MAIN ARTICLE

As simple as possible but not simpler: structural sensitivity testing of a dynamic model of adolescent overweight and obesity

Eduard Romanenko,[a]* Jack Homer[b] and Nanna Lien[a]

Abstract

We recently published results from an SD model of adolescent overweight and obesity using data from 31 European countries that participate in the Health Behavior in School-Aged Children (HBSC) study. During model development, we sought to identify a feedback structure with high explanatory power that avoided speculative relationships. Expert reviewers generally agreed with our modeling decisions, but two decisions did raise questions: (1) excluding the influences of food environment and built environment, for which HBSC provided no data; and (2) including five causal links that were supported statistically but might be considered disputable. To address the reviewers' questions, we created four possible model structures and performed automated calibration followed by intervention testing and ranking. We then compared the goodness-of-fit and intervention results. We discuss implications for how to move forward with the model, including through additional data gathering.

Syst. Dyn. Rev. **39**, 125–139 (2023)

Additional Supporting Information may be found online in the supporting information tab for this article.

Introduction

> We may assume the superiority, all other things being equal, of the demonstration which derives from fewer postulates or hypotheses.
>
> – Aristotle (384–322 BC).

> Plurality should not be posited without necessity…It is futile to do with more things that which can be done with fewer.
>
> – William of Ockham (c. 1287–1347).

> It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience…Everything should be as simple as possible but not simpler.
>
> – Albert Einstein (from "On the Method of Theoretical Physics," the Herbert.
>
> Spencer Lecture, Oxford, June 10, 1933; and attributed to Einstein, *New York Times*, January 8, 1950).

[a] Department of Nutrition, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway
[b] Homer Consulting and MIT Research Affiliate, Barrytown, New York, USA

* Correspondence to: Eduard Romanenko, Department of Nutrition, Institute of Basic Medical Sciences, University of Oslo, PO Box 1046, Blindern, N-0316, Oslo, Norway E-mail: eduard.romanenko@medisin.uio.no

Accepted by James Duggan, Received 13 December 2022; Revised 28 March 2023; Accepted 6 April 2023

System Dynamics Review
System Dynamics Review vol 39, No 2 (April/June 2023): 125–139
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/sdr.1732

Consistent with descriptions of the scientific enterprise going back to Aristotle and culminating with Einstein, system dynamics (SD) seeks to develop models that are adequate in complexity for addressing the problem at hand but also parsimonious enough to be comprehensible and well supported by the available evidence (Homer, 2014). Models often evolve from simpler to more complex in order to produce outputs that are more realistic or speak to particular policy concerns (see, e.g., Alfeld and Graham, 1976; Homer, 1996; Rahmandad, 2012; Rahmandad, 2022; Randers, 1973; Sastry, 1997). Yet one must also be careful not to clutter a model with excessive detail that undermines its clarity and explanatory power (Forrester, 1961).

Sensitivity testing is one of our most important tools, not only for model analysis but also for model improvement. Parametric sensitivity testing helps us understand the behavioral and policy implications of parameters of uncertain value, while structural sensitivity testing helps us understand the implications of variables or causal links of uncertain importance. As defined by Tank-Nielsen, structural sensitivity testing involves "an alteration of a causal relationship in the model… [which might be represented as the change in] a causal loop diagram" (1980, 192). Structural analysis includes both "boundary adequacy" and "structure assessment" testing (Sterman, 2000, 859-864) and allows us "to evaluate the impact of controversial or disputable relationships" (Tank-Nielsen, 1980, 192).

Such testing can help us decide what is essential to include in a model and how to proceed in gathering more evidence. If a variable or causal link lacks strong evidence (that is, weaker than the rest of the model) and does not affect policy findings, then one may consider excluding it (see, e.g. Mahamoud *et al.*, 2012). However, if such an uncertain variable or link does affect policy findings, one may include it conditionally — namely, on the condition that more evidence on it will be sought.

For all that has been written, the literature gives surprisingly little guidance on how to perform structural sensitivity testing (aside from the well-known use of on–off switches; Forrester, 1968) and how to balance the competing values of model adequacy and parsimony. In this article, we demonstrate how one may compare alternative models in a systematic way based on how they affect goodness-of-fit metrics as well as policy conclusions.

