Economic Modelling 119 (2023) 106127

Contents lists available at ScienceDirect

Economic Modelling

journal homepage: www.journals.elsevier.com/economic-modelling

The relevance of specification assumptions when analyzing the drivers of physical activity practice

Jaume García [a], María José Suárez [b,*]

[a] Department d'Economia i Empresa, Universitat Pompeu Fabra, C/Ramon Trías Fargas, 25-27, 08005, Barcelona, Spain
[b] Departamento de Economía, Universidad de Oviedo, Avenida del Cristo, s/n, 33006, Oviedo, Spain

ARTICLE INFO

JEL classification:
Z29, C25, C52

Keywords:
Physical activity participation
Frequency
Time
Econometric modeling

ABSTRACT

There is some heterogeneity in the economics literature about the determinants of physical activity practice and little evidence on the robustness of the results to the specification assumptions. Our contribution to the literature is to examine methodologically and empirically to what extent the econometric specification—the modeling of the zeros in particular—the definition of physical activity (i.e., participation, time, and frequency) and its functional form condition the conclusions. The estimates reveal substantial differences in the effects of some drivers between the specifications as well as the dimensions of physical activity. We conclude that two-equation models that separately analyze participation and time/frequency perform better than the rest, supporting theoretical models in which both variables are arguments of the objective function and account for different types of zeros. The empirical results highlight the importance of preferences and time constraints in explaining the behavior associated with physical activity decisions.

1. Introduction

In 2018, the World Health Organization approved the Global Action Plan on Physical Activity 2018–2030 to promote exercise given its multiple health benefits such as lowering the risk of contracting non-communicable diseases and other physical and mental problems. A more active lifestyle may also contribute to reducing pollution, which connects with the 2030 Agenda for Sustainable Development drawn up by the United Nations in 2015. Other social benefits of a more physically active population include well-being gains, health expenditure savings, and productivity increases due to the reduction in mortality risks, absenteeism, and presenteeism rates. Hafner et al. (2019) quantified the global economic benefits of encouraging people to be more physically active compared to the baseline scenario of no physical activity improvement. According to their simulations, the cumulative global gain in gross domestic product (GDP) due to the productivity effects could be between 125.2 and 453.6 United States dollars (US$) per adult by 2030 at 2019 present value, depending on the scenario considered, with annual GDP increases ranging between 0.1% and 0.37% through 2030.

The advantages of a physically active population have sparked the interest of economists, who have studied the determinants of such activity to guide policymakers in designing effective policies to promote exercise. The theoretical framework generally adopted in this literature is the neoclassical approach in which individuals maximize their utility subject to monetary and time constraints. Becker's allocation-of-time model (Becker, 1965) was the main starting point and the basis of the SLOTH[1] framework (Cawley, 2004), which is often the economic approach of reference. In this economic analysis, physical activity is included in the maximization problem in two ways: it either directly affects well-being as any other leisure activity or it is considered an input in the "production" of health with health being an argument of the individual utility function. On this matter, Downward and Rasciute (2010) pointed out that physical activity decisions should be connected to other leisure choices. Humphreys and Ruseski (2011) presented a model in which the participation decision and the time spent on physical activity are considered separate decisions and are both arguments of the utility function. Other theoretical approaches, sometimes called heterodox theories, underscore the importance of previous experience and social interaction in determining preferences and physical activity

* Corresponding author.
E-mail addresses: jaume.garcia@upf.edu (J. García), msuarezf@uniovi.es (M.J. Suárez).

[1] The acronym SLOTH stands for the main arguments of the utility function proposed by Cawley (2004): time spent sleeping (S), time at leisure (L), time devoted to occupation or paid work (O), time in transportation (T), and time spent in home production (H).

https://doi.org/10.1016/j.econmod.2022.106127
Received 20 September 2021; Received in revised form 22 November 2022; Accepted 25 November 2022
Available online 28 November 2022

participation.[2]

Most empirical analyses are based on previous theoretical frameworks. There is a fairly extensive literature on the determinants of physical activity, especially in developed countries, as demonstrated by the reviews performed by Cabane and Lechner (2015), Downward and Muñiz (2019), and Muñiz and Downward (2019). There is a lot of heterogeneity in the sign and/or the effect size of the usual covariates. These heterogeneous results may be due to factors associated with the specific datasets used in the exercises such as cultural and social differences between countries, different years of analysis, and/or the type and definition of the dependent variable. However, they may also be a consequence of how the zeros observed for the time and frequency variables are generated, how the dependent variable is defined, or the specific functional form chosen for the corresponding equation, what we refer to as the "specification assumptions." Based on previous theoretical frameworks, whether the zeros are corner solutions, can be associated with nonpotential participants, or are of both types matters for the consistency of the estimates. Conversely, the definition of the dependent variable, in terms of the dimension of physical activity (participation, time, and frequency) that is considered also matters because the effects of the explanatory variables could differ in size, sign, and significance depending on it. Finally, in analyzing the time devoted to physical activity, the choice of the functional form is relevant. For the first time in the literature, to the best of our knowledge, we propose the use of a Box–Cox transformation of the time variable that nests the standard linear and log-linear models.

This article sheds light on the relevance of the factors associated with the specification assumptions in explaining the heterogeneity in the conclusions regarding the determinants of physical activity behavior. Specifically, we examine, methodologically and empirically, the extent to which the modeling of the zeros, the definition of physical activity (participation, time, and frequency), and the functional form affect the conclusions. There has not yet been any comprehensive comparative analysis of the consequences of modeling assumptions, although some studies have compared different dimensions of physical activity or have examined several econometric specifications for a given dimension (e.g., Meltzer and Jena, 2010; Dawson and Downward, 2011; Borgers et al., 2016). From this perspective, the contributions of this paper may be relevant for the analysis of other leisure activities, such as attendance of cultural events, reading habits, and travel demand, and for all types of empirical analysis in which the dependent variable, either frequency or a continuous variable, can take the value of zero for a significant proportion of the observations.

In our empirical analysis, we use the 2015, 2016, and 2017 waves of the Mexican National Consumer Confidence Survey (Encuesta Nacional sobre Confianza del Consumidor or ENCO), which offers information about four dimensions of sports practice in the previous week (i.e., participation, time, frequency, and intensity).

Our main conclusion is that specification assumptions make a difference. We find disparities in the effect of some variables across dimensions of physical activity and across specifications within each dimension. In addition, the choice of the functional form for the time equation is relevant in terms of the effects of the drivers of physical activity. Therefore, the use of the correct specification is crucial to ensure the success of the public measures implemented to encourage people to become more physically active.

The remainder of the paper is structured as follows. Section 2 provides a summary of the empirical economics literature on mass sports participation, highlighting the diversity in both the definition of the dependent variable and the econometric model. Section 3 describes the data source, the variables, and the empirical models applied. Section 4 discusses the results, and Section 5 concludes.

[2] See Cabane and Lechner (2015) for a summary of the main theoretical explanations for engaging in physical activity.

2. Literature overview

The economics literature on the determinants of physical activity is quite extensive and heterogeneous. Physical activity can be measured in various ways that condition the empirical specification. In this section, we briefly review the different dimensions of physical activity and discuss the econometric models applied in the literature. To some extent, this review complements and updates those of Downward and Rasciute (2010), Downward and Muñiz (2019), and Muñiz and Downward (2019).

2.1. Dimensions of the dependent variable

When analyzing sports practice or physical activity, several issues must be considered as follows. Which activities should be included? How regularly is the activity done? How much time is allocated to the activity? How vigorously is the activity performed? What are the reasons for lack of participation? Some of these questions are related to the FITT principles that characterize physical activity, as defined by Rhodes et al. (2017): frequency (F), intensity (I), time (T), and type (T). The answers depend partially on the researcher's objective and partially the information available from the data source. Consequently, the empirical approaches in the economics literature on sports participation vary widely.

The type of physical activity is related to the first question posed above: which activities should be included? Economic studies usually focus on recreational sports practice or exercise. However, physical activity performed during daily tasks or transportation may sometimes be considered, and activities such as gardening and walking are often included in the study (e.g., Humphreys and Ruseski, 2011).

