Published in final edited form as:
Stat Med. 2015 November 30; 34(27): 3622–3636. doi:10.1002/sim.6559.

Numeric Score-Based Conditional and Overall Change-in-Status Indices for Ordered Categorical Data

Robert H. Lyles[1,*], **Lawrence L. Kupper**[2], **Huiman X. Barnhart**[3], and **Sandra L. Martin**[4]

[1] Department of Biostatistics and Bioinformatics, The Rollins School of Public Health of Emory University, 1518 Clifton Rd. N.E., Mailstop 1518-002-3AA, Atlanta, GA 30322

[2] Department of Biostatistics, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7420

[3] Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC 27710

[4] Department of Maternal and Child Health, Gillings School of Global Public Health, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7445

Abstract

Planned interventions and/or natural conditions often effect change on an ordinal categorical outcome (e.g., symptom severity). In such scenarios it is sometimes desirable to assign *a priori* scores to observed changes in status, typically giving higher weight to changes of greater magnitude. We define change indices for such data based upon a multinomial model for each row of a c×c table, where the rows represent the baseline status categories. We distinguish an index designed to assess conditional changes within each baseline category from two others designed to capture overall change. One of these overall indices measures expected change across a target population. The other is scaled to capture the proportion of total possible change in the direction indicated by the data, so that it ranges from −1 (when all subjects finish in the least favorable category) to +1 (when all finish in the most favorable category). The conditional assessment of change can be informative regardless of how subjects are sampled into the baseline categories. In contrast, the overall indices become relevant when subjects are randomly sampled at baseline from the target population of interest, or when the investigator is able to make certain assumptions about the baseline status distribution in that population. We use a Dirichlet-multinomial model to obtain Bayesian credible intervals for the conditional change index that exhibit favorable small-sample frequentist properties. Simulation studies illustrate the methods, and we apply them to examples involving changes in ordinal responses for studies of sleep deprivation and activities of daily living.

Keywords

Change score; Dirichlet; Multinomial; Ordinal data

[*] rlyles@sph.emory.edu.

1. INTRODUCTION

Paired data commonly arise from intervention studies with an ordinal outcome, e.g., based on assessing subjects before and after an intervention [1,2]. This typically results in a square table, with cell counts representing the numbers of subjects beginning and ending in each of the possible pairs of before and after categories. Such tables can arise when each category is inherently ordinal (e.g., 'mild', 'moderate', 'severe'), or when ordinal categories are defined by categorizing an underlying continuous response. One potential way of analyzing such ordinal data is to apply a weighted kappa coefficient [3,4], to assess the extent of agreement between 'before' and 'after' categories. Such agreement measures are not directly appropriate when the goal is to meaningfully evaluate the extent of change. Therefore, it can be valuable and intuitive to consider measures specifically designed to capture the magnitude of change in the response, while taking into account the ordinal structure of the data; for a recent example, see [5].

A source of debate regarding the characterization of ordinal data is the issue of whether and when numeric scores should be applied to categories. For example, one might assign scores such as (1, 2, 3) or (1, 2, 4) to 'mild', 'moderate', and 'severe', respectively. While concerns about this practice are long-standing given the difficulty in quantifying inter-categorical distances for ordinal data (e.g., [6]), the possibility of using numeric scores continues to be acknowledged in the modern literature. For example, Agresti [2] notes that doing so makes simple and interpretable quantitative measures available to the investigator, while cautioning that sensitivity analysis may be advisable to assess the consistency of conclusions across a reasonable range of scores. Along related lines, Podgor et al. [7] begin with several possible sets of scores for R×C tables with ordered row and column categories, and combine test statistics based on the different sets into a single efficient test of association. Another approach in the context of hypothesis testing is to let the data themselves determine scores (e.g., [8]), with common examples including midranks and ridits [2,9]. Some authors (e.g., [10]) caution against such data-driven scores, suggesting a preference for *a priori* choices based on researcher experience and subject-matter considerations.

In studies of psychology and physical function, the use of Likert scaling approaches [11] is a common approach. In such settings, researchers are generally reluctant to assign equal numeric score changes to step-by-step increases in functionality [12,13], especially when instruments for assessment involve a multitude of items that may have very different impacts upon daily life despite shared category descriptions (e.g., 'mildly' vs. 'moderately' impaired). Such concerns have motivated consideration of complex rank-based nonparametric measures of change for ordered categorical data [14], in order to avoid the assignment of numeric scores to categories. When hypothesis testing to compare treatments is the focus, other authors [15] propose nonparametric approaches to boost power by attributing greater weight to changes across multiple categories.

The approach taken in this article is to seek the simplicity and interpretability offered by numeric scores, assuming they are applied to *changes* rather than to the row and column categories themselves. We assume that the investigator has complete freedom to choose the score assigned to each type of change (e.g., from category j to category k), although these change scores may be varied in the interest of sensitivity analysis if desired. We stipulate that any change scores to be utilized should be determined *a priori*, i.e., prior to examining the data [10]. Although developed and motivated independently, one of our suggested indices of overall change has close connections to a recent proposal by Ferreira et al. [5]. We consider a number of issues and details that expand upon and mitigate concerns that were subsequently expressed [13,16] about the index developed in [5], while attempting to formalize and broaden the potential uses of such measures of change.

In what follows, we first define a conditional index representing the mean change score for subjects who begin in a particular ordinal baseline category. We discuss estimation of this index and its variance assuming a multinomial model for the cell counts in that baseline category, and we demonstrate the utility of a Bayesian approach to obtain credible intervals for the index that possess favorable frequentist properties in small samples. We then consider overall indices of change for the target population of interest. In particular, we define an index that captures the overall mean change score, and then we propose a second index scaled so as to represent the proportion of total possible change in the direction suggested by the data. We discuss how the process of sampling subjects into the baseline categories can impact the validity of estimators of these overall indices, and may suggest the need for sensitivity analysis or the incorporation of external data to permit estimation of baseline category prevalences. All the proposed indices are estimated using previously published real data examples, and we study the properties of these estimators and their proposed standard errors and confidence intervals via simulations.

2. METHODS

2.1. A change index conditional on the baseline category

Consider a c×c table, where the rows and columns (both numbered 1 to c) represent “before” and “after” ordinal categories, e.g., at baseline and after an intervention applied in the same manner to each subject. **Table 1** indicates this setup and establishes notation for probabilities associated with baseline category membership, cells, and transitions from before to after categories.

Table 2 establishes corresponding notation for cell counts, row marginal totals, and investigator-specified change scores corresponding to each transition from category j to category k.

The probabilities in Table 1 are defined as follows: ψ_j = Pr(“Before” category = j), $\pi_{k|j}$ = Pr(“After” category=k | “Before” category=j), and π_{jk} = Pr(“Before” category=j and “After” category=k) (j,k=1,..., c). These probabilities are subject to anticipated sum constraints, i.e., $\sum_{k=1}^{c} \pi_{k|j}=1$ for all j, and $\sum_{j=1}^{c}\sum_{k=1}^{c} \pi_{jk}=\sum_{j=1}^{c} \psi_j=1$. However, we assume that some of these probabilities may not be estimable depending on the sampling strategy employed. In particular, we expect in practice that sampling of subjects into the baseline categories will often be non-random (e.g., there may be oversampling in more or less severe categories). In such cases, one may require assumptions or external data to estimate the ψ_j's and π_{jk}'s, which has implications with regard to the two overall change indices that we propose (see

Section 2.3). What we do assume throughout is that the subjects whose data appear in row j (j=1,..., c) of Table 2 are representative of those in that baseline category, thus ensuring estimability of the conditional $\pi_{k|j}$ probabilities in Table 1.

