. Amongst the former group, 82 articles (85.4%) reported the presence of recalibration RS, 20 articles (20.8%) detected reprioritization RS, 4 articles (4.2%) detected reconceptualization RS, while 7 articles (7.3%) detected RS but did not report on the specific type of RS that was detected. When we stratified the results by the type of RS detection method, 55 articles (54.5%) adopted the *Then-test* method, while 57 articles (56.4%) adopted statistical methods, and 11 articles (10.9%) reported RS detection using both *Then-test* and statistical model-based methods. While 52 articles (91.2%) of the 57 articles that adopted the *Then-test* method detected RS, only 39 (68.4%) of the 57 articles that adopted statistical model-based methods detected RS. The majority of the investigated methods (e.g., *Then-test*, SEM, IRT, trajectory analysis, CART analysis, multiple linear regression

methods) are developed to test for recalibration response shift, hence the predominant detection of recalibration RS. Of the 20 articles that reported the detection of reprioritization RS, 12 articles (60%) adopted Oort's SEM, while the remaining articles adopted relative importance analysis, classification and regression trees, or random forest regression (Table 3). An examination of the effect size reporting practices showed that, of the 57 articles that adopted a RS detection method that allows for calculating RS effect sizes, 43 articles (75.4%) adopted the *Then-test* method, 18 articles (31.6%) adopted a SEM approach, while only 1 article (1.8%) applied an IRT approach.

Reporting of data characteristics and examination of statistical assumptions

Of the 101 articles included in the review, 15 (15.6%) articles reported on the method used to address missing data. The majority of these articles adopted imputation methods such as deterministic single imputation methods (e.g., personal mean score, overall mean score, or median imputation), probabilistic single imputation method (e.g., expectation–maximization algorithm, full information maximum likelihood), or multiple imputation methods. While the remaining 86 articles did not report on the pattern of missing data or method for dealing with missing data, most of these articles adopted complete case analysis with case-wise deletion of observations with incomplete data. Only a few articles examined the tenability of distributional assumptions and the robustness of the selected statistical procedure

Table 2 Uptake of RS detection methods, type of RS detected, and reporting of statistical assumptions ($n = 101$)

Data characteristics	Total number of articles ($N = 101$)	<i>Then-test</i> ($n = 55$)	Statistical model-based methods ($n = 57$)
Response shift method			
<i>Then-test</i>	55	55	0
Structural equation model (Oort)	26	0	26
Structural equation modeling (Schmidt)	5		5
Relative importance	4	0	4
Classification and regression trees	3	0	3
Random forest regression	2	0	2
Trajectory analysis of residuals	6		6
Mixed-effects regression	1	0	1
Item response theory (ROSALI)	3	0	3
Item response theory (Anota)	1		1
Multiple linear regression	5	0	5
Others ^a	4	0	4
Response shift detected (yes)	96	52	54
Recalibration	82	52	39
Reprioritization	20	0	20
Reconceptualization	4	0	4
Unspecified	7	0	7
No response shift detected	5	3	3
Missing data method reported (yes)	15	0	15
No imputation method			
Available case analysis	4	0	4
Deterministic single imputation method			
Mean median imputation	4	0	4
Probabilistic single imputation methods			
Expectation–maximization imputation	3	0	3
Full Information maximum likelihood	4	0	4
Multiple imputation	2	0	2
No missing data method reported	86	55	42
Effect size reporting (yes)	43 (57) ^b	43 (55)	14 (35)
Reporting of at least one Statistical assumption (yes)	25	6	19
Distributional assumption of PRO data ^c	12 (90)	5	7 (43)
Collinearity ^d	5 (51)	0	5 (51)
Multiple testing ^e	11 (87)	2	9 (43)
No reporting statistical assumptions	76	49	38

^aConsists of methods such as analysis of covariance, triangulation methods, and multivariate analysis of variance methods

^bThe denominators, in brackets, are the number of articles that adopted a RS detection method that allows for the reporting of RS effect sizes

^cThe denominators, in brackets, are the numbers of articles that adopted a RS detection method that relies on the assumption of normality

^dThe denominators, in brackets, are the number of articles that adopted a RS detection method that is sensitive to collinearity among the PRO domains