Frequency (How regularly is the activity done?) and time (How much time is allocated to the activity?) are largely determined by the survey information. Frequency is measured as the number of times or days that the activity is done over a certain period. Downward and Riordan (2007) and Muñiz et al. (2014) analyzed the number of times that individuals play a sport over 4 weeks; Borgers et al. (2016) defined frequency as the number of times per week; Oliveira-Brochado et al. (2017) studied the number of days a week that physical activity was undertaken; and Downward et al. (2011) and García and Suárez (2020) used qualitative information.

Time (How much time is allocated to the activity?) is defined as the number of minutes or hours spent practicing physical activity. García et al. (2011) considered the number of hours per day; Humphreys and Ruseski (2011) and Ruseski et al. (2011) examined the number of hours or minutes allocated per week; Eberth and Smith (2010) and Dawson and Downward (2011) analyzed time over a 4-week period; and Thibaut et al. (2017) considered the number of hours of practice over a 1-year period.

Intensity (How vigorously is it performed?) is the least analyzed dimension of physical activity practice in the economics field. It is usually measured via metabolic-equivalent tasks, as in Meltzer and Jena (2010).

The last question posed at the beginning of this section (What are the reasons for lack of participation?) is an important issue in the analysis of physical activity practice because of the high proportion of people who do not exercise. In fact, many authors focus on the factors associated with the probability of exercising, so the variable is defined as binary (i.e., yes or no), or establish a minimum level of participation in some cases. However, this dimension is generally jointly analyzed with time or frequency, particularly when two-equation models are used.

2.2. Empirical methods

Logit and Probit are the most common econometric models applied to the study of the participation probability (e.g., Kokolakakis et al., 2012; Dallmeyer et al., 2017). However, the participation decision is often examined together with the frequency or time spent practicing to the extent that the observed values for these variables are zero when an individual has not spent any time exercising during the reference period. The main issue in the literature is how to model nonparticipation (e.g., Dawson and Downward, 2011; Humphreys and Ruseski, 2011, 2015; Thibaut et al., 2017). There may be different reasons for observing a zero, and each one is associated with a particular model. Humphreys and Ruseski (2011) distinguished between genuine and nongenuine zeros following the definition of Jones (2000), who stated that a genuine zero corresponds to an actual choice of nonparticipation. According to the model specified by Humphreys and Ruseski (2011), in which both the decision to practice physical activity and the time devoted to it are based on the assumption of utility maximization, we could make a distinction between a zero associated with the participation decision (nonpotential participant) and a zero linked to the choice of time, which also results in an optimal decision and can be defined as a corner solution.

Given the specific characteristics of physical activity participation, most zeros could be associated with nonpotential participants, especially when the survey questions refer to regular practice. However, other surveys collect information about a particular period (e.g., last week or last month). Therefore, some zeros could come from potential participants who did not practice during the period of time due to their health condition or professional commitments, for example.

Two-part models and sample selection models can be associated with zeros that correspond to nonpotential participants, Tobit models with zeros coming from corner solutions, and double-hurdle models with both types of zeros that nest both the Tobit and the sample selection model in their general formulation (Jones, 2000).

The debate over the two-part versus the sample selection models that has been very intense in the health economics literature (Duan et al., 1984; Madden, 2008) and that applies here is worth mentioning. The sample selection model has been used in the limited dependent variable literature to deal with the estimation of an equation defined for positive observations only but not as it was originally introduced (i.e., to solve for missing data or an observability problem). In the case of physical activity participation, there does not seem to be a missing data problem and researchers are not generally interested in the expected value of the dependent variable for those who are not potential participants. By contrast, the two-part model is doubly attractive. It is simple to estimate and it can be understood as a Taylor series approximation to a more general expression of the conditional expected value of time. However, some distributional assumptions (such as log-normality, for instance) or some orthogonality conditions are required to determine the conditional and unconditional expected values of the dependent variable (Duan, 1983). These distributional requirements are also present in the sample selection model and, in this case, exclusion restrictions are sometimes necessary to obtain precise estimates because the correction term is almost linear for most observations. In fact, the two-part model can be interpreted as if the reference population is a potential participant.

All of the aforementioned models have been applied in the sports economics literature. Examples of Tobit estimates can be found in Ruseski et al. (2011) and Thibaut et al. (2017). In other cases, researchers use the Heckman's sample selection model (e.g., Downward and Riordan, 2007; García et al., 2011)[3] or the two-part model (Humphreys and Ruseski, 2011; García and Suárez, 2020). A double-hurdle model is used by Humphreys and Ruseski (2015).

Count data models have generally been applied to determine the number of times or days that one practices physical activity; zero-inflated count data specifications are sometimes estimated to take into account the two possible sources of optimal zeros (e.g., Dawson and Downward, 2011; Muñiz et al., 2014). In fact, a standard count data model is equivalent to the Tobit model when the dependent variable is a count variable, whereas the zero-inflated model is the counterpart of the double-hurdle model in its independent version.[4] Two-part count data models have been estimated in the literature (Oliveira-Brochado et al., 2017). When physical activity frequency is an ordinal (qualitative) variable, a standard ordered model or its zero-inflated version, which is a combination of a discrete choice model and an ordered model, is often used (Downward et al., 2011; Downward and Rasciute, 2015).

3. Data and methods

In this section, we describe the dataset used in the empirical analysis and the dependent and independent variables. We also provide some descriptive statistics. We briefly comment on the main characteristics of the econometric models applied in the empirical analysis.

3.1. Data and variables

The ENCO is a monthly database that is the result of a joint project between Mexico's Central Bank and the National Institute of Statistics and Geography. In our empirical analysis, we pooled the data from November questionnaires conducted between 2015 and 2017 that provide information about physical activity practice. Because the survey is a rotating panel, we selected a random subsample of all available households, so that no one appeared twice in the sample. The selected sample consists of adults (people of 18 years of age or older) who answered the survey module on sports practice and physical exercise and provided information on all of the variables included (N = 4299 observations).

The main advantage of this dataset is that it includes questions about several dimensions of physical activity, making it possible to use different econometric models and different measures of practice to check the importance of the specification assumptions for the results obtained. The specific questions asked were as follows: Do you do sports (soccer, basketball, karate, etc.), exercise (walking, cycling, aerobics, etc.) or both? How many days last week? About how many minutes did you do per day last week?

Consequently, three alternative dependent variables were used, all of them referring to the week prior to the survey: the participation decision (a dummy variable that is equal to 1 if the person did physical activity) and the minutes per week and the number of days per week allocated to physical activity. The descriptive statistics of these variables are included in Table 1 for the total sample and for a subsample of the participants.

About 40% of the individuals in the sample did some physical activity. The participants had done physical activity on 3–4 days over the previous week and allocated an average of 3 and one-half hours to this activity. It must be pointed out that only 51 of the 1785 individuals who stated that they did some physical activity in their leisure time had not done so during the last week. This is a feature that will be important for the econometric results discussed in the next section.

Turning to the covariates, we have included personal and family characteristics as well as income information. Gender is defined through a dummy variable that is equal to 1 for men (*Male*); age is measured in years; marital status is included as a dummy variable that is equal to 1 if married or in a cohabiting couple (*Married*); and family composition is measured as the number of children in the household under the age of 12 (*# Children < 12*). Education has four categories with corresponding dummies defined for primary, secondary, upper secondary, and higher

[3] Eberth and Smith (2010) estimated the sample selection model using flexible parametric forms based on a copula approach in which no normality assumption was required to define the joint cumulative distribution function of both dependent variables.

[4] As mentioned in Jones (2000), hurdle and two-part models are sometimes used as synonymous terms in the count data literature, but they are not synonymous.

Table 1
Descriptive statistics.

	Total sample		Physical activity participants	
	Mean	St. Deviation	Mean	St. Deviation
Participation	0.403	0.491	1.000	0.000
Time (minutes per *week)*	85.447	150.814	211.843	172.098
Frequency (days per *week)*	1.477	2.140	3.662	1.831
Male	0.434	0.495	0.5083	0.500
Age	43.269	16.350	41.389	15.833
Married	0.574	0.495	0.544	0.498
#Children < 12	0.630	0.917	0.563	0.882
Education				
Primary	0.224	0.417	0.158	0.365
Secondary	0.247	0.431	0.206	0.405
Upper Secondary	0.182	0.386	0.192	0.395
Higher	0.348	0.476	0.444	0.497
Worker	0.637	0.481	0.633	0.482
Individual earnings	1.065	1.398	1.217	1.720
Other earnings	1.349	1.829	1.477	2.163
# observations	4299		1734	

education. Employment status is included as a dummy variable (*Worker*) that is equal to 1 for workers. Two economic variables are also included: individual net weekly earnings (*Individual earnings*)[5] and the net weekly earnings of other household members (*Other earnings*), which are both measured in thousands of Mexican pesos.