We assume as a format convention for Table 2 that a move from before category j to after category k is a positive change ("improvement") when k > j and a negative change ("deterioration") when k < j. Also, assume that the change scores s_{jk} (j,k=1,..., c) are defined by the investigator *a priori* (i.e., before examining the data). These scores reflect the value attributed to a particular change, and will most likely be > 0 for positive changes and < 0 for negative changes, with increasing magnitude for larger changes. Commonly, s_{jj} may be 0 (j=1,..., c) when there is no change; however, the investigator is free to assign each score to reflect clinical or subject matter considerations. Thus, for example, a subject who begins and finishes in a "mild" category may earn a 0 or positive score, while one who begins and finishes in a "severe" category could be given a 0 or negative score. Accounting for the ordinal nature of the categories, we assume that $s_{j1} \le s_{j2} \le \ldots \le s_{jc}$.

We begin by defining a change index that is conditional on the baseline category (j):

$$\theta_{\mathrm{j}}=\sum_{\mathrm{k}=1}^{\mathrm{c}} \mathrm{s}_{jk}\pi_{\mathrm{k|j}}, \quad (1)$$

(j,k =1,..., c). Note that θ_j ranges from the least favorable score (s_{j1}) to the most favorable score (s_{jc}) in row j for the extreme cases where $\pi_{1|j}$ =1 and $\pi_{c|j}$=1, respectively. Defining the random variable S_j to represent the change score for a subject who begins in category j, we may also represent this conditional change index as follows:

$$\theta_{\mathrm{j}}=\mathrm{E}\left(\mathrm{S}_{\mathrm{j}}\right) \quad (2)$$

That is, θ_j is the expected change score for those in baseline category j.

Regardless of whether or not subjects are selected randomly from the target population or whether or not there is over-sampling or under-sampling of those in certain baseline categories, θ_j is easily estimable given our assumption of random sampling within each row of Table 2. Thus, we may estimate θ_j unbiasedly based on (1) as

$$\hat{\theta}_{\mathrm{j}}=\sum_{\mathrm{k}=1}^{\mathrm{c}} \mathrm{s}_{jk}\hat{\pi}_{\mathrm{k|j}}, \quad (3)$$

where $\hat{\pi}_{\mathrm{k|j}}=\mathrm{n}_{jk}/\mathrm{n}_{\mathrm{j}}$, or equivalently based on (2) as

$$\hat{\theta}_{\mathrm{j}}=\overline{\mathrm{S}}_{\mathrm{j}}=\mathrm{n}_{\mathrm{j}}^{-1}\sum_{\mathrm{k}=1}^{\mathrm{c}} \mathrm{n}_{jk}\mathrm{s}_{jk} \quad (4)$$

Expression (4) is the sample mean change score for subjects in row j. Note that expressions (3) and (4) suggest two alternative approaches for estimating the standard error of $\hat{\theta}_{\mathrm{j}}$ (see Section 2.4).

2.2. Overall indices of change

We propose two overall change indices, the first of which can be written as follows:

$$\theta_{\mathrm{A}}=\sum_{\mathrm{j}=1}^{\mathrm{c}}\psi_{\mathrm{j}}\theta_{\mathrm{j}}=\sum_{\mathrm{j}=1}^{\mathrm{c}}\psi_{\mathrm{j}}\left(\sum_{\mathrm{k}=1}^{\mathrm{c}}\mathrm{s}_{jk}\pi_{\mathrm{k}|\mathrm{j}}\right)\quad(5)$$

Note that θ_A is a weighted average of the row-specific conditional indices of change, with weights equal to the corresponding baseline category prevalances ψ_j. An equivalent representation of θ_A is as the overall expected change score for the target population, namely:

$$\theta_{\mathrm{A}}=\mathrm{E}\left(\mathrm{S}\right)=\sum_{\mathrm{j}=1}^{\mathrm{c}}\sum_{\mathrm{k}=1}^{\mathrm{c}}\mathrm{s}_{jk}\pi_{jk},\quad(6)$$

(j,k =1,..., c), where S represents the random change score for an arbitrary subject.

With the ψ_j's assumed known, random sampling of subjects within each baseline category yields an unbiased estimator of θ_A upon inserting $\hat{\theta}_{\mathrm{j}}$ from (3) or (4) in place of each unknown θ_j in (5). In practice, one may estimate θ_A unbiasedly based on an overall random sample from the target population by incorporating the $\hat{\theta}_{\mathrm{j}}$'s and also replacing ψ_j in (5) by $\hat{\psi}_{\mathrm{j}}=\mathrm{n}_{\mathrm{j}}/\mathrm{N}$, where $\mathrm{N}=\sum_{\mathrm{j}=1}^{\mathrm{c}}\mathrm{n}_{\mathrm{j}}$ is the total sample size. Equivalently, one could replace π_{jk} in (6) by $\hat{\pi}_{jk}=\mathrm{n}_{jk}/\mathrm{N}$ (j,k=1,..., c).

The second proposed index of overall change is a scaled version of θ_A, as follows:

$$\begin{gathered}\theta_{\mathrm{B}}=\omega\theta_{\mathrm{A}}=\omega\sum_{\mathrm{j}=1}^{\mathrm{c}}\sum_{\mathrm{k}=1}^{\mathrm{c}}\mathrm{s}_{jk}\pi_{jk}=\omega\sum_{\mathrm{j}=1}^{\mathrm{c}}\psi_{\mathrm{j}}\theta_{\mathrm{j}},\\ \text{where we define}\quad\omega=\left(\sum_{\mathrm{j}=1}^{\mathrm{c}}\mathrm{s}_{jc}\psi_{\mathrm{j}}\right)^{-1}\quad\text{if}\quad\theta_{\mathrm{A}}>0,\\ \omega=-\left(\sum_{\mathrm{j}=1}^{\mathrm{c}}\mathrm{s}_{\mathrm{j}1}\psi_{\mathrm{j}}\right)^{-1}\quad\text{if}\quad\theta_{\mathrm{A}}<0,\quad\text{and}\quad\omega=1\quad\text{if}\quad\theta_{\mathrm{A}}=0.\end{gathered}\quad(7)$$

The scaling factor ω ensures the desirable property that θ_B takes the value + 1 (− 1) if all subjects in the population finish in the most (least) favorable "after" category. It also lends an intuitive representation to θ_B, making it interpretable as the proportion of total possible directional change achieved by the population. The implication of "directional" here is that, overall, the population tends toward positive change when $\theta_A > 0$, and toward negative change when $\theta_A < 0$.

If the ψ_j's are assumed known, then the scaling factor ω is also known and we have the unbiased estimator $\hat{\theta}_{\mathrm{B}}=\omega\sum_{\mathrm{j}=1}^{\mathrm{c}}\psi_{\mathrm{j}}\hat{\theta}_{\mathrm{j}}$. More realistically, if $\hat{\psi}_{\mathrm{j}}=\mathrm{n}_{\mathrm{j}}/\mathrm{N}$ based on a random sample into the baseline categories, we have

$$\hat{\theta}_{\mathrm{B}}=\hat{\omega}\sum_{\mathrm{j}=1}^{\mathrm{c}}\hat{\psi}_{\mathrm{j}}\hat{\theta}_{\mathrm{j}}, \quad (8)$$

where now $\hat{\omega}$ is a stochastic scaling factor obtained by replacing ψ_j by $\hat{\psi}_j$ and θ_A by $\hat{\theta}_A$ in equation (7) and in its set of accompanying conditions. The estimator in equation (8) is similar to an index that was proposed previously [5], except the latter measured change only in a single direction (either positive or negative) rather than accounting for changes in both directions simultaneously. This feature of $\hat{\theta}_B$ helps to alleviate a primary concern that was raised with regard to the existing index of change [5,13].

2.3. Estimating overall change indices with non-random sampling into baseline categories

While random sampling into the baseline categories is an ideal design strategy for estimating the overall population parameters θ_A and θ_B, it is common for the data in Table 2 to arise in other ways. For example, many studies preferentially recruit subjects in “mild” or “severe” categories by design or for convenience. In such cases, estimators of the indices in (5) and (7) could be severely biased in reference to a target population if we replace the baseline prevalences ψ_j by $\hat{\psi}_j=n_j/N$.

