Evaluation of the link between the Guttman errors and response shift at the individual level

Yseulys Dubuy, Véronique Sébille, Marie Grall-Bronnec, Gaëlle Challet-Bouju, Myriam Blanchin, Jean-Benoit Hardouin

Title: Evaluation of the link between the Guttman errors and response shift at the individual level

-

Author manuscript

Published in final edited form:
Dubuy, Y., Sébille, V., Grall-Bronnec, M., Challet-Bouju, G., Blanchin, M., & Hardouin, J. B. (2022). Evaluation of the link between the Guttman errors and response shift at the individual level. *Quality of life research*, *31*(1), 61–73. https://doi.org/10.1007/s11136-021-03015-9

Authors:
Yseulys Dubuy[1], Véronique Sébille[1], Marie Grall-Bronnec[1,2], Gaëlle Challet-Bouju[1,2], Myriam Blanchin[1], Jean-Benoit Hardouin[1]

Affiliations:

[1] *INSERM U1246 SPHERE, University of Nantes, University of Tours, Nantes, France*

[2] *Addictive Medicine and Psychiatry Department, CHU Nantes, Nantes, France*

Corresponding author:
Yseulys Dubuy (Yseulys.Dubuy@univ-nantes.fr)

ORCiD:

Yseulys Dubuy https://orcid.org/0000-0001-8390-2285

Myriam Blanchin https://orcid.org/0000-0003-1318-7620

Acknowledgments:

We would like to warmly thank all the staff members of the EVALADD cohort.

Abstract

Purpose: Methods for response shift (RS) detection at the individual level could be of great interest when analyzing changes in patient-reported outcome data. Guttman errors (GEs), which measure discrepancies in respondents' answers compared to the average sample responses, might be useful for detecting RS at the individual level between two time points, as RS may induce an increase in the number of discrepancies over time. This study aims to establish the link between recalibration RS and the change in the number of GEs over time (denoted index *I*) via simulations and explores the discriminating ability of this index.

Methods: We simulated the responses of individuals affected or not affected by recalibration RS (defined as changes in the patients' standard of measurement) to determine whether simulated individuals with recalibration had a greater change in the number of GEs over time than individuals without recalibration. The effects of factors related to the sample, the questionnaire structure and recalibration were investigated. As an illustrative example, the change in the number of GEs was computed in patients suffering from eating disorders.

Results: Within simulations, simulated individuals affected by recalibration had, on average, a greater change in the number of GEs over time than did individuals without RS. Some of the parameters related to the questionnaire structure and recalibration magnitude appeared to have substantial effects on the values of *I*. Discriminating abilities appeared, however, globally low.

Conclusion: Some evidence of the link between recalibration and the change in GEs was found in this study. GEs could be a valuable nonparametric tool for RS detection at a more individual level, but further investigation is needed.

Keywords

Response shift · Guttman errors · Recalibration · Individual leve

Introduction

Patient-reported outcomes (PROs) are increasingly being used in longitudinal studies to take into account patients' perspectives on healthcare and to assess perceived health changes over time [1]. PROs are often investigated via questionnaires (directly completed by patients), including several items usually grouped into domains (e.g., physical, emotional, social functioning, etc.). The unobservable attributes targeted by these questionnaires (such as fatigue and anxiety) are assumed to be represented by nonobservable continuous variables known as "latent variables".

It is usually assumed that patients' perception of the concept of interest, the questions, and the response categories remain the same over time and that observed changes reflect changes in the latent variable (i.e., longitudinal measurement invariance). Hence, patients' responses at two different times are assumed to be directly comparable. However, the cognitive [2] and affective processes involved in questionnaires completion are complex, and PRO changes in longitudinal studies can be difficult to analyze and interpret. Moreover, the assumption of invariance may be questionable, particularly in the context of chronic diseases where patients regularly adapt to their life circumstances. Hence, there might be changes in the meaning of patients' self-evaluations of a target construct, referred to as response shift (RS) [3]. RS is usually assumed to have 3 manifestations: (1) recalibration (changes in the patient's internal standards of measurement), (2) reprioritization (changes in the relative importance a patient gives to a certain component of the target construct, e.g., social functioning, which can become more important than physical functioning) and (3) reconceptualization (changes in the patient's definition of what is being measured). It is essential to assess changes experienced by patients taking into account RS to avoid measurement bias[1] and to detect and quantify RS in a reliable and unbiased manner because of its possible association with patients' adaptation [3–5].

Several statistical methods have been proposed for RS detection. Until recently, these methods were all developed and applied at the domain level, which means that analyses are performed on the domain scores of a multidimensional scale. The most widely used method is Oort's procedure based on structural equation modeling (SEM) [6], which allows for the detection of the three manifestations of RS.

[1] i.e., nonrandom errors in the latent variable estimates

Recently, interest in exploring RS at the item level has increased [7]. Indeed, item-level methods could provide an interesting and complementary perspective when investigating RS, as domain-level analyses may sometimes not appropriately reflect what is occurring at the item level, especially if RS has opposite effects on different items. Among the item-level methods, ROSALI (RespOnse Shift ALgorithm at Item-level) based either on Item Response Theory (IRT) or Rasch Measurement Theory (RMT) has been proposed [8, 9]. IRT-based ROSALI aims at detecting RS between two measurement occasions by allowing the item parameters of a longitudinal generalized partial credit model to vary over time (i.e., item discrimination parameter and item difficulty parameter). Changes in discrimination parameters and difficulty parameters are assumed to be indicative of reprioritization and recalibration RS, respectively. RMT-based ROSALI follows the same algorithm as IRT-based ROSALI but relies on a longitudinal partial credit model; hence, it enables the detection of recalibration only. For SEM, Oort's procedure has also been applied at the item level in different ways to detect recalibration and reprioritization [10–14].

Most methods, such as Oort's procedure and ROSALI, assume homogeneous RS within the sample or subgroups of patients known in advance. This assumption is probably too restrictive. Indeed, RS is likely to occur at different times and have different manifestations and various effect sizes among patients. In addition, whether at the domain or item level, these methods are parametric and thus rely on assumptions (e.g., normal latent variable distribution, normally distributed item responses, and link functions) that might be too restrictive.

A method relaxing these assumptions and focusing on the item and individual levels could be of great practical value. At the item level, Blanchin *et al.* suggested that RS (by interfering with patients' internal standards of measurement and life priorities) might induce discrepancies in individuals' responses over time relative to sample responses [15]. Based on this assumption, they identified, in a real data application, two groups of patients: one with an approximately constant number of discrepancies over time (assumed unlikely to be affected by RS) and another with an increasing number of discrepancies (assumed likely to be affected by RS). The discrepancies were measured nonparametrically using the (weighted) Guttman errors (GEs), which were obtained by comparing the individual responses to the distribution of the sample responses. The ROSALI algorithm detected RS in the subgroup of patients identified as likely to be affected by RS and did not detect RS in the other subgroup. Hence, it was hypothesized that GE could be a useful nonparametric tool for item-level RS detection at a more individual level; however, this was an empirical example, and one may wonder whether RS actually leads

to an increase in the number of GE and to what extent RS has occurred if the number of GE increases over time. In addition, no information is currently available on the ability of the change in the number of GEs to discriminate patients affected by RS from the others.

We performed an exploratory simulation study aiming at 1) establishing the link between recalibration RS and the change in the number of GEs over time (i.e., determining whether recalibration comes with an increase in the number of GEs over time) in various scenarios representative of clinical research studies, 2) determining the simulation parameters associated with this change, and 3) providing some data on the discriminating ability of this GEs-based index in the case of recalibration RS.

Methods

We chose to ascribe our GEs-based index to the detection of recalibration in this first simulation study for the following reasons: 1) we wanted to focus on unidimensional scales; hence, reconceptualization could not be considered for now 2) the potential different meanings and interpretations of reprioritization at the item level from a methodological or conceptual perspective have already been raised [7], and Blanchin *et al.* went even further by questioning the pertinence of the concept of reprioritization at the item level [9].