These socioeconomic variables are common in the literature on this topic, and their inclusion may be justified based on the time allocation models discussed in Section 1. Apart from their potential influence on individual preferences, gender, marital status, family composition, and employment status may affect the physical activity choice because of their effect on time restrictions. Workers have less leisure time as do married women with children because of the unequal distribution of domestic tasks by gender. Young people do more physical activity, possibly because they have fewer time constraints, are in better physical condition, and due to their social relationships. Educational level often has a positive effect on physical activity practice; the main argument for its inclusion is that higher-educated people are more aware of the health benefits of physical activity. Regarding the economic variables, physical activity generally implies travel, clothing, sports equipment, and facility fees expenses. Therefore, on one hand, a higher income may have a positive impact on demand, but since labor earnings represent a leisure time opportunity cost, they may, on the other hand, also decrease the time allocated to this activity.

Table 1 provides the summary statistics of these variables. The proportion of men is higher among the participants, unlike married status. In addition, people who did physical activity were somewhat younger, had a higher educational level, had more earnings, and lived in households with fewer children than those who did not engage in this practice.

3.2. *Methods*

Various econometric specifications were implemented according to the dimensions of physical activity analyzed in this paper (participation, time, and frequency).[6] First, a Probit model was estimated to analyze the factors associated with the probability of doing physical activity that also constitutes the first stage of the two-part and sample selection models.[7]

Second, we estimated the Tobit, sample selection, two-part, and double-hurdle models to simultaneously deal with physical activity participation and the time devoted to this activity. All of these are two-equation models, with the exception of the Tobit model in which the same equation explains both participation in and the number of minutes devoted to physical activity.

The sample selection model (Heckman, 1979) can be understood as a simple way of overcoming the limitation of the Tobit model having only one equation to explain two dependent variables: participation and time. It assumes that the subsample of participants is not randomly selected from the population, so there may be a sample selection problem that is corrected by the specification of two equations (participation and time) the error terms of which may be correlated. The Heckman approach is appropriate in cases in which the dependent variable is not always observed and the researcher is interested in the mean response of the population. However, when a physical activity participation model is estimated, there is not a missing data problem but rather a problem of zeros being associated with nonparticipation. From an econometric point of view, Heckman's model is the same for dealing with a missing data problem or a problem of zeros, but the interpretation of the results is not exactly the same because the unconditional expectation of the dependent variable varies according to the problem. In the case of a zero problem, it is the product of the expectation of time conditional on being a participant—which includes a correction term—and the probability of participating. Thus, all of the variables in both equations could potentially affect the unconditional expectation of time devoted to physical activity.[8]

The two-part models exhibit some similarities to the sample selection model because both assume a two-stage decision procedure. First, individuals choose whether to do physical activity. Those who choose to participate then decide how much time to allocate to that activity, and this amount is positive. In the case of two-part models, the two equations are independent, so the likelihood function can be split into two components with each one associated with a dependent variable, allowing the separate estimation of each equation. The first stage is estimated via a Probit (Logit) model, whereas the second stage consists of estimating the time equation for a subsample of participants. In the time equation, the dependent variable is a positive random variable, and the distribution for this feature must be specified (i.e., log-normal, Gompertz, truncated normal at zero or gamma, among others).

The double-hurdle model is also a two-equation model in which two hurdles must be cleared to observe a positive value for the time variable. The first step consists of the potential participation decision, and the second stage refers to the amount-of-time decision. The main difference from the previous models is that here we may observe zeros among potential participants because some of them may not have exercised during the recorded period. Therefore, the double-hurdle model allows two types of zeros: those coming from people who would never participate (i.e., nonpotential participants) and those who are potential participants but who have not done any physical activity during the specific period under analysis (i.e., a corner solution). In this case, the probability of nonparticipation is equal to the probability of either being a nonpotential participant or being a potential participant but choosing zero minutes.

[5] We estimated an earnings equation with a subsample of individuals who offered information about this variable to impute earnings for the individuals who did not report it, taking into account potential sample selection problems.

[6] Although the Mexican survey contains information on how vigorous physical activity practice is, the results for this dimension are not included because the information is very basic and imprecise (i.e., whether the practice is vigorous or not). The results for a Probit model with sample selection are available on request. They indicate that males, workers, and young people are more likely to engage in vigorous physical activity.

[7] A Logit model could also be estimated, but the Probit estimates are reported because the normality assumption of the errors is assumed in most the two-equation models, and the Logit and Probit models do not differ very much, unless there are many observations in the tails of the distributions.

[8] In fact, the presence of this correction term in the time equation may justify the empirical findings of some imprecise estimates of the coefficients of this equation and the consideration of exclusion restrictions, although they are not strictly necessary.

The maximum likelihood estimation of all of these models requires some strong distributional assumptions. In particular, it is well known in the microeconometrics literature that normality and homoscedasticity, the two most imposed constraints, are necessary for the consistency of maximum likelihood estimates (Amemiya, 1984). However, the normality assumption is violated whenever the dependent variable has a skewed distribution, as is the case of the minutes-per-week variable used as the dependent variable in analyzing the time dimension. Sometimes this issue has been solved by using a logarithmic transformation of the dependent variable, but a more flexible approach has been applied in the empirical demand literature (Yen and Jones, 1996; Aristei and Pieroni, 2008; Artero et al., 2019) consisting of the following Box–Cox transformation (Box and Cox, 1964) of the dependent variable (y) as follows:

$$y_i^L = \begin{cases} \dfrac{y_i^\theta - 1}{\theta} \; for \; \theta > 0 \\ \ln(y_i) for \; \theta = 0 \end{cases} \tag{1}$$

The attractiveness of this transformation is that it nests, as particular cases, the two usual specifications of the time variable in the physical activity literature: the linear model ($\theta = 1$) and the log transformation ($\theta = 0$). This Box–Cox transformation of the dependent variable will also be used in our empirical exercise.

When estimating the frequency of participation, we applied the count data and ordered Probit models using the number of days on which an individual had done physical activity as the dependent variable. In particular, we used Poisson, Negative Binomial (NB), ordered Probit, and two-part models (consisting of a Probit in the first stage and a truncated Poisson or an ordered Probit for the subsample of participants in the second stage) and zero-inflated versions of the Poisson and NB models.

The Poisson, NB, and ordered Probit models are single-equation specifications that simultaneously explain participation and frequency, but they differ in terms of their distributional assumptions. In fact, the Poisson model is very restrictive because of the equidispersion property (i.e., the expected value and the variance are equal) whereas the NB model allows for overdispersion, but both the Poisson and NB models impose a particular structure for the probability of the dependent variable taking each value. The ordered model (McKelvey and Zavoina, 1975), which was not designed to deal with count data but rather with an ordered categorical variable, is more flexible in adjusting the structure of the probability because the cut-off points for the latent variable are estimated and not predetermined.

As previously mentioned, the previous models are the "Tobit" version for count variables, because one equation explains both participation and frequency. It is possible to define two-equation models similar to the two-part and double-hurdle models considered for the time variable. In the case of the two-part models, there is an equation for the participation decision that is estimated by applying a Probit (Logit) model in which the second stage consists of the estimation of the number of days of participation using a truncated (at zero) Poisson model. The models that are equivalent to the double-hurdle model are the so-called "zero-inflated count data models" (i.e., Poisson or NB) designed to deal with the empirical "excess of zeros" problem. The zero-inflated models assume independence between the two equations.

4. Results

In this section, we present the results of several models for different dimensions of physical activity and compare the estimated coefficients. Because the models are nonlinear, we also compute and discuss the average marginal effects of the covariates on the probability of engaging in physical activity, on the expected value of time/frequency, and on the expected time/frequency conditioned to participation. We will discuss the results for each dimension of physical activity (participation, time, and frequency), emphasizing the differences between the specifications.

