Received: 25 September 2023 | Revised: 30 October 2023 | Accepted: 31 October 2023

DOI: 10.1002/ajhb.24015

SPECIAL ISSUE ARTICLE

American Journal of Human Biology WILEY

Statistical analysis of the longitudinal fundamental movement skills data in the REACT project using the multilevel ordinal logistic model

Donald Hedeker[1] | **Sara Pereira**[2,3] | **Fernando Garbeloto**[2] | **Tiago V. Barreira**[4] | **Rui Garganta**[2] | **Cláudio Farias**[2] | **Go Tani**[5] | **Jean-Philippe Chaput**[6] | **David F. Stodden**[7] | **José Maia**[2] | **Peter T. Katzmarzyk**[8]

[1]Department of Public Health Sciences, University of Chicago, Chicago, Illinois, USA

[2]Centre of Research, Education, Innovation and Intervention in Sport (CIFI2D), Faculty of Sport, University of Porto, Porto, Portugal

[3]Research Center in Sport, Physical Education, and Exercise and Health (CIDEFES), Faculty of Physical Education and Sports, Lusófona University, Lisboa, Portugal

[4]Department of Exercise Science, Syracuse University, Syracuse, New York, USA

[5]Motor Behavior Laboratory, School of Physical Education and Sports, University of São Paulo, São Paulo, Brazil

[6]Healthy Active Living and Obesity Research Group, Children's Hospital of Eastern Ontario Research Institute, Ottawa, Ontario, Canada

[7]Department of Physical Education, University of South Carolina, Columbia, South Carolina, USA

[8]Pennington Biomedical Research Center, Baton Rouge, Louisiana, USA

Correspondence
Donald Hedeker, Department of Public Health Sciences, The University of Chicago, 5841 South Maryland Avenue, MC 2000, Chicago, IL 60637, USA.
Email: hedeker@uchicago.edu

Abstract

Objectives: The REACT project was designed around two main aims: (1) to assess children's growth and motor development after the COVID-19 pandemic and (2) to follow their fundamental movement skills' developmental trajectories over 18 months using a novel technological device (Meu Educativo®) in their physical education classes. In this article, our goal is to describe statistical analysis of the longitudinal ordinal motor development data that was obtained from these children using the multilevel ordinal logistic model.

Methods: Longitudinal ordinal data are often collected in studies on motor development. For example, children or adolescents might be rated as having poor, good, or excellent performance levels in fundamental movement skills, and such ratings may be obtained yearly over time to assess changes in fundamental movement skills levels of performance. However, such longitudinal ordinal data are often analyzed using either methods for continuous outcomes, or by dichotomizing the ordinal outcome and using methods for binary data. These approaches are not optimal, and so we describe in detail the use of the multilevel ordinal logistic model for analysis of such data from the REACT project. Our intent is to provide an accessible description and application of this model for analysis of ordinal motor development data.

Discussion: Our analyses show both the between-subjects and within-subjects effects of age on motor development outcomes across three timepoints. The between-subjects effect of age indicate that children that are older have higher motor development ratings, relative to thoese that are younger, whereas the within-subject effect of age indicates higher motor development ratings as a child ages. It is the latter effect that is particularly of interest in longitudinal

Funding information
Portuguese Foundation for Science and Technology, Grant/Award Number: PTDC/SAU-DES/2286/2021

studies of motor development, and an important advantage of using the multilevel ordinal logistic model relative to more traditional methods.

1 | INTRODUCTION

In many studies of motor development and health-related behaviors the outcome of interest is measured in a series of ordered categories. Such outcomes are termed "ordinal" and can represent a variety of graded responses such as ratings of skill level (e.g., poor, good, excellent), sleep quality (very bad, reasonably bad, good, very good), and screen time (e.g., not watching, less than 1, 1–2, 2–3, 4–5, and 5 h or more). In other cases, the outcome may represent a count (e.g., number of minutes exercised in a day) that has a large number of zero responses (i.e., no exercise), many values in the intermediate range, and a few extreme values. In these cases, an ordinal variable can be constructed with ordered categories of, say, 0, 1–15, 16–30, 31–45, 46–60, and more than 60 min of exercise in a day.

Researchers sometimes analyze ordinal outcomes assuming a normal (continuous) distribution for the outcome. However, treating the outcome as normal assumes that the intervals between the categories of the outcome are all equal, which is clearly a dubious assumption. Also, as will be described, the ordinal model takes into account the ceiling and floor effects of the dependent variable, whereas models for continuous data do not. For example, if the outcome is coded in categories 1 to 5, a model for normal data can easily yield estimates below 1 and above 5. In this case, as McKelvey and Zavoina (1975) point out, biased estimates of the regression slopes and incorrect conclusions can result. Furthermore, as Winship and Mare (1984) note, the advantage of ordinal models in accounting for ceiling and floor effects of the ordinal variable is most critical if the variable is highly skewed, which is often the case where many of the responses are observed in the lowest and/or highest category of the ordinal outcome. Bauer and Sterba (2011) conducted an extensive simulation study addressing these issues and concluded that continuous models were only reasonable when the ordinal outcome had seven or more response categories and its distribution was approximately normal.

Alternatively, researchers sometimes dichotomize an ordinal outcome and analyze it using (binary) logistic regression. Sankeya and Weissfeld (1998) provided a simulation study in which an ordinal outcome with 5 categories was dichotomized and observed rather large losses of precision and power resulting from this practice. Also, Strömberg (1996) showed that the regression estimates can be poorly estimated when dichotomizing an ordinal outcome in datasets of limited size. Since power is often a critical issue, it behooves researchers to analyze ordinal outcomes with ordinal models, rather than losing power and information by dichotomizing them.

The ordinal logistic regression model, described as the proportional odds model by McCullagh (1980), provides a useful approach for analyzing ordinal outcomes. For multilevel data, where observations are nested within clusters (e.g., classes, schools, clinics) or are repeatedly assessed across time within subjects, multilevel models (aka mixed-effects models) are often used to account for the dependency inherent in the data (Goldstein, 2011; Hedeker & Gibbons, 2006; Raudenbush & Bryk, 2002). Multilevel models for ordinal data have been developed for quite some time (Agresti & Natarajan, 2001; Hedeker & Gibbons, 1994; Tutz & Hennevogl, 1996), including software (Hedeker & Gibbons, 1996), making such analysis accessible to researchers. More recently, most of the major statistical packages (e.g., SAS, Stata, R) include multilevel ordinal models, making these methods even more accessible.

The purpose of this paper is to describe the application of the multilevel ordinal logistic regression model for data that are both longitudinal and clustered. In terms of the organization of this paper, the study design and fundamental movement skills (FMS) outcomes (e.g. kick, running, overhand throw) will be described in Section 2. The multilevel model for longitudinal ordinal data will be described in Section 3. We will begin with a 2-level longitudinal model in which observations (level-1) are nested within subjects (level-2). We will then present a 3-level model in which the subjects are nested within clusters (level-3). Section 4 will illustrate application of the 3-level model using the FMS total score as the outcome, accounting for the clustering of repeated observations within subjects and subjects within schools. Section 5 will present an analysis at the item level, considering the longitudinal FMS item responses from the 5 items as nested within subjects within schools. Finally, Section 6 will conclude with some discussion.

2 | STUDY DESIGN AND FUNDAMENTAL MOVEMENT SKILLS ASSESSMENT

This study, part of the REACT project, aimed to evaluate the growth and motor development of children after the COVID-19 pandemic. A new technological device called Meu Educativo® was used to track the developmental trajectories of children's FMS over 18 months in their physical education classes. For more information on the background, rationale, and methodology of the REACT

project, which was carried out in Matosinhos municipality, north of Portugal, see Pereira et al. (2023). The study included 849 children, aged approximately 6–10 years. These children were assessed every 6 months from the beginning of the study to the 12-month mark. We excluded children without permission from their parents or legal guardians and those with physical disabilities that limited their ability to complete all assessments.