There are at least three options to consider in this case. First, assume the investigator has knowledge of the sampling rates applied to recruit subjects into the c baseline groups, at least relative to an index category (e.g., category 1). Then, he or she could apply adjustments to the $\hat{\psi}_j$'s to be used in estimating θ_A and θ_B. Specifically, taking the first baseline group as the index category, assume we know the values ρ_{11}, ρ_{12}, ρ_{13},..., ρ_{1c}, where $\rho_{1j} = p_{s1}/p_{sj}$ (j=1,..., c) and p_{sj} is the probability that a subject is sampled from the target population given that this subject is in baseline category j (j=1,.., c). Then the observed row totals (n_j) in Table 2 can be used to estimate the true underlying baseline category prevalences (ψ_j), as follows:

$$\hat{\psi}_{\mathrm{j}}=\frac{\rho_{1\mathrm{j}}\mathrm{n}_{\mathrm{j}}}{\mathrm{n}^{*}} \quad (9)$$

(j=1,..., c), where $\mathrm{n}^{*}=\sum_{\mathrm{j}=1}^{\mathrm{c}}\rho_{1\mathrm{j}}\mathrm{n}_{\mathrm{j}}$. Note that this follows because, on average, we expect $n_j/p_{sj} = \rho_{1j}n_j/p_{s1}$ subjects to be selected into baseline category j under random sampling. One would estimate the row-specific change indices (θ_j) and corresponding standard errors and confidence intervals (CIs) in the usual way based on the original data in Table 2 (see next section). However, the adjusted $\hat{\psi}_j$'s in (9) would be used to compute estimates of the overall change indices (θ_A and θ_B) and in accompanying standard error calculations. We provide an example to illustrate this approach in Section 3.2.

Secondly, lacking knowledge of relative sampling rates, one could apply sensitivity analyses by varying the ψ_j's in (5) and (7) over plausible ranges to produce a sense of corresponding variation in the estimated overall indices. Again, we refer to Section 3.2 for a brief example.

Finally, one could incorporate estimated $\hat{\psi}_\mathrm{j}$'s based on an external sample from a comparable target population. Ideally, this sample would be random, with baseline category-specific cell counts ($\mathrm{n}_{\mathrm{j,ex}}$) available to permit estimating the variance-covariance matrix of the external $\hat{\psi}_\mathrm{j}$'s. Standard errors to accompany the estimated overall indices in this scenario, and under the ideal strategy of random sampling, are considered in the next section and in Appendix 1.

2.4. Standard errors and confidence intervals for proposed change indices

For the variance of the estimated conditional index $\hat{\theta}_\mathrm{j}$, we first consider expression (3) under a conditional multinomial model for the cell counts in row j of Table 2. That is, letting N_{jk} represent the random cell count occurring in column k, we assume

$$\mathbf{N}_\mathrm{j}=(\mathrm{N}_{\mathrm{j}1},\mathrm{N}_{\mathrm{j}2},\ldots,\mathrm{N}_{jc})^{'}\sim Multinomial\left(\mathrm{n}_\mathrm{j},\boldsymbol{\pi}_\mathrm{j}\right),$$

where $\boldsymbol{\pi}_\mathrm{j}^{'}=\left(\pi_{1|\mathrm{j}},\pi_{2|\mathrm{j}},\ldots,\pi_{\mathrm{c}|\mathrm{j}}\right)$ is a (1×c) row vector estimated as $\hat{\boldsymbol{\pi}}_\mathrm{j}^{'}=\left(\hat{\pi}_{1|\mathrm{j}},\hat{\pi}_{2|\mathrm{j}},\ldots,\hat{\pi}_{\mathrm{c}|\mathrm{j}}\right)$. It follows that $\mathrm{E}\left(\hat{\boldsymbol{\pi}}_\mathrm{j}\right)=\boldsymbol{\pi}_\mathrm{j}$ and $Var\left(\hat{\boldsymbol{\pi}}_\mathrm{j}\right)=\boldsymbol{\Sigma}_\mathrm{j}$ is the (c×c) matrix with kth diagonal element $\mathrm{n}_\mathrm{j}^{-1}\pi_{\mathrm{k}|\mathrm{j}}\left(1-\pi_{\mathrm{k}|\mathrm{j}}\right)$ and off-diagonal element $\left(\mathrm{k},\mathrm{k}^{'}\right)=-\mathrm{n}_\mathrm{j}^{-1}\pi_{\mathrm{k}|\mathrm{j}}\pi_{\mathrm{k}'|\mathrm{j}}$, for (k, k′) = 1,...,c . We then have

$$Var\left(\hat{\theta}_\mathrm{j}\right)=Var\left(\mathbf{s}_\mathrm{j}^{'}\hat{\boldsymbol{\pi}}_\mathrm{j}\right)=\mathbf{s}_\mathrm{j}^{'}\boldsymbol{\Sigma}_\mathrm{j}\mathbf{s}_\mathrm{j},$$

where $\mathbf{s}_\mathrm{j}^{'}=(\mathrm{s}_{\mathrm{j}1},\mathrm{s}_{\mathrm{j}2},\ldots,\mathrm{s}_{jc})$ is the (1×c) vector containing the change scores in row j of Table 2. This yields an initial standard error estimator, i.e.,

$$\mathrm{S}\hat{E}_\mathrm{I}\left(\hat{\theta}_\mathrm{j}\right)=\sqrt{\mathbf{s}_\mathrm{j}^{'}\hat{\boldsymbol{\Sigma}}_\mathrm{j}\mathbf{s}_\mathrm{j}},\quad (10)$$

with the $\hat{\pi}_{\mathrm{k}|\mathrm{j}}$'s inserted into Σ_j.

An alternative motivated by expression (4) is to simply calculate the usual standard error associated with the mean change score $\bar{\mathrm{S}}_\mathrm{j}$ in row j, i.e.,

$$\mathrm{S}\hat{E}_{II}\left(\hat{\theta}_\mathrm{j}\right)=\sqrt{\sum_{\mathrm{k}=1}^{\mathrm{c}}\mathrm{n}_{jk}(\mathrm{s}_{jk}-\bar{\mathrm{s}}_\mathrm{j})^2/\left[\mathrm{n}_\mathrm{j}\left(\mathrm{n}_\mathrm{j}-1\right)\right]}=\sqrt{\hat{\boldsymbol{\sigma}}_{S_j}^2/\mathrm{n}_\mathrm{j}},\quad (11)$$

where $\hat{\sigma}_{S_j}^2$ is the sample variance of the n_j change scores in row j. While we expect (10) and (11) to be equivalent for large n_j, they will differ in small samples. We compare these two standard error estimators empirically in Section 4.

A standard Wald-type confidence interval (CI) for θ_j is available using either standard error estimate, but we do not expect such a CI to perform well when the sample size in row j is small. This issue has been studied extensively in the case of estimating a binomial proportion [17,18], and one attractive option in that setting is a Bayesian credible interval based on a non-informative Jeffreys (beta) prior. The corresponding approach here is to assume a Dirichlet(½, ½,..., ½) prior for the c cell probabilities in row j, yielding the following posterior distribution for those probabilities:

$$\boldsymbol{\pi}_{\mathrm{j}}|\mathbf{N}_{\mathrm{j}}=\mathrm{n}_{\mathrm{j}} \sim Dirichlet\left(\mathrm{n}_{\mathrm{j1}}+\tfrac{1}{2}, \mathrm{n}_{\mathrm{j2}}+\tfrac{1}{2}, \ldots, \mathrm{n}_{jc}+\tfrac{1}{2}\right) \quad (12)$$

It is simple to obtain a large sample from this posterior by generating sequences of gamma random variables. For each such draw from the posterior distribution of the $\pi_{\mathrm{k|j}}$'s, we may then re-calculate $\theta_{\mathrm{j}}=\sum_{\mathrm{k}=1}^{\mathrm{c}} \mathrm{s}_{jk}\pi_{\mathrm{k|j}}$. The 2.5th and 97.5th sample quantiles of this large sample of θ_j's provides the desired credible interval, which we might expect to exhibit favorable frequentist properties (e.g., [18]). Such Dirichlet-multinomial extensions of the beta-binomial approach have previously been shown effective for interval estimation when targeting the multinomial proportions themselves [19]. In Section 4, we compare this approach with standard Wald-type CIs for the θ_j's calculated using the standard error estimator in eqn. (11).