4.1. Participation and time

Table 2 offers information about the estimated coefficients of the participation equation, which is also the first step in the two-part models, and the time equation. For the participation equation, we report the Probit estimates whereas for the time dimension, we provide the estimates of Tobit, Heckman, two-part, and double-hurdle models using the original time variable (minutes of exercise last week) and/or the logarithmic and Box–Cox transformations.

Starting with participation, the Probit and the first-stage Heckman estimates are practically the same, which is why the latter are not reported in the table.[9] All variables included have a significant effect in explaining participation. As usual in the literature, the likelihood of exercising is higher among males and highly educated individuals (Thibaut et al., 2017; Downward and Muñiz, 2019). Young people are also more likely to participate. Being married and the number of children under the age of 12 reduce the probability of participation, and workers are less likely to engage in physical activity. The negative effect of these variables, which has also been found in other studies (Cabane and Lechner, 2015; Downward and Muñiz, 2019), may be due to greater time constraints, since work and housework reduce the amount of time available for leisure activity. Both one's own and others' household earnings increase participation in line with the literature (Cabane and Lechner, 2015; Thibaut et al., 2017; Downward and Muñiz, 2019). Notice that the results of the Probit model are identical in sign and significance to those of the Tobit model, in which one equation explains both participation and time. However, the coefficients of the Probit model do not seem to satisfy the proportionality between the Probit and Tobit estimates when the Tobit model is the correct model.[10]

By contrast, most of the variables lose their significance in the first stage of the double-hurdle model. It is worth recalling that in this specification, the zeros come from two sources: people who do not want to engage in physical activity (first stage of the double-hurdle model) and people who do but whose optimal participation in the recorded period is zero. Because the participation equations are not equivalent in terms of the dependent variable, it may be that the nonsignificant variables in the first hurdle (i.e., gender, age, children, employment status, and others' household earnings) influence participation through their effect on the probability of participating. Only marital status, education, and earnings remain significant, and the sign of the marital status coefficient changes in comparison to the Probit model, which means that the zeros for married people are mostly generated by the second hurdle.

Regarding the estimates of the time equations, there is a considerable difference between the Tobit model (second column of Table 2) and the sample selection and two-part models (third and seventh column of Table 2) and also between the last two models mentioned. Thus, the econometric modeling may be one reason explaining the heterogeneous results in the literature. Neither gender nor weekly earnings are significant in the two-equation models, whereas employment status has a significant coefficient in the two-part model but not in the sample selection model. Additionally, when one considers the sample selection model with the time variable in logs (sixth column), the results change substantially in terms of sign and significance when the Heckman specification is compared to the original time variable (third column). This could be explained by the negative and significant estimate of the covariance between the error terms of the participation and time equations when the log transformation is used versus the nonsignificant covariance when the original time variable is used. Notice that in the

[9] The sample selection model was estimated by maximum likelihood, and this explains why the participation equation estimates are not numerically the same as in the Probit model.

[10] An intuitive (visual) test for the appropriateness of the Tobit specification is to check whether the coefficients of the Probit are those of the Tobit model divided by the standard deviation of the error term.

Table 2
Participation and time equations: The coefficients.

	Probit	Tobit	Heckman	Double-hurdle		Heckman	Two-part	Tobit	Heckman	Two-part
	Partic.	Time	Time	Pot. Partic.	Time	ln (Time)	ln (Time)	Box–Cox	Box–Cox	Box–Cox
Male	0.334[a]	81.471[a]	1.851	0.185	64.215[a]	−0.177[a]	−0.022	95.220[a]	−0.120	−0.037
	−0.006[a]	−2.063[a]	−0.877[a]	−0.007	−1.301[b]	0.000	−0.004[a]	−3.399[a]	−0.006[c]	−0.009[a]
Married	−0.070[c]	−29.973[a]	−29.590[a]	0.327[b]	−60.575[a]	−0.057	−0.090[b]	−42.680[a]	−0.170[b]	−0.193[b]
#Children < 12	−0.081[a]	−26.554[a]	−14.538[a]	0.111	−34.036[a]	0.000	−0.053[b]	−41.979[a]	−0.081	−0.112[b]
Education (ref.: Primary)										
Secondary	0.102	18.075	–	0.081	–	–	–	–	–	–
Upper secondary	0.264[a]	59.976[a]	–	0.565[b]	–	–	–	–	–	–
Higher	0.469[a]	113.153[a]	–	1.594[b]	–	–	–	–	–	–
Worker	−0.324[a]	−73.651[a]	−9.202	−0.292	−76.685[a]	0.058	−0.111[b]	−101.661[a]	−0.118	−0.204[b]
Individual earnings	0.067[a]	9.347[b]	−2.070	0.320[b]	7.543	−0.045[a]	−0.002	24.106[a]	−0.026	−0.005
Other earnings	0.024[b]	7.486[a]	4.059[b]	0.029	6.543[b]	0.004	0.017[c]	11.048[a]	−0.027	0.034[c]
Constant	−0.164	−15.874	278.799[a]	0.395	110.415[a]	5.863	5.390[a]	86.047[a]	8.815[a]	7.977
σ		288.580[a]	169.636[a]		270.265[a]	0.968[a]	0.772[a]	353.776[a]	1.523[a]	1.504[a]
			−0.032	−0.115		−0.752[a]			−0.255	
θ								1.038[a]	0.132[a]	0.134[a]
Log L	−2759.56	−13,735.78	−14,121.18	−13,723.43		−13,549.74	−13,558.85	−13,678.42	−13,544.67	−13,545.20
AIC	5541.11	27,495.56	28,284.36	27,488.87		27,141.48	27,157.70	27,376.84	27,133.34	27,132.40

σ is the standard deviation of the error term of the time equation.
ρ is the correlation coefficient between the error terms of the two-equation models.
θ is the parameter of the Box–Cox transformation.
AIC is the Akaike information criterion.
[a] $p < 0.01$.
[b] $p < 0.05$.
[c] $p < 0.10$.

two-equation models, the education variable is not included in the second equation (time). This is a convenient (but not strictly necessary) constraint imposed to identify the parameters when the Heckman selection model is used. In this particular exercise, education was discarded because when it was included, the coefficients were not significant and the remaining estimated coefficients did not change. In the case of the Tobit model, education dummies were included because the same equation explained both participation and time.

The estimates of the two-part model with the Box–Cox transformation shown in the last column of Table 2 have the same sign and significance as those corresponding to the log transformation. This is because the parameter θ, although significant, is very close to zero. This is not the case for the sample selection models, for which even with a θ coefficient close to zero, the results (ninth column of Table 2) differ substantially from those obtained when the dependent variable is measured in logs (sixth column of Table 2). This could be a consequence of the substantial change in the correlation coefficient —from −0.752 and significant when the log transformation is used to −0.255 and not significant when the Box–Cox transformation is used— which affects the precision of the time equation estimates. It is also important to point out that the results from the Tobit model with the Box–Cox transformation are in the same direction in terms of sign and significance as when the original variable is used. In fact, the estimate of θ (1.038), although significantly different from 1, is very close to this value, so the Box–Cox transformation does not imply a modification of the original variable unlike the two-part and the sample selection models.

The double-hurdle model deserves special consideration. The coefficients of the second equation are highly significant, with the exception of the weekly earnings variable, and they have the same signs as in the Tobit model. By contrast, the coefficients of the first hurdle (the potential participation equation) are quite imprecisely estimated, as previously mentioned. Thus, the second hurdle, which refers to the determinants of (potential) time, also seems to dominate the generation of the zeros, but this is a consequence of the time variable not being transformed. When a more flexible transformation is used, the interpretation of how the zeros are generated changes completely. In fact, although the Box–Cox transformation has also been applied to the double-hurdle model, no results are reported because it collapses into the sample selection model (i.e., all of the zeros generated in the first hurdle correspond to nonpotential participants).[11]

When looking at the goodness of fit of the different models by means of the Akaike information criterion (AIC), we can conclude that the Box–Cox transformation offers a substantial improvement in comparison to the models in which the original variable is not transformed. The lowest value of the AIC (the best fit) is associated with the two-part Box–Cox model, although it is very similar to that of the sample selection model, because the correlation between the error terms is not significantly different from zero. The two-equation models with the Box–Cox transformation clearly outperform the Tobit version, with substantially different implications about how the zeros are generated, and they also outperform the double-hurdle model without transforming the time variable. This last result is in agreement with what could be expected from the descriptive analysis of the participation variable, because only 51 out of 2565 observations can be considered potential participants (i.e., generated in the second equation or time equation).