^eThe denominators, in brackets, are the numbers of articles that adopted a RS detection method that simultaneously test multiple RS hypotheses. See Online Appendix B for a list of criteria for selection of number of articles that constitute the denominators for each assumption. For reported frequencies with no parenthesis, the denominator corresponds to the column total reported on the column headings

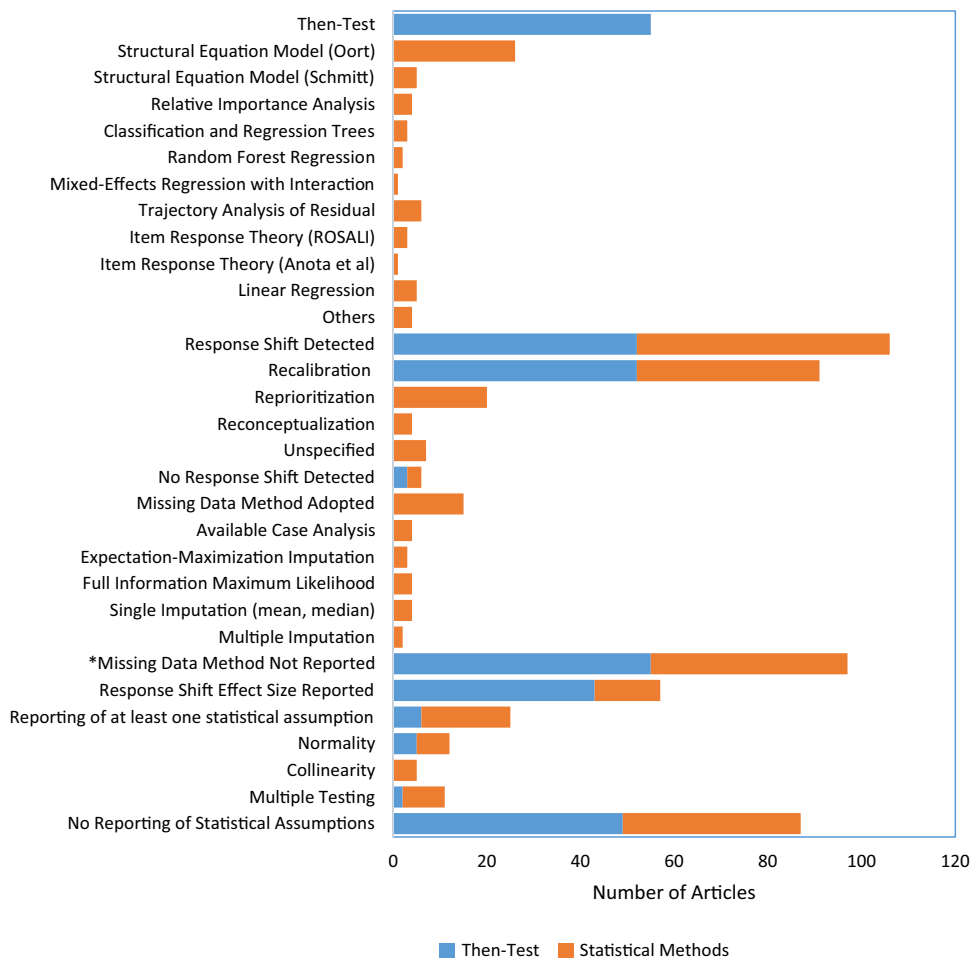
Table 3 Frequencies of articles by types of RS and RS detection methods

RS detection method	Recalibration	Reprioritization	Reconceptualization	Unspecified
<i>Then-test</i>	52	NA	NA	0
Structural equation model (Oort)	21	10	4	0
Structural equation modeling (Schmitt)	3	1	NA	0
Relative importance	NA	4	NA	0
Classification and regression trees	3	2	0	0
Random forest regression	NA	2	NA	0
Trajectory analysis of residuals	6	0	NA	0
Mixed-effects regression	1	0	NA	0
Item response theory (ROSALI)	2	1	NA	0
Item response theory (Anota)	1	0	NA	0
Multiple linear regression	1	NA	NA	4
Others ^a	1	NA	NA	4

^a Consists of methods such as analysis of covariance, triangulation methods, and multivariate analysis of variance methods