If subjects are randomly selected into the c baseline categories so that the c^2 cell counts in Table 2 may be viewed as a single multinomial sample, then one can utilize a simple standard error analogous to that in expression (11) in conjunction with $\hat{\theta}_{\mathrm{A}}$ estimated via (6). A Wald-type CI for θ_A is then available, as well as a CI based on the Dirichlet-multinomial approach described above.

However, for a number of reasons we prefer to recommend standard error estimation based on $\hat{\theta}_{\mathrm{A}}$ estimated via expression (5), treating data in the rows of Table 2 as a set of c independent multinomial samples with known or estimated baseline prevalences (ψ_j). First, if sampling into the baseline categories is non-random and one is forced to rely on sensitivity analyses in which the ψ_j's are varied over reasonable ranges, expressing variability in this way based on each assumed set of “known” baseline prevalences is natural. Secondly, if such non-random sampling is employed but with external estimates of the ψ_j's available, such an approach permits adjustments to properly account for the uncertainty in the external estimates. Finally, if sampling is completely random, then we propose an augmented approach that accounts for uncertainty in the ψ_j's and yields a standard error estimate for $\hat{\theta}_{\mathrm{A}}$ that will be very close in value to the analogue of expression (11) for the full table. This augmented approach involves imputing the row-specific sample sizes (n_j, j=1,..., c) and using a version of the well-known multiple imputation variance estimator [20] to accommodate the corresponding uncertainty in the ψ_j's. Details of this approach and slight modifications to estimate the standard error of $\hat{\theta}_{\mathrm{B}}$ are provided in **Appendix 1**.

3. REAL-DATA EXAMPLES

3.1 Activities of daily living

We first consider data from a Swedish study of aging, which investigated the development of dependence in activities of daily living (ADL) among subjects aged 70 and up [21]. The data considered here consist of cell counts indicating change in ADL status between the ages of 73 (‘before’) and 76 (‘after’). The levels of the ordinal ADL variable as assessed by an occupational therapist were fully independent (FI), dependent in instrumental ADL (DI), and dependent in both personal and instrumental ADL(DPI), where the activities included in the instrumental and personal categories are discussed in prior references [14,22]. Svensson [14] proposed rank-invariant nonparametric measures of change and used them to analyze these data.

Table 3 provides the cell counts in the format of Table 2, along with equally-spaced change scores chosen to illustrate the estimated indices considered in Section 2. This choice of scores makes the row specific θ_j's (j=1,2,3) interpretable as the expected number of ADL categories moved for subjects in each baseline status group, while θ_A captures the overall expected number of categories moved (assuming subjects were randomly sampled into the baseline groups). The rightmost column of the table provides estimated θ_j's and standard errors (SE) based on eqn. (11), along with approximate 95% CIs based on the Dirichlet-multinomial approach from Section 2.4 (see eqn. 12). These choices of SE and CI approaches are based on empirical studies, some of which are summarized in Section 4. The table also provides theoretical ranges for each row-specific measure of change (θ_j), which are useful when interpreting the magnitude of each corresponding estimate.

As seen in Table 3, the data reflect a very slight tendency toward improvement in ADL status among those who began in the most dependent category (DPI; $\hat{\theta}_1=0.13$). The corresponding tendency toward deterioration is noticeably greater in magnitude for those beginning in the independent category (FI; $\hat{\theta}_3=-0.33$), while those who began in the intermediate category experienced a small and non-significant tendency toward decline (DI; $\hat{\theta}_2=-0.16$).

If we assume random sampling of the participants into the baseline categories, the data in Table 3 yield the following estimates (SEs) and [CIs] for the two overall indices of change proposed in Section 2.2: $\hat{\theta}_{\mathrm{A}}=-0.293\,(0.031)\,[-0.354,-0.232]$; $\hat{\theta}_{\mathrm{B}}=-0.162\,(0.017)\,[-0.195,-0.129]$. These estimates indicate a significant overall tendency toward greater dependence as subjects aged, which is in qualitative agreement with a previous analysis of the same data using more complex nonparametric measures of change [14]. Our results suggest that subjects declined by approximately 0.3 ADL categories on average, and that the overall observed ADL movement represented approximately 16% of the total possible decline in the population. The Wald-type CIs reported in conjunction with $\hat{\theta}_{\mathrm{A}}$ and $\hat{\theta}_{\mathrm{B}}$ were computed using the standard errors reported with those estimates in Table 3, which were obtained via the approach described in Appendix 1.

3.2 Illustration: Sensitivity analysis and the incorporation of known baseline sampling rates

Using the data in Table 3, we first illustrate a simple sensitivity analysis that could be used in the event that sampling into baseline categories was non-random and there is no knowledge of the relative sampling rates or actual data to inform one about the true baseline prevalences. The goal of such an analysis is to see how the estimated overall θ_A and θ_B

indices and their standard errors vary over a range of assumed values for the vector ψ = (ψ_1, ψ_2, ψ_3). To compute the estimated indices, we insert each assumed set of ψ's into eqns. (5) and (7), while replacing the true row-specific θ_j's by their observed-data estimates in Table 3. Standard errors are then obtained by treating the rows as separate independent multinomial samples and taking square roots of the following expressions:

$$Var\left(\hat{\theta}_{\text{A}}\right)=\sum_{\text{j}=1}^{\text{c}}\psi_{\text{j}}^{2}Var\left(\hat{\theta}_{\text{j}}\right)\quad\text{and}\quad Var\left(\hat{\theta}_{\text{B}}\right)=\omega^{2}Var\left(\hat{\theta}_{\text{A}}\right), \qquad (13)$$

with the assumed ψ's treated as known and where we estimate $Var\left(\hat{\theta}_{\text{j}}\right)$ by applying eqn. (11) to the jth row of the observed table. **Table 4** provides a brief sensitivity analysis of this type.

Table 4 illustrates how the estimated overall indices vary as we move further away from the assumption that the observed row-specific prevalences in Table 3, i.e., (15, 45, 326)/386, were reflective of the true baseline prevalences. The last row indicates that if non-random sampling distorts the apparent prevalences enough, the directionality of the estimated indices can change. Note also that the standard errors in the top row are nearly identical to those we obtained based on Table 3 assuming a random sample, using the approach in Appendix 1 to account for uncertainty in the estimated ψ's. Nevertheless, we recommend the latter approach whenever the observed data permit the analyst to incorporate this uncertainty.

For a second illustration, suppose that sampling into the baseline categories in conjunction with Table 3 was non-random, but the investigator was in control of the relative sampling rates and knows that subjects in ADL categories DI and FI were selected respectively at 5 and 10 times the rate of subjects in the DPI category (row 1). That is, assume relative sampling rates (see Section 2.3) as follows: ρ_{11} = 1, ρ_{12} = 1/5, and ρ_{13} = 1/10. We adjust for such non-random sampling by using eqn. (9) to calculate adjusted estimates of the baseline prevalences, yielding the following: $\hat{\psi}_1=0.265$, $\hat{\psi}_2=0.159$, and $\hat{\psi}_3=0.576$. We can now obtain valid estimates of θ_A and θ_B that are adjusted for non-random sampling by using these new prevalence estimates along with the row-specific change indices $\left(\hat{\theta}_{\text{j}}\right)$ obtained directly from the data in Table 3 [e.g., for $\hat{\theta}_{\text{B}}$, see eqn. (8)]. In so doing, we obtain the following point estimates: $\hat{\theta}_{\text{A}}=-0.180$, $\hat{\theta}_{\text{B}}=-0.137$. These differ rather markedly from the values (−0.293 and −162, respectively) that we obtained when analyzing the data in Table 3 as if they had arisen via a random sample from the target population at baseline.