Guttman errors for recalibration RS detection

Let us consider a fictitious sample of individuals responding to a scale composed of 4 items, with 4 response categories, namely, 0, 1, 2 and 3, at two measurement occasions (t_1 and t_2). Let us assume that the following assumptions hold at both time points: 1) unidimensionality (i.e., all items in the questionnaire measure the same latent variable), 2) local independence (i.e., given the latent variable, item responses are independent), and 3) monotonicity (i.e., when the latent variable increases, the probability of obtaining at least score x on item j does not decrease). For simplicity, let us also assume that the underlying latent variable remains stable over time for all patients. The distributions of the sample responses at t_1 are given in Figure 1.

Guttman errors

To define GEs, we have to order all the response categories above 0 based on their difficulty. The term "difficulty" refers to how frequently a response category is endorsed: the less a category is endorsed, the more difficult it is considered to be. At t_1, all response categories greater than 0 can be ordered from the easiest to the most difficult. The difficulty of the response category x $(= 1, 2, 3)$ from item j $(= 1, 2, 3, 4)$ at t_1 is the proportion of patients scoring below x for item j at t_1.

In our example, we can observe that the easiest positive response category is "1" of item 4 since only 10% of the sample scored below this option. The second easiest category is "1" of item 3 (15% of the sample scored below this option). And so on, until the most difficult category which is "3" of item 1, since 80% of the sample scored below this option (i.e., 0, 1 or 2). Null categories are not included in this order since all patients endorse them. The order so defined is called the difficulty order observed at t_1. The term "difficulty" is sometimes referred to as "popularity" [16], easy and difficult response categories are then called "popular" and "unpopular", respectively.

GEs measure discrepancies in individual patient responses compared to the distribution of the sample responses. A GE occurs every time a patient endorses a response category for a given item, while he/she does not endorse an easier response category (for another item) [17]. For instance, a patient who responded item $1 = 2$, item $2 = 1$, item $3 = 2$ and item $4 = 2$ at t_1 has one GE according to the order defined at t_1. Indeed, he/she endorsed category "2" for item 1 but not category "2" for item 2. A formal definition of GE can be found in Emons' works [18].

Guttman errors and recalibration RS

Let us now assume that recalibration is observed on item 4 for half the sample between t_1 and t_2 (same manifestation and effect size for all affected patients) and that its three positive response categories' difficulties have increased at t_2.

At time t_2, patients without RS should give similar responses to those at t_1 since neither their latent variable nor their perception of the response categories have changed. Hence, among patients without RS, the GEs according to the order at t_1 is expected to remain the same over time. In contrast, at time t_2, responses for item 4 from patients affected by recalibration should deviate from the distribution of the sample responses observed at time t_1. The order observed

at t_1 should no longer fit their responses due to recalibration, inducing more discrepancies. Hence, the number of GE (based on the order at t_1) of patients affected by recalibration is expected to increase over time. For example, if the patient previously introduced was affected by the recalibration on item 4, he/she could have responded at the second time point: $\text{item } 1 = 2, \text{item } 2 = 1, \text{item } 3 = 2$ and $\text{item } 4 = 0$ (due to recalibration, the response categories of item 4 became more difficult to endorse). According to the order defined at t_1, he/she has eight GEs at t_2 (instead of one at t_1). His/her overall sum score (computed by summing item responses) has changed but reflects the occurrence of recalibration and not a latent variable change.

Hence, counting the number of GEs over time using the order observed at t_1 could help identify recalibration. Indeed, patients without recalibration should have an approximately constant number of GEs over time, while an increase should be observed among patients with recalibration. We introduced a GEs-based index: the change over time in the number of GEs computed using the difficulty order defined at t_1 (denoted I), defined as follows:

$$I = number\ of\ GE_{s_{order\ t_1}}^{(t_2)} - number\ of\ GE_{s_{order\ t_1}}^{(t1)}$$

where $number\ of\ GE_{s_{order\ t1}}^{(t^*)}$ denotes the number of GEs observed at t^* using the difficulty order defined at t_1. It should be noted that at the second time point, the Guttman errors are computed using the ordering defined at t_1, hence there are not Guttman errors in the conventional sense, but slight adaptations of Guttman errors. The index I can be computed for each individual. We can expect that $I \approx 0$ among patients without RS and $I \geq 0$ among those with recalibration.

The link between recalibration RS and GEs was explored using a simulation study. Different parameters commonly encountered in analyses of PRO data were explored to determine their effect on the values of I. In a second step, we assessed the ability of I to discriminate patients affected by recalibration from others.

Simulation study

Data simulation

We simulated the responses of N individuals to a unidimensional questionnaire composed of J polytomous items with M response categories, numbered from 0 to $M - 1$, at two different time points. Endorsing difficult response categories was assumed to be manifestations of a high latent variable level. The longitudinal partial credit model (LPCM) [19] was chosen to generate

data since it allowed modelling response category probabilities of polytomous items forming a unidimensional scale across time and provided a possibility to simulate recalibration for a changing proportion of individuals. All simulation parameters chosen for data generation are given in Table 1. Additional information on the simulation implementation is provided in Appendix 1.

Recalibration operationalization

Recalibration was operationalized as changes over time in the LPCM difficulty parameters [8, 9]. Recalibration may be uniform (UR: a change in all difficulty parameters of a given item in the same direction and to the same extent) or nonuniform (NUR: changes occur in various directions and intensities).

At the second time point t_2, recalibration was simulated as follows:
- Only one type of recalibration (UR or NUR) with the same size per data set was considered.
- The proportion p of the sample that was affected by recalibration was variable.
- The items affected by this recalibration were randomly selected (the same for all individuals affected by recalibration).

To generate the responses of the simulated patients affected by UR at t_2, all the difficulty parameters of the item(s) affected by recalibration decreased (-1), making the associated response categories easier. For simulated patients affected by NUR, difficulty parameters were differentially shifted at t_2 by values ranging from 0 to 2η, with $\eta = 1.8$. The first positive response category kept the same difficulty parameter over time, while other categories became more difficult. For simulated patients not affected by RS, the difficulty parameters remained constant over time.

We aimed to investigate the effect of recalibration RS-related factors (such as the number of items with recalibration J_{RS}, the proportion of the sample that was affected by recalibration p, and the recalibration type: UR and NUR) but also more global simulation parameters (the sample size N, the number of items in the questionnaire J, the number of response categories per item M, and the average change in the latent variable over time Δ). Simulation parameters were chosen to be representative of clinical research studies (Table 1).

The combination of all the simulation parameter values led to a total of 810 scenarios, and each of them was replicated 500 times.

Statistical analysis

Within the 500 replications of each scenario, index I was computed for each individual. For each scenario, the boxplots of the 500 mean values of index I obtained respectively among simulated patients affected and not affected by recalibration RS were plotted.

Over all replications, the discriminating abilities associated with the change in the number of GEs over time (i.e., index I) were estimated by the area under the receiver operating characteristic curve (AUROC), where response shift was the response variable and index I was the explanatory variable. Boxplots of the 500 AUROCs were also plotted (one per scenario). Stata software release 15 was used for data generation (*simirt* module) and statistical analyses (StataCorp, 2015). Graphics were realized using the 3.5.3 version of R software (R Core Team, 2019).

Results

Simulation study

A small subset of scenarios is selected to present a representative portrayal of the variability of index I across experimental conditions ($N = 200$, 25% of the sample affected by recalibration, negative average change in the latent variable over time $\Delta = -0.2$; these values were chosen to approach the empirical example). Of note, all results are available in Online Resource 1.

Boxplots of the 500 average values of I among simulated patients with/without recalibration obtained for every scenario where $N = 200$, 25% of the sample was affected by recalibration RS, $\Delta = -0.2$ (negative average change in the latent variable over time) and RS = uniform recalibration are given in Figure 2 according to the number of items affected by recalibration (J_{RS}), the number of items (J) and the number of response categories/item (M). Graphs under the same simulation conditions but for scenarios with nonuniform recalibration are given in Figure 3.

