Istenes, Brandon

Working Paper

Job allocation in the Levy Institute Microsimulation Model

Working Paper, No. 1079

Provided in Cooperation with:
Levy Economics Institute of Bard College

Suggested Citation: Istenes, Brandon (2025) : Job allocation in the Levy Institute Microsimulation Model, Working Paper, No. 1079, Levy Economics Institute of Bard College, Annandale-on-Hudson, NY

This Version is available at:
https://hdl.handle.net/10419/315959

Working Paper No. 1079

Job Allocation in the Levy Institute Microsimulation Model

by

Brandon Istenes
Levy Economics Institute

April 2025

I would like to thank Thomas Masterson for his generosity in sharing his code with me and taking the time to help me work with it, without which neither Istenes (2023) nor the present paper would have been possible; I am also deeply grateful for his patience with my questions and ideas that ensued. I would also like to thank Fernando Rios-Avila for his support and advice across the various drafts of this paper.

Levy Economics Institute
P.O. Box 5000
Annandale-on-Hudson, NY 12504-5000
http://www.levyinstitute.org

ISSN 1547-366X

ABSTRACT

The Levy Institute Microsimulation Model (LIMM) is a tool used for policy simulations to estimate ex-ante the employment and income effects of sectoral investments. In Istenes (2023), a simple implementation of the LIMM for New York State initially had difficulty producing realistic conditional distributions of allocated jobs. This paper identifies the sources of that problem, which produces significant distortions to the characteristic distributions of job recipients. Solutions to the problem are presented with theoretical and empirical analysis. The relevance of this problem to other LIMM-based models is discussed; while it is theoretically relevant, it is unlikely to have a substantial impact on results.

KEY WORDS: Employment Simulation, Statistical Matching, LIMM

JEL CLASSIFICATION: C53, C63, J16, J21

INTRODUCTION

The Levy Institute Microsimulation Model (LIMM) is a tool developed at the Levy Economics Institute for ex-ante analysis of sectoral investment policy. The model is used primarily to assess the total impact of policy on employment and time use. This has been used for work on social infrastructure in Ghana and Tanzania (Zacharias et al. 2019), and care investment simulations in Turkey (e.g. Kim, İlkkaracan, and Kaya 2019), Mexico (Masterson et al. 2022), and elsewhere.

The LIMM generally follows a macrosimulation which provides its inputs. The macrosimulation estimates the sectoral distributions of new income and employment from a hypothetical policy; these estimates are then used by the LIMM in order to estimate demographic distributions of this new income and employment. The macrosimulation generally involves using an input-output or social accounting matrix in order to estimate indirect and induced effects. The LIMM is used to model child care investment in New York State in Istenes (2023); in this instance, the macrosimulation is skipped. The macroeconomic effects of the policy on employment and income are treated as exogenous shocks. Only the direct policy effect on employment and income is used, and this is calculated arithmetically. This resulted in a very simple and small set of distributions of interest, which made it easy to detect inconsistencies in the relationships between them. The initial model had difficulty producing realistic conditional distributions of allocated jobs. This paper identifies the sources of that difficulty, a problem which produces significant distortions to the characteristic distributions of job recipients. Solutions to this issue are presented with theoretical and empirical analysis. It is explained why, given a set of observations with varying selection probabilities, simply selecting by highest selection probability results in a set that is distorted with respect to the selection probability distributions associated with the characteristics of those observations. The paper concludes with a discussion of the relevance of these results to other LIMM-based models.

GENERAL METHODOLOGY OF THE LIMM

This section explains how the LIMM works for employment simulation in general. It is largely based on the explanations in Masterson (2013), Zacharias, Masterson, and Kim (2009, 11), and

Zacharias et al. (2019, Appendix C.1). However, some details are filled in from the implementation code for Masterson et al. (2022).[1]

The first step of the LIMM is to identify the donor pool (i.e., the set of people who have jobs, from which information about newly allocated jobs will be derived) and the eligible pool (i.e., the potential recipients). Through statistical matching, the jobs of those in the donor pool will be copied over to the subset of the eligible pool who are selected as recipients. The donor pool is generally comprised of anyone with a job who meets the eligibility criteria for the allocation of a new job (aside from, of course, joblessness). If recipient eligibility is limited to people ages 18-65, then the donor pool will be limited to those who have jobs and are within that age range. The eligible pool, meanwhile, is restricted by relationship to the labor force. It may include all people without jobs within an age range, or only those who count themselves as in the labor force or who are looking for a job.