The Meu Educativo® technological platform was used to help the research team members to assess children's fundamental movement skills (FMS). Information on the validation process of the platform and reliability can be found elsewhere (Garbeloto et al., 2023). This device allows the assessment of 14 FMS. However, in the REACT project, we only considered 5 FMS associated with object control because they were part of the annual Physical Education program used in Matosinhos municipality for children aged 6 to 10 years:

1. Stationary dribbling (FMS1): the child must dribble (bounce) a ball at least four consecutive times without leaving the place.
2. Kick (FMS2): the child must kick a ball against the wall (or a goal) as hard as possible.
3. Overhand throw (FMS3): The child must throw a ball as hard as possible against the wall.
4. Catch the ball (FMS4): the child must receive a ball with both hands without leaving the place.
5. Underhand roll (FMS5): the child must roll a ball against the wall as hard as possible.

Before starting the assessments, team members participated in a specific training process following all procedures established by the Meu Educativo® Assessment Manual. During the training, raters needed to obtain at least an inter-rater and intra-rater agreement of 80%.

Following the Meu Educativo® assessment protocol, all participants were instructed to perform each skill at least twice, the first for familiarization and the second for scoring. If the rater had doubts regarding the child's performance level, a third and even fourth attempt was requested, but only one grade was assigned to the participant.

2.1 | Materials

The materials used were a cell phone with internet to access to the Meu Educativo® app., a tennis ball (or similar measurements) for the overhand throw and underhand roll, a soccer ball for kicking (or with similar measures to an official adult category ball), and a softball like a volleyball used by young players (or a ball with similar measurements).

2.2 | Performance measure

Each FMS assessed in the REACT project contains three (e.g., kicking and catching the ball) or four components (e.g., overhand throw, stationary dribbling, and underhand roll). The greater the number of components performed with proficiency, the better the performance level. For example, if a child performs all three components proficiently in kick, she/he is classified as a Wizard Climber representing the proficient level, that is, level 3. If she/he performs only two components proficiently, she/he will be classified as an Adventurous Climber representing the intermediate level, that is, level 2. Finally, if a child performs only one or none of the components proficiently, she/he will be classified as an Explorer Climber representing the beginner level, that is, level 1. Further information on the performance measure can be found in Pereira et al. (2023).

The five FMS items, each rated from 1 to 3, were added together to yield a sum or total score. A histogram of these scores (for the observations from all subjects across all timepoints) is presented in Figure 1. As can be seen, the distribution of the scores does not approximate a normal distribution. In fact, the modal response is the highest category of 15, rather than in the middle of the distribution. Clearly, assuming a normal distribution for this outcome is not reasonable according to the guidance provided by Bauer and Sterba (2011).

3 | MULTILEVEL LOGISTIC MODEL FOR LONGITUDINAL ORDINAL DATA

To begin, we will describe the multilevel model for longitudinal ordinal data, where repeated observations (level-1) are nested within subjects (level-2). To establish the notation, subjects are denoted as i (where $i = 1, ..., N$ subjects) and the repeated observations are denoted as j (where $j = 1..., n_i$). The number of repeated observations per subject is n_i, and so there is no assumption that each subject is measured on the same number of timepoints. In longitudinal studies, it is common to have incomplete data across time, so it is important that the model allows for this. Ordinal regression models often utilize cumulative comparisons of the categories. For this, define the cumulative probabilities for the C categories of the outcome Y as $P_{ijc} = Pr(Y_{ij} \le c) = \sum_{m=1}^{c} p_{ijm}$, where p_{ijm} represents the probability of response in category m. For example, with three categories, we would have $P_{ij1} = p_{ij1}$ as the probability of a response in category 1, and $P_{ij2} = p_{ij1} + p_{ij2}$ as the probability of a response in categories 1 and 2. The

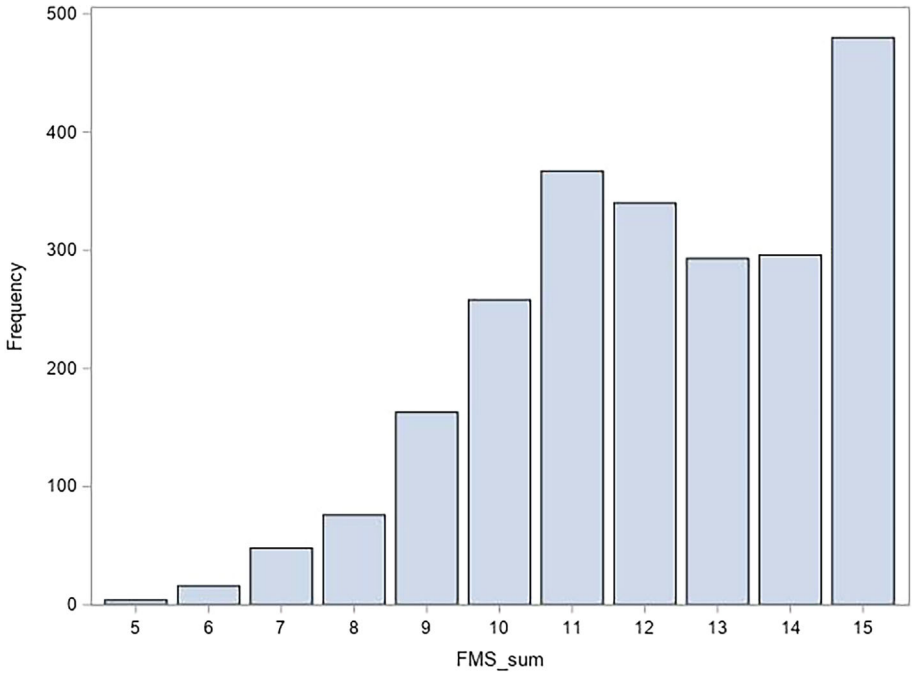

FIGURE 1 Histogram of the FMS total sum score.

probability of a response in category 3 would be obtained by subtraction as $p_{ij3} = 1 - P_{ij2}$.

3.1 | Random intercept model

The multilevel logistic regression model for the cumulative probabilities of subject i at timepoint j is given in terms of the $C-1$ cumulative logits as

$$log\left[\frac{P_{ijc}}{1-P_{ijc}}\right] = \gamma_c - \left[\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \upsilon_i\right], \quad (1)$$

with $C-1$ strictly increasing model thresholds γ_c. These thresholds are akin to intercepts and represent the cumulative logits when the covariates and random effects equal 0. Basically, the thresholds indicate how many responses are in the different categories (when the covariates and random effects equal 0), and are usually not of great interest. The distribution of responses in the ordered categories is completely arbitrary. As usual, $\boldsymbol{x}_{ij}$ are the covariates and $\boldsymbol{\beta}$ are the regression slopes (i.e., effects of the covariates). The covariates can be at level-1 (e.g., time) or level-2 (e.g., group), or could be cross-level interactions (e.g., group by time). The random effects υ_i reflect each subject's influence on their repeated observations. This model is referred to as a random-intercept model as the subject effects do not vary across time. These are assumed to be distributed in the population of subjects as $N\left(0, \sigma_\upsilon^2\right)$, and so the sample of subjects are thought to represent a population of subjects that one wants to make inferences about.

In terms of the effects of time on the repeated outcomes, typically the covariate(s) x_{ij} would include at least a linear effect of time. For example, suppose that subjects are measured at baseline, 6 and 12 months. Then, one of the covariates in $\boldsymbol{x}_{ij}$ might be a variable t_{ij} (and coded 0, 1, 2) to represent the linear effect of time (in 6 month intervals). With more timepoints, the model might also include quadratic effects to allow for curvilinear effects of time. That is, the response across time might be a decelerating or accelerating trend, rather than a simple linear trend. For this, one could include both t_{ij} and its square t_{ij}^2 to represent the linear and quadratic components of the trend across time. Alternatively, in some cases, it might be of interest to compare each follow-up to baseline and therefore to create dummy variables for each of the follow-ups treating baseline as the reference cell. Whether one uses polynomials for trends or dummy codes to represent the effects of time depends on the scientific questions of interest.