Model background

We recently completed the first phase of an SD study of adolescent overweight and obesity (AdOWOB) in Europe, based on survey data from the Health Behavior in School-Aged Children study (HBSC) from 31 countries and with particular emphasis on the five countries involved in the EU-funded CO-CREATE project (Romanenko *et al.*, 2022). Rising and persistently high AdOWOB prevalence in Europe has led to a growing recognition

DOI: 10.1002/sdr

of the importance of social, physical, and economic environments, including the effect of lower income, in shaping an individual's diet and activity behaviors, and thus health outcomes (Finegood *et al.*, 2010; Koplan *et al.*, 2005; Rutter *et al.*, 2017; Salas, 2015). However, no previously published dynamic model of AdOWOB has considered the wide range of personal and social factors described in the literature (Aguiar *et al.*, 2019). Chen *et al.* (2018) incorporated economic variables (employment and income distribution) into a simple model of weight distribution in the United States. Struben *et al.* (2014) introduced food industry variables. Our 2022 SD study was the first to integrate a wide range of behavioral and psychological factors, some directly reflecting the influences of family, friends, school, and the wider society.

In our 2022 study, we utilized a combination of literature review, statistical screening procedures (the analysis of probabilistic odds ratios, correlational analysis, and stepwise multivariate linear regressions), and SD modeling to build a strongly evidence-based model with only 12 major variables (8 of them endogenous and 4 exogenous) and 30 causal links (with corresponding strengths known as hazard ratios). The model variables represented population-level prevalences of adverse health behaviors (e.g., inadequate exercise), psychological conditions (e.g., feeling nervous), and social determinants that affect individuals (e.g., school pressure or computer overuse) that can affect AdOWOB directly or through other such variables. Automated calibration showed that the model could nicely reproduce HBSC data patterns from 24 different cases (differing by country, gender, and perceived wealth status) over the period 2002–14. For each case, we tested 10 potential points of intervention (starting in 2018) and ranked them by projected reduction of AdOWOB by 2026. Table 1 identifies the 24 cases and the 10 intervention points. We used our model-based findings to support or supplement the policies suggested by the adolescent participants who were part of CO-CREATE.

Our objective in model development was to identify a cluster of interrelated variables that demonstrated high explanatory power but was parsimonious with respect to available data—that is, a model that avoided speculative relationships. This approach had implications for which variables and causal links we did or did not include in the model. Public health experts involved in internal review of the model during the project generally agreed with our decisions, but two decisions did raise some questions among some experts.

The first of those two decisions was to exclude the food environment (FE, affecting dietary behaviors) and the built environment (BE, affecting physical activity). The literature points to the potential significance of FEBE as a factor affecting adolescent obesity (Elbel *et al.*, 2020; Gilliland *et al.*, 2012; Malacarne *et al.*, 2022), but neither HBSC nor any other multicountry European survey to date includes items related to FEBE. We excluded FEBE

DOI: 10.1002/sdr

Table 1. Twenty-four cases and 10 intervention points in the adolescent obesity modeling analysis

	Less well-off		More well-off	
Twenty-Four Cases	Boys	Girls	Boys	Girls
Avg31	LWOB_AV	LWOG_AV	MWOB_AV	MWOG_AV
England	LWOB_EN	LWOG_EN	MWOB_EN	MWOG_EN
Netherlands	LWOB_NL	LWOG_NL	MWOB_NL	MWOG_NL
Norway	LWOB_NO	LWOG_NO	MWOB_NO	MWOG_NO
Poland	LWOB_PL	LWOG_PL	MWOB_PL	MWOG_PL
Portugal	LWOB_PT	LWOG_PT	MWOB_PT	MWOG_PT
Six behavioral intervention points				
Inadequate exercise		Inadequate breakfast		
Inadequate fruit		Inadequate vegetables		
Dieting		Computer overuse		
Four psychological intervention points				
Feel low		Feel nervous		
School pressure		Life dissatisfaction		

Notes: (a) "Well-off" is based on the response to an HBSC study question on one's perceived household wealth. (b)The five named countries are those in the CO-CREATE study. (c) The 10 intervention points are also the 10 variables in the model other than the two variables for overweight and obesity (AdOWOB, OWOB Age 10–11). (d) Seven of these 10 variables are involved in feedback loops; the other 3 (Computer overuse, School pressure, and Life dissatisfaction) are exogenous.

because we had no data, not even proxies or trend data, to estimate it. Lack of sufficient evidence can be a valid scientific reason to exclude certain variables from SD models, even if the literature or stakeholders suggest possible causality (Homer, 2014; Sterman, 2018).