Among simulated patients not affected by recalibration, the means of I fluctuated around values close to 0, regardless of the scenario considered. A slight increase in the means of I could, however, be observed when J and M increased. Among simulated patients with UR, the means

of I fluctuated around positive values. These values remained low for scenarios with $M = 4$; however, they rose sharply when M increased. This rise was larger when J and J_{RS} were large. For simulated patients affected by NUR ($\eta = 1.8$), similar effects as those observed among simulated individuals affected by UR were observed for the average index values, but the trends were much less pronounced.

For each scenario, the dispersion of the means of I increased with J and M; the larger the overall number of response categories was, the wider the range of possible values for the number of GEs. This dispersion decreased logically as N increased.

Boxplots of the 500 AUROCs obtained for every scenario where $N = 200$, 25% of the sample was affected by recalibration RS, and $\Delta = -0.2$ (negative average change in the latent variable over time) are given in Figure 4 (for uniform recalibration) and Figure 5 (for nonuniform recalibration) according to the number of items affected by recalibration (J_{RS}), the number of items (J) and the number of response categories/item (M).
For UR, the discriminating abilities of I appeared to be low over all scenarios, particularly when $M = 4$. Indeed, in these cases, the average AUROC remained under 0.60. A slight increase could nonetheless be observed with increasing M and J_{RS}. For instance, the scenario with 3 items affected by recalibration, $J = 7$ and $M = 10$ resulted in an average AUROC close to 0.70. For NUR, the same phenomena were observed, but the AUROC values were even lower.

Illustrative example

To illustrate these results, we used a longitudinal study called EVALADD, which takes place at the Addictive Medicine and Psychiatry Department of Nantes University Hospital (France). The EVALADD cohort follows patients starting treatment for a behavioral addiction in order to assess the determinants of addictive disorders and, consequently, to improve therapies and preventive strategies. For this analysis, we focused on patients suffering from eating disorders (EDs) included between September 2012 and October 2016 (ED diagnoses were established according to the DSM-IV criteria [20] and explored via the French version of the Mini International Neuropsychiatric Interview [21, 22]). Patients completed self-reported questionnaires, including the Eating Disorder Inventory 2 (EDI-2) [23], at the initiation of nutritional and psychotherapeutic care (t_1) and one year later (t_2). The EDI-2 is an 11-domain scale translated into French and validated by Archinard *et al.* [24]. We focused on the "Drive for thinness" domain since clinicians felt that the corresponding items could potentially be

affected by recalibration after care (recalibration being part of the goals of care). “Drive for thinness” includes 7 items with a six-point Likert response scale ranging from “never” to “always” (1. *I eat sweets and carbohydrates without feeling nervous*; 2. *I think about dieting*; 3. *I feel extremely guilty after overeating*; 4. *I am terrified of gaining weight*; 5. *I exaggerate or magnify the importance of weight*; 6. *I am preoccupied with the desire to be thinner*; 7. *If I gain a pound, I worry that I will keep gaining*). We only considered patients who fully completed the scale at both time points (209 patients of the 210 who attended both visits). We computed the change in the number of GEs (index *I*) for each patient. Then, we assessed the association between index *I* and the covariates collected at baseline that seemed relevant to clinicians involved in the cohort (i.e. sociodemographic data, ED characteristics, psychiatric comorbidities [20–22], and character traits measured by the Temperament and Character Inventory [25–27]). We assessed the associations between index *I* and the other covariates with Mann-Whitney and Kruskal-Wallis tests (for categorical covariates) and Spearman's rank correlation coefficient r (for quantitative covariates). Due to the low discriminating ability of index *I*, no strong association was expected.

Across the sample, the average age at baseline was 24.2 ($\mathrm{sd} = 8.8$), and most patients were women (93%). Of the 209 patients, 31% suffered from restricting anorexia nervosa, 13% from binge eating/purging anorexia nervosa, 25% had bulimia nervosa, 6% displayed binge eating disorder and 24% had eating disorders not otherwise specified (i.e., did not meet the criteria for other diagnoses). The ED had, on average, started 7.3 years before ($\mathrm{sd} = 8.2$), and the average BMI was 18.7 kg/m² ($\mathrm{sd} = 5.4$).

Index *I* was associated with baseline “Drive for thinness” score ($r = 0.28$, p-value < 0.001) and ideal BMI ($r =$ -0.17, p-value = 0.017). These results indicated that higher values of index *I* are associated with greater concern about body image, weight, and shape. In addition, patients with current mood disorders at baseline showed lower distribution of index *I* than patients without (median = 0 *versus* 1.5, p-value = 0.036). No other significant association was noticed (Table 2). These results may suggest several clinical hypotheses related to RS. Indeed, patients without mood disorders might be more receptive to interventions targeting cognitive distortions related to body image, and they may therefore be more prone to RS. In addition, among patients with high concern about body image, weight, and shape, care might tend to focus more on deconstructing cognitive distortions and thus induce RS. However, these associations remain globally weak in our sample.

Discussion

Main results

Some evidence of the link between recalibration RS and the change over time in the number of GEs was found in our simulation study. Indeed, as expected, the GEs-based index I remained on average close to 0 among simulated patients not affected by RS, while its average values increased among simulated patients with recalibration. However, the performance of I depended on M, J_{RS}, and the type of recalibration.

The best results were obtained within scenarios with UR. Indeed, when $M > 4$, substantial differences between the means of I among simulated patients with/without recalibration were noticed. These differences were larger when J and J_{RS} were large. When $M = 4$, the index I had lower performance: differences between the means of I among simulated patients with/without RS were small or even nonexistent.

This might be due to the difficulty parameters of the LPCM used to simulate data. Indeed, when $M = 4$, they were widely spaced, and shifts over time (among individuals with recalibration) did not impact the ordering of difficulty parameters. Thus, the difficulty order observed at t_1 still fitted the responses at t_2 of patients with recalibration, resulting in a stable number of GEs over time. However, within scenarios with $M = 7$ or 10, gaps between the difficulty parameters narrowed. Therefore, in these scenarios, shifts over time among individuals with recalibration did impact the ordering of the difficulty parameters (leading to an increase in the number of Guttman errors for these simulated patients). Size of UR used for the simulations might not be detectable with index I when $M = 4$. This issue is problematic and limiting since several domains within QoL questionnaires (SF-36, QLQ-C30…) are composed of items with four or fewer response categories [30, 31]. Additional UR sizes should hence be explored.

When NUR was simulated, differences between the means of I among simulated patients with and without recalibration RS were less marked than with UR. Trends noticed with UR were still observed but less pronounced. Several reasons might explain these results. First, when $M = 4$, the argument evoked for UR concerning the widely spaced difficulty parameters also applies. In addition, we operationalized NUR by differentially shifting the difficulty parameters of the response categories above "1". Thus, the response category "1" of items affected by recalibration kept the same difficulty parameter over time, and some of the shifts for categories above "1" were very small. It might also have hampered the generation of discrepancies at t_2.

In addition, unfortunately, the random selection of items affected by recalibration led to shift response categories among the most difficult to make them even more difficult. Again, this has probably hampered the generation of discrepancies at t_2. The effect of the position of the response categories affected by recalibration should hence be explored.

Limitations and perspectives

In the simulation study, we decided to focus on recalibration to determine whether index *I* is sensitive to this RS manifestations. However, shifts in GEs could be the result of other types of RS (*i.e.* reprioritization and reconceptualization), thus further investigations are needed to determine if index *I* is sensitive or insensitive to other RS manifestations.

In addition, we assumed that changes over time in the number of GEs were due to RS, but phenomena other than RS can also interfere in the real world. For instance, a change in the individual latent variable level can impact the range of the possible values for the number of GEs and hence potentially interfere with index *I*. Within the simulation study, three configurations were considered regarding the mean change in the latent variable level over time (no change in the average latent variable level, an average decrease of 0.2 in the latent variable level and an average increase of 0.2). The results were very similar among these 3 conditions, but it would be worth investigating larger size of change. The normed number of GEs (i.e., the number of GEs divided by the maximum number of GEs that was achievable given the patient's score and the difficulty order considered) [18] could also be a path to follow to take into account changes in the latent variable at the individual level; the index would hence be the change in the normed number of GEs (denoted I_{norm}). The results for this index within the scenarios emphasized in this article are given in Online Resource 2. It is important to note that we remained under the situation where the questionnaire was still adapted to the population at the second time point. If it turns out that the questionnaire is no longer adapted to the studied population at the second time point, this method would not be adequate (the same would be true for other RS detection methods).