Interactions with the time effects are usually of interest in longitudinal models in order to assess, for example, the degree to which trends vary across groups of subjects. So, if there is a grouping variable G_i, say coded 0 for a control group and 1 for an intervention group, and one simply included a linear effect of time, the following model might be posited:

$$\log\left[\frac{P_{ijc}}{1-P_{ijc}}\right] = \gamma_c - \left[\beta_1 T_{ij} + \beta_2 G_i + \beta_3\left(G_i \times T_{ij}\right) + \upsilon_i\right]. \quad (2)$$

Here, β_2 represents the group difference when T_{ij} equals 0, and β_3 indicates how the group difference varies with

time. Or, β_1 represents the time trend for the control group (when G_i equals 0), and β_3 represents the difference in the trend for the intervention group relative to the control group. Thus, testing the significance of β_3 is of great interest as it represents how the trends differ between the two groups.

3.1.1 | Intraclass correlation

For a random-intercept model, it is often of interest to express the subject variance in terms of an intraclass correlation (ICC). The ICC indicates the proportion of unexplained variance that is at the subject level, and is given by $\text{ICC} = \sigma_\upsilon^2/(\sigma_\upsilon^2 + \sigma^2)$, where σ_υ^2 is the subject or level-2 variance and σ^2 is the level-1 variance. For a logistic regression model (either binary or ordinal), the level-1 variance, which is not estimated, equals the variance of the standard logistic distribution, namely $\pi^2/3$ (Agresti, 2002). Note that the ICC has a minimum of 0 and a maximum of 1, with higher values indicating a greater proportion of the unexplained variance in the outcome that is at the subject level. Because the ICC reflects the unexplained variance (i.e., the variance that is not explained by model covariates), it can change depending on the covariates that are included in the model.

3.2 | Random intercept and trend model

Thus far, the model only includes a single random subject effect υ_i and assumes that a subject's effect on their responses is the same across all timepoints. This is often an unreasonable assumption because subjects often vary in their trends across time. To permit this, we can extend the model by including a random subject trend:

$$\log\left[\frac{P_{ijc}}{1-P_{ijc}}\right] = \gamma_c - \left[\beta_1 T_{ij} + \beta_2 G_i + \beta_3 (G_i \times T_{ij}) + \upsilon_{0i} + \upsilon_{1i} T_{ij}\right]. \quad (3)$$

Here, υ_{1i} is essentially an interaction of subject by time, indicating the degree to which subjects have different time trends. In this model, υ_{0i} represents the subject effect when T_{ij} equals 0, and υ_{1i} indicates how a subject's effect varies with time. Subjects have different time trends to the extent that the υ_{1i} parameters are non-zero. Both random effects are usually assumed to be normally distributed in the population of subjects with variances $\sigma_{\upsilon_0}^2$ and $\sigma_{\upsilon_1}^2$, respectively. The covariance between a subject's intercept and trend, $\sigma_{\upsilon_{01}}$, indicates the degree to which a subject's starting point is associated with their trend.

Notice that the random-intercept model in Equation (2) is a special case of the random trend model in Equation (3). By not including the random time effect υ_{1i}, the random intercept model assumes that these are all zero and thus that the variance $\sigma_{\upsilon_1}^2$ and covariance $\sigma_{\upsilon_{01}}$ both equal zero. Thus, comparison of the two models via a likelihood ratio test can be performed to test whether these two co(variance) parameters ($\sigma_{\upsilon_1}^2$ and $\sigma_{\upsilon_{01}}$) equal zero. If the test is non-significant, then the simpler random-intercept model is supported and there is no appreciable subject heterogeneity in the time trends (other than the random intercept υ_{0i}). Alternatively, if this test is significant it indicates that subjects do vary in their trends, and the simpler random-intercept model would be rejected in favor of the random trend model.

3.2.1 | Time-varying covariates

In some studies, there might be time-varying covariates which are thought to influence the ordinal outcome. In this case, the model might be

$$\log\left[\frac{P_{ijc}}{1-P_{ijc}}\right] = \gamma_c - \left[\beta_1 T_{ij} + \beta_2 X_{ij} + \upsilon_{0i} + \upsilon_{1i} T_{ij}\right], \quad (4)$$

where X_{ij} represents the time-varying covariate. One might also examine whether there is an interaction of X_{ij} with time, by including the product term $X_{ij} \times T_{ij}$ into the model, which would suggest that the relationship between the covariate and the outcome varies with time.

When time-varying covariates are included in the model, as in Equation (4), an assumption is made that the between and within-subjects effects of the covariate are equal (Hedeker & Gibbons, 2006; van de Pol & Wright, 2009). To see this, express the time-varying covariate X_{ij} as $X_{ij} = \overline{X}_i + (X_{ij} - \overline{X}_i)$, where $\overline{X}_i$ is the mean of the time-varying covariate (averaged across time) for each subject (i.e., a between-subjects variable). The term $(X_{ij} - \overline{X}_i)$ represents the subject's deviation at timepoint j around their mean (i.e., a within-subjects variable). Including both of these terms into the model yields:

$$\log\left[\frac{P_{ijc}}{1-P_{ijc}}\right] = \gamma_c - \left[\beta_1 T_{ij} + \beta_2 \overline{X}_i + \beta_3 (X_{ij} - \overline{X}_i) + \upsilon_{0i} + \upsilon_{1i} T_{ij}\right], \quad (5)$$

The total effect of X_{ij}, $\beta_2 \overline{X}_i + \beta_3 (X_{ij} - \overline{X}_i)$, is partitioned into its between- and within-subjects effects (i.e., β_2 and β_3, respectively). The between-subjects part indicates the degree to which the subject's average covariate level is related to their average outcome level, averaging across time. The within-subjects component represents the degree to which change in a subject's covariate level at a particular timepoint is associated with

change in their outcome at that timepoint (i.e., a within-subject change). For example, in the analyses presented below, we will partition the effect of age in this way. The between-subjects component reflects differences between children of different ages (i.e., a cohort effect), whereas the within-subjects component reflects differences as a child ages (i.e., an aging effect). If these two are equal ($\beta_3 = \beta_2$), then the effect is exactly as in Equation (4). Thus, Model (4) makes the assumption that the within- and between-subjects effects of the covariate are the same. This assumption can be assessed by comparing the models specified by (4) and (5) via a likelihood ratio test. If these two models are significantly different, then the assumption is rejected and the more general Model (5) is preferred; whereas if the models are not significantly different then the assumption is reasonable and Model (4) can be used. In some situations, one wants to disentangle these two effects, regardless of the results of this test. In this case, one wants to obtain estimates of the "pure" within-subjects effect and the "pure" between-subjects effect of the time-varying covariate, and then Model (5) is clearly preferred.

3.2.2 | Intraclass correlation

For a random intercept and trend model, there is no longer a single ICC, as the subject variance varies with the time variable T_{ij}. However, one can still calculate the ICC for particular values of T_{ij}. This can be useful to examine the degree to which the proportion of (unexplained) variance that is at the subject level varies with time. In this case, the subject (level-2) variance equals $\sigma^2_{\upsilon_0} + T^2_{ij}\sigma^2_{\upsilon_1} + 2T_{ij}\sigma_{\upsilon_{01}}$. Here, $\sigma^2_{\upsilon_0}$ is the variance of the random subject intercepts, $\sigma^2_{\upsilon_1}$ is the variance of the random subject time trends, and $\sigma_{\upsilon_{01}}$ is the covariance of the two random subject effects. The level-1 variance is $\pi^2/3$ (for the standard logistic distribution), and the ICC is still the level-2 (subject) variance divided by the sum of the level-1 and level-2 variances.

