Qual Life Res
DOI 10.1007/s11136-015-0938-2

Overall performance of Oort's procedure for response shift detection at item level: a pilot simulation study

Antoine Vanier · Véronique Sébille · Myriam Blanchin · Alice Guilleux · Jean-Benoit Hardouin

Accepted: 4 February 2015

Abstract
Objective This simulation study was designed to provide data on the performance of Oort's procedure (OP) for response shift (RS) detection (regarding type I error, power, and overall performance), according to sample characteristics, at item level. A specific objective was to assess the impact of using different information criteria (IC), as alternatives to the LRT (likelihood-ratio test), for global assessment of RS occurrence.
Methods Responses to five binary items at two times of measurement were simulated. Thirty-six combinations of sample characteristics [sample size (n), "true change," correlations between the two latent variables and presence/absence of uniform recalibration RS (ur)] were considered. A thousand datasets were generated for each combination. RS detection was performed on each dataset following OP. Type I error and power of the global assessment of RS occurrence, as well as overall performance of the OP, were assessed.
Results The estimated type I error was close to 5 % for the LRT and lower than 5 % for the IC. The estimated power was higher for the LRT as compared to the AIC, which was the highest among the other IC. For the LRT, the estimated power for $n = 100$ and for the combination of $n = 200$ and ur $= 1$ item was below 80 %. Otherwise, for other combinations of sample characteristics, the estimated power was above 90 %.
Conclusion For the LRT, higher values of power were estimated compared to IC with appropriate values of type I error. These results were consistent with Oort's proposal to use the LRT as the criterion to assess global RS occurrence.

Keywords Response shift · Patient-reported outcomes · Structural equation modeling · Methodology · Simulation study

A. Vanier · V. Sébille · M. Blanchin · A. Guilleux · J.-B. Hardouin
EA 4275 Biostatistics Pharmacoepidemiology and Subjective Measures in Health Sciences, LUNAM, University of Nantes, Nantes, France

A. Vanier
Department of Biostatistics, Sorbonne Universités, UPMC Univ. Paris 06, Paris, France

A. Vanier (✉)
Department of Biostatistics Public Health and Medical Informatics, AP-HP, University Hospitals Pitié-Salpêtrière Charles-Foix, 47-83 Boulevard de l'Hôpital, 75651 Paris Cedex 13, France
e-mail: antoine.vanier@psl.aphp.fr

V. Sébille · J.-B. Hardouin
Unit of Biostatistics and Methodology, University Hospital of Nantes, Nantes, France

Introduction

When assessing changes observed over time on a score resulting from a patient-reported outcome (PRO) instrument, the need to detect potential response shift (RS) effects (i.e., a change in the meaning of one's self-evaluation of a target construct over time [1]) that may obfuscate "true change" assessment is well established [2, 3]. To do so, various methods have been developed since the late 1990s [4, 5]. One of the most attractive methods to detect RS is Oort's procedure (OP) [6]. OP is based on structural equation modeling (SEM), a statistical modeling technique for testing and estimating different types of causal relations using a combination of quantitative data (i.e., covariance ± mean structures) and

Published online: 11 February 2015

qualitative hypotheses [7]. One of the strengths of SEM is the ability to construct latent variables (i.e., variables that are not observed directly, but inferred from several measured variables) [7].

OP allows detection of all forms of RS (non-uniform and uniform recalibration, reprioritization, and reconceptualization) without the need of a specific design [6]. Nonetheless, it implies analyses at group level [6].

OP relies on an operationalization of the different forms of RS as change(s) in the value of SEM parameters between two times of measurement. This (these) change(s) is (are) the value of error variances for non-uniform recalibration, intercepts for uniform recalibration, and factor loadings for reprioritization [6]. Reconceptualization corresponds to a change in the pattern of factor loadings [6].

OP is an algorithm including four major steps [6]. Each of these steps is associated with a particular longitudinal confirmatory factor analysis (CFA) model. The first step consists in establishing an appropriate measurement model (Model 1) of observed scores at two times of measurement. The second step is a global assessment of RS occurrence. To do so, a model verifying the hypothesis of no RS (Model 2) is constructed, and its fit is compared with Model 1 by testing whether the difference between the χ^2 values of the two models is statistically significant [χ^2 difference test, also known as likelihood-ratio test (LRT)]. If the abovementioned LRT is significant, the fit of Model 2 is worse than Model 1, which is interpreted as a global presence of RS, and the procedure continues. The third step is performed using an iterative process (by relaxing one constraint at a time) starting from Model 2. It is dedicated to detect all forms of RS on all potentially affected items (Model 3). A final model is estimated, in which differences in factor means are indicative of "true change" after accounting for RS (Model 4).

Since its publication, OP has been successfully used to detect RS on several clinical datasets, usually at domain level (i.e., with continuous scores as observed variables) [8–14]. However, the performance of the algorithm, regarding the type I error and statistical power of the global assessment of RS occurrence, or the overall behavior of the procedure (its ability to detect only truly existing RS) remains quite unknown. If the performance of SEM to detect measurement bias has already been investigated in previous studies [15–17], the procedures assessed in these studies, although sharing some similarities with OP, are not strictly equivalent. In addition, nothing is known about the performance of OP in the context of detecting RS at item level (i.e., with categorical responses as observed variables). Lastly, some methodological choices, like the use of the LRT as a global assessment of RS occurrence, can be questioned. Indeed, global assessment of RS occurrence could be achieved using information criteria (IC) instead. IC are designed to help model selection, by summarizing in one numeric value a balance between the information explained by a model and its complexity (parsimony principle). The lowest the value of the IC is, the more parsimonious the model is [18–20]. Therefore, a global presence of RS would be reflected by an increase in the value of the IC in Model 2 compared to Model 1. Assessing the probabilistic performance of a statistical procedure can be approached by estimating the results that it produces on a large number of simulated datasets, as the values of the parameters (i.e., the values of the sample characteristics) used to generate these datasets are fully determined, and therefore known.