Moreover, phenomena other than RS, such as differential item functioning and violation of the local independence assumption, can also interfere with the index. Indeed, if some patients perceive items differently than the majority of the sample at t_1 (interpreted as differential item functioning, DIF [32, 33]), their responses might result in numerous GEs from the very first measurement occasion. In this case, we may wonder if the changes in the number of GEs is due to RS, DIF, or another phenomenon. In addition, we assumed within the simulation study that

the assumption of local independence (at a time point and across time t_1 and time t_2) holds. However, several forms of violation can occur in the real world. For instance, at one time point, two types of violations can occur. First, the targeted latent variable alone may not be sufficient for explaining the correlations among some subsets of items. This violation is referred to as *trait dependence* and is a type of dimensionality issue [34] because additional unmodeled latent variables are involved. Second, the response to one item may depend on the response given to another item. Such an issue is a violation of statistical independence and is referred to as *response dependence* [34]. Both phenomena might impact the number of GEs but in opposite directions. Response dependence, by increasing the similarity of the individuals' responses, might induce a decrease in the number of GEs. In contrast, trait dependence, as an additional source of variation, might induce an increase in the number of GEs [35]. However, these violations of local independence are likely to occur at both time points, limiting the impact on the index. Several diagnosis and detection methods for local dependence exist within Rasch measurement theory (see for instance [35–37]), item response theory ([38–44]) and the nonparametric item response theory framework ([45]). In addition, across measurement occasions, the correlations among an individual's responses might be more important than what the latent variable can explain. For instance, it can be easier to endorse an item when it has already been endorsed before. This phenomenon is also a violation of local independence (response dependence across time points). Olsbjerg and Christensen argued that such a violation could lead to spurious evidence of recalibration and reprioritization (designated by the term "item parameter drift" in their work) [46]. Local dependence across time points has been operationalized as changes in item difficulty parameters over time, depending on the responses given at the first time point [46, 47]. Following this operationalization, violation of local independence could also lead to changes in the number of GEs, resulting in the same manifestation as that for RS. SAS macros, which are available to test the assumption of local independence across time points and item parameter invariance over time within IRT and RMT models at the sample level (based on likelihood ratio tests) [48, 49], can be used to test these assumptions.

Intermittent missing data (MD) were left out of this simulation study as the number of GEs cannot be determined for patients who did not respond to all items. Intermittent MDs are, however, commonly encountered in clinical research and psychometrics. In this case, the normed number of GEs could be computed on the subset of nonmissing items for each patient.

Finally, we simulated samples that were partly affected by only one type of recalibration at a time (UR or NUR) with the same size of RS for all affected patients. The individual nature of this phenomenon was neglected. Simulations with subgroups of patients affected by different sizes and types of recalibration should be explored.

Alone, the change in the number of GEs has a low discriminating ability. However, it could be used as a preliminary analysis (when RS occurrence is suspected) to identify covariates associated with RS or possibly a subgroup of patients more likely to present RS. Therefore, index I may guide the choice of the adequate method for identifying RS and estimate its size (by introducing a covariate or conducting the analysis at the subgroup level). However, the methodology for defining the threshold to classify individuals must still be developed. Indeed, we have shown in our simulation study that patients without RS had a value of I that fluctuated on average around approximately 0, yet some variability was observed, notably when J and M increased. This variability generated an overlap in the distributions of index I for patients with and without RS. This phenomenon makes it difficult to define a threshold for the index (which would likely be a function of J and M). To overcome the effect of the questionnaire structure on the threshold, the normed number of GEs could be used instead of the number of GEs since it takes into account the maximum number of GEs reachable for each individual given the questionnaire structure.

Conclusion

Some evidence of the link between RS and the change in GEs was found in this study. GEs could be a valuable nonparametric tool for RS detection at a more individual level, but further investigation is needed.

References

1. Basch, E. (2017). Patient-Reported Outcomes - Harnessing Patients' Voices to Improve Clinical Care. *The New England Journal of Medicine*, *376*(2), 105–108. https://doi.org/10.1056/NEJMp1611252
2. Schwartz, C. E., Finkelstein, J. A., & Rapkin, B. D. (2017). Appraisal assessment in patient-reported outcome research: methods for uncovering the personal context and meaning of quality of life. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *26*(3), 545–554. https://doi.org/10.1007/s11136-016-1476-2
3. Sprangers, M. A. G., & Schwartz, C. E. (1999). Integrating response shift into health-related quality of life research: a theoretical model. *Social Science & Medicine*, *48*(11), 1507–1515. https://doi.org/10.1016/S0277-9536(99)00045-3
4. 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. Leplege, S. Briancon, E. Spitz, & J. Coste (Eds.), *Perceived Health and Adaptation in Chronic Disease* (1st ed.). Abingdon, Oxon ; New York, NY : Routledge, 2017.: Routledge.
5. Schwartz, C.E., Sprangers, M.A., and Fayers, P.M. (2005). Response shift: You know it's there, but how do you capture it? Challenges for the next phase of research. In *Assessing quality of life in clinical trials* (2nd Edition.). Oxford; New-York, NY: Oxford University Press.
6. Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *14*(3), 587–598.

7. Schwartz, C. E. (2016). Introduction to special section on response shift at the item level. *Quality of Life Research, 25*(6), 1323–1325. https://doi.org/10.1007/s11136-016-1299-1
8. Guilleux, A., Blanchin, M., Vanier, A., Guillemin, F., Falissard, B., Schwartz, C. E., … Sébille, V. (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. https://doi.org/10.1007/s11136-014-0876-4
9. Blanchin, M., Guilleux, A., Hardouin, J.-B., & Sébille, V. (2020). Comparison of structural equation modelling, item response theory and Rasch measurement theory-based methods for response shift detection at item level: A simulation study. *Statistical Methods in Medical Research, 29*(4), 1015–1029. https://doi.org/10.1177/0962280219884574
10. Vanier, A., Sébille, V., Blanchin, M., Guilleux, A., & Hardouin, J.-B. (2015). Overall performance of Oort's procedure for response shift detection at item level: a pilot simulation study. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *24*(8), 1799–1807. https://doi.org/10.1007/s11136-015-0938-2
11. Nolte, S., Mierke, A., Fischer, H. F., & Rose, M. (2016). On the validity of measuring change over time in routine clinical assessment: a close examination of item-level response shifts in psychosomatic inpatients. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *25*(6), 1339–1347. https://doi.org/10.1007/s11136-015-1123-3
12. Gandhi, P. K., Schwartz, C. E., Reeve, B. B., DeWalt, D. A., Gross, H. E., & Huang, I.-C. (2016). An item-level response shift study on the change of health state with the rating of asthma-specific quality of life: a report from the PROMIS(®) Pediatric Asthma Study. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 25*(6), 1349–1359. https://doi.org/10.1007/s11136-016-1290-x

13. 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: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *25*(6), 1361–1383. https://doi.org/10.1007/s11136-015-1195-0
14. 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: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *23*(9), 2421–2430. https://doi.org/10.1007/s11136-014-0699-3
15. Blanchin, M., Sébille, V., Guilleux, A., & Hardouin, J.-B. (2016). The Guttman errors as a tool for response shift detection at subgroup and item levels. *Quality of Life Research*, *25*(6), 1385–1393. https://doi.org/10.1007/s11136-016-1268-8
16. Meijer, R. R., Niessen, A. S. M., & Tendeiro, J. N. (2016). A Practical Guide to Check the Consistency of Item Response Patterns in Clinical Research Through Person-Fit Statistics: Examples and a Computer Program. *Assessment*, *23*(1), 52–62. https://doi.org/10.1177/1073191115577800
17. Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, Calif.: SAGE.
18. Emons, W. H. M. (2008). Nonparametric Person-Fit Analysis of Polytomous Item Scores. *Applied Psychological Measurement*, *32*(3), 224–247. https://doi.org/10.1177/0146621607302479
19. Fischer, G. H., & Ponocny, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika*, *59*(2), 177–192. https://doi.org/10.1007/BF02295182