Once the donor and eligible pools are identified, three models are produced. The first model is fitted only to the donor pool, and predicts propensity to work in each occupation and each industry based on demographic characteristics. This is typically done using a multinomial logit or probit model, which also may be run in separate cells of the donor pool (see, for example, Zacharias et al. [2019, 164]). The model is then used to produce propensity scores for each occupation and each industry for everyone in the combined donor and eligible pools. These propensity scores are then used to rank the likeliest occupations and industries for each person.

Second, a model of employment likelihood is fitted to the combined donor and recipient set. It predicts, based on demographic characteristics, the probability that a person will be in the donor pool—that is, employed. This is typically done with a probit model, which may be run in separate cells divided by sex or age group. The model is then used to predict employment probability for the combined donor and recipient pools, so that everyone receives an employment propensity score.

The third model—actually a pair of models—is fitted only to the donor pool, and predicts working hours and the logarithm of wages based on demographic characteristics. These are subject to selection bias; thus, naively running linear and log-linear models to predict these

[1] The implementation code is in Stata and was generously made available to me by Thomas Masterson.

would bias the results upward. In order to correct for this, the Inverse Mills Ratio (IMR) is calculated. This is produced by obtaining a probit estimation of labor force participation, the predicted values of which are used to produce the IMR for each observation. Once the IMR has been obtained, working hours can be estimated as a linear function of demographic characteristics and the IMR. Wages can be estimated as a log-linear function of the same regressors. Details can be found in Zacharias et al. (2019, 165), but these last models are not particularly relevant to this discussion.

Once these propensities and predictions have been computed, the job allocation process begins. The general idea is to assign each recipient their likeliest industry and occupation using the predicted likelihoods (Masterson 2013). For each likelihood ranking of industry and occupation, the algorithm iterates through the set of all industries and occupations. Jobs are allocated exhaustively per industry and occupation likelihood ranking, sub-ordered according to employment probability. If there are 200 jobs in industry A and occupation X to allocate, and 2000 people have A and X as their likeliest industry and occupation, the 200 of those people with the highest employment likelihoods will be allocated jobs. It is important to count people as represented by the sample weights; jobs are allocated by sample weights, not observations. If there are 1000 jobs in industry B and occupation Y to allocate, and 800 people have B and Y as their likeliest industry and occupation, all 800 will receive those jobs regardless of their likelihood of working in general. After all other industry-occupation combinations are allocated to individuals who have that combination as their likeliest, the next 200 jobs in industry B and occupation Y will be allocated among those for whom B and Y were the second-likeliest industry and occupation. Within each likelihood ranking, job allocation runs as follows. Match members of the current likelihood rank for the current industry and occupation are statistically matched to donors who actually have that industry and occupation. Imputation using hot decking or multiple imputation using hot decking is used to transfer wage and hour characteristics to the recipients, constituting a “job offer.” The statistical matching in the hot decking is based again on demographic characteristics, as well as predicted wages and hours. It is ensured that recipients are matched with donors whose actual industry and occupation match the recipient’s imputed industry and occupation.

Once the job offers are made for a given occupation and industry, recipients choose whether to accept or reject the job based on their prior year earnings. For example, the criterion may be that if the offered wage is at least 75 percent of the person's prior year earnings, they will accept the job. Rejected job offers leave the job available for eligible recipients in the next round (i.e., those in the next ranking group for that occupation and industry). After all jobs are assigned (or the pool of eligible recipients is exhausted), the simulated distributions are examined. The next section provides technical detail on the methodology used in the simulation for Istenes (2023), which is largely a simplified version of the more general LIMM methodology presented here.

METHODOLOGY OF THE NEW YORK MODEL

The model used in Istenes (2023) closely follows the general LIMM procedure described above. Data from the 2021 five-year American Communities Survey (obtained from IPUMS) is used. The only jobs which are allocated are childcare jobs. Specifically, these are jobs in the childcare services industry with occupation childcare workers ("workers"), education and childcare administrators ("administrators"), or teaching assistants ("assistants"). The number of jobs to allocate is determined arithmetically based on policy goals. The existing number of jobs for each of these three occupations, respectively, are 46,741 workers, 5,141 administrators, and 11,583 assistants, and the numbers of new jobs for each occupation are 138,205 workers, 15,201 administrators, and 34,248 assistants.