The second decision that raised questions was the inclusion of five causal links: from school pressure to AdOWOB and inadequate vegetable consumption; from life dissatisfaction to inadequate vegetable and fruit consumption; and from nervousness to AdOWOB. These links were supported by statistical screening but were deemed "indirect," meaning that their support from the literature required assumptions about an unmeasured intermediate variable, specifically high-calorie snacking. For example, the statistical screening suggested a link from nervousness to AdOWOB, which required explanation in two steps: from nervousness to snacking and from snacking to AdOWOB. Our model includes several other dietary behavior variables, but it does not include snacking. Optional snacking questions were part of the HBSC survey, but only data from the mandatory questions of the survey were available to us through open access. We described snacking to the experts as a hidden variable in the model and kept implicit for lack of data, and we made the point that all models include implicit variables (see Alfeld and Graham, 1976); some experts still questioned this approach.

DOI: 10.1002/sdr

Reflecting on our first phase of modeling, we realized that we might use structural sensitivity testing to address the experts' questions. First, perhaps we could find a way to infer trends in FEBE despite the lack of direct data on it. Might the inclusion of such trends affect our policy conclusions? Second, what would happen if we eliminated all five of the "indirect" causal links? Might such elimination affect our policy conclusions?

In this article, we describe this two-fold structural sensitivity analysis of the existing model and its implications for future data needs.

Structural testing procedures

Alternative model structures

We started by making two types of modifications to the original model structure. One was to incorporate the concept of FEBE through three assumed linear trends (switchable on or off) affecting the variables of inadequate exercise, inadequate fruit, and inadequate vegetables. Despite the lack of data on FEBE, we reasoned that if (a) the inclusion of such trends (after optimized calibration) allowed for a better overall goodness of fit and (b) they ended up altering the policy conclusions, then we could justify the inclusion of these trends in the model. Each linear effect was formulated as a ramp starting in 2002 with two parameters to be optimized: End Year and End Change (that is, the ramp's percentage change from 2002 to End Year).

Another modification was to allow reduction of the model by selectively (switchable on or off) excluding the five "indirect" causal links that implicitly go through high-calorie snacking. Two of these links (from School Pressure and Feel Nervous) bypass behavioral variables on the way to AdOWOB. The other three links capture the effect of environmental variables (School Pressure and Life Dissatisfaction) on fruit and vegetables consumption. If we found that excluding these "disputable" links (to use the Tank-Nielsen term) did not alter the model's policy conclusions, then, by the logic of parsimony, we might safely eliminate them from the model. If, on the other hand, they did alter policy conclusions, then we would lean toward the original model but on the condition that we could find more evidence to support the disputable links.

The structural sensitivity analysis was performed through the testing of four possible model configurations: (1) the original model including the five disputable links but excluding FEBE ("Full_noFEBE"); (2) a model including both the five links and FEBE ("Full_FEBE"); (3) a model excluding the five links as well as FEBE ("Reduced_noFEBE"); and (4) a model excluding the five links but including FEBE ("Reduced_FEBE").

Figure 1 is an interpretive sector diagram of the original model (with both explicit and implicit links), showing the five disputable links going through