Thus, the main objective of this study was to provide for the first time data on the performance of OP (regarding type I error, power, and overall behavior), at item level with binary items, via a simulation study. A specific objective was to assess the impact of using different IC, as alternatives to LRT, for global assessment of RS occurrence.

Materials and methods

Simulated datasets

Responses to five binary items, at two times of measurement (t_0 and t_1), were simulated. As we chose to investigate the OP at item level, it appeared to be suited to simulate these responses via a model related to item response theory. So, these responses were generated, as a function of a latent trait and item difficulties (for each times of measurement), using a longitudinal Rasch model (which has good measurement properties and is commonly used when modeling responses to dichotomous items using the IRT framework) [21]. Thus, the general form of the longitudinal CFA measurement model which was defined to fit Model 1 is of five binary items loading on one latent variable at two times of measurement (Fig. 1).

As this study was a pilot simulation study and as we chose to simulate data with a Rasch model, when RS on an item was simulated, it was uniform recalibration only, operationalized as a one-unit decrease in item difficulty between t_0 and t_1.

Four types of sample characteristics could vary according to different fixed levels:

1. n (sample size) could be fixed at 100, 200, or 300;
2. α (changes in latent trait mean level between the two times or "true change") could be fixed at 0 (no "true change") or -0.2 (a decrease in latent trait mean level between t_0 and t_1);
3. r (correlation between latent traits between the two times) could be fixed at 0.4 (moderate correlation) or 0.9 (very strong correlation);

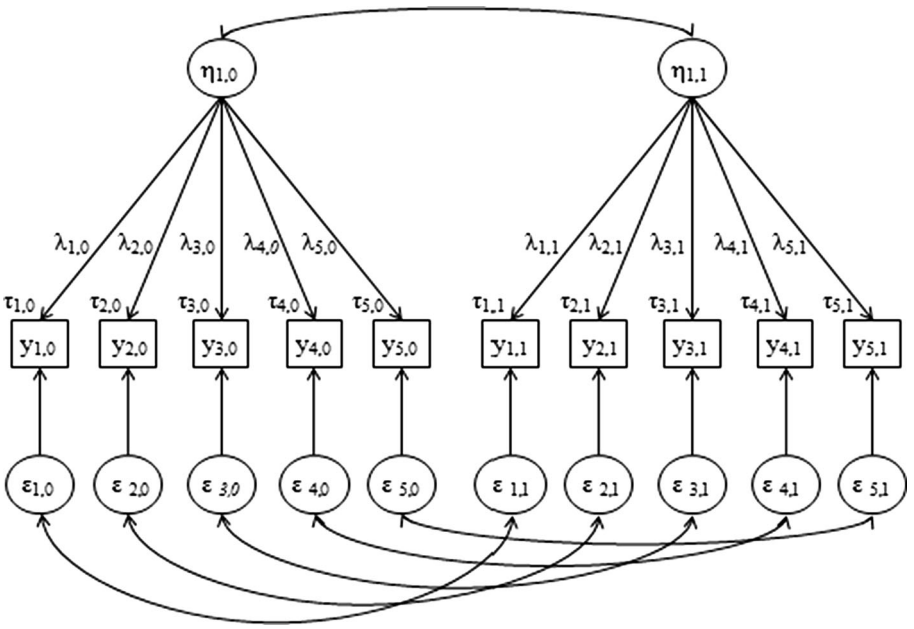


Fig. 1 Graphical representation of the general form of the measurement model (Model 1) fitted on the data. *Notes Circles* represent latent variables and *squares* represent observed variables. The measurement model for the observed scores (responses to items) of an arbitrary subject i at time t may be given by: $y_{i,\,t} = \tau + \Lambda\eta_{i,\,t} + \varepsilon_{i,\,t}$, where y are a vector of observed scores, η a vector of unobserved common factor scores and ε a vector of unobserved residual factor scores. Matrix Λ contains factor loadings ($\lambda_{j,\,t}$ with j the jth item at time t), and vector τ contains intercepts ($\tau_{j,\,t}$). In the figure, the indicates for y, λ, τ and ε are of the form j, t with j the jth item at time t. For η, the indices are of the form k, t, with k the kth common factor score at time t.

4. ur (occurrence of uniform recalibration) could be fixed at 0 item, 1 item (on the third item), or 2 items (on the second and fourth items).

The sample size values were chosen in accordance with sizes usually reported in studies investigating RS [2]. A small negative effect of the catalyst on latent trait mean level (−0.2) was chosen to reflect plausible effect sizes frequently observed in clinical research. As we hypothesized that the correlation between latent traits between the two times would have a negligible impact on RS detection, we chose a moderate (0.4) and an extreme value (0.9) to test this hypothesis. A one-unit decrease in item difficulty was chosen to simulate uniform recalibration, because we had previously showed in another simulation study (aiming at studying the power of the test of group effect in a Rasch model) that the degree of uncertainty of the item difficulty parameters had to be high (a one-unit difference), to observe a moderate impact on power [22].

Thirty-six combinations of the levels of the sample characteristics were investigated. A thousand datasets have been simulated for each combination.