The data is cleaned as follows. Earned income is dropped if it is below $5,000 per year or over $300,000 per year. Weeks worked per year is obtained directly from wkswork1 or imputed from the categorical variable wkswork2 using the mode of each wkswork1 bucket corresponding to the ranges of the categories in wkswork2. Hourly earnings are then obtained by dividing earned income by hours worked times weeks worked. Education is categorized as "Less than HS/GED," "Completed HS/GED," or "Completed bachelor's." The age variable consists of those under 25, ages 26-40, ages 41-65, and 66 and older. The variables race and hispan are combined to produce four race groups: White not Hispanic, Black not Hispanic, Other not Hispanic, and Hispanic. Marital status is categorized as Married, Separated/Widowed/Divorced, or Single. Number of children under five is collected as either zero, one, or more than one.

People are considered eligible for a job if they are between ages 18 and 74 (this is determined heuristically from the age distribution of childcare workers), and indicated that they are available for work, looking for work, or part of the labor force. People are considered donors if they currently work in a childcare job, as defined above.

Instead of computing likelihood of employment in general for this model, likelihood of employment in a childcare job is computed. Since there is only one industry counted for childcare jobs, this likelihood is equivalent to the industry likelihood. As we will see, using only the likelihood of employment in childcare rather than employment in general is highly consequential for the results of the simulation. This likelihood is computed separately for men and women. A probit model with robust standard errors estimates the likelihood of having a child care job as a function of age, education, race, marital status, number of children under five, and whether the person lives in New York City (NYC).

A nearly identical pair of probits is run in order to calculate the IMR. For each sex, these estimate the probability of working in childcare based on education, labor force status, age, and number of children under five. Note that this is a point of some divergence with the general LIMM approach. Here, labor force status is a regressor and employment is the dependent variable, rather than having labor force status be the dependent variable. Nevertheless, this detail is not relevant to the issues discussed below. These two probit models are run using what might be called a naive lasso-style approach, where if the model fails to converge after 50 iterations or a variable is multi-collinear with another, a variable is dropped and it tries again. All of the variables in these regressions are categorical, and the final result of this operation is a fit probit model for which some of them have been dropped. The coefficients of this final probit are used to produce a linear prediction. The IMR is then produced from this linear prediction.

A single multinomial logit with robust standard errors is used to compute each person's probability of working in each of the three childcare occupations. It is fit to the donor set. The independent variables used are sex, age, education, race, marital status, number of children under five, and whether the person lives in NYC.

An OLS regression estimates the logarithm of wages as a function of the ungrouped age variable and its square, the (again grouped) education, number of children under five, occupation

likelihoods, and IMR. Another OLS regression estimates hours worked as a function of the same variables, with the addition of the predicted wage from the first regression. These two regressions are run in four separate cells, based on sex and whether the person lives in NYC.

With these regressions computed, the job allocation process can begin. Each person's occupation likelihoods are transformed into a ranking of the three childcare occupations. This is a crucial difference from the usual LIMM model, which considers occupations of every sort; as we will see, this is highly consequential for the results of the simulation. For each of the three occupations, jobs are offered to the eligible recipients who have that occupation as their likeliest. As described above, the job offer is made using imputation using hot-decking to conduct statistical matching between the eligible recipients being considered and the set of donors who have that occupation. Individuals accept or reject job offers based on their current earnings. Those who accept the job offer are assigned a job in order of their likelihood of employment in childcare, until either jobs or recipients are exhausted. If jobs remain after this first round, then for each occupation, jobs are allocated to the eligible recipients who have that occupation as their second-likeliest, and so on.

PROBLEMS WITH THE DISTRIBUTION OF NEW JOBS

Using the methodology described above for New York State, two problems are encountered. First, the distribution of education and childcare administrator jobs conditioned on sex is highly unrealistic. The true sex distribution of administrators is shown in the first two columns of Table 1. The distribution produced by the simulation described above is shown in the latter two columns. The simulation suggests that the proportion of childcare administrators who are men will increase from less than 10 percent to more than 60 percent. Subjectively, this seems like an unrealistically large increase.

There are a few explanations for this possibility which would not indicate a problem with the matching algorithm. First, it could be due to randomness in the process of statistical matching. However, if the true statistic is 9.26 percent, then an estimation of 63 percent should be nearly impossible, and in fact a similar number is produced with every run. Another possibility is that it could reflect differences between the already-employed and the unemployed populations. This is

a key rationale for using the LIMM—that it allows these differences to be taken into account when simulating job creation due to investment. However, this would only make sense if the unemployed were overwhelmingly male. As we can see in Table 2, this is not the case. Among those who do not have jobs and would be eligible to receive a job in the simulation, men are only a slight majority.