3.3 | Three level extension

Thus far, we have considered repeated observations (level-1) within subjects (level-2), and have presented two-level models. In some cases, the subjects might be nested within clusters (e.g., schools, hospitals, workplaces), which then requires a three-level model. Three-level models for ordinal data have been developed and described by Raman and Hedeker (2005) and Liu and Hedeker (2006). For this, consider clusters to be denoted as i (where $i = 1,...,N$ clusters), subjects to be denoted as j (where $j = 1,...,n_i$ subjects in cluster i), and the repeated observations are denoted as k (where $k = 1...,n_{ij}$ for subject j in cluster i). The number of subjects per cluster and repeated observations per subject are not assumed to be equal. Then, we can generalize (5) by including a random cluster effect υ_i as:

$$\log\left[\frac{P_{ijkc}}{1-P_{ijkc}}\right] = \gamma_c - \Big[\beta_1 T_{ijk} + \beta_2 \overline{X}_{ij} + \beta_3\left(X_{ijk} - \overline{X}_{ij}\right) + \upsilon_i + \upsilon_{0ij} + \upsilon_{1ij} T_{ijk}\Big], \quad (6)$$

Here, the random cluster effects υ_i are assumed to be normally distributed in the population (of clusters) with mean 0 and variance $\sigma^2_{\upsilon_{(3)}}$. Notice that these random cluster effects are in addition to the random subject effects in the model. They represent the effect of a given cluster on an outcome (from a subject at a given timepoint) within the same cluster, over and above the effect of a subject on their repeated outcomes. In this way, the random cluster effects account for the correlation in the outcomes of subjects within the same cluster (over and above the influence of the subject on their repeated outcomes).

3.3.1 | Intraclass correlation

For a three level model, one can calculate both the level-3 and level-2 ICCs, which indicate the degree of unexplained variance that is attributable to the cluster (level-3) and to the subject (level-2). Again, if there are random subject intercepts and trends, then one can calculate the subject variance for particular values of T_{ij}. The ICC for clusters (level-3) is then given by:

$$\text{ICC}_{(3)} = \frac{\sigma^2_{(3)}}{\sigma^2_{(3)} + \sigma^2_{(2)} + \pi^2/3}, \quad (7)$$

and the ICC for subjects (level-2) is:

$$\text{ICC}_{(2)} = \frac{\sigma^2_{(3)} + \sigma^2_{(2)}}{\sigma^2_{(3)} + \sigma^2_{(2)} + \pi^2/3}, \quad (8)$$

where $\sigma^2_{(3)}$ is the cluster (level-3) variance, and $\sigma^2_{(2)}$ is the subject (level-2) variance (which again, can vary by T_{ij} in models with random time trends). The reason that the level-3 variance appears in the numerator of the level-2 ICC is that subjects are nested within clusters.

3.4 | Proportional odds assumption

Models for ordinal outcomes often include the proportional odds assumption for model covariates. For an ordinal response with C categories, this assumption states that the effect of the covariate is the same across the $C-1$ cumulative logits of the model (or proportional across the cumulative odds). The idea is that if one did dichotomize the ordinal outcome and used a (binary) logistic regression model, the regression slopes would be equal, regardless of how one did the dichotomization (e.g., for an ordinal variable with 3 categories there are two possible dichotomizations: 1 vs. 2 & 3, and 1 & 2 vs. 3). In previous papers (Hedeker & Mermelstein, 1998, 2000), we have described an extension to allow for non-proportional odds for the covariates. While this extension is useful, especially for categorical covariates, one should be cautious in using non-proportional odds models if the model covariates are continuous variables. The reason for this is that in the non-proportional odds model the trend lines for the cumulative logits (y-axis) versus the continuous covariate (x-axis) are not parallel, and therefore cross each other at some value of the covariate. For example, the cumulative logits of 1 vs. 2 & 3 and 1 & 2 vs. 3 cross each other. If this crossing happens within the range of values for the continuous covariate, it would imply, for example, that the probability of a 1 response exceeds the probability of a 1 & 2 response, for covariate values below the crossing, which is clearly impossible. Thus, for continuous covariates, it is generally reasonable to maintain the proportional odds assumption. More information about this can be found in Hedeker et al. (1999) and Fullerton and Xu (2016).

4 | 3-LEVEL ANALYSIS OF THE FMS TOTAL SCORE

As seen in Figure 1, the distribution of the FMS total score does not approximate a normal distribution, and so here we present an ordinal analysis of this outcome. As can be seen in the figure, the scores ranged from a minimum value of 5 to a maximum value of 15, for a total of 11 ordinal categories. Boys ($N = 411$) and girls ($N = 438$) were analyzed in separate models, treating the repeated observations (level-1) as nested within subjects (level-2) who were nested within schools (level-3; $N = 25$). Each subject had measurements at three timepoints at which their FMS, Age, and body mass index (BMI) were recorded. In total, there were 1120 observations for the 411 boys (an average of 2.725 observations per boy), and 1192 observations for the 438 girls (an average of 2.721 observations per girl). Here, the repeated FMS scores are the longitudinal outcomes, and Age and BMI are time-varying covariates. As indicated above, for these time-varying covariates, we decomposed them in terms of their WS and BS effects. As recommended in McArdle (2006) and others, we use age instead of study wave as our time variable, thus the WS version of Age is our time variable. This permits us to examine how subjects change in their FMS total score as they got older in age. Across all observations, the minimum age value was 5.57 and the maximum was 11.23 (mean = 8.40). This variation in age was also observed at each study wave: wave 1 (mean = 7.92, min = 5.57, max = 10.63), wave 2 (mean = 8.50, min = 6.39, max = 11.19), and wave 3 (mean = 8.88, min = 6.62, max = 11.23). For BMI, the minimum was 10.4 and the maximum was 42.92 (mean = 17.85). Finally, we also included the school size (number of children) as a school-level covariate (min = 77, max = 370, mean = 183.8). Our model is then:

$$\log\left[\frac{P_{ijkc}}{1-P_{ijkc}}\right] = \gamma_c - \Big[\beta_1 \text{SchoolSize}_i + \beta_2 \overline{\text{Age}}_{ij} + \beta_3\left(\text{Age}_{ijk} - \overline{\text{Age}}_{ij}\right) + \beta_4 \overline{\text{BMI}}_{ij} + \beta_5\left(\text{BMI}_{ijk} - \overline{\text{BMI}}_{ij}\right) + \upsilon_i + \upsilon_{0ij} + \upsilon_{1ij}\left(\text{Age}_{ijk} - \overline{\text{Age}}_{ij}\right)\Big]. \quad (9)$$

We used the program Supermix (Hedeker et al., 2008) for all analyses. Supermix provides estimates that are both conditional (adjusting for the random effects) and marginal (averaging over the random effects). The latter er are often called "population-averaged" effects, and are of interest when inference is to be made about the population, whereas the former are sometimes called "subject-specific" effects, and are useful when interest is on inference for individual subjects (Hu et al., 1998). Here, we present the marginal or "population-averaged" estimates. More information about the difference between the conditional and marginal effects can be found in Hedeker et al. (2018).

The estimates of the model covariates are presented in Table 1, which provides the logit estimates, odds ratios (OR), and 95% confidence limits for the ORs for the separate analyses for boys and girls. Note, that if the 95% confidence limit does not include 1, this would reflect a statistically significant effect at the $\alpha = 0.05$ level.

As can be seen from Table 1, age has a very significant positive effect for both boys and girls, and both in terms of the BS and WS effects. Thus, older boys and girls have, on average, higher FMS total scores (BS effects), and subjects increase their FMS total score as they age

Parameter	Logit estimate	Odds ratio (OR)	OR 95% confidence interval	
			Lower	Upper
Boys				
SchoolSize	−0.006	0.994	0.984	1.004
WS age	0.726	2.067	1.501	2.848
BS age	1.005	2.733	1.868	3.999
WS BMI	−0.041	0.960	0.857	1.075
BS BMI	−0.082	0.922	0.850	0.999
Girls				
SchoolSize	−0.002	0.998	0.994	1.002
WS age	1.483	4.408	2.293	8.474
BS age	1.515	4.547	3.527	5.862
WS BMI	−0.114	0.892	0.758	1.050
BS BMI	−0.032	0.969	0.906	1.036

TABLE 1 3-level ordinal analysis of FMS total score for boys and girls. Logit estimates, odds ratios (OR), and OR 95% confidence intervals.

Abbreviations: BS, between-subject effect; WS, within-subject effect.

Parameter	Estimate	Standard error	*z* value	*p* value
Boys				
Subject int var	11.722	1.214	9.659	.001
Subject WS age/int cov	2.561	0.288	8.892	.001
Subject WS age var	0.645	0.077	8.378	.001
School int var	0.660	0.388	1.703	.089
Girls				
Subject int var	3.950	0.591	6.685	.001
Subject WS age/int cov	0.022	0.334	0.067	.946
Subject WS age var	0.772	0.598	1.290	.197
School int var	0.256	0.171	1.497	.135

TABLE 2 3-level ordinal analysis of FMS total score for boys and girls. Random effect variance estimates, standard errors, *z* values, and *p* values.