RS detection

RS detection was performed on each datasets following the four steps of OP [6]. SEM models were fitted using robust maximum-likelihood estimator with a Satorra–Bentler correction (MLM) [23], with lavaan package 0.5–13 [24] for R software 3.0.1 [25].

A root-mean-square error of approximation (RMSEA) close to 0.05 (p of close fit >0.05) and comparative fit index (CFI) $\geq$ 0.95 were used as indicators of good fit for Model 1 and Model 4 [26]. Both of these fit indices were computed using Satorra–Bentler corrected χ^2 values.

Global assessment of RS occurrence (step 2) was performed with two different strategies:

1. A Satorra–Bentler scaled χ^2 difference test (which will be thereafter referred as LRT for simplicity) between Model 2 and Model 1 considered significant if the estimated p value was below 0.05 [6];
2. An increase in the value of the IC in Model 2 compared to Model 1 [three common IC were investigated in this study: Akaike information criterion (AIC), Bayesian information criterion (BIC), and sample size adjusted BIC (SABIC)] [27].

If there was global evidence of RS, untenable constraints on RS parameters were relaxed one at a time, starting from Model 2 (step 3). Relaxing the constraints on error variances (non-uniform recalibration) was performed first, followed by intercepts (uniform recalibration) and factor loadings (reprioritization), thus following a hierarchy in testing the different forms of RS proposed in two previous studies [28,

29]. At each time in step 3, the constraint that was proposed to be relaxed was the one leading to a model with the lowest corrected χ^2 value. Each time, the relevance of relaxing a constraint was tested using an LRT, which was considered significant if the estimated *p* value was below 0.05 [30]. Step 3 was performed until relaxing a proposed constraint that led to a nonsignificant LRT.

Statistical analyses

The type I error regarding the global assessment of RS occurrence was estimated as the proportions of datasets where global RS was evidenced among datasets *where no RS was simulated*.

Power of the global assessment of RS occurrence was estimated as the proportions of datasets where global RS was evidenced among datasets *where RS was simulated*.

Overall behavior of the procedure was estimated by means of two indicators:

1. *Overall behavior indicator 1 (OBI1)*: The assessment of the proportion of datasets for which the whole OP *had properly detected uniform recalibration RS on only truly affected item(s)* (after a significant LRT ascertaining global RS occurrence), *disregarding any false detections of RS on these or one of the other items*, and considering only datasets *where RS was simulated;*
2. *Overall behavior indicator 2* (*OBI2*): This indicator was nearly identical as OBI1, but with an additional requirement of *no false detections of RS on any item*(*s*).

Confidence intervals at a 95 % level ($CI_{95\ \%}$) were estimated for all the aforementioned proportions.

Results

Number of analyzed datasets

As illustrated in Fig. 2, analyses were restricted to 25,134 (69.8 %) of the 36,000 datasets initially generated. Three main reasons could cause a dataset to be discarded from analyses: (1) the non-convergence of the estimation algorithm when fitting any model of the whole OP; (2) Model 1 or Model 4 estimated with poor fitting criterion; (3) Model 1 or Model 4 estimated with any odd parameter(s) (negative error variance) (Fig. 2). Most of the 10,866 datasets discarded from analyses were excluded because Model 1 fit (87.0 % of these 10,866 datasets), or Model 4 fit (8.9 %), was not satisfactory.

Type I error of the global assessment of RS occurrence (Model 2 vs. Model 1)

Table 1 shows estimated type I error using different strategies for global assessment of RS occurrence. Overall, regardless of the value of *n*, α, or *r*, estimated type I error for the LRT was close to 5 % [5 % was included in every $CI_{95\ \%}$, except for one combination ($n = 300$, $\alpha = 0$, $r = 0.4$)]. At $n = 100$, type I error estimated for SABIC was close to that estimated for LRT. Otherwise, for all the IC (AIC, BIC, and SABIC) and combinations of sample characteristics, type I error estimated for IC ranged from 0.0 to 1.4 %.

Power of the global assessment of RS occurrence (Model 2 vs. Model 1)

Table 2 shows estimated power using different strategies for global assessment of RS occurrence. For $n = 100$, estimated power for SABIC was slightly higher than that estimated for LRT. Otherwise, regardless of the value of α, *r*, or ur, estimated power was higher for LRT than that estimated for AIC, which was the highest among the other IC.

Two sample characteristics were associated with a substantial increase in estimated power, regardless of the assessed criteria (LRT or IC): an increase in sample size (*n*) and an increase in the number of items affected by uniform recalibration (ur). For LRT, an increase in r was associated with a slight increase in estimated power, especially for $n = 100$.

For all assessed criteria, estimated power for $n = 100$ and the combination of $n = 200$ and ur = 1 was below 80 %. Otherwise, for other combinations of sample characteristics, estimated power for LRT was above 90 %. Estimated power for BIC was always below 5 % and for most of the sample characteristics combinations close to 0 %.