For standard errors to accompany these new overall change index estimates, one option would be to treat the adjusted $\hat{\psi}_{\text{j}}$'s as known and to utilize them in eqn. (13) as we proposed for sensitivity analysis. However, while the effect may often be slight, such an approach tends to underestimate the true variability due to estimating the true ψ_j's. We provide details in **Appendix 2** for estimating the variance-covariance matrix associated with the set of $\hat{\psi}_{\text{j}}$'s obtained by eqn. (9). This estimated variance-covariance matrix can then be used directly in the standard error estimation approach outlined in Appendix 1. Utilizing this strategy, we obtain the following estimates (standard errors) [Wald-type 95% CIs] for the overall indices upon adjusting for non-random sampling: $\hat{\theta}_{\mathrm{A}} = -0.180\,(0.042)\,[-0.262, -0.098]$, $\hat{\theta}_{\mathrm{B}} = -0.137\,(0.027)\,[-0.190, -0.084]$.

3.3 Time to sleep among insomnia patients

Our second example is based on data from a randomized clinical trial in which investigators compared a placebo to an active drug among patients with insomnia [23]. Outcome data consisted of self-reported time (in minutes) to fall asleep at a baseline and follow-up occasion, where subjects were selected from the four baseline categories < 20, 20-30, 30-60, or > 60 minutes. Here we take an approach suggested by prior authors [1], attributing midpoints (10, 25, 45, 75) to each category but the last. For change scores, we take the differences between these values from baseline to follow-up. **Table 5** shows the corresponding scores and cell counts for both groups in the format of Table 2, where k > j represents improvement (less time to fall asleep), and k < j indicates deterioration.

Using difference scores makes the proposed approach for estimation and inference about the row-specific θ_j's analogous to a paired t test, except for the recommended Dirichlet-multinomial model-based CIs. As seen in Table 5, the θ_j point estimates suggest better average time to sleep changes in the Active group (vs. Placebo) for subjects who began in the 1st, 2nd, and 4th baseline categories, while Placebo group subjects fared somewhat better than Active subjects in the 20-30 minute baseline category. Wald tests suggest that Active subjects in the first two baseline categories of Table 5 experienced significantly better improvement than Placebo subjects in those categories, while there was no significant difference between groups in the final two baseline categories. To avoid confusion, note that positive estimated θ_j values in Table 5 indicate improvement in the sense of a *decrease* in time to sleep (e.g., the estimate of 38.83 in the first row reflects that many minutes less on average to fall asleep at the follow-up occasion).

Assuming random sampling of subjects into the baseline categories, the data in Table 5 yield estimates (SEs) and [CIs] for the overall indices of change for the Active group, as follows: $\hat{\theta}_{\mathrm{A}}=22.18$ minutes (2.61) [16.21, 28.16]; $\hat{\theta}_{\mathrm{B}}=0.555$ (0.044) [0.469, 0.641]. Corresponding results for the Placebo group are: $\hat{\theta}_{\mathrm{A}}=12.96$ (2.19) [8.97, 16.95]; $\hat{\theta}_{\mathrm{B}}=0.321$ (0.048) [0.232, 0.410]. Thus, both groups experienced significantly improved times, but the Active group saw a markedly greater change that represented a higher proportion of their total possible improvement. Two-sided Wald tests for equality of θ_A ($p=0.007$) and θ_B ($p<0.001$) across the two groups support these conclusions.

4. SIMULATIONS

We conducted several simulation studies, primarily to evaluate the standard error estimators and CI procedures discussed in Section 2.4. In each case, we generated row-specific sample sizes (n_j) randomly from a multinomial distribution, and then generated multinomial cell counts within each row. **Table 6** shows the average cell counts targeted in our first simulation study, representing a case with a moderate overall approximate sample size of

N=100. **Table 7** provides the results based on assuming equally-spaced *a priori* change scores (s_{jk} = k-j), for a total of 2500 replications.

As expected, Table 7 reflects the unbiasedness of the estimators $\hat{\theta}_j$ (j=1,..., c). Note the superior overall coverage performance achieved by Dirichlet-multinomial-based CIs (Section 2.4) as compared to Wald-type CIs for the corresponding row-specific parameters, particularly for row 3 (which had the smallest average sample size). The Wald-type CIs were calculated using $\mathrm{S}\hat{E}_{II}\left(\hat{\theta}_j\right)$ in eqn.(11) as opposed to $\mathrm{S}\hat{E}_{\mathrm{I}}\left(\hat{\theta}_j\right)$ in eqn.(10), as we see from Table 7 that the latter multinomial-based SE estimates tend to be slightly optimistic in finite samples while the former match closely on average to the empirical SDs of the θ_j estimates. Finally, the table also reflects a virtual match between the empirical SDs and mean estimated SEs corresponding to the estimated overall change index parameters θ_A and θ_B. These SEs were calculated using the method described in Appendix 1, which is appropriate when subjects are randomly sampled into baseline categories (as simulated here), or when they are not but valid external data are available to provide information about the row-specific proportions (ψ_j). We observe excellent coverage for the corresponding Wald-type CIs employing the proposed MI-type standard errors.

Table 8 summarizes simulations under the same conditions, except where the average count in each cell in Table 6 is divided by 2 (thus corresponding to a small overall average sample size of N=50). Note that the conclusions based on this table are very similar to those based on Table 7, except that Table 8 highlights even more strongly the benefit of Dirichlet-based CIs for the θ_j's when working with small row-specific samples. Nevertheless, Wald-type CIs for the overall indices θ_A and θ_B continue to perform well when based on SEs obtained as described in Appendix 1.

We obtained similar qualitative conclusions based on further simulation scenarios (not summarized here), including those closely mimicking the conditions reflected in the observed example data in Tables 3 and 5.

5. DISCUSSION

We have developed conditional and overall indices of change for paired ordinal data that can be represented by a c×c matrix as in Table 2. The conditional index applies to a given row of the table, and captures change among subjects in a particular “before” (or baseline) category. In contrast, the two proposed overall change indices respectively capture expected movement and the proportion of total possible change achieved in the target population.

We caution that the proposed indices of change depend on assigned *a priori* change scores. The indices should thus be used only if the investigator is comfortable with his or her assigned scores for each possible transition (from category j to k), and/or as part of sensitivity analyses in which these scores are varied across reasonable ranges. In such cases, the general approach taken here offers clear benefits in terms of accessibility, implementation and interpretability. While the need to specify change scores can be viewed as a drawback, it is noteworthy that the approach allows a flexible application of clinical or subject-matter judgment to gauge the relative magnitudes of each possible transition, without necessitating the assignment of scores to each ordinal category. In the absence of strong clinical motivation, the use of equally-spaced change scores is often natural [10], allowing one to interpret the proposed change indices in terms of expected numbers of categories moved. While not detailed here, it is also worth noting that these methods extend with little difficulty to scenarios of r×c tables similar to Table 2. For example, 'after' categories may include 'before' categories (e.g., 'mild', 'moderate', 'severe') as a subset, in addition to new ones (e.g., 'healed' or 'severe with adverse reaction').

The two proposed overall indices of change (θ_A and θ_B) invoke different interpretations that may make one or the other more appealing in a given situation. In particular, θ_A measures average change across subjects in a target population; hence, it can be used to assess the expected benefit to be experienced by a randomly selected individual subsequent to an intervention or shift in conditions. The scaled index θ_B measures the proportion of total possible change achieved by a population. Thus, θ_B may be especially useful, for instance, when there is a need to determine which of two or more target populations will benefit most from making an intervention widely available, or which of two or more programs should yield maximum benefit for a given population. θ_B recalls a similar index found in recent literature [5], but offers greater flexibility for the choice of change scores along with the advantage of capturing movement in either direction.