Abbreviations: cov, covariance; int, intercept; var, variance.

(WS effects). In terms of ORs, the BS effects for boys and girls are equal to 2.7 and 4.5, respectively, meaning that the odds of a higher value on the FMS total score increase by a factor of 2.7 (boys) and 4.5 (girls) for each year of age. Since this is the BS effect, it is comparing boys and girls of different average ages. The WS effects indicate that boys have increased odds of a higher FMS total score by a factor of approximately 2 per year as they age, whereas for girls this factor is 4.4. Clearly, subjects significantly improve on their FMS total score as they age. BMI has no significant effect for girls, however the BS effect of BMI is significant and negative for boys. Thus, comparing boys of different average BMI values (averaged across time) indicates that the odds of a higher FMS total score are decreased by a factor of 0.92 with each unit increase of (average) BMI. Thus, boys with higher (average) BMI have lower (average) FMS total scores. Finally, school size does not have a significant effect on the FMS total scores in both boys and girls.

Table 2 lists the estimates of the variance parameters associated with the random subject and random school effects, separately for boys and girls. The random subject intercept effect variances are highly significant for both boys and girls. Thus, there is clear evidence of heterogeneity in FMS total scores across subjects. In terms of the random subject WS age variances, there is also heterogeneity in the trends across age for boys, but not for girls. Taken together, subjects differ from each other in their average FMS scores, while boys also differ from each other in the change in FMS scores as they age. Based on the covariances, the two random effects are not significantly associated in girls ($p = .946$), and positively related in boys ($p < .001$). Expressed as a correlation, the covariance of 2.561 for boys is a correlation of

$2.561/\left(\sqrt{11.722}\times\sqrt{0.645}\right)=0.931$, reflecting a strong positive association. For boys there is a significant positive association between their average FMS total score and their trend in FMS total score as they age. Boys that have steeper trend lines have higher average FMS total scores. Finally, the *p*-values for the school variances do not reach the .05 level in Table 2. However, these are two-sided *p*-values. For variances, which cannot be negative, it is more appropriate to have one-sided *p*-values (Snijders & Bosker, 2012), which would be equal to $p=.089/2=.045$ for boys and $p=.135/2=.0675$ for girls. Thus, there is a significant clustering effect attributable to schools on the FMS total scores for boys, but not quite for girls.

Using the ICC equations presented earlier, we can calculate the amount of variance that is attributable to both schools and subjects based on the variance estimates in Table 2. Since there are random subject time trends in these models, the ICCs can be calculated for any value of the time variable $\left(\text{Age}_{ijk}-\overline{\text{Age}}_{ij}\right)$. For simplicity, we will do this when this variable equals 0, namely for when a subject is at their average age. Using the approximate value of $\pi^2/3=3.290$, we get:

School ICC for boys

$$\text{ICC}=\frac{0.660}{0.660+11.722+3.290}=0.042$$

Subject ICC for boys

$$\text{ICC}=\frac{0.660+11.722}{0.660+11.722+3.290}=0.790$$

School ICC for girls

$$\text{ICC}=\frac{0.256}{0.256+3.950+3.290}=0.034$$

Subject ICC for girls

$$\text{ICC}=\frac{0.256+3.950}{0.256+3.950+3.290}=0.561$$

These ICC calculations show that the amount of variation in FMS total scores that is at the subject level is quite large, especially for boys. The amount of variation that is at the school level is approximately 4% (boys) and 3% (girls), which are in a similar range as what has been reported in other school-based studies of physical activity (Kristensen et al., 2013; Murray et al., 2004; Pereira et al., 2020; Steenholt et al., 2018).

5 | ANALYSIS AT THE ITEM LEVEL

The previous analyses used the FMS total scores as the outcome, however it is also possible to analyze at the item level. This approach allows one to test whether, for example, the effect of age is the same or different on the items that comprise the FMS total score. Here, there were 5 items, each rated from 1 to 3, as described above in Section 2. Figures 2 and 3 provide the category proportions of the ratings to the 5 items across time for boys and girls, respectively. From Figure 2, one can see that for boys category 3 is the most frequent rating for most items at the first timepoint, and for all items at the second and third timepoints. For girls, from Figure 3, category 3 is rated highest only for item 4 (catch the ball). From both figures, one can see the overall improvement in ratings across the three timepoints.

An analysis considering items nested within times within subjects within schools was first considered, however the time variance went to zero in the estimation. Thus, a 3-level analysis considering items by time (15 observations) within subjects within schools was used, separately for boys and girls. Here, *i* is for schools, *j* is for subjects, *k* is for time, *l* is for items, and *c* is for the cumulative logit of the ordinal items (here $c=1,2$ for the two cumulative logits of the 3-category items). The model is given by:

$$\begin{aligned}\log\left[\frac{P_{ijklc}}{1-P_{ijklc}}\right]=\gamma_{lc}-\Big[&\beta_1\text{SchoolSize}_i+\beta_2\overline{\text{Age}}_{ij}\\&+\beta_3\left(\text{Age}_{ijk}-\overline{\text{Age}}_{ij}\right)+\beta_4\overline{\text{BMI}}_{ij}\\&+\beta_5\left(\text{BMI}_{ijk}-\overline{\text{BMI}}_{ij}\right)+\upsilon_i+\upsilon_{0ij}\\&+\upsilon_{1ij}\left(\text{Age}_{ijk}-\overline{\text{Age}}_{ij}\right)\Big],\end{aligned}\tag{10}$$

Since the thresholds (γ parameters) have the *l* subscript, the distribution in the three categories is allowed to vary across the 5 items. Figures 2 and 3 show that this is generally the case, as, for example, item 4 is a much easier item (i.e., more responses in the higher categories) than the other items for both boys and girls.