Overall performance of the OP

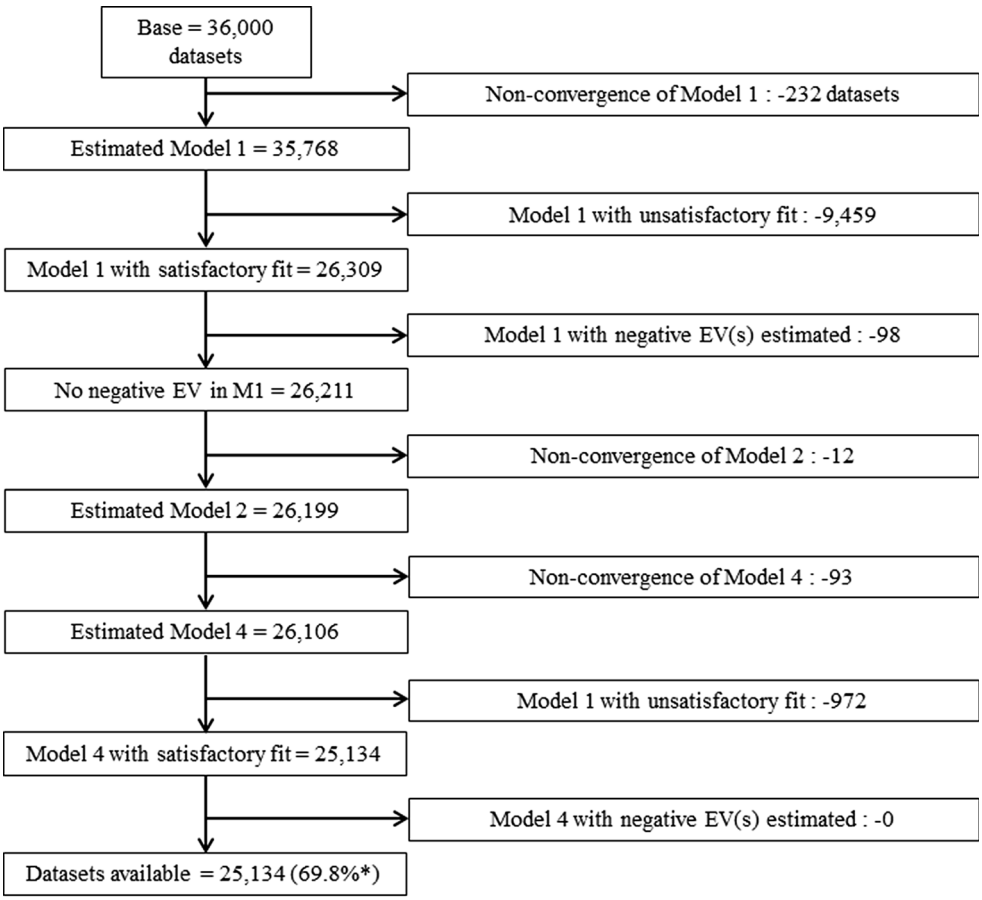
OBI1

Table 3 shows estimated OBI1 (with LRT as the only strategy investigated for global assessment of RS occurrence) according to the different combinations of sample characteristics.

The estimated proportion of datasets for which the whole OP had properly detected uniform recalibration, either on only the third item (ur = 1), or only on the second and fourth items (ur = 2), ranged from 21.9 to 75.5 %, mostly according to sample size (*n*). Indeed, that proportion increased as sample size increased for both ur = 1 and ur = 2. The increase in estimated OBI1 was moderately lower when ur = 2 and $n = 200$ or 300 compared to ur = 1 and $n = 200$ or 300.

OBI2

Table 3 shows estimated OBI2 according to the different combinations of sample characteristics.



Notes: EV: error variance, M1: Model 1, *denominator = 36,000

Fig. 2 Flow chart of datasets discarded from final statistical analyses. *Notes*: *EV* error variance, *M1* model 1, *denominator = 36000

Table 1 Estimated type I error for different strategies for global assessment for RS occurrence (Model 2 vs. Model 1)

n	α	*r*	ur	LRT ($p < 0.05$)		AIC2 > AIC1		SABIC2 > SABIC1		BIC2 > BIC1	
				%	$CI_{95\ \%}$	%	$CI_{95\ \%}$	%	$CI_{95\ \%}$	%	$CI_{95\ \%}$
100	0	0.4	0	4.5	[3.1–6.7]	0.8	[0.3–1.9]	5.1	[3.5–7.3]	0.0	[0.0–0.7]
		0.9	0	3.7	[2.4–5.5]	0.2	[0.0–0.9]	4.0	[2.7–5.9]	0.0	[0.0–0.6]
	−0.2	0.4	0	5.9	[4.1–8.2]	0.4	[0.1–1.4]	7.0	[5.1–9.6]	0.0	[0.0–0.7]
		0.9	0	5.9	[4.2–8.1]	0.7	[0.3–1.8]	6.6	[4.8–8.9]	0.0	[0.0–0.7]
200	0	0.4	0	4.0	[2.8–5.8]	0.4	[0.2–1.3]	0.1	[0.0–0.8]	0.0	[0.0–0.6]
		0.9	0	3.6	[2.5–5.2]	0.8	[0.4–1.7]	0.4	[0.1–1.1]	0.0	[0.0–0.5]
	−0.2	0.4	0	4.0	[2.8–5.7]	0.3	[0.1–1.0]	0.3	[0.1–1.0]	0.0	[0.0–0.5]
		0.9	0	5.4	[4.1–7.3]	1.4	[0.8–2.5]	1.2	[0.6–2.2]	0.0	[0.0–0.5]
300	0	0.4	0	3.0	[2.0–4.4]	0.0	[0.0–0.5]	0.0	[0.0–0.5]	0.0	[0.0–0.5]
		0.9	0	4.5	[3.3–6.1]	0.5	[0.2–1.2]	0.0	[0.0–0.4]	0.0	[0.0–0.4]
	−0.2	0.4	0	5.8	[4.4–7.7]	0.9	[0.4–1.8]	0.0	[0.0–0.5]	0.0	[0.0–0.5]
		0.9	0	4.4	[3.3–6.0]	0.5	[0.2–1.2]	0.0	[0.0–0.4]	0.0	[0.0–0.4]

Overall, the estimated proportion of datasets for which the whole OP had properly detected uniform recalibration on affected item(s) and had appropriately not indicated occurrence of whatever other form(s) of RS on any item(s), ranged from 0.6 to 18.3 %. That estimated proportion was substantially lower than that estimated via OBI1 indicator. Estimated proportion via OBI2 indicator decreased as the number of simulated items affected by uniform recalibration (ur) increased.