DOI: 10.1002/sdr

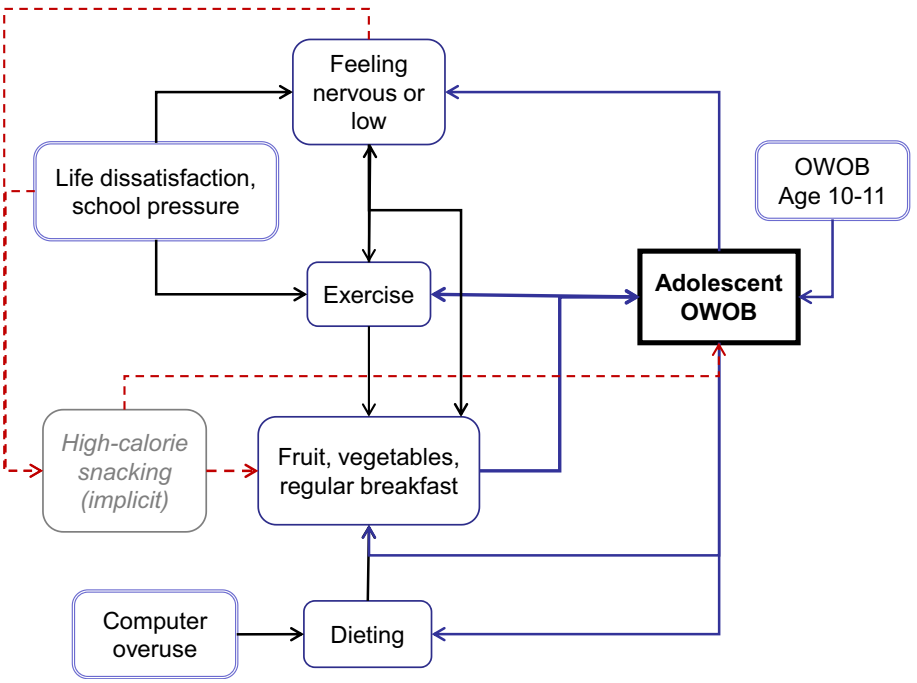

Fig. 1. Sector diagram of the original model, showing disputable links (dashed red) going through the implicit (unmodeled) variable of high-calorie snacking. (A standard causal-loop diagram, absent snacking, is presented in Romanenko *et al.*, 2022). [Color figure can be viewed at wileyonlinelibrary.com]

the implicit (unmodeled) variable of high-calorie snacking. Figure 2 extends this diagram to include the possible FEBE influences. The model documentation is reported in the online supporting information (Tables S1–S3 report summary information about the model; Table S4 provides a complete listing of model equations).

Model calibration and testing

For each model configuration, we repeated the analysis we performed in our original study (Romanenko *et al.*, 2022). First, we used Powell optimization to calibrate each of the alternative structures to the HBSC data for each of our 24 country-gender-wealth cases (the optimization specifications are reported in the online supporting information, including Table S5 on computational costs of the optimization experiments). Next, we calculated two types of goodness-of-fit statistics for the cases, for all eight of the model's endogenous variables: (1) the mean absolute percentage error (MAPE) between simulated output and data and (2) a customized R-squared measure ("R2i", range 0 to 1) of how well the model predicts changes away from the initial data point in 2002.

We recorded these statistics for AdOWOB (the main variable of interest in the model and the ultimate target for intervention testing), as well as averaged across all eight endogenous variables (hereafter "All8," of which

DOI: 10.1002/sdr

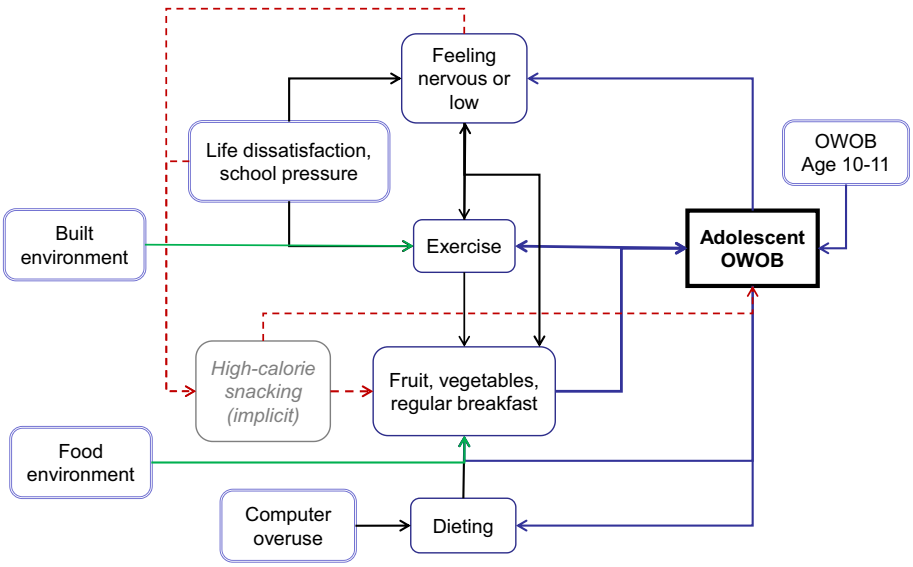

Fig. 2. Extended sector diagram, including possible Food Environment and Built Environment influences (solid green). [Color figure can be viewed at wileyonlinelibrary.com]

AdOWOB is one). This resulted in four goodness-of-fit measures (AdOWOB MAPE, All8 MAPE, AdOWOB R2i, and All8 R2i) for each model and each case. To facilitate comparison between the four models and in line with the literature supporting the appropriateness of combining multiple goodness-of-fit measures when comparing alternative models (Mehdiyev *et al.*, 2016; Xu and Ouenniche, 2012), we transformed MAPE into a continuous 0–1 index ("MAPE index") in which 0% MAPE is 1 (best) and ≥20% MAPE is 0 (worst). We averaged the two MAPE indices with the two R2i measures, weighting all four equally, to produce a combined index of model adequacy. We did this for each case and then averaged across all 24 cases for an overall average of model adequacy.