NA not applicable

Fig. 4 Uptake of RS detection methods, Type of RS detected and reporting of statistical assumptions and data characteristics. Denominators for the reported frequencies are found in Table 2; asterisk means number of articles added up to more than 101 because some of the articles adopted both *Then-test* and statistical model-based methods



to data characteristics. Of the 101 articles included in this review, 25 (24.8%) articles reported on statistical assumptions that underlie the RS detection method adopted. Specifically, 12 articles (13.3%; out of possible 90) assessed the assumption of multivariate normality of the PRO data and/or adopted an alternative strategy to deal with the non-normal distribution of the data. While regression-based models are the most sensitive to multicollinearity, only 5 articles (out of a possible 51 articles [9.8%]) that used a regression-based RS detection method examined the presence of multicollinearity and/or adopted a strategy to deal with it. Finally, of the 87 articles that adopted RS detection methods that allow for testing of multiple RS hypotheses, only 11 (12.6%) examined or adopted a strategy to control familywise Type I error.

Discussion

This study investigated the uptake of RS detection methods in PRO literature and described the current reporting practices and methodological research gaps in RS literature. Our scoping review revealed that there has been an increasing number of empirical investigations of RS in longitudinal PRO studies between 1999 and 2016 (see Fig. 2). The *Then-test* method was the most commonly adopted method for detecting RS. Another one-third of the articles adopted either the Oort or Schmitt SEM procedure. We also found inconsistent reporting practices with respect to the type and magnitude of the RS effect and the testing and adjustment for violations of statistical assumptions.

There are a few points to note when interpreting these findings. First, not all RS detection methods were developed to test all forms of RS. Oort's SEM method is the most versatile, so it is not surprising that it has been used in majority of studies. Secondly, the majority of the existing RS detection methods are developed to test recalibration and/or reprioritization RS. But, only a few methods are available to test for reconceptualization RS. There is need for further research on new methods to test for reconceptualization RS in PRO studies. Third, the reviewed RS detection methods are not equally sensitive to a variety of data characteristics (e.g., multicollinearity, missing data, non-normal distributions). For example, PRO data are generally characterized by skewed or heavy-tailed distributions, incomplete data, and strong between-domain/item correlations [41]. Relative importance analysis based on logistic regression is sensitive to strong between-domain correlations but robust to departures from the assumption of multivariate normality. On the other hand, conventional SEM based on maximum likelihood estimation is reasonably robust to multicollinearity but sensitive to violations of the distributional assumption of multivariate normality [42, 43].

Our review showed that the majority of the articles published on RS did not report on whether any of the assumptions underlying each RS detection method are examined or not, suggesting that there is less emphasis and limited research on the consequences of violations of these statistical assumptions. The International Society for Quality of Life (ISOQOL) RS Special Interest Group (SIG) has spearheaded research investigations into some of these issues. For example, a series of papers that investigated various approaches for detecting RS in incomplete longitudinal was published as a special issue in *Quality of Life Research* [44–46]. More recently, another series of methodological studies about novel statistical approaches, such as IRT and SEM based on weighted least squares with mean and variance adjusted estimators and robust maximum likelihood estimators, for testing for RS in item-level data, which are typically characterized by non-normal distributions, has been published [28, 29, 47, 48]. Further methodological research is needed to guide the choice of appropriate RS detection methods in longitudinal PRO data that violate conventional distributional assumptions.

There was inconsistent reporting of RS effect sizes observed in our review; this can be partly attributed to the fact that not all the existing RS detection methods can produce a measure of effect size. For example, while RS effect sizes can be estimated from *Then-test*, SEM, and IRT methods [20, 49], the magnitude of RS effects cannot be estimated in recently developed RS detection methods such as relative importance analysis, CART, and random forest regression methods. This lack of reporting on RS effect sizes has limited meta-analytic investigations of RS effects across multiple RS detection methods. Further research is needed to develop appropriate measures of effect size for recently developed RS detection methods. This will aid future meta-analytic investigation of RS literature and inform research, clinical, and/or policy decisions about RS.

Recommendations for the International Society for Quality of Life (ISOQOL)

The results of this scoping review shed light on several important gaps in research that must be addressed to advance the RS research field. We accordingly provide the following recommendations for the ISOQOL. First, our review suggests that RS has been mostly investigated in cancer, stroke, and multiple sclerosis patient populations, but there is limited investigation of RS in more prevalent chronic diseases such as diabetes, hypertension, and other cardiovascular diseases. The ISOQOL needs to devise knowledge translation strategies to engage researchers and clinicians in investigating RS and its implications on the measurement of change in PRO in these prevalent diseases conditions.