Table 2 Estimated power for different strategies for global assessment for RS occurrence (Model 2 vs. Model 1)

n	α	r	ur	LRT ($p < 0.05$)		AIC2 > AIC1		SABIC2 > SABIC1		BIC2 > BIC1	
				%	$CI_{95\ \%}$	%	$CI_{95\ \%}$	%	$CI_{95\ \%}$	%	$CI_{95\ \%}$
100	0	0.4	1	36.4	[32.3–40.7]	12.7	[10.1–15.9]	39.1	[35.0–43.4]	0.0	[0.0–0.7]
			2	57.7	[53.1–62.2]	28.8	[24.8–33.1]	60.4	[55.8–64.8]	0.2	[0.0–1.2]
		0.9	1	37.8	[34.0–41.8]	12.1	[9.7–15.0]	39.3	[35.5–43.3]	0.0	[0.0–0.6]
			2	61.2	[57.1–65.1]	33.6	[29.8–37.6]	64.4	[60.4–68.3]	0.2	[0.0–1.0]
	−0.2	0.4	1	36.6	[32.7–40.8]	15.6	[12.8–18.8]	38.8	[34.8–43.0]	0.0	[0.0–0.7]
			2	60.0	[55.6–64.2]	28.0	[24.2–32.1]	63.6	[59.3–67.7]	0.0	[0.0–0.8]
		0.9	1	41.0	[37.1–45.1]	15.0	[12.3–18.1]	43.0	[39.0–47.0]	0.0	[0.0–0.7]
			2	61.1	[57.1–65.1]	30.2	[26.6–34.1]	63.6	[59.6–67.5]	0.2	[0.0–1.0]
200	0	0.4	1	72.8	[69.4–76.0]	45.0	[41.3–48.7]	39.5	[35.9–43.1]	0.1	[0.0–0.8]
			2	94.6	[92.6–96.1]	77.8	[74.5–80.8]	71.5	[68.0–74.8]	0.3	[0.1–1.1]
		0.9	1	75.1	[71.8–78.0]	44.4	[40.9–48.0]	38.3	[34.9–41.8]	0.1	[0.0–0.8]
			2	94.9	[93.1–96.3]	79.9	[76.9–82.6]	74.5	[71.2–77.5]	0.5	[0.2–1.4]
	−0.2	0.4	1	75.0	[71.7–78.1]	44.3	[40.6–48.0]	37.6	[34.1–41.3]	0.1	[0.0–0.8]
			2	92.2	[89.9–94.0]	74.4	[71.0–77.5]	68.5	[64.9–71.9]	0.7	[0.3–1.7]
		0.9	1	76.2	[73.1–79.2]	46.4	[42.9–50.0]	41.1	[37.6–44.6]	0.0	[0.0–0.5]
			2	94.3	[92.5–95.8]	80.7	[77.7–83.3]	75.6	[72.4–78.5]	1.1	[0.5–2.1]
300	0	0.4	1	92.3	[90.3–94.0]	75.4	[72.3–78.2]	50.4	[47.0–53.9]	0.1	[0.0–0.7]
			2	99.6	[98.9–99.9]	95.5	[93.8–96.7]	85.8	[83.2–88.1]	3.1	[2.1–4.6]
		0.9	1	94.1	[92.3–95.5]	78.3	[75.5–81.0]	52.5	[49.2–55.8]	0.1	[0.0–0.7]
			2	99.6	[98.9–99.9]	95.8	[94.2–97.0]	86.9	[84.5–89.1]	4.6	[3.3–6.2]
	−0.2	0.4	1	92.2	[90.1–93.8]	74.7	[71.6–77.6]	49.8	[46.3–53.3]	0.0	[0.0–0.5]
			2	99.5	[98.7–99.8]	96.8	[95.3–97.8]	86.3	[83.7–88.5]	3.9	[2.7–5.4]
		0.9	1	92.3	[90.3–93.9]	73.3	[70.3–76.1]	47.1	[43.8–50.4]	0.2	[0.1–0.8]
			2	99.8	[99.2–99.9]	96.6	[95.1–97.6]	88.5	[86.3–90.5]	4.2	[3.1–5.8]

Discussion

Regarding global assessment of RS occurrence, the main results of this study were as follows:

- Overall, estimated type I error for the LRT was close to 5 % but substantially lower for IC (except for SABIC at $n = 100$ for which estimated type I error was close to that estimated for LRT);
- Estimated power for LRT was below 80 % for $n = 100$ and for the combination of $n = 200$ and ur = 1, otherwise power was above 90 %;
- Overall, estimated power for LRT was higher than for IC (except for SABIC at $n = 100$, for which estimated power was moderately higher).

Regarding the overall performance of the procedure, the main results of this study were as follows:

- The whole OP properly detected uniform recalibration on only affected item(s) (OBI1) on 21.9–75.5 % of the datasets that proportion increased mostly according to sample size (n);
- Overall, the whole OP properly detected uniform recalibration only on appropriate item(s) and did not indicate occurrence of whatever other form(s) of RS on any item(s) (OBI2), on 0.6–18.3 % of the datasets.