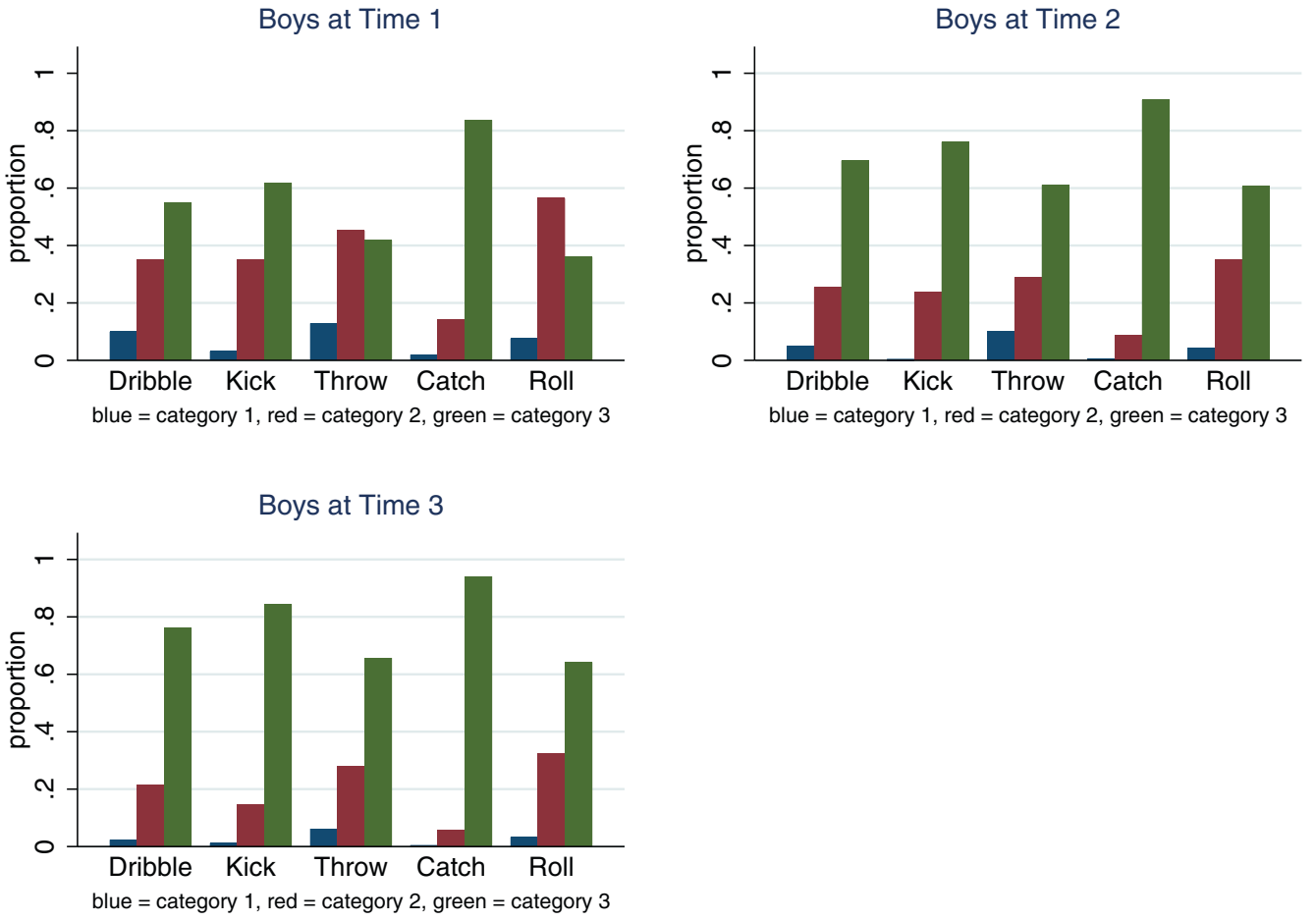

FIGURE 2 Response proportions for the five FMS items across time for boys.

Table 3 lists the estimates of the covariates for the analysis of boys and girls. Age, both BS and WS, has a significant positive effect for both boys and girls. In terms of the WS effects, the odds ratio for boys is estimated to be 2.2, and for girls it is 2.1. Thus, the odds of a response in a higher category on these items increases by a factor of 2.2 (2.1) as a boy (girl) ages. Similarly, the BS effects of age yield an estimated odds ratio of 2.0 for boys and 2.2 for girls. This indicates that comparing boys (girls) of different ages, a 1 year difference in age corresponds to an increased odds of a higher rating by a factor of 2.0 (2.2). Taken together, the odds of a higher rating on these items increases significantly as a child ages, and also comparing children of different ages. Finally, the covariate effects of BMI and School size are not significant, either for boys or girls, except for the BS effect of BMI for boys. As in the analysis of the FMS total score, this indicates that boys with higher average BMI have lower average ratings on the FMS items.

The above analyses assumed that the age effects were equal across the 5 items. However, it could be that age has differential effects across these 5 items. For this, we can include item by age interactions, both BS and WS, to assess this assumption via a likelihood-ratio test. This involves comparing the model deviances (−2 log likelihood values) for the model assuming equal age effects versus the model that relaxes this assumption (i.e., the model that adds in the item by age interactions). If the deviances for these two models are statistically similar, then we accept the assumption of equal age effects across the 5 items, whereas a significant difference would indicate a rejection of this assumption. Here, for boys, this equals $\chi^2_8 = 5941.952 - 5919.175 = 22.777, p = .004$, and for girls: $\chi^2_8 = 8615.619 - 8583.724 = 31.895, p < .001$. These tests are based on 8 degrees of freedom because the model with equal age effects included 2 age effects (BS and WS), whereas the model allowing for differential age effects on the 5 items included 10 age effects (BS and WS for each of the 5 items). The degrees of freedom is the difference in these two numbers of parameters. Based on these likelihood-ratio test results, the

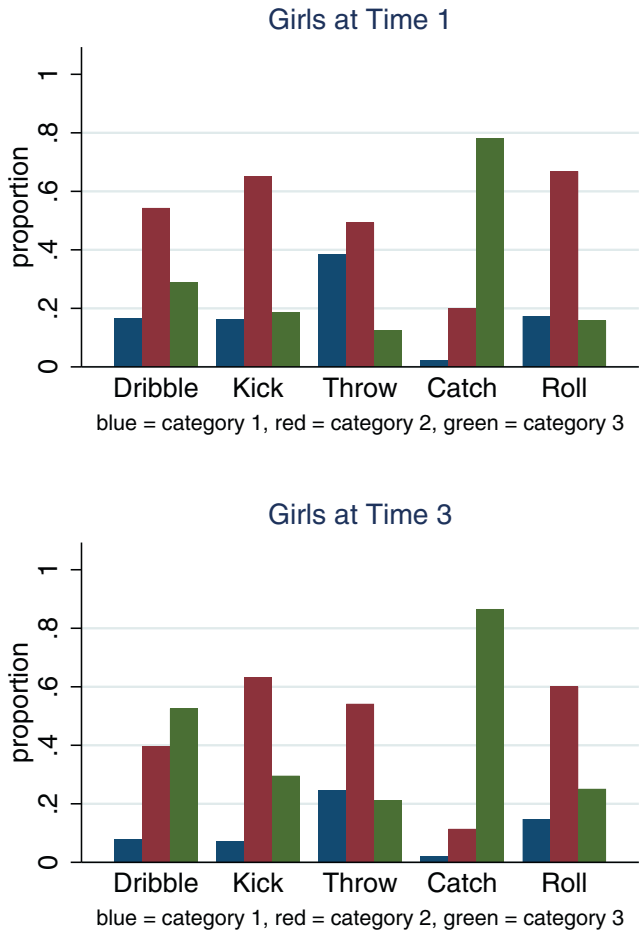

FIGURE 3 Response proportions for the five FMS items across time for girls.

TABLE 3 3-level ordinal analysis of FMS item responses for boys and girls. Logit estimates, odds ratios (OR), and OR 95% confidence intervals.

			OR 95% confidence interval	
Parameter	**Logit estimate**	**Odds ratio (OR)**	**Lower**	**Upper**
Boys				
SchoolSize	−0.001	0.999	0.996	1.003
WS age	0.794	2.211	1.662	2.941
BS age	0.708	2.030	1.660	2.483
WS BMI	−0.035	0.965	0.895	1.041
BS BMI	−0.070	0.933	0.888	0.980
Girls				
SchoolSize	−0.001	0.999	0.997	1.001
WS age	0.756	2.129	1.764	2.568
BS age	0.792	2.208	1.942	2.511
WS BMI	−0.063	0.939	0.865	1.019
BS BMI	−0.019	0.982	0.946	1.018

Abbreviations: BS, between-subject effect; WS, within-subject effect.

assumption of equal age effects is rejected, and so the effect of age on the item ratings varies significantly for both boys and girls.

Table 4 lists the estimates for the age by item interactions for boys and girls (from the model that added in these interactions). From Table 4, while all age effects are

TABLE 4 3-level ordinal analysis of FMS item responses for boys and girls. Logit estimates, odds ratios (OR), and OR 95% confidence intervals for the WS and BS effects of age on each item.