It is worth paying attention to the models in which the time variable is not transformed, because this is the most common specification in the literature (Columns 1–5 of Table 2). The double-hurdle model outperforms the Tobit and sample selection models, which are nested in the former, and the Tobit model has a better fit than the sample selection model. This clearly illustrates how important the choice of the functional form for the dependent variable is in terms of both the selection of the "best" model and the implications for the behavior of the individuals (i.e., how the zeros are generated).

As previously mentioned, all of these models are highly nonlinear, so the coefficients are not informative about the size of the effects of the explanatory variables or, in some cases, about the sign of the effects. Thus, it is important to compute the marginal effects, which are different for each individual because they depend on the values of the explanatory variables.

Table 3 shows the mean values of the marginal effects of each

[11] This is because with the Box–Cox transformation, the condition associated with being a potential participant depends on $(1/\theta)$, where θ is the parameter of the Box–Cox transformation. Because the estimate of this parameter is very small (around 0.13), it has a huge influence on the probability of being a potential (or a nonpotential) participant.

explanatory variable (x) calculated for all individuals. In particular, we compute the change in the probability of practicing physical activity ($\partial \Pr(y > 0)/\partial x$), in the unconditional expected time ($\partial E(y)/\partial x$) and in the expected value of the time conditional on being positive ($\partial E(y/y > 0)/\partial x$). The marginal effects reported correspond to the Tobit, sample selection, and double-hurdle models included in the first five columns of Table 2. We chose these models because in all of them, the dependent variable (time) is not transformed, as is usual in the empirical literature on the topic. Table 3 shows the marginal effects for the two-part models with both the log and Box–Cox transformations.[12] The sample selection model with the Box–Cox transformation was discarded because it was statistically equivalent to the two-part model, and the Tobit versions were discarded because the evidence previously presented does not support this particular specification.

Focusing on the two-part model with the Box–Cox transformation, which was the preferred specification (last column in Table 3), the interpretation of the marginal effects is as follows. Being a male increases the probability of practicing physical activity by 12.4 percentage points (pp) on average, increases the number of minutes per week devoted to sports practice by 24.1 min, and decreases the expected number of minutes conditional on practicing physical activity by 3.7 min, although the last marginal effect is not significantly different from zero because the coefficient of the gender variable (*Male*) in the second equation of the two-part Box–Cox model is not significant. For quantitative variables such as age and earnings, the marginal effects measure approximate changes in the corresponding dependent variable when the covariate increases by one unit. According to this model, the variables with the greatest impact on the probability of doing sports are gender, educational level, and labor situation. This probability is around 12 pp higher for men, 18 pp higher for highly educated people, and 12 pp lower for workers. Regarding the time allocated to sports by participants, marital and labor status are the variables with the highest effect, reducing it by 20 min a week approximately, followed by the number of children in the household, which reduces the time spent on sports by participants by around 11 min per child. These results highlight the importance of preferences that may vary by gender and education and the time barriers generated by housework and work when deciding whether to play sports and for how long.

As reported in Table 3, there are no substantial differences in the effects of the independent variables on the probability of being a participant between specifications. Regarding time, we can observe differences in some marginal effects depending on whether the variable is transformed or not. This is the case for gender: the marginal effect on the conditional expectation is small and not significant when the Box–Cox transformation is used, but it is positive and highly significant when the Tobit or double-hurdle model is used. This is a consequence of the significance of the gender variable in the second equation of the models. The marginal effects of gender on the unconditional expectation are much more similar, but they are higher in the models in which the dependent variable is not transformed. In the case of weekly earnings, different signs of the marginal effects on the conditional expectation are found depending on the transformation of the dependent variable. For the remaining covariates, the main differences are in the size of the effects. Finally, the effect of individual earnings is generally greater in absolute value, when significant, than that of other household earnings.

4.2. Frequency

Table 4 provides information about the coefficients of the models estimated for the number of days per week that people do physical activity. Because this is a count variable, we applied standard versions of count data models (i.e., Poisson and NB). Since the dependent variable can only take eight values (0–7), we also estimated a standard ordered Probit, treating the number of days as an ordered categorical variable. Finally, we estimated the two-part versions of these models, with a Probit model for the participation equation, as well as the zero-inflated versions.[13]

In the standard Poisson, NB, and ordered Probit models, we find that the participation equation dominates the results again in terms of sign and significance of the coefficients. Moreover, the coefficients of these three models have the same sign as the Probit (and Tobit) model in Table 2, but the precision of the estimates of the NB model is much lower than that of the Poisson model, as usual. Notice that there is a huge improvement in the log-likelihood when we allow for overdispersion, and the coefficient (α) for the parameterization of the variance in terms of the expected value is highly significant.[14] This is reflected neither in the coefficients nor in the marginal effects on the expected number of days, to a certain extent, as we will see later. Finally, the ordered Probit model provides a much better fit because it imposes a much more flexible structure for the probabilities of the values of the dependent variable.

We also estimated the two-part versions of the standard models in which the number of days only takes a positive value in the second equation (the truncated model). Additionally, since the frequency variable refers to the week, we also take into account the upper truncation in 7. In Table 4, we report the estimates for the double-truncation version (values from 1 to 7) of the Poisson and NB models and for the ordered Probit model. The results of the first stage are those corresponding to the Probit model presented in the first column of Table 2 and discussed in Subsection 4.1.

There is a substantial improvement in the value of the log-likelihood function when the two-part versions are used, in particular, for the count data models. However, the ordered model still has a better fit. As in the two-equation models for time, this huge improvement in the fit compared to the standard versions translates into different effects of the covariates in both equations.

In particular, the effect of the number of children younger than age 12 and the individual earnings variables lose their significance in the second step of most two-part models. However, employment status is significant in all cases. According to the Probit estimates of the first equation, workers are less likely to participate, and those who do, allocate fewer days to sports practice. Downward and Rasciute (2015) obtained the same effect of this variable on the frequency of practice. The rest of the covariates show differences in the results depending on the specification. As in the analysis of time, when participation and frequency are modeled as separate decisions, some variables have a different effect on the probability of doing sports and on the number of days. Specifically, men and young people are more likely to do sports, but they practice it fewer days per week than women and older people, respectively; individual and others' earnings do not significantly affect frequency, a result similar to that of Oliveira-Brochado et al., 2017 who found a nonsignificant effect of social class, but it increased participation. Even when we focus on the second stage of the two-part models,

[12] The calculation of the marginal effects on $\partial E(y/y > 0)/\partial x_j$ with the Box–Cox transformation would require numerical integration. Instead, we use the proposal by Abrevaya (2002) based on a flexible estimator (a smearing estimator) as proposed by Duan (1983).

[13] A linear regression model for the number-of-days-per-week variable was also used, as it is sometimes found in the empirical literature. Although this model does not take into account the nonnegative, integer, and limited features of the dependent variable, the results are the same in terms of the sign of the effect of the explanatory variables as those for the standard count data models. However, the significance of the estimated coefficients is not as clear as in the standard Poisson model.

[14] The NB model estimated is the Type II version, where $var(y) = E(y) + \alpha[E(y)]^2$. Conversely, the presence of overdispersion is evident when we look at the sample mean and the sample variance of the dependent variable (which are 1.48 and 4.58, respectively).

Table 3
Participation and time equations: Marginal effects.