20. American Psychiatric Association, & American Psychiatric Association (Eds.). (2009). *Diagnostic and statistical manual of mental disorders: DSM-IV-TR* (4. ed., text revision, 13. print.). Arlington, VA: American Psychiatric Assoc.
21. Lecrubier, Y., Sheehan, D., Weiller, E., Amorim, P., Bonora, I., Harnett Sheehan, K., … Dunbar, G. (1997). The Mini International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: reliability and validity according to the CIDI. *European Psychiatry*, *12*(5), 224–231. https://doi.org/10.1016/S0924-9338(97)83296-8
22. Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., … Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of Clinical Psychiatry, 59 Suppl 20*, 22-33;quiz 34-57.
23. Garner, D. M. (1991). *Eating disorder inventory-2. Professional manual.* Odessa, Florida: Psychological Assessment Research.
24. Archinard, M., Rouget, P., Painot, D. Liengme, C. (2002). Inventaire des troubles alimentaires 2 [Eating Disorder Inventory 2]. In M. Bouvard & J. Cottraux (Eds.), *Protocoles et échelles d'évaluation en psychiatrie et en psychologie [Protocols and evaluation scales in psychiatry and psychology]* (3rd ed., pp. 249–251). Paris: Masson.
25. Cloninger, C. R., Przybeck, T. R., & Svrakic, D. M. (1994). *The temperament and character inventory (TCI) a guide to its development and use*. St. Louis, Mo.: Center for Psychobiology of Personality, Washington University.
26. Pélissolo, A., & Lépine, J.-P. (1997). Traduction française et premières études de validation du questionnaire de personnalité TCI. [Validation study of the French version of the TCI.]. *Annales Médico-Psychologiques*, *155*(8), 497–508.
27. Chakroun-Vinciguerra, N., Faytout, M., Pélissolo, A., & Swendsen, J. (2005). Validation française de la version courte de l'Inventaire du Tempérament et du Caractère (TCI-125).

Journal de Thérapie Comportementale et Cognitive, *15*(1), 27–33. https://doi.org/10.1016/S1155-1704(05)81209-1

28. Cooper, P. J., Taylor, M. J., Cooper, Z., & Fairbum, C. G. (1987). The development and validation of the body shape questionnaire. *International Journal of Eating Disorders*, *6*(4), 485–494. https://doi.org/10.1002/1098-108X(198707)6:4<485::AID-EAT2260060405>3.0.CO;2-O
29. Rousseau, A., Knotter, A., Barbe, P., Raich, R., & Chabrol, H. (2005). [Validation of the French version of the Body Shape Questionnaire]. *L'Encephale*, *31*(2), 162–173. https://doi.org/10.1016/s0013-7006(05)82383-8
30. Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Medical Care*, *30*(6), 473–483.
31. Aaronson, N. K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N. J., … de Haes, J. C. (1993). The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute*, *85*(5), 365–376. https://doi.org/10.1093/jnci/85.5.365
32. Holland, P. W., & Wainer, H. (Eds.). (1993). Differential item functioning. *Differential item functioning.*, xv, 453–xv, 453.
33. Osterlind, S., & Everson, H. (2009). *Differential Item Functioning*. 2455 Teller Road, Thousand Oaks California 91320 United States of America: SAGE Publications, Inc. https://doi.org/10.4135/9781412993913
34. Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement*, *9*(3), 200–215.

35. Christensen, K. B., Kreiner, S., & Mesbah, M. (Eds.). (2013). *Rasch models in health*. London: ISTE [u.a.].
36. Andrich, D., & Kreiner, S. (2010). Quantifying Response Dependence Between Two Dichotomous Items Using the Rasch Model. *Applied Psychological Measurement*, *34*(3), 181–192. https://doi.org/10.1177/0146621609360202
37. Andrich, D., Humphry, S. M., & Marais, I. (2012). Quantifying Local, Response Dependence Between Two Polytomous Items Using the Rasch Model. *Applied Psychological Measurement*, *36*(4), 309–324. https://doi.org/10.1177/0146621612441858
38. Yen, W. M. (1984). Effects of Local Item Dependence on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement*, *8*(2), 125–145. https://doi.org/10.1177/014662168400800201
39. Chen, W.-H., & Thissen, D. (1997). Local Dependence Indexes for Item Pairs Using Item Response Theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265. https://doi.org/10.2307/1165285
40. Hoskens, M., & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods*, *2*(3), 261–277. https://doi.org/10.1037/1082-989X.2.3.261
41. Douglas, J., Kim, H. R., Habing, B., & Gao, F. (1998). Investigating Local Dependence with Conditional Covariance Functions. *Journal of Educational and Behavioral Statistics*, *23*(2), 129–151. https://doi.org/10.2307/1165318
42. Ip, E. H. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika*, *66*(1), 109–132. https://doi.org/10.1007/BF02295736
43. Ip, E. H. (2002). Locally dependent latent trait model and the dutch identity revisited. *Psychometrika*, *67*(3), 367–386. https://doi.org/10.1007/BF02294990

44. Edwards, M. C., Houts, C. R., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods*, *23*(1), 138–149. https://doi.org/10.1037/met0000121

45. Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2016). Using Conditional Association to Identify Locally Independent Item Sets. *Methodology*, *12*(4), 117–123. https://doi.org/10.1027/1614-2241/a000115

46. Olsbjerg, M., & Christensen, K. B. (2015). Modeling local dependence in longitudinal IRT models. *Behavior Research Methods*, *47*(4), 1413–1424. https://doi.org/10.3758/s13428-014-0553-0

47. Marais, I. (2009). Response dependence and the measurement of change. *Journal of applied measurement, 10*, 17–29.

48. Olsbjerg, M., & Christensen, K. B. (n.d.). LIRT: SAS macros for longitudinal IRT models, 49.

49. Olsbjerg, M., & Christensen, K. B. (2015). %lrasch_mml : A *SAS* Macro for Marginal Maximum Likelihood Estimation in Longitudinal Polytomous Rasch Models. *Journal of Statistical Software*, *67*(Code Snippet 2). https://doi.org/10.18637/jss.v067.c02

Appendix 1: Simulation Implementation

Longitudinal Partial Credit Model

The longitudinal Partial Credit Model (LPCM) was chosen to generate data since it allowed modelling response categories probabilities of polytomous items forming a unidimensional scale across time, and provided a possibility to simulate RS for a changing proportion of patients. The probability of patient n to answer m $(= 0, \dots, M-1)$ on item j at time t under the LPCM is given by:

$$P\left(X_{nj}^{(t)} = m \middle| \theta_n^{(t)}, \delta_{j1}^{(t)}, \dots, \delta_{jM-1}^{(t)}\right) = \frac{\exp\left(m.\theta_n^{(t)} - \sum_{p=1}^{m} \delta_{jp}^{(t)}\right)}{\sum_{l=0}^{M-1} \exp\left(l.\theta_n^{(t)} - \sum_{p=1}^{l} \delta_{jp}^{(t)}\right)}$$

Where:

$X_{nj}^{(t)}$ denotes the response to the item $j = 1, \dots, J$ of the individual n at time t

$\theta_n^{(t)}$ stands for the latent variable level of the individual n at t (realization of the random variable Θ).

$$\begin{pmatrix}\Theta^{(t_1)} \\ \Theta^{(t_2)}\end{pmatrix} \sim N\left(\begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}, \Sigma = \begin{bmatrix}\sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2\end{bmatrix}\right)$$

$\delta_{jm}^{(t)}$ is the difficulty of the response category $m = 1, \dots, M-1$ from item j at the time point t. If $\delta_{jm}^{(t)}$ is low, the proportion of patients scoring m or more to item j will be high: m is hence an easy response category (vice versa for difficult response categories). Null response categories do not have a difficulty parameter.