Parameter	Logit estimate	Odds ratio (OR)	OR 95% confidence interval	
			Lower	Upper
Boys				
FMS1 WS age	0.822	2.276	1.621	3.194
FMS2 WS age	0.831	2.295	1.484	3.549
FMS3 WS age	0.616	1.851	1.395	2.455
FMS4 WS age	0.957	2.603	1.528	4.435
FMS5 WS age	0.809	2.247	1.551	3.254
FMS1 BS age	0.753	2.123	1.686	2.673
FMS2 BS age	0.595	1.814	1.401	2.348
FMS3 BS age	0.499	1.647	1.343	2.020
FMS4 BS age	1.095	2.989	2.146	4.163
FMS5 BS age	0.762	2.143	1.697	2.707
Girls				
FMS1 WS age	1.107	3.025	2.282	4.009
FMS2 WS age	0.735	2.085	1.531	2.839
FMS3 WS age	0.613	1.846	1.454	2.343
FMS4 WS age	0.787	2.196	1.482	3.254
FMS5 WS age	0.491	1.635	1.229	2.174
FMS1 BS age	0.997	2.709	2.280	3.220
FMS2 BS age	0.630	1.877	1.573	2.239
FMS3 BS age	0.664	1.942	1.664	2.265
FMS4 BS age	0.947	2.578	2.035	3.267
FMS5 BS age	0.758	2.134	1.797	2.534

Abbreviations: BS, between-subject effect; WS, within-subject effect.

highly significant (since the CIs do not include 1), they do vary across the 5 items, and vary somewhat for boys and girls. For boys, in terms of both the WS and BS effects, FMS4 shows the strongest age effects (OR = 2.6 and 2.99, respectively), while FMS3 shows the weakest (OR = 1.85 and 1.65). Thus, as a boy ages and comparing boys of different ages, there is clear improvement in FMS4 (catch the ball), and less improvement in FMS3 (overhand throw). For the girls WS effects, FMS1 shows the strongest age effect (OR = 3.03), while FMS5 shows the weakest (OR = 1.64). FMS1 also shows the strongest BS age effect (OR = 2.71), while item 2 shows the least BS age effect (OR = 1.88). Thus, as a girl ages and comparing girls of different ages, there is clear improvement in FMS1 (dribbling). Also, as a girl ages, there is more modest improvement in FMS5 (underhand roll), and comparing girls of different ages, there is less of a difference in terms of FMS2 (kicking). Analysis at the item level therefore gives more specific information on the items, and how age effects the responses to these items, than the analysis of the total FMS score.

6 | DISCUSSION

This article has focused on describing application of the multilevel ordinal logistic model for analysis of motor development data in a practical and accessible way. Models have been described for both clustered and longitudinal motor development data. In our applications, students were clustered within schools and measured longitudinal over three timepoints. For clustered data, random cluster effects characterize the dependency of subjects' responses from the same cluster (i.e., school). For longitudinal data, random subject effects account for the dependency of the responses within the same subject, and give information about the heterogeneity in trends across time at the subject level. The 3-level ordinal model included both random effects for schools and subjects. Based on the ICC estimates, the dependency attributable to schools was approximately 5% (boys) and 8% (girls), while the dependency attributable to subjects was approximately 80% (boys) and 60% (girls). We presented models using both the total FMS score and also the individual FMS items.

Our models included time-varying covariates of age and BMI. For both variables, we decomposed the covariate effect in terms of the BS and WS effects (Hedeker & Gibbons, 2006; van de Pol & Wright, 2009). The separation of these effects allowed us to examine whether subjects of different average ages (and BMI values) had different average FMS total scores (BS effect), and also if as subjects age across time (or change BMI values) their FMS total scores changed (WS effect). Our results indicated highly significant BS and WS positive effects of age for both boys and girls. For BMI, a negative BS effect was observed only for boys. These effects were also observed at the item level, which showed that the age effects (both BS and WS) on items varied across the items. An additional school-level covariate, school size, was included, but found to be statistically non-significant.

In some studies, researchers have compared subjects of different age groups in terms of ordinal ratings of motor development outcomes using cross-sectional statistical methods such as the $L-$statistic (Lane et al., 2018) and item response theory (IRT) methods Sacko et al. (2021). While these can be useful for between-subjects comparisons of age groups, here we have focused on within-subject changes in motor development outcomes as subjects age (Roberton et al., 1980). Since between-subjects effects of age can always be confounded by cohort effects, the ability to simultaneously estimate both the between- and within-subjects effects of age provides a more powerful and informative approach. This is especially the case for analyzing the motor development of children and young adolescents. Also, other time-varying covariates (e.g., BMI) can be included and decomposed in terms of their between- and within-subjects effects on the time-varying motor development outcomes. As such, the multilevel ordinal logistic model provides a more comprehensive approach for the analysis of motor development data.

Another area of application for ordinal models is for time to event data in which the timing is not known precisely but only within time periods. For example, one might be interested in modeling time until subjects reach some value on the FMS total score (say, 15) in the students who are measured at the three timepoints. Here, the ordered outcome is the timepoint in which the FMS total score equals 15. We have described such multilevel survival analysis using the ordinal modeling approach (Hedeker et al., 2000; Hedeker & Mermelstein, 2011). Rather than using a logit link function, these survival models typically use a complementary log-log link function in order to yield a proportional hazards interpretation. Also, in this scenario one needs to consider the possibility of right-censoring in which the time of the event is unknown beyond a certain timepoint.

Certainly, researchers are more familiar with normal models and software, and so often treat ordinal outcomes as normal outcomes. One might wonder about whether this is a reasonable practice or not. In this regard, a comprehensive examination of this practice was performed by Bauer and Sterba (2011). They examined the performance of mixed normal and ordinal models to ordinal outcomes with 3 to 7 categories, and distributions that were symmetric, skewed, and polarized. In terms of bias, these authors concluded that the multilevel normal model only gave reasonable results if there were 7 categories and the distribution was symmetric. In all other cases, the multilevel normal model yielded unduly biased estimates of regression coefficients. In comparison, the multilevel ordinal model (i.e., the same model as presented in the current paper) produced unbiased estimates regardless of the number or shape of the distribution across the ordered categories. Furthermore, as described by Harrell (2015) and Liu et al. (2017), the ordinal model can also be used for continuous outcomes that do not follow a normal distribution, in order to obtain robust inferences.

Another consideration is statistical power, particularly for smaller datasets. In this regard, Armstrong and Sloan (1989) ordinalized a continuous outcome and reported relative efficiency (i.e., power) of 94% to 99% for 4 to 9 categories, respectively, as compared to the continuous outcome. Thus, even if the outcome is continuous, there is little efficiency loss, especially as the number of categories is increased. Conversely, if one dichotomizes an ordinal outcome, there can be appreciable loss in statistical power. Strömberg (1996) dichotomized an ordinal outcome with 5 categories, and for which the power level was 78%. The dichotomized outcomes had power levels between 38% to 68% depending on the cutpoint chosen. Thus, blindly dichotomizing an ordinal outcome can severely reduce power.

In conclusion, this article has attempted to describe the ordinal model clearly and in relatively non-technical terms. Ordinal models are probably not as popular as use of normal and binary models, despite the fact that ordinal outcomes are often obtained. Given that the methods and software are widely available, hopefully this situation will change as researchers become more familiar with application of the ordinal model.

AUTHOR CONTRIBUTIONS

Conceptualization: Donald Hedeker and José Maia; Funding acquisition: José Maia; Writing—original draft preparation: Donald Hedeker; Writing—review and editing: Sara Pereira, Fernando Garbeloto, Tiago V. Barreira, Rui Garganta, Cláudio Farias, Go Tani, Jean-Philippe

Chaput, David F. Stodden, José Maia, and Peter T. Katzmarzyk.

ACKNOWLEDGMENTS

We want to acknowledge the following people who helped us in the tremendous task of implementing REACT, and a special word goes to Prof. Henrique Calisto, who suggested the idea of this study to us. First, we want to thank the city-hall mayor, Dra. Luísa Salgueiro, the city councilors Prof. António Correia Pinto and Dr. Vasco Jorge Pinho. Second, to the city-hall education officers, Drs. Lília Pinto, Hugo Cruz, and Joana Aguiar. Third, to the Physical Education coordinating team, João Begonha, Ana Cunha, and Ricardo Ferreira. Fourth, to the Physical Education teachers, Ana Cunha, Ana Sousa, Ana Melo, Ana Santos, Ana Almeida, André Azevedo, Carlos Nogueira, Cátia Rodrigues, Filipe Silva, Frederico Meneses, Hélia Cardoso, Joana Brito, João Begonha, João Costa, Luís Machado, Nuno Pereira, Patrícia Rocha, Ricardo Ferreira, Ricardo Jesus, Ricardo Oliveira, Rui Correia, Rui Costa, Pedro Madureira, Solange Pereira, Tanya Poças, Tiffany Poças, Telmo Ribeiro e Rute Poças. Fifth, the assessment team, Renata Lucena, Ricardo Santos, Priscyla Praxedes, Catarina Ferreira, Patrícia Soares, José Guerra. Sixth, the twenty-five school coordinators. Finally, and most importantly, to all participating children and their families goes our greatest recognition for their willingness to be part of the study.