		Tobit	Heckman	Double-hurdle	Two-part	Two-part
		Time	Time	Time	ln (Time)	Box–Cox
Male	∂Pr (y > 0)/∂x	0.106	0.124	0.096	0.124	0.124
	∂E(y)/∂x	33.136	27.168	29.905	24.573	24.137
	∂E (y/y > 0)/∂x	25.897	3.081	23.453	−4.621	−3.661
Age	∂Pr (y > 0)/∂x	−0.003	−0.002	−0.002	−0.002	−0.002
	∂E(y)/∂x	−0.825	−0.858	−0.675	−0.882	−0.838
	∂E (y/y > 0)/∂x	−0.650	−0.901	−0.486	−0.916	−0.852
Married	∂Pr (y > 0)/∂x	−0.039	−0.026	−0.038	−0.026	−0.026
	∂E(y)/∂x	−12.074	−17.563	−16.518	−13.517	−13.172
	∂E (y/y > 0)/∂x	−9.473	−29.850	−20.260	−19.342	−18.968
#children < 12	∂Pr (y > 0)/∂x	−0.034	−0.030	−0.029	−0.030	−0.030
	∂E(y)/∂x	−10.620	−12.140	−10.977	−10.968	−10.598
	∂E (y/y > 0)/∂x	−8.361	−14.836	−11.654	−11.351	−10.966
Education (ref.: Primary)						
Secondary	∂Pr (y > 0)/∂x	0.022	0.036	0.013	0.036	0.036
	∂E(y)/∂x	6.144	7.552	2.936	7.734	7.486
	∂E (y/y > 0)/∂x	5.135	0.393	0.444	–	–
Upper secondary	∂Pr (y > 0)/∂x	0.076	0.096	0.086	0.096	0.096
	∂E(y)/∂x	21.979	20.306	19.396	20.588	19.927
	∂E (y/y > 0)/∂x	17.811	1.004	2.956	–	–
Higher	∂Pr (y > 0)/∂x	0.147	0.175	0.185	0.175	0.175
	∂E(y)/∂x	45.401	37.165	41.274	37.447	36.246
	∂E (y/y > 0)/∂x	35.582	1.742	6.497	–	–
Worker	∂Pr (y > 0)/∂x	−0.095	−0.118	−0.122	−0.118	−0.118
	∂E(y)/∂x	−30.480	−29.177	−37.369	−35.940	−33.468
	∂E (y/y > 0)/∂x	−23.750	−10.381	−28.271	−23.938	−20.243
Individual earnings	∂Pr (y > 0)/∂x	0.012	0.024	0.042	0.024	0.024
	∂E(y)/∂x	3.739	4.357	10.479	5.019	4.851
	∂E (y/y > 0)/∂x	2.943	−1.823	3.865	−0.523	−0.523
Other earnings	∂Pr (y > 0)/∂x	0.010	0.009	0.011	0.009	0.009
	∂E(y)/∂x	2.994	3.501	3.295	3.325	3.182
	∂E (y/y > 0)/∂x	2.357	4.148	2.429	3.523	3.328

Note: ∂Pr(y > 0)/∂x represents the marginal effect on the probability of practicing physical activity; ∂E(y)/∂x the marginal effect on the unconditional expectation of time; and ∂E(y/y > 0)/∂x the marginal effect on the expected value of time conditional to being positive.

Table 4
Frequency equations: The coefficients.

	Poisson	Negative binomial	Ordered Probit[a]	Two-part (Second step)			ZIP	
	#days	#days	#days	Truncated Poisson (1–7)	Truncated NegBin (1–7)	Ordered Probit[a]	Pot. Partic.	#days
Male	0.204***	0.195***	0.229***	−0.127***	−0.202***	−0.228***	0.353***	−0.101***
Age	−0.002**	−0.003	−0.003**	0.005***	0.008***	0.008***	−0.007***	0.004***
Married	−0.073***	−0.063	−0.064*	−0.063*	−0.100*	−0.093*	−0.066	−0.050*
#children < 12	−0.095***	−0.096***	−0.073***	−0.028	−0.043	−0.053*	−0.079***	−0.024
Education (ref.: Primary)								
Secondary	0.104**	0.086	0.088	–	–	–	0.101	–
Upper secondary	0.261***	0.238**	0.220***	–	–	–	0.269***	–
Higher	0.495***	0.477***	0.415***	–	–	–	0.478***	–
Worker	−0.341***	−0.348***	−0.283***	−0.141***	−0.223***	−0.250***	−0.325***	−0.113***
Individual earnings	0.023***	0.032	0.029*	−0.010	−0.012	−0.011	0.074***	−0.008
Other	0.025***	0.023	0.023**	0.014*	0.020	0.022*	0.024**	0.009
Constant	0.372***	0.406***		1.308***	1.416***		−0.128	1.266***
α		2.735***			−1.700***			
Log L	−8943.10	−6779.30	−6000.23	−6051.02	−6020.68	−5910.49	−6123.32	
AIC	17,908.20	13,582.60	12,034.46	12,140.04	12,081.36	11,868.98	12,284.64	

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.
ZIP is the zero-inflated Poisson model.
AIC is the Akaike information criterion.
[a] No constant term is reported for the ordered model, since (6–7) cut-off points are estimated to define the intervals associated with each category. α is the over-dispersion parameter of the negative binomial model.

some differences arise depending on the specification assumptions. For example, the number of children reduces the frequency of physical activity practice according to the ordered Probit model, but it is not significant in the other specifications; the earnings of other household members increased the frequency of practice in some cases.

Turning now to the zero-inflated models, the zero-inflated NB model did not converge for the same reasons pointed out previously in the discussion about the two-part models. On the other hand, the zero-inflated version of the ordered Probit model is not reported because it converges to the two-part model (i.e., all of the zeros are generated in the first equation).[15] The fit of the zero-inflated Poisson model (ZIP) reported in Table 4 is much better than that of the standard model. This is because allowing the zeros to be generated by a different model matters. When we look at the coefficients, the estimates of the potential participation equation are very similar to those of the participation equation in Table 2.

According to the goodness of fit, we can conclude that a two-equation model is the most appropriate for the analysis of frequency, as in the case of time. Additionally, the two-part structure in which the zeros are treated as corresponding to nonpotential participants seems to be more appropriate, and the ordered models capture the empirical frequency patterns observed for this dataset better than the standard count data specifications.

Given the nonlinear nature of the models, it is advisable to compute the marginal effects. In Table 5, we report the same type of marginal effects that we calculated in the previous subsection for the time variable, and the interpretation is similar except that now the variable under study is the number of days of physical activity practice per week.

Comparing one-equation versus two-equation models (i.e., two-part and zero-inflated specifications), the most notable difference is that the variables for which the coefficient has a different sign and/or different significance in the two equations in the latter case (i.e., gender, age, and weekly earnings) have a marginal effect on the conditional expectation that differs from that of the one-equation models.

Moreover, the marginal effect on the unconditional expectation is very much dominated by the effect on the probability of practicing physical activity. However, even in the cases in which the sign of the marginal effect on the unconditional and conditional expectations is the same, there are still considerable differences in magnitude. This happens, for instance, with the marital status variable, the negative effect of which is higher in absolute value in the two-equation models.

There are no substantial differences between the marginal effects of the two-equation models reported in Table 5. This is surprising because the fit of the two-part ordered model seems to be much better than that of the rest, but we can better appreciate the different performance of these models when we compare the adjusted probabilities with the sample frequencies. In the case of the two-part ordered Probit model, the average of the adjusted probabilities for each value of the dependent variable (from 0 to 7) is almost equal to the observed frequencies, unlike what happens with the count data models. In particular, the average adjusted probability for 1 day is 0.328 for the Poisson model and 0.161 for the NB model, when the sample frequency is 0.046. In the case of 7 days, these numbers are 0.002, 0.013, and 0.051, respectively. Given the specific features of these models, much more attention should be paid to the probabilities instead of focusing only on the expected values.

According to the marginal effect of the two-part models, the variables with the highest impact on the frequency of sports practice conditioned to participate are gender and labor status. On average, the number of days of weekly practice decreases between 0.33 and 0.39 in the case of males and between 0.37 and 0.43 for workers. Being married is another factor with a relevant negative effect on the conditional sports frequency, being that its marginal effect is about one-half that of gender.

5. Conclusions

The benefits to society of a more physically active population include reductions in mortality and morbidity rates, well-being gains, lower health expenditures, and productivity increases. Consequently, the analysis of sports and exercise decisions is important for the design of public policies to promote physical activity. In this regard, there is extensive economic literature on the drivers of physical activity participation with heterogeneous empirical evidence. Among other things, this heterogeneity can be explained by the economic specification derived from the theoretical models and the different dimensions of physical activity that are used to define the dependent variable. The goal of this paper is to contribute to this literature by making a broad comparative analysis of the various econometric methodologies and dimensions of physical activity. We also deal with the issue of functional form and consider the Box–Cox transformation, which allows for a flexible transformation of the dependent variable, encompassing the two usual specifications in the literature (i.e., linear and log-linear). For these purposes, we use an adult sample from the 2015–2017 waves of the Mexican National Consumer Confidence Survey known as ENCO.