Number of analyzed datasets

In this study, 29.2 % of the datasets were discarded from analyses because of unsatisfactory fit, or occurrence of negative error variance(s). Nonetheless, there are solutions in practical setting, to deal with these issues when analyzing a real dataset (these solutions were not implemented in our simulation framework, due to the huge number of datasets to analyze). For example, adding correlation paths between some residual factors, if, for instance, the hypothesis of local independence does not hold, can greatly improve model fit. In addition, dealing with negative error variance(s) can be done by choosing different starting values.

Global assessment of RS occurrence

In this study, estimated type I error for the LRT was close to 5 %. With normally distributed continuous variables, the test statistic of an LRT between two nested SEM models is

Table 3 Estimated OBI1 and OBI2 (see "Materials and Methods" for definition) as a function of sample characteristics

n	α	r	ur	OBI1		OBI2	
				%	$CI_{95\ \%}$	%	$CI_{95\ \%}$
100	0	0.4	1	26.2	[22.6–30.2]	4.5	[3.0–6.7]
		0.9	1	27.1	[23.6–30.8]	6.6	[4.8–8.8]
	−0.2	0.4	1	27.5	[23.9–31.4]	6.6	[4.8–9.0]
		0.9	1	28.3	[24.8–32.2]	6.6	[4.9–8.9]
200	0	0.4	1	58.4	[54.7–62.0]	11.2	[9.1–13.7]
		0.9	1	59.7	[56.2–63.2]	13.9	[11.6–16.5]
	−0.2	0.4	1	60.5	[56.8–64.1]	16.7	[14.1–19.6]
		0.9	1	60.9	[57.4–64.4]	18.5	[15.9–21.5]
300	0	0.4	1	75.5	[72.4–78.4]	11.0	[9.0–13.3]
		0.9	1	74.5	[71.5–77.3]	13.6	[11.5–16.1]
	−0.2	0.4	1	75.2	[72.1–78.1]	16.1	[13.7–18.8]
		0.9	1	75.5	[72.5–78.2]	18.3	[15.9–21.0]
100	0	0.4	2	21.9	[18.3–25.9]	2.4	[1.4–4.3]
		0.9	2	27.2	[23.7–31.0]	3.7	[2.5–5.6]
	−0.2	0.4	2	28.2	[24.4–32.3]	4.0	[2.6–6.1]
		0.9	2	28.8	[25.2–32.7]	3.4	[2.2–5.2]
200	0	0.4	2	52.6	[48.8–56.4]	1.3	[0.7–2.5]
		0.9	2	57.7	[54.2–61.2]	2.1	[1.3–3.4]
	−0.2	0.4	2	57.4	[53.6–61.0]	6.5	[4.9–8.6]
		0.9	2	64.9	[61.5–68.2]	4.9	[3.5–6.6]
300	0	0.4	2	62.5	[59.0–65.8]	0.6	[0.3–1.5]
		0.9	2	69.8	[66.6–72.8]	0.8	[0.4–1.7]
	−0.2	0.4	2	69.2	[65.9–72.3]	2.7	[1.8–4.1]
		0.9	2	72.5	[69.5–75.4]	2.5	[1.7–3.8]

OBI1: the proportion of datasets for which the whole OP *had properly detected uniform recalibration RS on only truly affected item(s), disregarding any false detections of RS on these or one of the other items,* OBI2: this indicator was nearly identical as OBI1, but with an additional requirement of *no false detections of RS on any item(s).* Global assessment of RS occurrence (Model 2 versus Model 1) was performed using a Satorra–Bentler scaled χ^2 difference test

assumed to follow a χ^2 distribution under the null hypothesis, with a number of degrees of freedom (df) equal to the difference in freely estimated parameters between the two models [26]. Here, we have worked with binary items, but we have corrected the test statistic according to Satorra–Bentler proposal [23]. If this correction was adequate, as the LRT was considered significant if the estimated p value was below 0.05, it was expected to observe type I error for the LRT close to 5 % [31]. The results have matched this expectation.

Estimated type I error using IC was lower compared to LRT. Comparison of the IC between two SEM models does not constitute statistical hypothesis testing in a formal way [32]. Therefore, in theory, it was not expected that the estimated type I error using IC had to be around any specific value (and especially 5 %). Model 2 is formally a simpler model than Model 1: It is nested in Model 1, and in this study, Model 2 has 13 more df than Model 1. As stated before, IC are criteria designed to help evaluating model parsimony [18–20]. When no RS was simulated, it was consistent that the value of Model 2 IC was lower than of Model 1, for almost every datasets. Indeed, in that case, Model 2 adequately respected the parsimony principle. Estimated type I error was the lowest for BIC. This result was consistent with the fact that compared to AIC and SABIC, BIC is constructed to penalize complexity the most [19, 32]. Estimated type I error for SABIC was the highest among the IC for $n = 100$. Again, this was consistent with the fact that at $n = 100$, the added penalty for each supplementary freely estimated parameter to the log-likelihood of an SEM model is lower for SABIC than for AIC and BIC [20, 32]. However, at $n = 200$ and $n = 300$, for SABIC, this aforementioned penalty is between AIC and BIC [20, 32].

In this study, estimated power for LRT was below 80 % for $n = 100$, and above 90 % when the sample size was at least equal to 200 (with ur at least equal to 2 when $n = 200$). Except for SABIC when $n = 100$, the aforementioned estimated power for LRT was the highest, compared to IC. This result could reflect a tendency of IC to be too conservative compared to LRT for a global assessment of RS occurrence via SEM. As Model 2 is the simplest in terms of number of freely estimated parameters, it was more often considered as the most appropriate fitting model when comparing Model 2 and Model 1 using IC as compared to LRT. This seemed to be particularly the case for BIC, with estimated power close to 0 % for most of the sample characteristic combinations assessed in the study.