To supplement the interpretation and estimation of the conditional and overall change indices, we have provided a thorough treatment of standard errors and confidence interval (CI) procedures. This treatment brings to light a number of findings, such as the small-sample benefits of the Dirichlet-multinomial approach to CI estimation when applied to conditional change within baseline categories and the need to accommodate non-random sampling into those categories. With regard to the latter, we treat standard errors in a unified fashion (Appendices 1 and 2) so as to take full advantage of a broad scope of realistic scenarios in which the overall change indices may be estimable. With respect to conditional changes, the Bayesian approach presented is tangential to the purpose of estimating the conditional change index itself; however, it performs very well as an inferential tool in support of that purpose. Specifically, for this and for related problems [17-19], Bayesian credible intervals tend to behave better in terms of frequentist coverage properties than do standard confidence intervals when sample sizes are small. We hope that such sampling and inferential considerations will prove useful not only in the current context, but in other settings in which similar statistical challenges arise.

ACKNOWLEDGEMENTS

This research was supported in part by grants from the National Institute of Environmental Health Sciences (5R01ES012458-07), and the National Center for Advancing Translational Sciences (UL1TR000454).

Appendix 1: Standard Error Estimation Procedure to Accompany θ^A and θ^B

Define **N** = (N_1, N_2,..., N_c) as a random vector of baseline category-specific sample sizes, of which **n**=(n_1, n_2,..., n_c) is the observed realization (see Table 2). We assume **N** ~

multinomial(N, ψ), where ψ=(ψ$_1$, ψ$_2$,..., ψ$_c$), $\hat{\psi}=\mathbf{n}/\mathrm{N}$, and N is the total sample size. Note that **n** and N are obtained directly from the observed data in Table 1 in the event that subjects were randomly sampled into the baseline categories, in which case the $\hat{\psi}_{\mathrm{j}}$'s used to compute $\hat{\theta}_{\mathrm{A}}$ and $\hat{\theta}_{\mathrm{B}}$ are derived accordingly from that table. Otherwise, we assume that **n**$_{ex}$ = (n$_{1,ex}$, n$_{2,ex}$,..., n$_{c,ex}$) and N come from an external sample from the same (or a comparable) target population for which this was the case, and the $\hat{\psi}_{\mathrm{j}}$'s used to compute $\hat{\theta}_{\mathrm{A}}$ and $\hat{\theta}_{\mathrm{B}}$ also come from this sample.

To calculate standard errors, we first generate multiple realizations of the vector **n** based on the approximation that $\hat{\boldsymbol{\psi}} \overset{\bullet}{\sim} MVN\left(\boldsymbol{\psi}, \boldsymbol{\Sigma}\right)$, where Σ is the c×c variance-covariance matrix with multinomial structure corresponding to $\hat{\psi}=\mathbf{n}/\mathrm{N}$. Specifically, we generate ψ$_m$ from $MVN\left(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\Sigma}}\right)$ using a random MVN generator from standard statistical software where the singularity of the $\hat{\boldsymbol{\Sigma}}$ matrix ensures that the elements of ψ$_m$ are constrained to sum to 1. The vector **n**$_m$=(n$_{m1}$, n$_{m2}$,..., n$_{mc}$) is then obtained as Nψ$_m$, for a total of M replications, where N is the total sample size in Table 1.

Upon each such replication of the set of row-specific sample sizes (n$_j$, j=1,..., c) we recalculate the estimated indices $\hat{\theta}_{\mathrm{A}}$ and $\hat{\theta}_{\mathrm{B}}$ as follows:

$$\hat{\theta}_{\mathrm{A,m}}=\sum_{\mathrm{j=1}}^{\mathrm{c}}\psi_{\mathrm{j,m}}\sum_{\mathrm{k=1}}^{\mathrm{c}}\mathrm{s}_{jk}\hat{\pi}_{\mathrm{k|j}} \quad \text{and} \quad \hat{\theta}_{\mathrm{B,m}}=\hat{\omega}_{\mathrm{m}}\hat{\theta}_{\mathrm{A,m}},$$
$$\text{where} \quad \hat{\omega}_{\mathrm{m}}=\left(\sum_{\mathrm{j=1}}^{\mathrm{c}}\mathrm{s}_{jc}\psi_{\mathrm{j,m}}\right)^{-1} \quad \text{if} \quad \hat{\theta}_{\mathrm{A,m}}>0,$$
$$\hat{\omega}_{\mathrm{m}}=-\left(\sum_{\mathrm{j=1}}^{\mathrm{c}}\mathrm{s}_{\mathrm{j}1}\psi_{\mathrm{j,m}}\right)^{-1} \quad \text{if} \quad \hat{\theta}_{\mathrm{A,m}}<0, \quad \text{and} \quad \hat{\omega}_{\mathrm{m}}=1 \quad \text{if} \quad \hat{\theta}_{\mathrm{A,m}}=0.$$

Each replicated value $\hat{\theta}_{\mathrm{A,m}}$ is computed using the $\hat{\pi}_{\mathrm{k|j}}$'s derived from Table 1. The resulting set of M replicated estimates of each index is then used to compute adjusted standard errors, using a slightly modified version of the variance estimator proposed in [20]. In the case of θ$_A$, we compute

$$\bar{\hat{\theta}}_{\mathrm{A}}=\mathrm{M}^{-1}\sum_{\mathrm{m=1}}^{\mathrm{M}}\hat{\theta}_{\mathrm{A,m}} \quad \text{and} \quad \mathrm{V\hat{a}r}\left(\hat{\theta}_{\mathrm{A}}\right)=\mathrm{U}+\left(1+1/\mathrm{M}\right)\mathrm{B},$$

where

$$\mathrm{U}=\mathrm{V\hat{a}r}\left(\hat{\theta}_{\mathrm{A}}|\hat{\boldsymbol{\psi}}=\boldsymbol{\psi}\right) \quad \text{and} \quad \mathrm{B}=\mathrm{M}^{-1}\sum_{\mathrm{m=1}}^{\mathrm{M}}\left(\hat{\theta}_{\mathrm{A,m}}-\bar{\hat{\theta}}_{\mathrm{A}}\right)^{2}.$$

Note conceptually that U is a conditional variance estimate based on the observed data in Table 1 in conjunction with the original $\hat{\psi}_j$ estimates, and the addition of B accounts for added variability due to uncertainty about the true ψ_j's. Specifically, we compute

$$U = \hat{Var}\left(\hat{\theta}_A | \hat{\boldsymbol{\psi}} = \boldsymbol{\psi}\right) = \sum_{j=1}^{c} \hat{\psi}_j^2 \hat{\sigma}_{S_j}^2 / n_j,$$

where again the $\hat{\psi}_j$'s come directly from Table 1 if sampling was random into the baseline categories, and otherwise from the external sample. Note that in the preceding expression for U, we apply a straightforward variance estimator for each row-specific $\hat{\theta}_j$ as reflected in equation (10). The same MI-type variance calculation is applied to derive the standard error to accompany the estimate of θ_B, except in that case we compute

$$U = \hat{Var}\left(\hat{\theta}_B | \hat{\boldsymbol{\psi}} = \boldsymbol{\psi}\right) = \hat{\omega}^2 \sum_{j=1}^{c} \hat{\psi}_j^2 \hat{\sigma}_{S_j}^2 / n_j,$$

where $\hat{\omega}$ is obtained by inserting the $\hat{\psi}_j$'s into the expression for ω that follows equation (7).

The above approach was applied to obtain standard errors for $\hat{\theta}_A$ and $\hat{\theta}_B$ in all example and simulation scenarios presented in the text, except for the illustrations accompanying Tables 3 and 4 (see Section 3.2). When applying the MI-type approach, we performed a total of 10 imputations per dataset (i.e., M=10).