Next came intervention testing, again done exactly as before (Romanenko *et al.*, 2022). We tested the 10 potential intervention points using identical 25% effect sizes starting in 2018, and we ranked the interventions in terms of their ability to reduce projected AdOWOB in 2026. To facilitate comparison of the models, we counted the number of times each intervention point appeared in the Top 4 ranking of 10 interventions, across all 24 cases.

Results

Optimized ramp change parameters for FEBE

The optimization of the *FEBE* models (see the online supporting information for optimization specifications) resulted in all ramps (for both *Full_FEBE* and *Reduced_FEBE*

DOI: 10.1002/sdr

Table 2. Summary statistics for the end-change percentages for the two models with food and built environment (FEBE) influences optimized for the 24 cases. These models allow for three exogenous ramps (each with its own optimized end time and end-change percentage) affecting the prevalence fractions of Inadequate Exercise, Inadequate Fruit, and Inadequate Vegetables, respectively. A positive end-change percentage indicates a worsening trend (more inadequacy), while a negative end-change percentage indicates an improving trend (less inadequacy)

	Full_FEBE			Reduced_FEBE		
	Inad Ex	Inad Fruit	Inad Veg	Inad Ex	Inad Fruit	Inad Veg
Mean	0.2%	1.0%	−1.4%	−0.3%	0.9%	−3.3%
SD	2.2%	5.0%	3.2%	2.7%	6.1%	5.2%
Min	−5.4%	−11.9%	−14.5%	−7.7%	−14.1%	−19.2%
Max	5.5%	12.9%	2.7%	4.1%	13.9%	4.5%

and for all 24 cases) having an End Year close to 2010. Table 2 reports the summary statistics for the estimates of the ramp end-change percentages across the 24 cases (Table S6 in the online supporting information reports the results by case).

The table shows that the optimized end-change percentages are generally of modest size. The parameters for Exercise are always less than 8% in either direction, with means of about zero for both *Full_FEBE* and *Reduced_FEBE*. The parameters for Fruit are always less than 15% in either direction, with means of about +1% (worsening trend) for both models. The parameters for Vegetables are always less than 20% in either direction, with a mean of −1% for *Full_FEBE* and −3% for *Reduced_FEBE* (both improving trends).

Goodness of fit

Table 3 summarizes goodness-of-fit statistics for the four tested model configurations, with each measure averaged across the 24 cases (Tables S7–S9 in the online supporting information provide details by case). The individual fit statistics (the first four rows) do not vary by much from one model to another. The combined adequacy measure (the last row) is similarly tight, with the largest model, *Full_FEBE* (at 61.2%) providing only a slightly better fit than the smallest model, *Reduced_noFEBE* (at 59.0%).

Although the differences are not great, this table does provide some information about the types of contribution coming from (a) the five disputable links in the *Full* models and (b) the three new ramp effects in the *FEBE* models.

The clearest benefit of *Full* is in the two MAPE indices (rows 1 and 2), where *Full* beats *Reduced* by 1.0% to 2.6%. This fact suggests that the five disputable links (four of which come from the exogenous variables of life dissatisfaction and school pressure) give the model a greater ability to follow turning points in the data. It must be that some of the ups and downs in

DOI: 10.1002/sdr

Table 3. Goodness-of-fit statistics for the four tested model versions, averaged across the 24 cases. MAPE is Mean Absolute Percentage Error; R2i is a novel R-squared measure (see Romanenko *et al.*, 2022), All8 refers to all eight endogenous variables

	Model			
Fit statistic	Full_noFEBE	Full_FEBE	Reduced_noFEBE	Reduced_FEBE
MAPE index – AdOWOB	53.7%	54.3%	52.0%	51.7%
MAPE index – All8	62.3%	63.3%	60.7%	62.3%
R2i – AdOWOB	66.6%	66.9%	66.3%	66.3%
R2i – All8	58.0%	60.5%	57.0%	59.7%
Combined Adequacy (*equal weighting*)	*60.1%*	*61.2%*	*59.0%*	*60.0%*

these exogenous variables help to explain corresponding ups and downs in AdOWOB and the other endogenous variables.

The clearest benefit of *FEBE* is in the "All8" fit statistics (rows 2 and 4), where *FEBE* beats *NoFEBE* by 1.0%–2.7%. The addition of the exogenous ramps for exercise, fruit, and vegetables improves the model's fit to those variables, but it does not improve the fit to AdOWOB.