Overall, as estimated type I error for LRT was indeed close to the theoretically expected 5 % and as estimated power for LRT was the highest, these results are consistent with Oort's proposal to use LRT between Model 2 and Model 1 as the criterion to assess global RS occurrence [6], rather than IC.

Overall performance of the OP

In this study, the estimated proportion of datasets for which the whole OP had properly detected uniform recalibration, either on only the third item (ur = 1) or only on the second and fourth item (ur = 2) (OBI1 indicator), ranged from 21.9 to 75.5 %. That estimated proportion increased, mostly with sample size. These results seem to indicate that as long as the LRT between Model 2 and Model 1 is significant, the procedure correctly detects uniform recalibration on appropriate item(s) in most of the cases. However, when we consider the fact that the procedure not only must detect uniform recalibration on appropriate item(s), but also should avoid detecting other form(s) of RS

on any item(s) (OBI2 indicator), the resulting estimated proportion decreased compared to OBI1 indicator and ranged from 0.6 to 18.3 %. For ur = 1, the procedure had detected non-uniform recalibration on only one item in 30.5 to 56.9 % of the datasets according to sample characteristics. In most cases, the item detected was the same that the item on which uniform recalibration was simulated. For ur = 2, the procedure had detected non-uniform recalibration on at least 1 item in 51.6 to 92.5 % of the datasets. Again, in most cases, this (those) item(s) was (were) the one(s) on which uniform recalbration was simulated. A first explanation to this phenomenon can be linked to the simulation process of uniform recalibration coupled with the fact that this was a work on binary items. Indeed, uniform recalibration was simulated using a longitudinal Rasch model by a change in the value of difficulties across time. On a binary item, this can be equated with a change in the proportion of positive responses (P). In SEM models, non-uniform recalibration is detected by a change in the value of error variance. Error variances are linked with the variance of the item, which is represented for binary items by $P(1-P)$. Therefore, when uniform recalibration was simulated on binary items, it seems plausible that non-uniform recalibration might have been simulated too. Nonetheless, another explanation to the aforementioned phenomenon can reflect the issue regarding the need to introduce, or not, at step 3 of the procedure, a hierarchy in testing for different forms of RS [28, 29]. In this study, a hierarchy proposed in two previous studies was followed [28, 29], which consists in testing non-uniform recalibration first, followed by uniform recalibration and finally reprioritization. This hierarchy was derived from measurement invariance studies [33]. So, in SEM operationalization, we can hypothesize that when an item is affected by uniform recalibration, it also sometimes operationalizes as contingent non-uniform recalibration, which is detected first when the abovementioned hierarchy is followed. If this hypothesis holds, it could advocate against the need to impose a hierarchy. Indeed, if uniform recalibration was allowed to be detected first, then maybe it would correct for the risk of detecting contingent non-uniform recalibration. Thus, if the aforementioned hierarchy had not been imposed, perhaps estimated OBI2 indicator would have been higher.

Limits

This study suffered from some limits. The main one is the method used to estimate SEM parameters. Theoretically, working with binary items requires estimating matrices of tetrachoric correlations alongside with the use of robust diagonally weighted least squares (DWLS) estimator [34]. However, this method imposes to estimate thresholds instead of intercepts and requires more identifiability constraints (known as delta or theta parameterizations) [35]. Currently, the operationalization of the RS detection (especially for non-uniform and uniform recalibration) used in the OP is not adapted to work with DWLS [6]. Thus, we used covariance analyses with robust maximum-likelihood. So, although we performed a Satorra–Bentler correction, which seemed to have corrected for the risk of a biased LRT (as illustrated by the fact that the type I error is consistent with the 5 % theoretically expected), SEM parameters are probably somehow biased which could have affected OBI1 and OBI2 values.

The second main limit is the scope of the study. This study was a pilot simulation study, and simulation studies are usually consuming in terms of computational resources. Therefore, we have chosen to restrict our work to binary items, we also have only simulated uniform recalibration RS and a unique simple structure (five items loading on one dimension). Thus, if the results give some clues about how OP behaves at item level with unidimensional model when detecting uniform recalibration RS, they cannot be easily extrapolated to other settings (polytomous items or continuous scores, other types of RS). In particular, the results cannot be easily generalized to other practical settings in HRQL measurement, like multidimensional instruments with many items.

In addition, we have simulated, using Rasch models, uniform recalibration always with the same magnitude. Although we have empirical data to support the fact that this value was of a sufficient magnitude to represent a significant uniform recalibration effect [22], it remains an uncertainty about what this value represents in SEM. For instance, if it was too low to simulate such effect, it could have a negative impact on the results.

Lastly, we did not investigate in this study other relevant issues related to the OP: like the aforementioned issue of the need, or not, of a hierarchy in step 3, or the need, or not, to correct for multiple hypothesis testing [36].

Conclusion

This study proposed for the first time results on the probabilistic behavior of OP at item level, in terms of type I error, power, and overall performance via a simulation study. The results were consistent with Oort's proposal to use the LRT as a criterion for global assessment of RS occurrence. However, several issues about the most efficient way to conduct RS detection via OP can still be discussed. Moreover, the results of this study are limited by some choices that were made. New simulation studies could be performed to investigate the aforementioned limits. Lastly, this study also emphasizes the need to properly adapt the OP to item-level analyses.

Acknowledgments This study was supported by the Institut National du Cancer (France), under reference "INCA_6931."