Finally, if subjects were sampled non-randomly into the baseline category groups but according to known relative sampling rates, we recommend this same procedure for estimating standard errors except with alterations to the calculations of $\hat{\boldsymbol{\psi}}$ and $\hat{\boldsymbol{\Sigma}}$ that are used to generate the ψ_m replications and the conditional variance estimates (U). Specifically, $\hat{\boldsymbol{\psi}}$ is computed via equation (9) and $\hat{\boldsymbol{\Sigma}}$ is computed as described in Appendix 2.

Appendix 2: Variance-Covariance Matrix for ψ^j's Assuming Known Relative Sampling Rates

Assume the vector of row-specific cell counts (n_j) in Table 2 arises through non-random sampling into the baseline categories as discussed in Sections 2.3 and 3.2, and that this vector of cell counts is distributed as multinomial(N, ψ^o), with N the total sample size and the superscript "o" denoting "observed" The vector ψ^o is estimated as usual based on the proportions in each row of Table 2, and the estimated c×c variance-covariance matrix $\hat{\boldsymbol{\Sigma}}^o$ associated with the vector of row totals has diagonal elements $N\hat{\psi}_j^o\left(1-\hat{\psi}_j^o\right)$ and off-diagonal elements $-N\hat{\psi}_j^o\hat{\psi}_{j'}^o, \quad \left(j, j'\right) = 1, \ldots, c$.

Accounting for the relative sampling rates (ρ_{1j}, j=1,...,c), the estimated vector (ψ) of true baseline prevalences contains the individual $\hat{\psi}_j$'s defined in eqn. (9), which are nonlinear functions of the observed row totals (n_j). To obtain a c×c estimated variance-covariance matrix $\hat{\Sigma}$ for $\hat{\psi}$, we can apply the multivariate delta method based on the following:

$$\partial\hat{\psi}_{\mathrm{j}}/\partial \mathrm{n}_{\mathrm{j}}=\rho_{1\mathrm{j}}\left(1-\mathrm{n}_{\mathrm{j}}\rho_{1\mathrm{j}}/\mathrm{n}^{*}\right)/\mathrm{n}^{*}\quad(\mathrm{j}=1,\ldots,\mathrm{c})$$

and

$$\partial\hat{\psi}_{\mathrm{j}}/\partial \mathrm{n}_{j'}=-\mathrm{n}_{\mathrm{j}}\rho_{1\mathrm{j}}\rho_{1\mathrm{j}'}/\mathrm{n}^{*2}\quad\left(\mathrm{j}\neq\mathrm{j}'\right)$$

Defining the 1×c vectors $\hat{\mathbf{D}}_{\psi_{\mathrm{j}}}=\left(\partial\hat{\psi}_{\mathrm{j}}/\partial \mathrm{n}_1,\quad \partial\hat{\psi}_{\mathrm{j}}/\partial \mathrm{n}_2,\ldots,\quad \partial\hat{\psi}_{\mathrm{j}}/\partial \mathrm{n}_{\mathrm{c}}\right)$, we obtain

$$\mathrm{V\hat{a}r}\left(\hat{\psi}_{\mathrm{j}}\right)=\hat{\mathbf{D}}_{\psi_{\mathrm{j}}}\hat{\mathbf{\Sigma}}^{\mathrm{o}}\hat{\mathbf{D}}_{\psi_j}^{'}\quad\text{and}\quad\mathrm{C\hat{o}v}\left(\hat{\psi}_{\mathrm{j}},\hat{\psi}_{\mathrm{j}}^{'}\right)=\hat{\mathbf{D}}_{\psi_{\mathrm{j}}}\hat{\mathbf{\Sigma}}^{\mathrm{o}}\hat{\mathbf{D}}_{\psi_{\mathrm{j}'}}^{'},$$

(j, j′) = 1,...,c. The resulting estimated variance-covariance matrix $\hat{\Sigma}$ may then be used directly along with the vector $\hat{\psi}$ of adjusted prevalence estimates within the procedure described in Appendix 1, to obtain appropriate standard errors to accompany the overall change index estimates $\hat{\theta}_{\mathrm{A}}$ and $\hat{\theta}_{\mathrm{B}}$.

REFERENCES

1. Agresti A, Natarajan R. Modeling clustered ordered categorical data: A survey. International Statistical Review. 2001; 69:345–371.
2. Agresti, A. Analysis of Ordinal Categorical Data. 2nd edn. Wiley; New York: 2010.
3. Cohen J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin. 1968; 70:213–220. [PubMed: 19673146]
4. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement. 1973; 33:613–619.
5. Ferreira MLP, Almeida RMVR, Luiz RR. A new indicator for the measurement of change with ordinal scores. Quality of Life Research. 2013; 22:1999–2003. [PubMed: 23435665]
6. Stevens SS. On the averaging of data. Science. 1955; 121:113–116. [PubMed: 13225751]
7. Podgor MJ, Gastwirth JL, Mehta CR. Efficiency robust tests of independence in contingency tables with ordered categories. Statistics in Medicine. 1996; 15:2095–2105. [PubMed: 8896142]
8. Davis, CS. Proceedings of the 13th Annual SAS User's Group International Conference. SAS Institute, Inc.; Cary, NC: 1988. Estimation of row and column scores in the linear-by-linear association model for two way ordinal contingency tables.; p. 946-951.
9. Bross IDJ. How to use ridit analysis. Biometrics. 1958; 14:18–38.
10. Graubard BI, Korn EL. Choice of column scores for testing independence in ordered 2×k contingency tables. Biometrics. 1987; 43:471–476. [PubMed: 3607207]
11. Likert R. A technique for the measurement of attitudes. Archives of Psychology. 1932; 140:5–55.
12. Stucki G, Daltroy L, Katz JL, Johannesson M, Liang MH. Interpretation of change scores in ordinal clinical scales and health status measures: The whole may not equal the sum of the parts. Journal of Clinical Epidemiology. 1996; 49:711–717. [PubMed: 8691219]

13. Varlas A. Commentary on “A new indicator for the measurement of change with ordinal scores”. Quality of Life Research. 2013; 22:2005–2007. [PubMed: 23435666]
14. Svensson E. Ordinal invariant measures for individual and group changes in ordered categorical data. Statistics in Medicine. 1998; 17:2923–2936. [PubMed: 9921610]
15. Bajorski P, Petkau J. Nonparametric two-sample comparisons of changes on ordinal responses. Journal of the American Statistical Association. 1999; 94:970–978.
16. Almeida RMVR, Luiz RR. On “A new indicator for the measurement of change with ordinal scores”. Quality of Life Research. 2013; 22:2009.
17. Agresti A, Coull BA. Approximate is better than “exact” for interval estimation of binomial proportions. The American Statistician. 1998; 52:119–126.
18. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. Statistical Science. 2001; 16:101–133.
19. Sangeetha U, Subbiah M, Srinivasan MR. Estimation of confidence intervals for Multinomial proportions of sparse contingency tables using Bayesian methods. International Journal of Scientific and Research Publications. 2013; 3:1–7.
20. Rubin, DB. Multiple Imputation for Nonresponse in Surveys. Wiley; New York: 1987.
21. Rinder L, Roupe S, Steen B, Svanborg A. Seventy-year-old people in Gothenburg. A population study in an industrialised Swedish city. I. General presentation of the study. Acta Medica Scandinavica. 1975; 198:397–407. [PubMed: 1081814]
22. Sonn U, Grimby G, Svanborg A. Activities of daily living studied longitudinally between 70 and 76 years of age. Disability and Rehabilitation. 1996; 18:91–100. [PubMed: 8869511]
23. Francom SF, Chuang-Stein C, Landis JR. A log-linear model for ordinal data to characterize differential change among treatments. Statistics in Medicine. 1989; 8:571–582. [PubMed: 2727476]

Table 1

Notation for Baseline Prevalence, Transition, and Cell Probabilities[*]