Intervention testing

Table 4 reports, for each of the model configurations, the percentage reductions in AdOWOB (in 2026 relative to no intervention) averaged across the 24 cases for each of the 10 interventions separately and for all 10 combined (Tables S10–S13 in the online supporting information report the results by case). The interventions vary greatly in terms of their impact on AdOWOB, even when averaged across the cases. For all model configurations, the most impactful interventions include Fruit, Exercise, Breakfast, Life Dissatisfaction, and Vegetables (in roughly that order). Yet, there are also differences between the models. First,

Table 4. Percentage reduction of AdOWOB in 2026 for the 10 interventions, for the four tested model versions, averaged across the 24 cases

	Model			
Intervention	Full_noFEBE	Full_FEBE	Reduced_noFEBE	Reduced_FEBE
Inad exercise	3.6%	5.1%	4.4%	3.9%
Inad fruit	5.0%	5.2%	4.7%	4.6%
Inad veg	1.6%	2.0%	1.9%	1.6%
Inad breakfast	1.8%	2.5%	2.1%	2.1%
Dieting	0.3%	0.5%	0.5%	0.5%
Feel nervous	2.5%	3.2%	0.9%	1.0%
Feel low	1.7%	1.6%	0.9%	0.9%
Life dissatisfaction	2.4%	3.8%	2.0%	1.6%
School pressure	3.7%	3.7%	0.8%	0.6%
Computer overuse	0.2%	0.2%	0.2%	0.1%
All 10 Combined	*13.5%*	*16.4%*	*12.6%*	*12.3%*

DOI: 10.1002/sdr

Table 5. Counts of Top 4 ranking for the 10 interventions, for the four tested model versions, across the 24 cases (maximum count of 24). Note that, for a given model, some cases may have fewer than four ranked interventions (i.e., interventions with any simulated impact on AdOWOB), and some cases may have more than four "Top 4" ranked interventions (due to ties in percentage impact to the first decimal point). As a result, columns in this table do not sum to 96 (=24 × 4)

	Model			
Intervention	Full_noFEBE	Full_FEBE	Reduced_noFEBE	Reduced_FEBE
Inad exercise	18	19	20	21
Inad fruit	17	17	18	18
Inad veg	8	9	12	13
Inad breakfast	10	12	12	15
Dieting	1	0	1	0
Feel nervous	6	9	4	5
Feel low	3	4	3	6
Life dissatisfaction	16	19	11	10
School pressure	11	13	3	4
Computer overuse	0	0	0	0

for the *Full* models (unlike the *Reduced* models), Feel Nervous and School Pressure are additional interventions with good impact. These two variables account for three of the five disputable links included in the *Full models*. Second, the inclusion of the three *FEBE* ramps tends to boost the impact of the leading interventions in combination with *Full* (i.e., *Full_FEBE*), but it does not do so in combination with *Reduced* (i.e., *Reduced_FEBE*).

Table 5 reports the overall counts of "top 4" ranking for each of the 10 interventions for the four model configurations, summing across the 24 cases (Tables S14–S17 in the online supporting information report the results by case). It is evident that including *FEBE* has no real effect on intervention priorities, whether the starting point is the *Full* or *Reduced* models. In contrast, the inclusion of the five disputable links in the *Full* models does clearly affect the intervention rankings. The *Full* models elevate the rankings of School Pressure and Life Dissatisfaction (and also, somewhat, Feel Nervous), and they demote the ranking of Vegetables (and also, somewhat, Breakfast).

Discussion

We used structural sensitivity testing to evaluate two decisions made in the development of the AdOWOB model: (1) excluding the influences of FEBE, for which we had no data, and (2) including five causal links that were supported statistically but which some public health experts considered indirect and disputable.

DOI: 10.1002/sdr

Our analysis showed, first, that exogenous linear trends representing FEBE and affecting exercise and fruit and vegetable consumption could improve the model's fit to those variables but did not improve the fit to AdOWOB itself and had no effect on intervention priorities. This is not to say that we do not recognize the importance of FEBE as a type of intervention in its own right. On the contrary, exercise, fruit, and vegetables are all important intervention points in our model and, in the real world, may be influenced by FEBE interventions. However, our results showed that allowing for possible historical trends in FEBE did not alter optimized hazard ratios enough to change intervention rankings. Therefore, we concluded that FEBE trends did not add value to the original model and, according to the logic of evidence and parsimony, could be safely excluded.

Our testing also showed that the inclusion of the five disputable causal links in the *Full* model configurations provided a somewhat better fit to all variables including AdOWOB (by the MAPE criterion) and affected the intervention rankings, elevating the priority of school pressure and life dissatisfaction. This policy sensitivity suggests that one should be cautious about eliminating the links in question and rather lean toward the original model on the condition that more evidence to support these links could be found.