At the first measurement occasion, difficulty parameters were chosen to be spaced along the latent variable continuum (assumed normally distributed, with a zero mean and a standard deviation equaled to 1). For each item j, the difficulty parameter of the first positive response category (denoted $\delta_{j1}^{(t1)}$) equaled the $\frac{j}{J+1}th$ quantile from a $N(0,1)$. Difficulty parameters of the following response categories were then regularly shifted from the first one: $\delta_{jm}^{(t_1)} = \delta_{j1}^{(t_1)} + (m-1) \times \frac{2}{M-2}$. Finally, difficulty parameters of all items were centered on the mean $\bar{\delta} = \frac{\sum_{j,m} \delta_{jm}^{(t_1)}}{J(M-1)}$ so that difficulty parameters were centered on the mean of the latent variable distribution (*i.e.* 0). It hence corresponded to the situation where the questionnaire is suitable

for a population with a latent variable following a standard normal distribution. At the first measurement occasion, the model is a rating scale model.

Recalibration operationalization

To simulate the responses of patients affected by UR at t_2, we choose to shift by -1 all the difficulty parameters of the item(s) affected by recalibration, making all response categories easier. For patients affected by NUR, difficulty parameters were differentially shifted by values ranging 0 to 2η, with $\eta = 1.8$: the first positive response category kept the same difficulty parameter over time, while other categories became more difficult. Finally, we kept the difficulty parameters constant over time to simulate the responses of patients not affected by RS.

$$\text{for all } m \text{ in } \{1, \dots, M-1\},\ \delta_{jm}^{(t_2)} = \begin{cases} \delta_{jm}^{(t_1)} + \eta_m & \text{for individuals affected by RS} \\ \delta_{jm}^{(t_1)} & \text{for individuals not affected by RS} \end{cases}$$

For UR, $\eta_m^{UR} = -1$ for all m in $\{1, \dots, M-1\}$

$$\text{For NUR, } \eta_m^{NUR} = \begin{cases} \frac{(m-1)\eta}{m} & \text{if} \quad 1 \leq m < \frac{M}{2} \\ \eta & \text{if} \quad m = \frac{M}{2} \\ \frac{(M-m+1)\eta}{M-m} & \text{if} \quad \frac{M}{2} < m \leq \text{M-1} \end{cases} \quad \text{where } \eta = 1.8$$

Fig. 1 Percent stacked barchart showing the proportion of patients choosing each response category (0, 1, 2, 3) for each item (item 1, item 2, item 3, item 4) at t_1 (fictitious example)

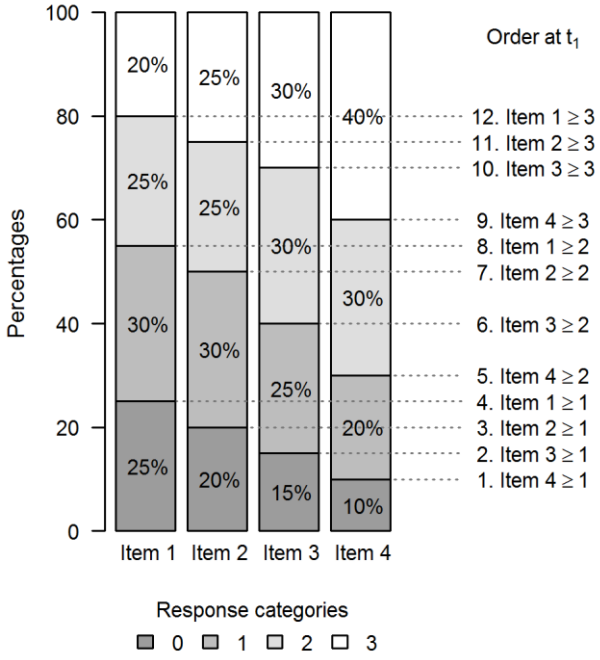

Fig. 2 Boxplots of the 500 mean values of index I obtained for each scenario among simulated patients affected by response shift (in white) and among simulated patients not affected by response shift (in grey). Each pair of boxplots corresponds to one scenario. Subset of scenarios considered: N = 200 (sample size), p =25% (proportion of patients affected by response shift), $\Delta = -0.2$ (average change in the latent variable over time); uniform recalibration (UR).

RS: Response shift; J: number of items; M: number of response categories per item**.**

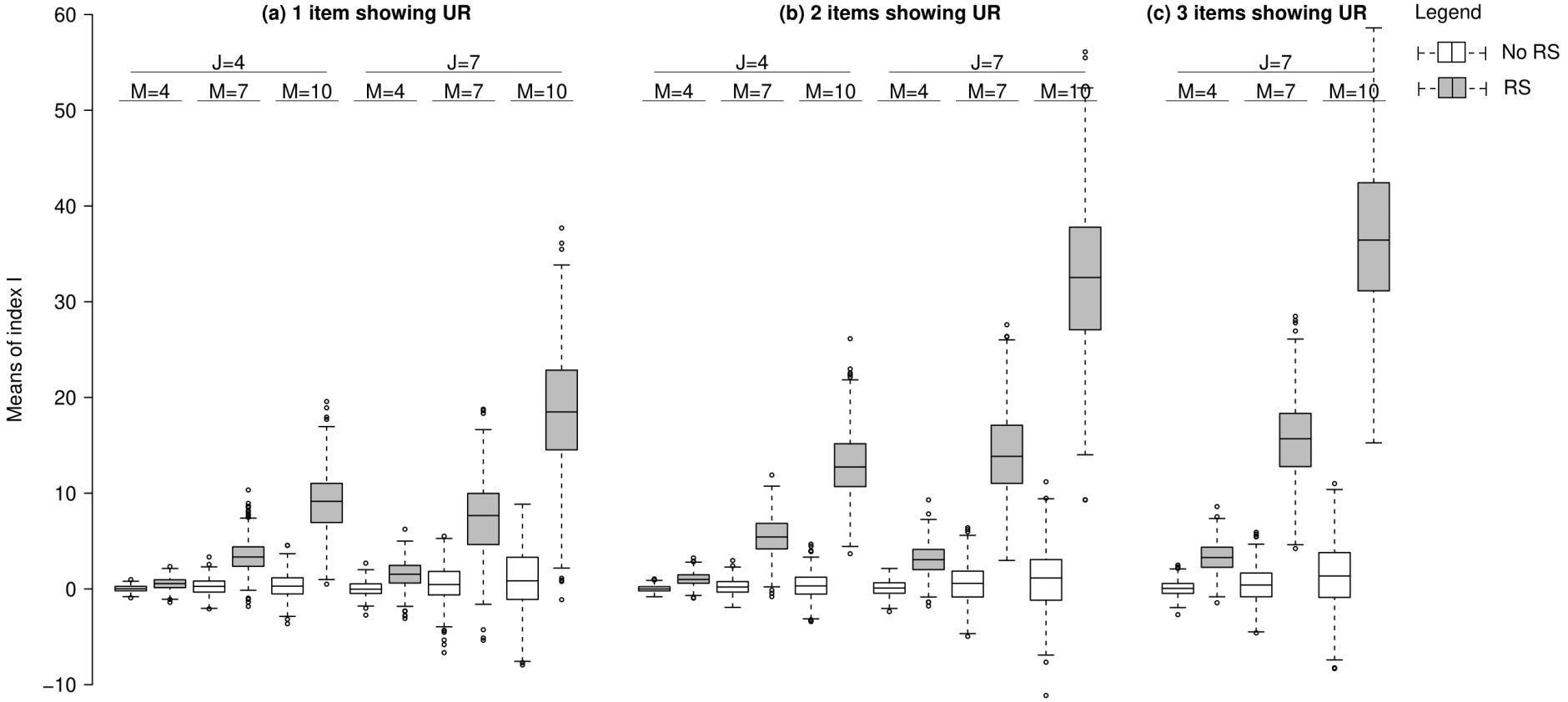

Fig. 3 Boxplots of the 500 mean values of index I obtained for each scenario among simulated patients affected by response shift (in white) and among simulated patients not affected by response shift (in grey). Each pair of boxplots corresponds to one scenario. Subset of scenarios considered: N = 200 (sample size), p =25% (proportion of patients affected by response shift), $\Delta = -0.2$ (average change in the latent variable over time); nonuniform recalibration (NUR).