Second, our review highlights the limited reporting of investigations about distributional assumptions underlying RS detection methods. The detection of a RS effect is at least partially dependent on the method's robustness to data analytic conditions. There is a need for ongoing methodological work to address issues pertaining to comparison of methods with respect to their statistical power to detect RS effects in PRO data characterized by skewed or heavy-tailed distributions, missing data, and other characteristics. This will ensure that robust methodologies are available to detect RS in longitudinal data despite violations of underlying distributional assumptions. Several studies have examined the convergent validity of different RS detection methods (*Then-test*, SEM, group-based trajectory analysis of residuals, and relative importance analysis) in secondary analyses of longitudinal data and reported conflicting results about the presence of RS [25, 34–37]. These inconsistencies in the detection of RS are not entirely surprising given that these methods operationalize RS differently, possess different statistical properties, and are not equally robust to violations of distributional assumptions. Simulation analysis provides a valuable approach for comparing the statistical properties of existing RS detection methods by artificially generating pseudo datasets that mimic various data analytic conditions often observed in real-life PRO datasets. Although computer simulations have been used to evaluate the performance of SEM [29] and ROSALI algorithm [21, 45], further research that uses simulation methods to compare statistical properties of across multiple RS detection methods is needed.

Although our review has focused on methodological considerations, it is important to recognize that the notion of RS is ultimately not predominantly a statistical matter, but rather a theoretical matter that pertains to measurement validation and the processes by which individuals respond to measurement items [10–12, 50–55]. The choice of an appropriate RS detection method seems to be primarily guided by technical considerations, such as the design of a study, sample size, and variable distributions. However, the various statistical methods represent different aspects of response processes and measurement validation. As is further explained by Sawatzky et al [10], “it is, therefore, important that decisions about which method to utilize be informed by theoretical considerations related to the purposes of using PRO information and the type of decisions that can be made based on the PRO information.” For example, methods that allow for item-level analyses [e.g., 21, 44] should be utilized when the purpose is to understand people's interpretation and response to the items [17, 18]. Readers are referred to previously published work on cognitive processes underlying patients' response to

measurement items [8, 50–52] and modern perspectives on measurement validation as it relates to RS theory [53–55].

Our review also highlights variations in reporting of statistical results related to RS effect sizes, statistical methods and their underlying assumptions, and study data characteristics, across most of the articles investigated. These inconsistencies and variations in reporting practices could constrain scientific advances in the field, including reduced opportunities for meta-analysis of RS effects. Standardized reporting guidelines can help to ensure uniform reporting of RS results. Although existing reporting guidelines, such as STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) and Consolidated Standards of Reporting Trials (CONSORT) guidelines, can be adapted for use in reporting about observational studies and randomized controlled trials, additional guidelines are needed to ensure reporting of comparable and consistent information about RS in different studies [56, 57].

Finally, despite methodological advancements in the use of modern statistical methods for detecting RS, more education about these methods through webinars and workshops, is needed along with their automation of recently developed methods in existing statistical software packages [58–61] to facilitate easier and accessible uptake of these methods.

Conclusion

In summary, RS research continues to be an integral part of PRO research field. While our review highlights the methodological advancements made over the last two decades, *Then-test* and SEM are the two commonly used to assess RS in PRO studies. There is a need for more research on the development of innovative methods, standardized reporting guidelines, and knowledge translation activities to expand RS research frontiers into new areas.

Acknowledgements The Canadian Institutes of Health Research provided support (Grant # MOP-142404) to Drs Sajobi, Lix, Zumbo, and Sawatzky in this research. Dr. Sajobi is supported by the O'Brien Institute for Public Health; Dr. Lix is supported by the Manitoba Research Chair; and Dr. Sawatzky holds a Canada Research Chair in Patient-Reported Outcomes at Trinity Western University, Langley, British Columbia. We are grateful for the support in conducting literature searches provided by Duncan Dixon, health sciences librarian at Trinity Western University. This research was initiated during Dr. Sajobi's visit to Trinity Western University in 2014.

References

1. Revicki, D. A. (1989). Health-related quality of life in the evaluation of medical therapy for chronic illness. *Journal of Family Practice*, 29(4), 377–380.