Table 1: Sex Distribution of Childcare Administrator Jobs, Actual vs Simulated

Sex	Actual count	Percent of actual	Simulated count	Percent of simulated
Male	476	9.26	12,799	63.01
Female	4,665	90.74	7,513	36.99
Total	5,141	100.00	20,312	100.00

Table 2: Sex Distribution of Unemployed People Meeting Eligibility Criteria

Sex	Count	Percent
Male	405,083	52.63
Female	364,642	47.37
Total	769,725	100.00

Finally, it is theoretically possible that the reason so many childcare administrators are women actually has little to do with sex (or gender), but is in fact much more strongly related to some other variable which is highly correlated with being a woman in the childcare worker population but is correlated with being a man in the unemployed population. Given the stylized facts about childcare employment and occupational gender segregation and the list of independent variables, we rule out this final possibility.

What remains is the recognition that the job allocation algorithm described above produces conditional distributions of occupation with respect to sex which are not adequately realistic, even granting the empirical limitations of the data. The issue occurs due to the treatment of

occupation and employment propensities. Recall that jobs are allocated by first identifying the group of people who would be most likely to work in that occupation, given (counterfactually) that they are working. Among men who work in childcare, the proportion who work as administrators is much larger than the proportion who work as childcare workers or teaching assistants. The proportion of men in childcare who work as administrators, shown in the first two columns of Table 3, is 12.3 percent. For women, shown in the latter two columns of the same table, that figure is 7.8 percent.

Table 3: Occupation Distribution of Men and Women Working in Child Care, Actual

Occupation	Male count	Percent of males	Female count	Percent of females
Administrator	476	12.29	4,665	7.83
Assistant	377	9.73	11,206	18.80
Worker	3,021	77.98	43,720	73.37
Total	3,874	100.00	59,591	100.00

As a result, the occupation likelihoods in the model bias men toward administrator jobs. This is evident from comparing histograms of childcare administrator occupation likelihoods between men and women, as is shown in Figure 1.

This distribution of childcare administrator likelihoods produces the likelihood rankings shown in Table 4. All of the eligible recipients who have childcare administrator as their most likely job in the childcare sector are men. This is an anomaly caused by exclusively considering occupations within the childcare sector. If all occupations were being considered, childcare administrator would probably not be the most likely occupation for many men.

Figure 1: Likelihood of Working as a Childcare Administrator, Given Employment in Child Care