The five links were identified as disputable because they all go through high-calorie snacking, an intermediate variable which is not included in the mandatory questions of the HBSC study. A logical direction for gathering further supporting evidence is to collect data on individual snacking or daily caloric intake. Obtaining the data on snacking from optional HBSC questions (asked by a subset of countries) could be useful. However, the HBSC measures only frequency of snacking and not the type or amount. Better data could be obtained, for example, by applying diet checklists for snacking behavior over a sufficiently long time period, probably two weeks or more as described by the DAPA Measurement Toolkit (https://www.dapa-toolkit.mrc.ac.uk/diet/subjective-methods/diet-checklist). Data from even just a few countries, perhaps some of those in the five-country CO-CREATE project, could help determine whether the disputable links can be supported by direct evidence.

We believe that our work here could contribute to SD modeling practice, in at least two ways.

First, structural sensitivity testing has long been described as an important part of building confidence in SD models, yet the literature gives little guidance on how to do it (aside from the well-known use of on–off switches). Here, we have demonstrated how one may compare alternative models based on their goodness of fit and their effect on policy conclusions.[i]

[i]We recognize there is one aspect of structural sensitivity testing we did not demonstrate, namely, determining whether a model can produce all relevant or problematic modes of behavior, such as oscillation or overshoot-and-decline. The only behavioral pattern we saw clearly in the HBSC data, across the 24 cases, was AdOWOB adjusting in a goal-seeking (decelerating) fashion to perturbations in other variables. This behavior pattern was produced by all four of the alternative model configurations we considered.

DOI: 10.1002/sdr

Second, the existing literature gives little guidance on how to balance the competing values of model adequacy and parsimony. Here, we have offered the following approach:

1. evaluate the strength of the evidence for a causal link in question based on the literature, expert knowledge, and available data;
2. for a causal link with weaker evidence (an uncertain or disputable link), evaluate whether including the link improves the model's explanatory power or affects policy findings;
3. eliminate the uncertain link if it does not add value to the model;
4. if the uncertain link does add value to the model, include it on the condition that more evidence will be sought to confirm or reject the link.

We believe that structural sensitivity testing is an important tool that could allow SD modelers to be more scientific and show that their models are "as simple as possible but not simpler."

Funding Information

The CO-CREATE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 774210. The content of this article reflects only the authors' views and the European Commission is not liable for any use that may be made of the information it contains.

Conflict Of Interest

The authors declare no conflicts of interest.

Biographies

Eduard Romanenko is a researcher in the Public Health Nutrition group at the University of Oslo, Norway. He has been working as an SD modeler for the CO-CREATE project described here since 2020. He graduated from the European Master program in SD in 2014 and received a PhD in SD from the University of Bergen, Norway in 2022.

Jack Homer is a system dynamics modeling consultant and has directed Homer Consulting since 1989. Before that, he taught at the University of Southern California and received a PhD from MIT and BS and MS degrees from Stanford University. He received the SD Society's Forrester Award in

DOI: 10.1002/sdr

1997 for his work studying cocaine use in the United States and was lead modeler for the team that won the Applications Award in 2011 for the "PRISM" model of cardiovascular disease for the US Centers for Disease Control and Prevention. He is author of the books "Models That Matter" (2012) and "More Models That Matter" (2017).

Nanna Lien is a professor at the Department of Nutrition, University of Oslo, Norway and leads the research group in Public Health Nutrition. She has more than 15 years of experience in school-based intervention research in nutrition and obesity prevention in Norway and on European funded projects. In the CO-CREATE project, she leads the work package on "Evaluation of Co-Created policy interventions and the methodology." She is a fellow of the International Society of Behavioral Nutrition and Physical Activity (ISBNPA) and a deputy editor of the International Journal of Behavioral Nutrition and Physical Activity (IJBNPA).

References

Aguiar A, Gebremariam M, Kopainsky B, Savona N, Allender S, Lien N. 2019. *Review of Existing System Dynamics Models on Overweight/Obesity in Children and Adolescents*. University of Oslo: Oslo. https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5c8d1c9d0&appId=PPGMS. Accessed June 1, 2022.

Alfeld LE, Graham AK. 1976. *Introduction to Urban Dynamics*. MIT Press: Cambridge, MA (Preface, pp. xiii-xv.).

Chen H-J, Xue H, Liu S, Huang TTK, Sang YC, Wang Y. 2018. Obesity trend in the United States and economic intervention options to change it: A simulation study linking ecological epidemiology and system dynamics modeling. *Public Health* **161**: 20–28.