2. Berzon, R., Hays, R. D., & Shumaker, S. A. (1993). International use, application and performance of health-related quality of life instruments. *Quality of Life Research*, 2(6), 367–368.
3. Deshpande, P. R., Rajan, S., Sudeepthi, B. L., & Abdul Nazir, C. P. (2011). Patient-reported outcomes: A new era in clinical research. *Perspectives in Clinical Research*, 2(4), 137–144.
4. Golembiewski, R. T., Billingsley, K., & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *The Journal of Applied Behavioral Science*, 12, 133–157.
5. Howard, G. S., Ralph, K. M., Gulanick, N. A., Maxwell, S. E., Nance, S. W., & Gerber, S. K. (1979). Internal invalidity in pretest-posttest self-report evaluations and reevaluation of retrospective pretests. *Applied Psychological Measurement*, 3, 1–23.
6. Breetvelt, I. S., & Van Dam, F. S. (1991). Underreporting by cancer patients: the case of response-shift. *Social Science and Medicine*, 32, 981–987.
7. Sprangers, M. A., & Schwartz, C. E. (1999). Integrating response shift into health-related quality of life research: a theoretical model. *Social Science and Medicine*, 48(11), 1507–1515.
8. Taminiau-Bloem, E. F., van Zuuren, F. J., Visser, M. R. M., Tishelman, C., Schwartz, C. E., Koeneman, M. A., et al. (2011). Opening the black box of cancer patients' quality-of-life change assessments: a think-aloud study examining the cognitive processes underlying responses to transition items. *Psychol Health*, 26(11), 1414–1428.
9. Schwartz, C. E., Andresen, E. M., Nosek, M. A., & Krahn, G. L. & RRCT Expert Panel on Health Status Measurement. (2007). Response shift theory: Important implications for measuring quality of life in people with disability. *Archives of Physical Medicine and Rehabilitation*, 88(4), 529–536.
10. Schwartz, C. E. (2010). Application of response shift theory and methods to participation measurement: A brief history of a young field. *Archives of Physical Medicine and Rehabilitation*, 91(9), S38–S43.
11. Rapkin, B. D., & Schwartz, C. E. (2004). Toward a theoretical model of quality-of-life appraisal: Implications of findings from studies of response shift. *Health Qual Life Outcomes*, 2, 14.
12. Schwartz, C. E., Bode, R., Repucci, N., Becker, J., Sprangers, M. A., & Fayers, P. M. (2006). The clinical significance of adaptation to changing health: A meta-analysis of response shift. *Quality of Life Research*, 15(9), 1533–1550.
13. Sawatzky, R., Sajobi, T. T., Brahmabhatt, R., Chan, E. K. H., Lix, L. M., & Zumbo, B. D. (2017). Longitudinal change in response processes: A response shift perspective. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 251–276). New York: Springer.
14. Balanin, B., Ennis, O., Kanis, G., Siinghal, R., Roberts, S. N., Rees, D., et al. (2009). Response shift in self-reported functional scores after knee microfracture for full thickness cartilage lesions. *Osteoarthritis and Cartilage*, 17(8), 1009–1013.
15. Sajobi, T. T., Fiest, K. M., & Wiebe, S. (2014). Changes in quality of life after epilepsy surgery: the role of reprioritization response shift. *Epilepsia*, 55(9), 1331–1338.
16. Vanier, A., Falissard, B., Sébille, V., & Hardouin, J. B. (2017). The complexity of interpreting changes observed over time in health-related quality of life: A short overview of 15 years of research on response shift theory. In F. Guillemin, A. Leplège, S. Briançon, E. Spitz & J. Coste (Eds.), *Perceived health and adaptation in chronic disease. Stakes and future challenge* (pp. 202–230). New York, NY: CRC Press.
17. Bray, J. H., Maxwell, S. E., & Howard, G. S. (1984). Methods of analysis with response-shift bias. *Educational and Psychological Measurement*, 44(4), 781–804.