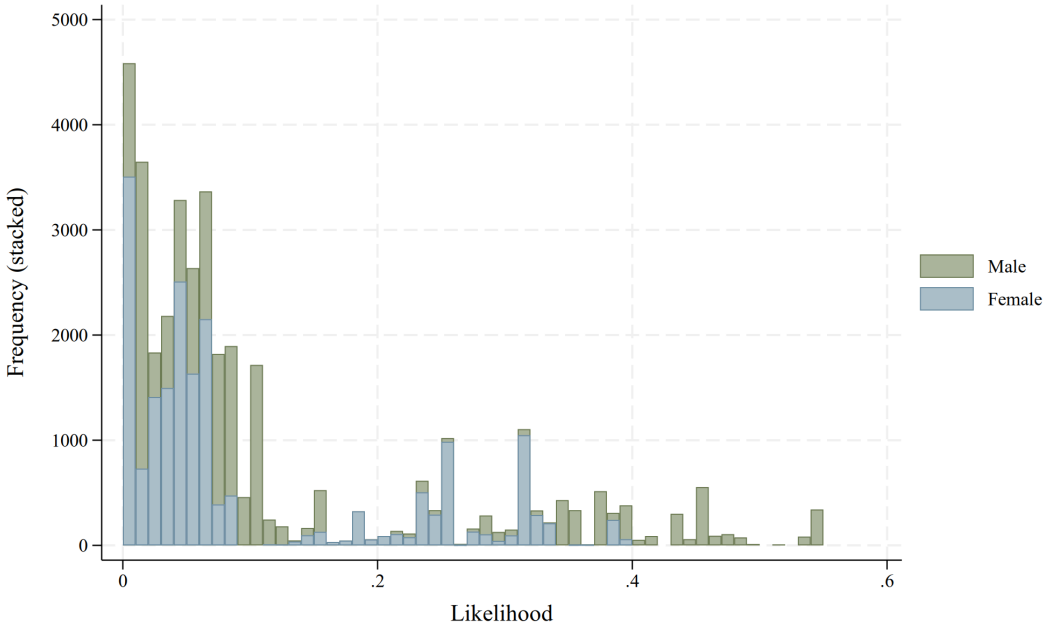
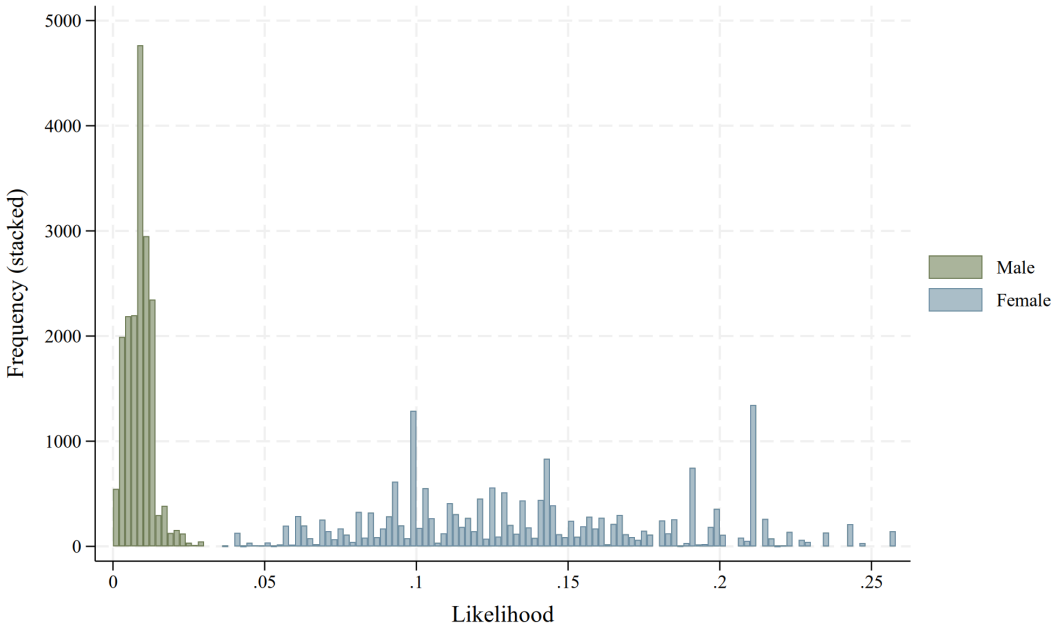


Table 4: Likelihood Ranking of the Administrator Occupation, by Sex

Sex	1st rank	2nd rank	3rd rank	Total
Male	13,557	144,804	246,722	405,083
Female	0	75,111	289,531	364,642
Total	13,557	219,915	536,253	769,725

There are 15,201 administrator jobs to allocate. Recall that the job allocation algorithm, when allocating administrator jobs, first makes offers for all of the individuals with administrator as their most likely childcare occupation. For those who accept the offer, jobs are allocated in order of employment likelihood. Figure 2 shows that the predicted likelihood of men to work in childcare is much lower than that of women.

Figure 2: Likelihood of Working in Child Care, by Sex



The composition of the administrator first rank is nearly as many men as there are jobs—with zero women falling in this category. Thus, the distribution of employment likelihood is irrelevant for the allocation of jobs in this first round. Most of the childcare administrator jobs will be allocated to men, regardless of the likelihood that they would work in childcare at all.

Another pathological case is one in which, for a given occupation, the distribution of likelihoods of working in that occupation is similar for each sex, but employment likelihood is very different between the sexes. In this case, although the top likelihood ranking will have both men and women to choose from, the fact that jobs are allocated strictly in order of decreasing employment likelihood means that the group with greater employment likelihood will—if it is sufficiently large in comparison with the number of jobs—take all of the jobs. This is exactly what happens with the childcare worker occupation. In this case, the distributions of occupation likelihoods are quite comparable, as seen in Figure 3.

This continues to hold for the likelihood rankings, shown in Table 5, which have very large groups of both men and women for whom childcare worker is the most likely occupation in the childcare industry.