Elbel B, Tamura K, McDermott ZT, Wu E, Schwartz AE. 2020. Childhood obesity and the food environment: A population-based sample of public school children in new York City. *Obesity* **28**(1): 65–72.

Finegood DT, Merth TD, Rutter H. 2010. Implications of the foresight obesity system map for solutions to childhood obesity. *Obesity (Silver Spring)* **18**(n1s): S13–S16. https://doi.org/10.1038/oby.2009.426.

Forrester JW. 1961. *Industrial Dynamics*. MIT Press. (Appendix O, "Beginners' Difficulties": Cambridge, MA; 449–456.

Forrester JW. 1968. Market growth as influenced by capital investment. *MIT Sloan Management Review* **9**(2): 83–105.

Gilliland JA, Rangel CY, Healy MA, Tucker P, Loebach JE, Hess PM, He M, Irwin JD, Wilk P. 2012. Linking childhood obesity to the built environment: A multi-level analysis of home and school neighbourhood factors associated with body mass index. *Canadian Journal of Public Health* **103**(9 Suppl 3): eS15–eS21.

Homer JB. 1996. Why we iterate: Scientific modeling in theory and practice. *System Dynamics Review* **12**(1): 1–19.

DOI: 10.1002/sdr

Homer J. 2014. Levels of evidence in system dynamics modeling. *System Dynamics Review* **30**: 75–80.

Koplan JP, Liverman CT, Kraak VA. 2005. Institute of Medicine, committee on prevention of obesity in children and youth. In *Preventing Childhood Obesity: Health in the Balance*. National Academies Press: Washington DC.

Mahamoud A, Roche B, Homer J. 2012. Modelling the social determinants of health and simulating short-term and long-term intervention impacts for the city of Toronto, Canada. *Social Science & Medicine* **93**: 247–255.

Malacarne D, Handakas E, Robinson O, Pineda E, Saez M, Chatzi L, Fecht D. 2022. The built environment as determinant of childhood obesity: A systematic literature review. *Obesity Reviews* **23**(S1): e13385.

Mehdiyev N, Enke D, Fettke P, Loos P. 2016. Evaluating forecasting methods by considering different accuracy measures. *Procedia Comp Sci* **95**: 264–271.

Rahmandad H. 2012. Impact of growth opportunities and competition on firm-level capability development trade-offs. *Organization Science* **23**(1): 138–154.

Rahmandad H. 2022. Behavioral responses to risk promote vaccinating high-contact individuals first. *System Dynamics Review* **38**(3): 246–263.

Randers J. 1973. *Conceptualizing Dynamic Models of Social Systems: Lessons from a Study of Social Change*. PhD dissertation. MIT Sloan School of Management: Cambridge, MA.

Romanenko E, Homer J, Fismen AS, Rutter H, Lien N. 2022. Assessing policies to reduce adolescent overweight and obesity: Insights from a system dynamics model using data from the health behavior in school-aged children study. *Obesity Reviews* **24**(1): e13519.

Rutter H, Bes-Rastrollo M, De Henauw S *et al*. 2017. Balancing upstream and downstream measures to tackle the obesity epidemic: A position statement from the European Association for the Study of obesity. *Obesity Facts* **10**(1): 61–63. https://doi.org/10.1159/000455960.

Salas XR. 2015. The ineffectiveness and unintended consequences of the public health war on obesity. *Canadian Journal of Public Health* **106**(2): e79–e81. https://doi.org/10.17269/cjph.106.4757.

Sastry MA. 1997. Problems and paradoxes in a model of punctuated organizational change. *Admin Science Quarterly* **42**: 237–275.

Sterman JD. 2000. Truth and beauty: Validation and model testing. In *Business Dynamics: Systems Thinking and Modeling for a Complex World*. Irwin McGraw-Hill. Ch. 21: Boston, MA; 845–891.

Sterman J. 2018. System dynamics at sixty: The path forward. *System Dynamics Review* **34**(1–2): 5–47.

Struben J, Chan D, Dubé L. 2014. Policy insights from the nutritional food market transformation model: The case of obesity prevention. *Annals of the New York Academy of Sciences* **1331**(1): 57–75.

Tank-Nielsen C. 1980. Sensitivity Analysis in System Dynamics. Chapter 9. In *Elements of the System Dynamics Method*, Randers J (ed). MIT Press: Cambridge, MA; 185–202.

Xu B, Ouenniche J. 2012. Performance evaluation of competing forecasting models: A multidimensional framework based on MCDA. *Expert Sys with Applic* **39**: 8312–8324.

DOI: 10.1002/sdr

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's website.

Data S1: Supporting Information

DOI: 10.1002/sdr