We examine three different dimensions of physical activity: participation, time (the number of minutes per week), and frequency (the number of days a week) allocated to this activity. Several specifications were considered and discussed depending on the nature of the dependent variable, including its definition and the presence of a significant proportion of zeros. A Probit model was applied to estimate the probability of participation. Regarding time, Tobit, Heckman, double-hurdle, and several two-part models were estimated. In the case of frequency, Poisson, NB, ordered Probit, zero-inflated, and some two-part models were estimated.

The results indicate that the effects of some covariates differ depending on the specification within each dimension of physical activity—specifically, time and frequency. Therefore, the econometric specification is one reason for some of the heterogeneous results in the literature. In particular, the models that separate participation and time/frequency reveal the uneven effect of some covariates on each decision. This is the case with gender and individual earnings. Both variables increase the likelihood of participation, but they either do not affect or have a negative influence on time and frequency of practice according to most of the two-equation specifications. In general, the AIC benefits the specifications that separate participation from time or frequency, and the Box–Cox transformation from the usual specifications in the literature when the time equation is modeled. There are also substantial disparities in the effect of some variables across dimensions.

In sum, our first conclusion is that the adequacy of the specification, particularly, how the zeros are generated, matters in terms of the consistency of the estimates. Secondly, the dimension associated with the dependent variable is relevant in terms of the effects of the drivers of physical activity participation. Finally, the choice of the functional form is also important, and transformations such as the Box–Cox transformation that nest the standard linear and log-linear models, are preferred.

The specific results of the empirical exercise for Mexico favor the specification of a two-part model in which the zeros are associated with nonpotential participants. Additionally, the effects of the main drivers for the different dimensions of physical activity are consistent with the conclusion that the main obstacles to practicing sports or increasing its practice have to do with preferences and time restrictions that individuals face rather than monetary constraints.

As recommendations for researchers studying the correlates of physical activity practice, it is important to consider which is the variable of interest (i.e., participation, time, frequency, or intensity), how it is defined (period of reference), and the potential reasons for not engaging in this activity (whether they may be corner solutions, infrequency of practice, or lack of interest) because the appropriate model varies accordingly. Information criteria (such as the AIC) can be used to compare the nonnested models estimating a different number of parameters. It is also advisable to perform robustness checks of the specification and the covariates included.

This analysis is also relevant for policymakers designing policies to foster physical activity because the target collectives may vary according to the objective, whether that may be to increase the number of active people, to increase the amount of time that they practice, or to

[15] In fact, when we look at the estimates of the zero-inflated ordered Probit model, the first cut-off point estimate is −5.847. This means that the probability of a zero being generated in the second equation is almost negligible for any individual in the sample.

Table 5
Frequency equations: Marginal effects.

		Poisson	Neg. binomial	Ord. Probit	Two-part models		ZIP
					Trunc. Poisson (1–7)	Ord. Probit	
Male	∂Pr (y > 0)/∂x	0.066	0.032	0.086	0.124	0.124	0.123
	∂E(y)/∂x	0.305	0.292	0.405	0.320	0.296	0.317
	∂E (y/y > 0)/∂x	0.227	0.414	0.223	−0.331	−0.388	−0.332
Age	∂Pr (y > 0)/∂x	−0.001	−0.000	−0.001	−0.002	−0.002	−0.002
	∂E(y)/∂x	−0.003	−0.004	−0.005	−0.003	−0.004	−0.004
	∂E (y/y > 0)/∂x	−0.002	−0.006	−0.003	0.013	0.013	0.012
Married	∂Pr (y > 0)/∂x	−0.024	−0.010	−0.024	−0.026	−0.026	−0.026
	∂E(y)/∂x	−0.109	−0.093	−0.113	−0.163	−0.160	−0.162
	∂E (y/y > 0)/∂x	−0.081	−0.132	−0.062	−0.165	−0.157	−0.165
#children < 12	∂Pr (y > 0)/∂x	−0.031	−0.016	−0.027	−0.030	−0.030	−0.029
	∂E(y)/∂x	−0.140	−0.142	−0.128	−0.139	−0.145	−0.141
	∂E (y/y > 0)/∂x	−0.104	−0.202	−0.071	−0.074	−0.089	−0.081
Education (ref.: primary)							
Secondary	∂Pr (y > 0)/∂x	0.037	0.014	0.032	0.036	0.036	0.035
	∂E(y)/∂x	0.124	0.103	0.137	0.133	0.133	0.130
	∂E (y/y > 0)/∂x	0.086	0.149	0.079	–	–	0.000
Upper secondary	∂Pr (y > 0)/∂x	0.091	0.039	0.081	0.096	0.096	0.096
	∂E(y)/∂x	0.336	0.308	0.360	0.355	0.355	0.355
	∂E (y/y > 0)/∂x	0.239	0.444	0.204	–	–	0.000
Higher	∂Pr (y > 0)/∂x	0.164	0.077	0.157	0.175	0.175	0.174
	∂E(y)/∂x	0.722	0.700	0.731	0.648	0.648	0.644
	∂E (y/y > 0)/∂x	0.532	0.997	0.402	–	–	0.000
Worker	∂Pr (y > 0)/∂x	−0.109	−0.056	−0.106	−0.118	−0.118	−0.121
	∂E(y)/∂x	−0.530	−0.542	−0.508	−0.599	−0.624	−0.607
	∂E (y/y > 0)/∂x	−0.399	−0.766	−0.278	−0.373	−0.430	−0.380
Individual earnings	∂Pr (y > 0)/∂x	0.008	0.005	0.011	0.025	0.025	0.026
	∂E(y)/∂x	0.035	0.047	0.050	0.080	0.083	0.086
	∂E (y/y > 0)/∂x	0.026	0.067	0.028	−0.026	−0.018	−0.025
Other earnings	∂Pr (y > 0)/∂x	0.008	0.004	0.009	0.009	0.009	0.009
	∂E(y)/∂x	0.036	0.033	0.040	0.047	0.047	0.044
	∂E (y/y > 0)/∂x	0.027	0.047	0.022	0.036	0.037	0.029

Note: The same as in Table 3, but y refers to the number of days instead of time.

foster more regular practice throughout the week.

One of the limitations of our study is that the database does not offer information about some of the covariates that are often considered in the literature such as sports facilities. In addition, the sample size is not very large compared to the samples used in empirical papers in the literature, although we used a pooled cross-section from three consecutive years. For future research, it would be interesting to consider multi-level structures or replicate the analysis with data from other countries to check whether the conclusions are maintained. Finally, the availability of panel data would help better identify the causal effects of some relevant variables.

Declarations of interest: None.

Declaration of competing interest

None.

Data availability

The authors do not have permission to share data.

Acknowledgments

We are grateful to the Editor, Associate Editor, Copy Editor and two anonymous referees for their constructive comments and suggestions on a previous version of this paper. We also wish to thank the participants in the XI Conference of the European Sport Economics Association (ESEA), which was held in Gijón in May 2019, for their comments on the preliminary estimates. Jaume García acknowledges financial support from project PID 2020-114231RB-I00 and María José Suárez from project PAPI-21-GR-2015-0007. The usual disclaimer applies.

References

Abrevaya, J., 2002. Computing marginal effects in the Box-Cox model. Econom. Rev. 21 (3), 383–393. https://doi.org/10.1081/ETC-120015789.

Amemiya, T., 1984. Tobit models: a survey. J. Econom. 24 (1–2), 3–61. https://doi.org/10.1016/0304-4076(84)90074-5.

Aristei, P., Pieroni, L., 2008. A double-hurdle approach to modelling tobacco consumption in Italy. Appl. Econ. 40 (19), 2463–2476.

Artero, I., Bandrés, E., García, J., Rodríguez, P., 2019. Demand of professional sports: attendance and audience. In: García, J. (Ed.), Sports (And) Economics, vol. 7. FUNCAS Social and Economic Studies, Madrid, pp. 183–218.