	After Category (k)				
Before Category (j)	**1**	**2**	**...**	**c**	**Baseline prevalence**
1	$\pi_{1\|1}$ π_{11}	$\pi_{2\|1}$ π_{12}	...	$\pi_{c\|1}$ π_{1c}	ψ_1
2	$\pi_{1\|2}$ π_{21}	$\pi_{2\|2}$ π_{22}	...	$\pi_{c\|2}$ π_{2c}	ψ_2
⋮	⋮	⋮	...	⋮	⋮
c	$\pi_{1\|c}$ π_{c1}	$\pi_{2\|c}$ π_{c2}	...	$\pi_{c\|c}$ π_{cc}	ψ_c

[*] $\pi_{k|j}$ denotes transition probability to category k from category j; π_{jk} denotes population probability associated with before category j and after category k; ψ_j denotes population probability associated with before category j (j,k =1,..., c)

Table 2

Notation for Cell Counts and Change Scores[*]

	After Category (k)				
Before Category (j)	**1**	**2**	**...**	**c**	**Row totals**
1	n_{11} (s_{11})	n_{12} (s_{12})	...	n_{1c} (s_{1c})	n_1
2	n_{21} (s_{21})	n_{22} (s_{22})	...	n_{2c} (s_{2c})	n_2
⋮	⋮	⋮	...	⋮	⋮
c	n_{c1} (s_{c1})	n_{c2} (s_{c2})	...	n_{cc} (s_{cc})	n_c

[*] n's represent cell counts; numbers in parentheses (s_{jk}) represent assigned change scores for transitions from category j to category k

Table 3

Summary of ADL Example With Cell Frequencies (%) and Equally-Spaced *a priori* Change Scores (s_{jk} = k-j)

		After ADL category (age 76)				
		DPI	DI	FI	Row totals (n_j's)	$\hat{\theta}_j$ (SE) [95% CI] {Range}*
Before ADL category (age 73)	DPI	13 (0)	2 (1)	0 (2)	15	0.13 (0.09) [0.05, 0.48] {0, 2}
	DI	13 (–1)	26 (0)	6 (1)	45	–0.16 (0.09) [–0.33, 0.03] {–1, 1}
	FI	23 (–2)	62 (–1)	241 (0)	326	–0.33 (0.03) [–0.40, –0.27] {–2, 0}

*Standard errors (SE) obtained via eqn. (11); Approx. 95% CIs based on Dirichlet-multinomial approach leading to eqn. (12); Range indicates feasible theoretical lower and upper bounds for θ_j's

Table 4

Sensitivity Analysis Illustrated Using ADL Example Data

Assumed baseline prevalences	$\hat{\theta}_A$ **(SE)**	$\hat{\theta}_B$ **(SE)**
ψ = (15, 45, 326)/386	–0.293 (0.030)	–0.162 (0.017)
ψ = (1, 2, 7)/9	–0.277 (0.035)	–0.156 (0.020)
ψ = (3, 3, 3)/9	–0.118 (0.045)	–0.118 (0.045)
ψ = (7, 2, 1)/9	0.032 (0.074)	0.018 (0.042)

Table 5

Summary of Sleep Time Example with Difference Scores for Active and Placebo Groups[*]

ACTIVE DRUG GROUP (frequency and *a priori* change score)							
		After					
Before		75 (> 60)	45 (30-60)	25 (20-30)	10 (< 20)	Row totals (n_j's)	$\hat{\theta}_j$ (SE) [95% CI] {Range}[†]
	75 (>60)	8 (0)	13 (30)	17 (50)	9 (65)	47	38.83 (3.14) [32.54, 44.42] {0, 65}
	45 (30-60)	1 (–30)	3 (0)	23 (20)	13 (35)	40	22.13 (2.03) [16.69, 25.04] {–30, 35}
	25 (20-30)	2 (–50)	2 (–20)	5 (0)	11 (15)	20	1.25 (4.66) [–9.88, 7.30] {–50, 15}
	10 (< 20)	0 (–65)	1 (–35)	4 (–15)	7 (0)	12	–7.92 (3.23) [–20.13, –4.64] {–65, 0}
PLACEBO GROUP (frequency and *a priori* change score)							
		After					
Before		75 (> 60)	45 (30-60)	25 (20-30)	10 (< 20)	Row totals (n_j's)	$\hat{\theta}_j$ (SE) [95% CI] {Range}[*]
	75 (>60)	22 (0)	14 (30)	11 (50)	4 (65)	51	24.12 (3.27) [18.47, 30.88] {0, 65}
	45 (30-60)	2 (–30)	18 (0)	9 (20)	6 (35)	35	9.43 (2.85) [3.72, 14.62] {–30, 35}
	25 (20-30)	0 (–50)	1 (–20)	5 (0)	14 (15)	20	9.50 (2.14) [0.88, 11.71] {–50, 15}
	10 (< 20)	1 (–65)	2 (–35)	4 (–15)	7 (0)	14	–13.93 (5.17) [–26.64, –7.67] {–65, 0}

[*] Standard errors (SE) obtained via eqn. (11); Approx. 95% CIs based on Dirichlet-multinomial approach leading to eqn. (12); Range indicates feasible theoretical lower and upper bounds for θ_j's

Table 6

Average Cell Counts for Simulation Under Moderate Sample Size (N=100)

	After Category				
Before Category		1	2	3	Total
	1	10	14	26	50
	2	2	8	20	30
	3	2	2	16	20

Table 7

Results of 2500 Simulations Under Moderate Sample Size (N=100)

Parameter (True value)	Mean estimate (Empirical SD)	Mean estimated SE *	95% CI coverage
θ_1 (1.32)	1.319 (0.113)	0.110, 0.111	93.8%, 94.8% †
θ_2 (0.60)	0.597 (0.110)	0.110, 0.111	93.8%, 95.7% †
θ_3 (–0.30)	–0.298 (0.144)	0.136, 0.140	88.2%, 95.0% †
θ_A (0.78)	0.779 (0.094)	0.096	94.4% ‡
θ_B (0.60)	0.599 (0.057)	0.057	95.1% ‡

[*] First mean SE value based on $\mathrm{S\hat{E}_I}\left(\hat{\theta}_\mathrm{j}\right)$ in eqn.(10); Second value based on $\mathrm{S\hat{E}}_{II}\left(\hat{\theta}_\mathrm{j}\right)$ in eqn.(11)

[†] First value for Wald-type CI using $\mathrm{S\hat{E}}_{II}\left(\hat{\theta}_\mathrm{j}\right)$ in eqn.(11); Second value for Dirichlet-multinomial CI based on eqn. (12)

[‡] CIs for θ_A and θ_B calculated using Mi-type procedure described in Appendix

Table 8

Results of 2500 Simulations Under Small Sample Size (N=50)

Parameter (True value)	Mean estimate (Empirical SD)	Mean estimated SE	95% CI coverage
θ_1 (1.32)	1.317 (0.161)	0.155, 0.158[*]	92.6%, 94.6% [†]
θ_2 (0.60)	0.599 (0.161)	0.151, 0.157[*]	90.4%, 95.4% [†]
θ_3 (–0.30)	–0.308 (0.196)	0.186, 0.198[*]	83.9%, 96.2% [†]
θ_A (0.78)	0.775 (0.134)	0.136	94.1% [‡]
θ_B (0.60)	0.596 (0.081)	0.081	94.0% [‡]

[*] First mean SE value based on $\hat{\mathrm{SE}}_{\mathrm{I}}\left(\hat{\theta}_{\mathrm{j}}\right)$ in eqn.(10); Second value based on $\hat{\mathrm{SE}}_{II}\left(\hat{\theta}_{\mathrm{j}}\right)$ in eqn.(11)

[†] First value for Wald-type CI using $\hat{\mathrm{SE}}_{II}\left(\hat{\theta}_{\mathrm{j}}\right)$ in eqn.(11); Second value for Dirichlet-multinomial CI based on eqn. (12)

[‡] CIs for θ_A and θ_B calculated using Mi-type procedure described in Appendix 1