The authors thank William R. Leonard, Editor-in-Chief, and an anonymous reviewer for helpful comments that led to an improved article.

FUNDING INFORMATION

The REACT project was funded by the Portuguese Foundation for Science and Technology (FCT) under the reference: PTDC/SAU-DES/2286/2021.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data supporting this study's findings are available from José Maia (jmaia@fade.up.pt) upon reasonable request as well as the Faculty of Sports, University of Porto, policies regarding data sharing.

ORCID

Donald Hedeker https://orcid.org/0000-0001-8134-6094
Fernando Garbeloto https://orcid.org/0000-0001-8024-909X
Jean-Philippe Chaput https://orcid.org/0000-0002-5607-5736
Peter T. Katzmarzyk https://orcid.org/0000-0002-9280-6022

REFERENCES

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Wiley.

Agresti, A., & Natarajan, R. (2001). Modeling clustered ordered categorical data: A survey. *International Statistical Review*, *69*, 345–371.

Armstrong, B. G., & Sloan, M. (1989). Ordinal regression models for epidemiologic data. *American Journal of Epidemiology*, *129*(1), 191–204.

Bauer, D. J., & Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological Methods*, *16*, 337–390.

Fullerton, A. S., & Xu, J. (2016). *Ordinal regression models*. CRC Press.

Garbeloto, F., Pereira, S., Tani, G., Chaput, J.-P., Stodden, D. F., Garganta, R., Hedeker, D., Katzmarzyk, P. T., & Maia, J. (2023). Validity and reliability of Meu Educativo®: A new tool to assess fundamental movement skills in school-aged children. *American Journal of Human Biology* (In press).

Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). Wiley.

Harrell, F. E. (2015). Ordinal logistic regression. In *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (pp. 311–325). Springer International Publishing.

Hedeker, D., du Toit, S. H. C., Demirtas, H., & Gibbons, R. D. (2018). A note on marginalization of regression parameters from mixed models of binary outcomes. *Biometrics*, *74*, 354–361.

Hedeker, D., & Gibbons, R. D. (1994). A random effects ordinal regression model for multilevel analysis. *Biometrics*, *50*, 933–944.

Hedeker, D., & Gibbons, R. D. (1996). MIXOR: A computer program for mixed-effects ordinal probit and logistic regression analysis. *Computer Methods and Programs in Biomedicine*, *49*, 157–176.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Wiley.

Hedeker, D., Gibbons, R. D., du Toit, M., & Cheng, Y. (2008). *Supermix: Mixed efects models*. Scientifc Software International.

Hedeker, D., & Mermelstein, R. J. (1998). A multilevel thresholds of change model for analysis of stages of change data. *Multivariate Behavioral Research*, *33*, 427–455.

Hedeker, D., & Mermelstein, R. J. (2000). Analysis of longitudinal substance use outcomes using ordinal random-effects regression models. *Addiction*, *95*(Supplement 3), S381–S394.

Hedeker, D., & Mermelstein, R. J. (2011). Multilevel analysis of ordinal outcomes related to survival data. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of multilevel analysis* (pp. 115–136). Routledge.

Hedeker, D., Mermelstein, R. J., & Weeks, K. A. (1999). The thresholds of change model: An approach for analyzing stages of change data. *Annals of Behavioral Medicine*, *21*, 61–70.

Hedeker, D., Siddiqui, O., & Hu, F. B. (2000). Random-effects regression analysis of correlated grouped-time survival data. *Statistical Methods in Medical Research*, *9*, 161–179.

Hu, F., Goldberg, J., Hedeker, D., Flay, B., & Pentz, M. (1998). Comparison of population-averaged and subject-specific approaches

for analyzing repeated binary outcomes. *American Journal of Epidemiology*, *147*, 694–703.

Kristensen, P. L., Olesen, L. G., Ried-larsen, M., Grøntved, A., Wedderkopp, N., Froberg, K., & Andersen, L. B. (2013). Between-school variation in physical activity, aerobic fitness, and organized sports participation: A multi-level analysis. *Journal of Sports Sciences*, *31*(2), 188–195.

Lane, A., Molina, S., Tolleson, D., Langendorfer, S., Goodway, J., & Stodden, D. (2018). Developmental sequences for the standing long jump landing: A pre-longitudinal screening. *Journal of Motor Learning and Development*, *6*, 114–129.

Liu, L. C., & Hedeker, D. (2006). A mixed-effects regression model for longitudinal multivariate ordinal data. *Biometrics*, *62*, 261–268.

Liu, Q., Shepherd, B. E., Li, C., & Harrell, F. E. (2017). Modeling continuous response variables using ordinal regression. *Statistics in Medicine*, *36*(27), 4316–4335.

McArdle, J. J. (2006). Latent curve analyses of longitudinal twin data using a mixed-efects biometric approach. *Twin Research and Human Genetics*, *9*(3), 343–359.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society, Series B*, *42*, 109–142.

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, *4*, 103–120.

Murray, D. M., Catellier, D. J., Hannan, P. J., Treuth, M. S., Stevens, J., Schmitz, K. H., Rice, J. C., & Conway, T. L. (2004). School-level intraclass correlation for physical activity in adolescent girls. *Medicine and Science in Sports and Exercise*, *36*(5), 876–882.

Pereira, S., Katzmarzyk, P. T., Hedeker, D., Barreira, T. V., Garganta, R., Farias, C., Garbeloto, F., Tani, G., Chaput, J. P., Stodden, D. F., & Maia, J. (2023). Background, rationale, and methodological overview of the REACT project-return-to-action on growth, motor development, and health after the COVID-19 pandemic in primary school children. *American Journal of Human Biology*. https://doi.org/10.1002/ajhb.23968

Pereira, S., Reyes, A., Moura-Dos-Santos, M., Santos, C., Gomes, T. N., Tani, G., Vasconcelos, O., Barreira, T. V., Katzmarzyk, P. T., & Maia, J. (2020). Why are children different in their moderate-to-vigorous physical activity levels? A multilevel analysis. *Jornal de Pediatria*, *96*(2), 225–232.

Raman, R., & Hedeker, D. (2005). A mixed-effects regression model for three-level ordinal response data. *Statistics in Medicine*, *24*, 3331–3345.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Sage.

Roberton, M. A., Williams, K., & Langendorfer, S. (1980). Prelongitudinal screening of motor development sequences. *Research Quarterly for Exercise and Sport*, *51*(4), 724–731.

Sacko, R. S., Utesch, T., Cordovil, R., De Meester, A., Ferkel, R., True, L., Gao, Z., Goodway, J., Bott, T. S., & Stodden, D. F. (2021). Developmental sequences for observing and assessing forceful kicking. *European Physical Education Review*, *27*(3), 493–511.

Sankeya, S. S., & Weissfeld, L. A. (1998). A study of the effect of dichotomizing ordinal data upon modeling. *Communications in Statistics – Simulation and Computation*, *27*, 871–887.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis* (2nd ed.). Sage.

Steenholt, C. B., Pisinger, V. S. C., Danquah, I. H., & Tolstrup, J. S. (2018). School and class-level variations and patterns of physical activity: A multilevel analysis of danish high school students. *BMC Public Health*, *18*, 255.

Strömberg, U. (1996). Collapsing ordered outcome categories: A note of concern. *American Journal of Epidemiology*, *144*, 421–424.

Tutz, G., & Hennevogl, W. (1996). Random effects in ordinal regression models. *Computational Statistics and Data Analysis*, *22*, 537–557.

van de Pol, M., & Wright, J. (2009). A simple method for distinguishing within-versus between-subject effects using mixed models. *Animal Behaviour*, *77*(3), 753–758.

Winship, C., & Mare, R. D. (1984). Regression models with ordinal variables. *American Sociological Review*, *49*, 512–525.

How to cite this article: Hedeker, D., Pereira, S., Garbeloto, F., Barreira, T. V., Garganta, R., Farias, C., Tani, G., Chaput, J.-P., Stodden, D. F., Maia, J., & Katzmarzyk, P. T. (2023). Statistical analysis of the longitudinal fundamental movement skills data in the REACT project using the multilevel ordinal logistic model. *American Journal of Human Biology*, e24015. https://doi.org/10.1002/ajhb.24015