18. Schmitt, N. (1982). The use of analysis of covariance structures to assess beta and gamma change. *Multivariate Behavioral Research*, 17, 343–358.
19. Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research*, 14(3), 587–598.
20. Oort, F. J., Visser, M. R. M., & Sprangers, M. A. G. (2005). An application of structural equation modeling to detect response shifts and true change in quality of life data from cancer patients undergoing invasive surgery. *Quality of Life Research*, 14(3), 599–609.
21. Anota, A., Bascoul-Mollevi, C., Conroy, T., Guillemin, F., Velten, M., Jolly, D., et al. (2014). Item response theory and factor analysis as a mean to characterize occurrence of response shift in a longitudinal quality of life study in breast cancer patients. *Health Qual Life Outcomes*, 12, 32.
22. Guilleux, A., Blanchin, M., Vanier, A., Guillemin, F., Falissard, B., Schwartz, C. E., et al. (2015). RespOnse Shift ALgorithm in Item response theory (ROSALI) for response shift detection with missing data in longitudinal patient-reported outcome studies. *Quality of Life Research*, 24(3), 553–564.
23. Lowy, A., & Bernhard, J. (2004). Quantitative assessment of changes in patients' constructs of quality of life: an application of multilevel models. *Quality of Life Research*, 13(7), 1177–1185.
24. Mayo, N. E., Scott, S. C., Dendukuri, N., Ahmed, S., & Wood-Dauphinee, S. (2008). Identifying response shift statistically at the individual level. *Quality of Life Research*, 17(4), 627–639.
25. Lix, L. M., Sajobi, T. T., Sawatzky, R., Liu, J., Mayo, N. E., Huang, Y., et al. (2013). Relative importance measures for reprioritization response shift. *Quality of Life Research*, 22(4), 695–703.
26. Li, Y., & Rapkin, B. (2009). Classification and regression tree uncovered hierarchy of psychosocial determinants underlying quality-of-life response shift in HIV/AIDS. *Journal of Clinical Epidemiology*, 62(11), 1138–1147.
27. Boucekine, M., Loundou, A., Baumstarck, K., Minaya-Flores, P., Pelletier, J., Ghattas, B., et al. (2013). Using the random forest method to detect a response shift in the quality of life of multiple sclerosis patients: a cohort study. *BMC Medical Research Methodology*, 13, 20.
28. Verdam, M. G. E., Oort, F. J., & Sprangers, M. A. G. (2016). Using structural equation modeling to detect response shifts and true change in discrete variables: an application to the items of the SF-36. *Quality of Life Research*, 25(6), 1361–1383.
29. Vanier, A., Sebille, V., Blanchin, M., Guilleaux, A., & Hardouin, J. (2015). Overall performance of Oort's procedure for response shift detection at item-level. *Quality of Life Research*, 24(8), 1799–1807.
30. Ahmed, S., Sawatzky, R., Levesque, J. F., Ehrmann-Feldman, D., & Schwartz, C. E. (2014). Minimal evidence of response shift in the absence of a catalyst. *Quality of Life Research*, 23(9), 2421–2430.
31. Joyce, C. R. B., O'Boyle, C., & McGee, H. (1999). *Individual quality of life: Approaches to conceptualization and assessment*. Amsterdam: Harwood Academic Publishers.
32. Schwartz, C. E., & Sprangers, M. A. (1999). Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Social Science and Medicine*, 48(11), 1531–1548.
33. Ahmed, S., Mayo, N. E., Wood-Dauphinee, S., Hanley, J. A., & Cohen, R. S. (2005). Using the patient generated index to evaluate response shift post-stroke. *Quality of Life Research*, 14(10), 2247–2257.
34. Ahmed, S., Mayo, N. E., Wood-Dauphinee, S., Hanley, J. A., & Cohen, R. S. (2005). The structural equation modeling technique did not show a response shift, contrary to the results of the then