References

1. Sprangers, M. A. G., & 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.
2. Schwartz, C. E., Bode, R., Repucci, N., Becker, J., Sprangers, M. A. G., & 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.
3. The SAMSI Psychometric Program Longitudinal Assessment of Patient-Reported Outcomes Working Group, Swartz, R. J., Schwartz, C., Basch, E., Cai, D. L., Fairclough, B., & Rapkin, L. (2011). The king's foot of patient-reported outcomes: Current practices and new developments for the measurement of change. *Quality of Life Research, 20*(8), 1159–1167.
4. Schwartz, C. E., & Sprangers, M. A. G. (1999). Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Social Science and Medicine, 48*(11), 1531–1548.
5. Barclay-Goddard, R., Epstein, J. D., & Mayo, N. E. (2009). Response shift: A brief overview and proposed research priorities. *Quality of Life Research, 18*(3), 335–346.
6. Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research, 14*(3), 587–598.
7. Raykov, T. (2006). *A first course in structural equation modeling* (2nd ed.). Mahwah: Lawrence Erlbaum Associates, Publishers.
8. 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.
9. Visser, M. R. M., Oort, F. J., & Sprangers, M. A. G. (2005). Methods to detect response shift in quality of life data: a convergent validity study. *Quality of Life Research, 14*(3), 629–639.
10. Barclay-Goddard, R., Lix, L. M., Tate, R., Weinberg, L., & Mayo, N. E. (2011). Health-related quality of life after stroke: Does response shift occur in self-perceived physical function? *Archives of Physical Medicine and Rehabilitation, 92*(11), 1762–1769.
11. King-Kallimanis, B. L., Oort, F. J., Nolte, S., Schwartz, C. E., & Sprangers, M. A. G. (2011). Using structural equation modeling to detect response shift in performance and health-related quality of life scores of multiple sclerosis patients. *Quality of Life Research, 20*(10), 1527–1540.
12. Nagl, M., & Farin, E. (2012). Response shift in quality of life assessment in patients with chronic back pain and chronic ischaemic heart disease. *Disability and Rehabilitation, 34*(8), 671–680.
13. Fokkema, M., Smits, N., Kelderman, H., & Cuijpers, P. (2013). Response shifts in mental health interventions: An illustration of longitudinal measurement invariance. *Psychological Assessment, 25*(2), 520–531.
14. Gandhi, P. K., Ried, L. D., Huang, I.-C., Kimberlin, C. L., & Kauf, T. L. (2013). Assessment of response shift using two structural equation modeling techniques. *Quality of Life Research, 22*(3), 461–471.
15. Barendse, M. T., Oort, F. J., Werner, C. S., Ligtvoet, R., & Schermelleh-Engel, K. (2012). Measurement bias detection through factor analysis. *Structural Equation Modeling A Multidisciplinary Journal, 19*(4), 561–579.
16. Barendse, M. T., Oort, F. J., & Garst, G. J. A. (2010). Using restricted factor analysis with latent moderated structures to detect uniform and nonuniform measurement bias; a simulation study. *AStA Advances in Statistical Analysis, 94*(2), 117–127.
17. Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement, 35*(5), 339–361.
18. Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.
19. Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464.
20. Sclove, L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52*, 333–343.
21. Fischer, G., & Molenaar, I. (1995). *Rasch models: Foundation, recent developments, and applications*. New-York: Springer.
22. Sébille, V., Hardouin, J.-B., Le Neel, T., Kubis, G., Boyer, F., Guillemin, F., & Falissard, B. (2010). Methodological issues regarding power of classical test theory and IRT-based approaches for the comparison of patient-reported outcome measures-a simulation study. *BMC Medical Research Methodology*, 10–24.
23. Satorra, A., & Bentler, P. (1994). Corrections to test statistics and standards errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks: Sage.
24. Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36.
25. R Development Core Team. (2013). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
26. Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness of fit measures. *Methods of Psychological Research Online, 8*(2), 23–74.
27. Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.
28. Nolte, S., Elsworth, G. R., Sinclair, A. J., & Osborne, R. H. (2009). Tests of measurement invariance failed to support the application of the "then-test". *Journal of Clinical Epidemiology, 62*(11), 1173–1180.
29. Ahmed, S., Bourbeau, J., Maltais, F., & Mansour, A. (2009). The Oort structural equation modeling approach detected a response shift after a COPD self-management program not detected by the Schmitt technique. *Journal of Clinical Epidemiology, 62*(11), 1165–1172.
30. Bryant, F. B., & Satorra, A. (2012). Principles and practice of scaled difference chi square testing. *Structural Equation Modeling A Multidisciplinary Journal, 19*(3), 372–398.
31. Lehmann, E. L. (2008). *Testing statistical hypotheses* (3rd ed.). New York: Springer.
32. Hu, L., & Bentler, P. (1995). Evaluating model fit. In *Structural equation modeling. Concepts, issues, and applications* (p. 76–99). London: Sage.
33. Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care, 44*(11 Suppl 3), S78.
34. Finney, S. J., & DiStefano, C. (2013). Non-normal and categorical data in structural equation modeling. In Structural Equation (Ed.), *Modeling: A second course* (pp. 439–492). Charlotte: IAP, Information Age Publ.
35. Beaujean, A. (2014). Models with dichotomous indicator variables. In *Latent variable modeling using R. A step-by-step guide* (p. 93–113). New-York, NY: Taylor and Francis.
36. Barclay-Goddard, R., Lix, L. M., Tate, R., Weinberg, L., & Mayo, N. E. (2009). Response shift was identified over multiple occasions with a structural equation modeling framework. *Journal of Clinical Epidemiology, 62*(11), 1181–1188.