RS: Response shift; J: number of items; M: number of response categories per item**.**

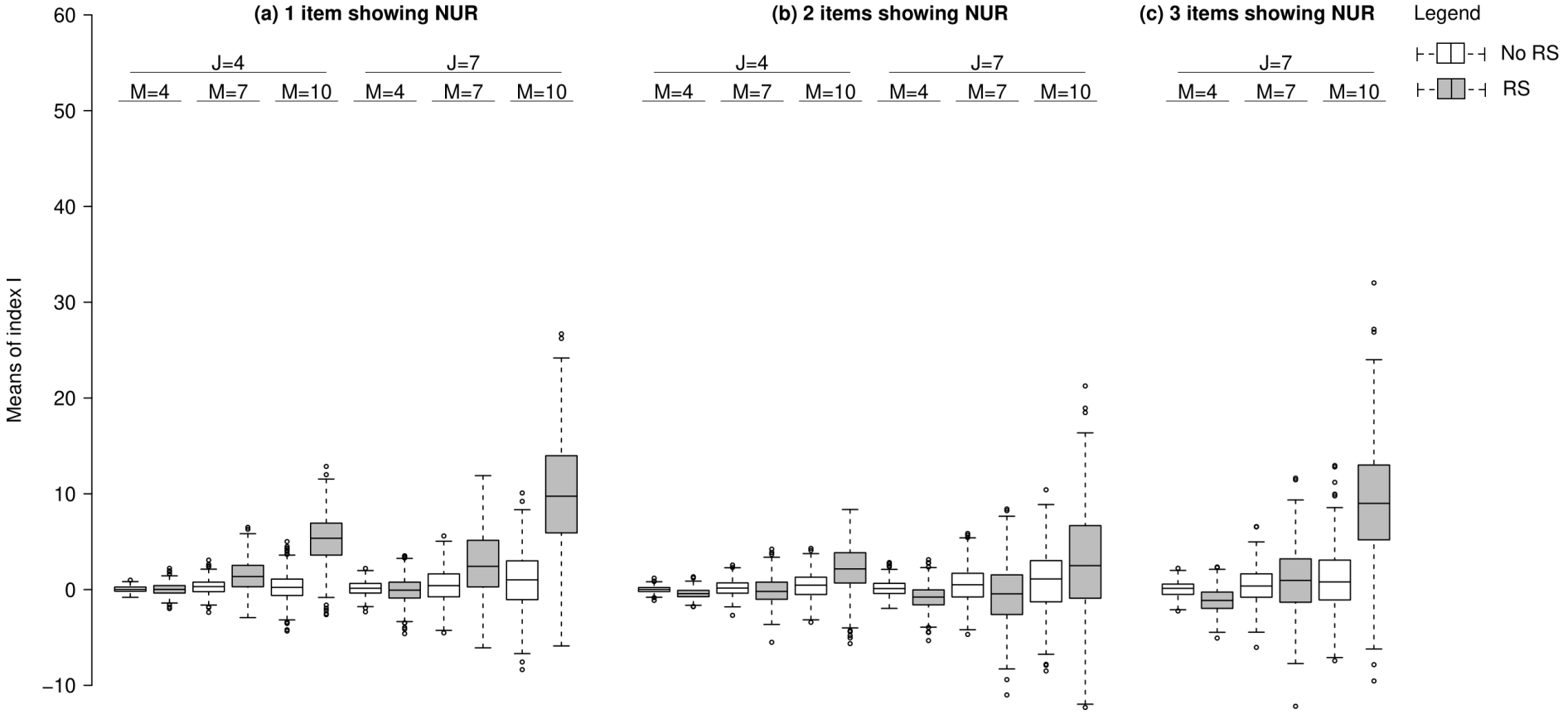

Fig. 4 Boxplots of the 500 AUROCs associated with I obtained for each scenario. Subset of scenarios considered: N = 200 (sample size), $p = 25\%$ (proportion of patients affected by response shift), $\Delta = -0.2$ (average change in the latent variable over time), uniform recalibration (UR). AUROC: area under the receiver operating characteristic curve; J: number of items; M: number of response categories per item.

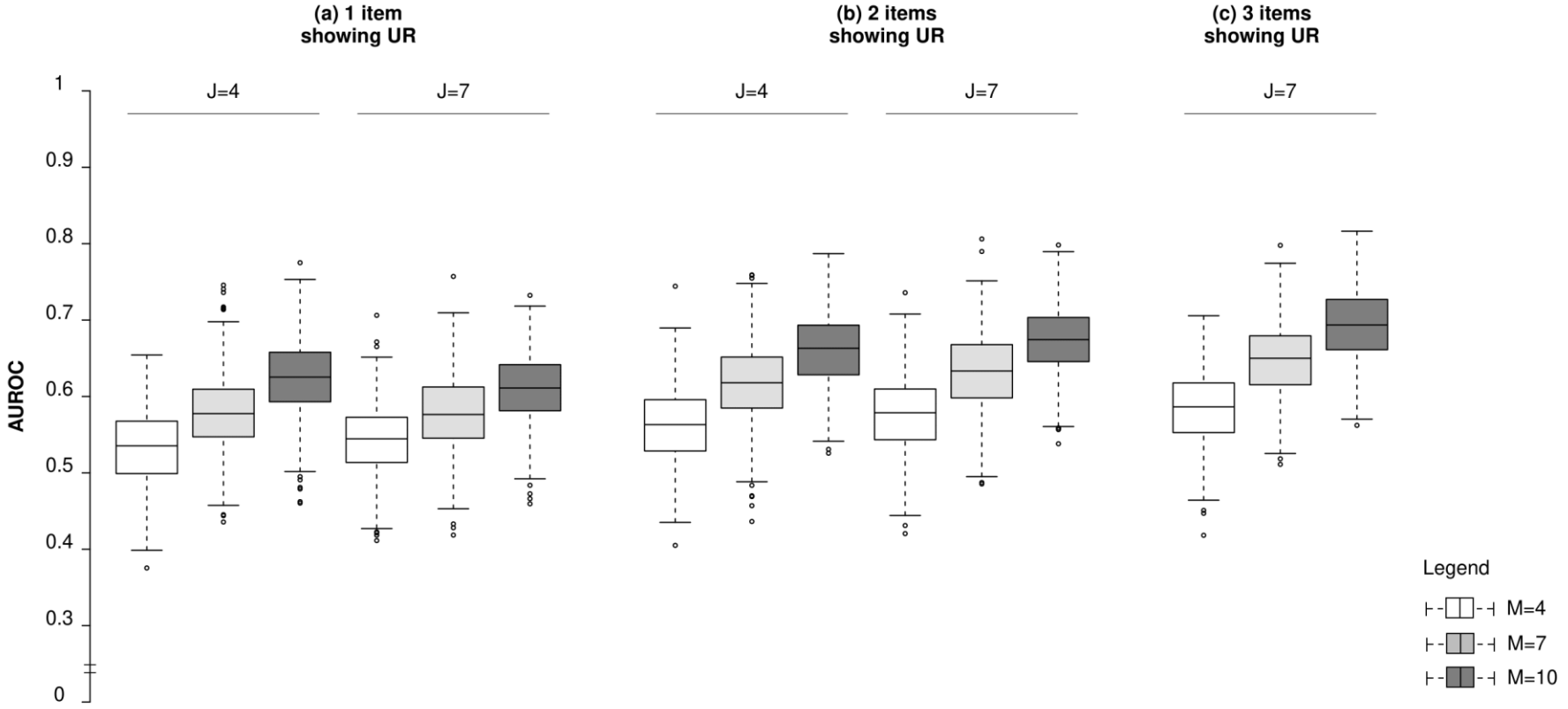

Fig. 5 Boxplots of the 500 AUROCs associated with I obtained for each scenario. Subset of scenarios considered: N = 200 (sample size), $p = 25\%$ (proportion of patients affected by response shift), $\Delta = -0.2$ (average change in the latent variable over time), nonuniform recalibration (NUR). AUROC: area under the receiver operating characteristic curve; J: number of items; M: number of response categories per item.