- test and the individualized approaches. *Journal of Clinical Epidemiology*, 58(11), 1125–1133.
35. Visser, M. R., Oort, F. J., & Sprangers, M. A. (2005). Methods to detect response shift in quality of life data: A convergent validity study. *Quality of Life Research*, 14(3), 629–639.
 36. Mayo, N. E., Scott, S. C., Bernstein, C. N., & Lix, L. M. (2015). How are you? Do people with inflammatory bowel disease experience response shift on this question? *Health Qual Life Outcomes*, 13, 52.
 37. Lix, L. M., Chan, E. K. H., Sawatzky, R., Sajobi, T. T., Liu, J., Hopman, W., et al. (2016). Response shift and disease activity in inflammatory bowel disease. *Quality of Life Research*, 25(7), 1751–1760.
 38. Schwartz, C. E., Ahmed, S., Sawatzky, R., Sajobi, T., Mayo, N., Finkelstein, J., et al. (2013). Guidelines for secondary analysis in search of response shift. *Quality of Life Research*, 22(10), 2663–2673.
 39. Ware, J. E. Jr., Kosinski, M., & Keller, S. D. (1994). *SF-36 physical and mental scales: A user's manual*. Boston: The Health Institute, New England Medical Center.
 40. Herdman, M., Gudex, C., Lloyd, A., Jansen, M., Kind, P., Parkin, D., et al. (2011). Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). *Quality of Life Research*, 20(10), 1727–1736.
 41. Beaumont, J. L., Lix, L. M., Yost, K. J., & Hahn, E. A. (2006). Application of robust statistical methods for sensitivity analysis of health-related quality of life outcomes. *Quality of Life Research*, 15(3), 349–356.
 42. Grewal, R., Cote, J. A., & Baumgartner, H. (2004). Multicollinearity and measurement error in structural equation models: Implications for theory testing. *Marketing Science*, 23(4), 519–529.
 43. Muthén, B., & Kaplan, D. (1985). A comparison of methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, 38(1), 171–189.
 44. Schwartz, C. E., Sajobi, T. T., Verdam, M. G., Sebille, V., Lix, L. M., Guilleux, A., et al. (2015). Method variation in the impact of missing data on response shift detection. *Quality of Life Research*, 24(3), 521–528.
 45. Sajobi, T. T., Lix, L. M., Singh, G., Lowerison, M., Engbers, J., & Mayo, N. E. (2015). Identifying reprioritization response shift in a stroke caregiver population: A comparison of missing data methods. *Quality of Life Research*, 24(3), 529–540.
 46. Verdam, M. G., Oort, F. J., van der Linden, Y. M., & Sprangers, M. A. (2015). Taking into account the impact of attrition on the assessment of response shift and true change: A multigroup structural equation modeling approach. *Quality of Life Research*, 24(3), 541–551.
 47. Blanchin, M., Sebille, V., Guilleux, A., & Hardouin, J. (2016). The Guttman errors as a tool for response shift detection at subgroup and item levels. *Quality of Life Research*, 25(6), 1385–1393.
 48. Schwartz, C. E. (2016). Introduction to special section on response shift at the item level. *Quality of Life Research*, 25(6), 1323–1325.
 49. Verdam, M. G. E., Oort, F. J., & Sprangers, M. A. G. (2017). Structural equation modeling-based effect-size indices were used to evaluate and interpret the impact of response shift effects. *Journal of Clinical Epidemiology*, 85, 37–44.
 50. Tourangeau, R., Rips, L. J., & Rasinski, K. A. (2000). *The psychology of survey response*. Cambridge: Cambridge University Press.
 51. Jobe, J. B. (2003). Cognitive psychology and self-reports: Models and methods. *Quality of Life Research*, 12, 219–227.
 52. Sawatzky, R., Chan, E. K. H., Zumbo, B. D., Ahmed, S., Bartlett, S. J., Bingham, C. O., et al. (2016). Challenges and opportunities in patient-reported outcomes validation. *Journal of Clinical Epidemiology*. <https://doi.org/10.1016/j.jclinepi.2016.12.002>.
 53. Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research*, 103(2), 219–230.
 54. Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741–749.
 55. Zumbo, B. D., & Chan, E. K. H. (2014). Reflections on validation practices in the social, behavioral, and health sciences. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (Vol. 54, pp. 321–327). New York: Springer International Publishing.
 56. von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P., STROBE Initiative. (2008). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Journal of Clinical Epidemiology*, 61(4), 344–349.
 57. Schulz, K. F., Altman, D. G., & Moher, D., for the CONSORT Group. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *Annals of Internal Medicine*, 152(11), 726–732.
 58. StataCorp. (2015). *Stata Statistical Software: Release 14*. College Station, TX: StataCorp LP.
 59. SAS Institute Inc. (2014). *SAS/STAT version 9.4*. Cary, NC: AS Institute Inc.
 60. R Core Team. (2012). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
 61. Muthén, L. K., & Muthén, B. O. (2011). *Mplus user's guide* (6th Edn.). Los Angeles, CA: Muthén & Muthén.