Becker, G.S., 1965. A theory of the allocation of time. Econ. J. 75 (299), 493–517. https://doi.org/10.2307/2228949.

Borgers, J., Breedveld, K., Tiessen-Raaphorst, A., Thibaut, E., Vandermeerschen, H., Vos, S., Scheerder, J., 2016. A study on the frequency of participation and time spent on sport in different organisational settings. Eur. Sport Manag. Q. 16 (5), 635–654. https://doi.org/10.1080/16184742.2016.1196717.

Box, G.E.P., Cox, D.R., 1964. An analysis of transformations (with Discussion). J. R. Stat. Soc. Ser. B-Stat. Methodol. 26 (2), 211–252. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x.

Cabane, C., Lechner, M., 2015. Physical activity of adults: a survey of correlates, determinants and effects. Journal of Economics and Statistics 235 (4–5), 367–402. https://doi.org/10.1515/jbnst-2015-4-504.

Cawley, J., 2004. An economic framework for understanding physical activity and eating behaviors. Am. J. Prev. Med. 27 (3S), 117–125. https://doi.org/10.1016/j.amepre.2004.06.012.

Dallmeyer, S., Wicker, P., Breuer, C., 2017. Public expenditure and sport participation: an examination of direct, spillover, and substitution effects. Int. J. Sport Financ. 12 (3), 244–264.

Dawson, P., Downward, P., 2011. Participation, spectatorship and media coverage in sport: some initial insights. In: Andreff, W. (Ed.), Contemporary Issues in Sports Economics: Participation and Professional Team Sports. Edward Elgar, Cheltenham, pp. 15–42.

Downward, P., Lera-López, F., Rasciute, S., 2011. The Zero-Inflated ordered probit approach to modelling sports participation. Econ. Modell. 28 (6), 2469–2477. https://doi.org/10.1016/j.econmod.2011.06.024.

Downward, P., Muñiz, C., 2019. Sports participation. In: Downward, P., Frick, B., Humphreys, B.R., Pawlowski, T., Ruseski, J.E., Soebbing, B.P. (Eds.), The SAGE Handbook of Sports Economics. SAGE Publications, London, pp. 33–44.

Downward, P., Rasciute, S., 2010. The relative demand for sports and leisure in England. Eur. Sport Manag. Q. 10 (2), 189–224. https://doi.org/10.1080/16184740903552037.

Downward, P., Rasciute, S., 2015. Exploring the covariates of sport participation for health: an analysis of males and females in England. J. Sports Sci. 33 (1), 67–76. https://doi.org/10.1080/02640414.2014.924056.

Downward, P., Riordan, J., 2007. Social interactions and the demand for sport: an economic analysis. Contemp. Econ. Pol. 25 (4), 518–537. https://doi.org/10.1111/j.1465-7287.2007.00071.x.

Duan, N., 1983. Smearing estimate: a nonparametric retransformation method. J. Am. Stat. Assoc. 78 (383), 605–610.

Duan, N., Manning, W.G., Morris, C.N., Newhouse, J.P., 1984. Choosing between the sample-selection and the multi-part model. J. Bus. Econ. Stat. 2 (3), 283–289. https://doi.org/10.2307/1391711.

Eberth, B., Smith, M.D., 2010. Modelling the participation decision and duration of sporting activity in Scotland. Econ. Modell. 27 (4), 822–834. https://doi.org/10.1016/j.econmod.2009.10.003.

García, J., Lera-López, F., Suárez, M.J., 2011. Estimation of a structural model of the determinants of the time spent on physical activity and sport: evidence for Spain. J. Sports Econ. 12 (5), 515–537, 10.1177%2F1527002510387080.

García, J., Suárez, M.J., 2020. Organised and non-organised physical activity among children in Spain: the role of school-related variables. Eur. Sport Manag. Q. 20 (2), 171–188. https://doi.org/10.1080/16184742.2019.1594329.

Hafner, M., Yerushalmi, E., Phillips, W., Pollard, J., Deshpande, A., Whitmore, M., Millard, F., Subel, S., van Stolk, C., 2019. The Economic Benefits of a More Physically Active Population. An International Analysis. Rand Corporation.

Heckman, J., 1979. Sample selection bias as a specification error. Econometrica 47 (1), 153–161.

Humphreys, B.R., Ruseski, J.E., 2011. An economic analysis of participation and time spent in physical activity. B.E.J. Econ. Anal. 11 (1) https://doi.org/10.2202/1935-1682.2522 article 47.

Humphreys, B.R., Ruseski, J.E., 2015. The economic choice of participation and time spent in physical activity and sport in Canada. Int. J. Sport Financ. 10 (2), 138–159.

Jones, A.M., 2000. Health econometrics. In: Culyer, A.J., Newhouse, J.P. (Eds.), Handb. Health Econ. ume 1a (chapter 6), 265–344. https://doi.org/10.1016/S1574-0064(00)80165-1. North Holland.

Kokolakakis, T., Lera-López, F., Panagouleas, T., 2012. Analysis of the determinants of sports participation in Spain and England. Appl. Econ. 44 (19–21), 2785–2798. https://doi.org/10.1080/00036846.2011.566204.

Madden, D., 2008. Sample selection versus two-part models revisited: the case of female smoking and drinking. J. Health Econ. 27 (2), 300–307. https://doi.org/10.1016/j.jhealeco.2007.07.001.

McKelvey, R.D., Zavoina, W., 1975. A statistical model for the analysis of ordinal level dependent variables. J. Math. Sociol. 4 (1), 103–120. https://doi.org/10.1080/0022250X.1975.9989847.

Meltzer, D.O., Jena, A.B., 2010. The economics of intense exercise. J. Health Econ. 29 (3), 347–352. https://doi.org/10.1016/j.jhealeco.2010.03.005.

Muñiz, C., Downward, P., 2019. The outcomes related to sport and physical activity: a better understanding health, social, labor and academic impacts. In: García, J. (Ed.), Sports (And) Economics, FUNCAS, Social and Economic Studies, vol. 7, pp. 393–423.

Muñiz, C., Rodríguez, P., Suárez, M.J., 2014. Sports and cultural habits by gender: an application using count-data models. Econ. Modell. 36, 288–297. https://doi.org/10.1016/j.econmod.2013.09.053.

Oliveira-Brochado, A., Quelhas Brito, P., Oliveira-Brochado, F., 2017. Correlates of adults' participation in sport and frequency of sport. Sci. Sports 32 (6), 355–363. https://doi.org/10.1016/j.scispo.2017.03.005.

Rhodes, R.E., Janssen, I., Bredin, S.S.D., Warburton, D.E.R., Bauman, A., 2017. Physical activity: health impact, prevalence, correlates and interventions. Psychol. Health 32 (8), 942–975. https://doi.org/10.1080/08870446.2017.1325486.

Ruseski, J.E., Humphreys, B.R., Hallmann, K., Breuer, C., 2011. Family structure, time constraints, and sport participation. Eur. Rev. Aging Phys. Act. 8 (2), 57–66. https://doi.org/10.1007/s11556-011-0084-y.

Thibaut, E., Eakins, J., Vos, S., Scheerder, J., 2017. Time and money expenditure in sports participation: the role of income in consuming the most practiced sports activities in Flanders. Sport Manag. Rev. 20 (5), 455–467. https://doi.org/10.1016/j.smr.2016.12.002.

Yen, S.T., Jones, A.M., 1996. Individual cigarette consumption and addiction: a flexible limited dependent variable approach. Health Econ. 5 (2), 105–117. https://doi.org/10.1002/(SICI)1099-1050(199603)5:2%3C105::AID-HEC188%3E3.0.CO;2-I.

Jaume Garcia is professor of Applied Economics at Universitat Pompeu Fabra in Barcelona (Spain) since 1991. He was former president of the Spanish Statistics Office (2008–2011). His research interests are applied microeconometrics, sports economics, health economics, housing economics and labour economics.

María José Suárez is Associate Professor of Economics at the University of Oviedo (Spain) where she got her PhD in Economics. She was former Associate Dean for International Accreditation and Student Mobility at the School of Economics and Business of the University of Oviedo (2011–2014). Her research interests cover labour, cultural and sports economics.