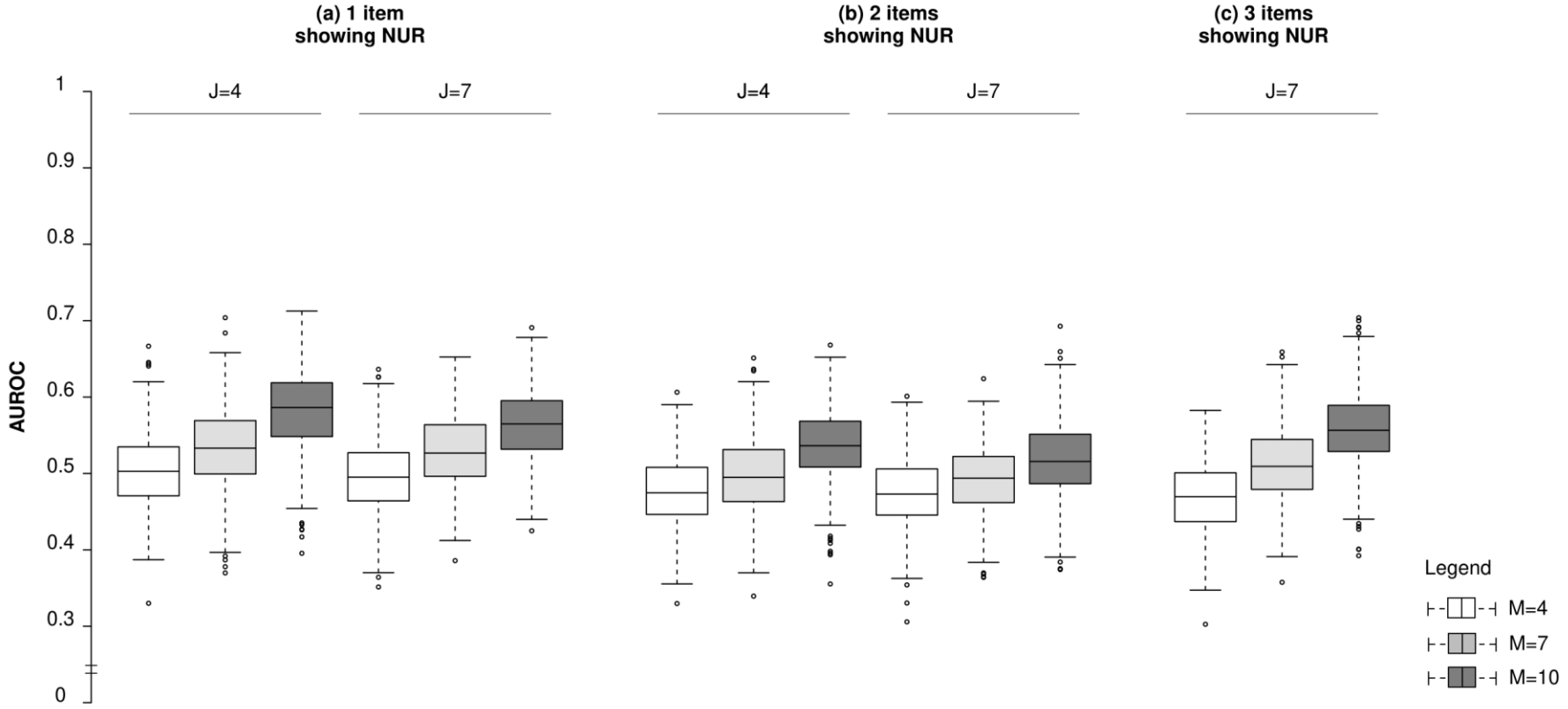

Table 1 Simulation parameters

Sample and questionnaire	
Sample size (N)	$N = 100; 200; 300$
Proportion of patients affected by recalibration (p)	$p = 0.25; 0.5; 0.75$
Number of items (J)	$J = 4; 7$
Number of response categories/item (M)	$M = 4; 7; 10$
Latent variable (θ)	
Mean at time t_1 (μ_1)	$\mu_1 = 0$
Mean change ($\Delta = \mu_2 - \mu_1$)	$\Delta = -0.2; 0; 0.2$
Variance (σ_1^2, σ_2^2)	$\sigma_1^2 = \sigma_2^2 = 1$
Covariance between the two measurement occasions ($\sigma_{1,2}$)	$\sigma_{1,2} = 0.6$
Recalibration response shift size (η)	
UR	$\eta = -1$
NUR	$\eta = 1.8$
Items selected to show recalibration	
$J = 4$	
1 item affected	Item 3
2 items affected	Items 3 and 4
$J = 7$	
1 item affected	Item 5
2 items affected	Items 6 and 7
3 items affected	Items 4, 6 and 7

Table 2 Association between index I (the change in the number of Guttman errors) and baseline characteristics (N = 209)

	Index *I* *Categorical variable: median (Q1 ; Q3)* *Quantitative variable: Spearman's r*	**NA**	**p-value**
Socio-demographic:			
Gender		0	0.770
Female ($n = 195$)	1.2 (-11.0 ; 17.0)		
Male ($n = 14$)	2.0 (-28.0 ; 19.5)		
Age (years)	-0.01	0	0.865
Eating disorder characteristics:			
Type		0	0.930
AN-R ($n = 65$)	0.0 (-14.0 ; 20.0)		
AN-BP ($n = 27$)	2.0 (-8.0 ; 17.0)		
BN ($n = 53$)	5.0 (-11.0 ; 15.0)		
BED ($n = 13$)	-7.0 (-8.0 ; 16.0)		
EDNOS ($n = 51$)	2.0 (-12.5 ; 16.5)		
Duration (years)	-0.06	0	0.368
Lowest BMI (kg/m²)	-0.05	0	0.439
Ideal BMI (kg/m²)	-0.17	8	0.017
Current BMI (kg/m²)	-0.04	1	0.554
Drive for thinness score [a]	0.28		<0.001
Body shape concerns [b]		0	0.070
No to moderate body shape concern ($n = 122$)	-4.0 (-15.5 ; 16.5)		
Marked body shape concerns ($n = 87$)	5.0 (-5.5 ; 18.0)		
Psychiatric comorbidities:			
Current anxiety disorder		0	0.914
No ($n = 137$)	0.0 (-13.5 ; 20.5)		
Yes ($n = 72$)	2.0 (-11.0 ; 15.0)		
Current mood disorder		0	0.036
No ($n = 86$)	0.0 (-12.0 ; 17.0)		
Yes ($n = 123$)	1.5 (-11.5 ; 17.5)		
Self-reported character trait [c]***:***			
Cooperativeness	0.01	0	0.869
Self-transcendence	0.03	0	0.671
Self-Directedness	0.03	0	0.710

NA: number of missing observations, *Q1*: first quartile, *Q3*: third quartile, *n:* number of patients
AN-R: anorexia nervosa restricting subtype, *AN-BP:* anorexia nervosa binge eating or purging subtype, *BN:* bulimia nervosa, *BED:* binge eating disorder, *EDNOS:* eating disorders not otherwise specified
BMI: body mass index

[a] A high score indicates a strong search for thinness
[b] Evaluated by the BSQ: Body Shape Questionnaire [28, 29]
[c] Measured by the Temperament and Character Inventory, a high score indicates a more pronounced character trait

Declarations

Funding:

Y. Dubuy received a national grant from the French Ministry of Higher Education, Research and Innovation. The EVALADD cohort is sponsored by Nantes University Hospital (CHU Nantes).

Conflicts of interest:

Authors declare that they have no conflict of interest.

Availability of data and material:

Modules, scripts and an extract of the simulated data used in the paper are available at the Open Science Framework via the link:
https://osf.io/h9nyd/?view_only=b196db78f31c4e9fbb07013342a133a2

Compliance with ethical standards :

The EVALADD cohort (Investigator: M. Grall-Bronnec) was approved by the local Research Ethics Committee (Groupe Nantais d'Ethique dans le Domaine de la Santé), by the CCTIRS (Comité Consultatif sur le Traitement de l'Information en matière de Recherche dans le domaine de la Santé) and by the CNIL (Commission Nationale de l'Informatique et des Libertés). All participants provided written informed consent (for under 18-year-olds, a legal representative provided informed consent), in accordance with the Helsinki declaration.