Figure 3: Likelihood of Being a Childcare Worker, Given Employment in Child Care

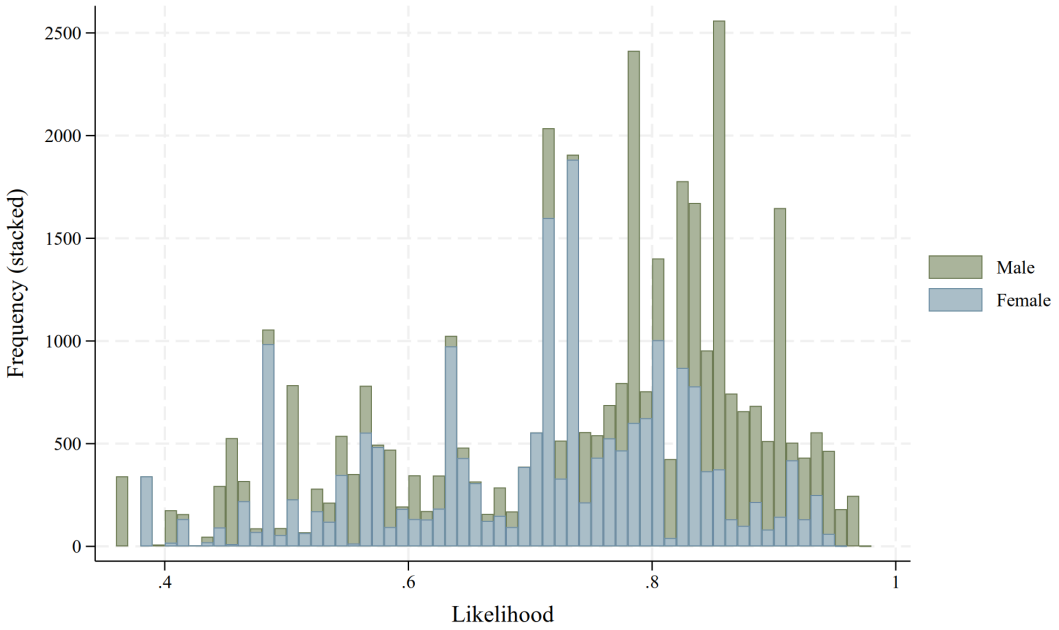


Table 5: Likelihood Ranking of the Childcare Worker Occupation, by Sex

Sex	1st rank	2nd rank	Total
Male	391,526	13,557	405,083
Female	364,334	308	364,642
Total	755,860	13,865	769,725

Despite this, one hundred percent of new childcare jobs are allocated to women. This is shown by comparing the first two columns of Table 6, which show the actual distribution of jobs, with the latter two columns of the same table, which show the simulated data. After all new jobs are allocated, the only men who are childcare workers are the 3,021 men who already were childcare workers. This is because of the stark difference in childcare employment likelihood seen in Figure 2. Recall that in the New York model, childcare employment likelihood is used in place of employment likelihood, as childcare jobs are the only sector being allocated. When all eligible recipients who have childcare worker as their most likely occupation are sorted by likelihood of

employment in childcare, the queue effectively becomes sorted by sex, and the available jobs are exhausted before any man is allocated a job.

Table 6: Sex Distribution of Childcare Worker Jobs, Actual vs Simulated

Sex	Actual count	Percent of actual	Simulated count	Percent of simulated
Male	3,021	6.46	3,021	1.63
Female	43,720	93.54	181,908	98.37
Total	46,741	100.00	184,929	100.00

These two problems compound to severely distort the distribution of jobs among the sexes in the New York model. This is due to the fact that childcare employment likelihood is used, rather than the likelihood of any employment; and because only the three childcare occupations are ranked and allocated, rather than all occupations. As explained in the subsequent sections, it is theoretically possible for highly attenuated versions of these problems to occur in other implementations of the LIMM, due to the similar way rankings and probabilities are handled during allocation. Due to the differences in how those rankings and probabilities are computed compared to the New York model, however, they are unlikely to be pathological. That is, the computed rankings and probabilities are unlikely to cause substantial distortions in the allocation phase of other LIMM implementations. The mathematical basis for this conclusion is established in the following section, which identifies solutions to these two problems. Implications for other LIMM implementations are explored further in the discussion section.

RECOMMENDATIONS FOR IMPROVED JOB ALLOCATION

The first problem is that employment likelihood is only taken into account if the distribution of occupation likelihoods is favorable. Occupation likelihood is a probability that is conditional on the counterfactual employment of the person under consideration. When allocating jobs based primarily on occupation likelihood and considering employment likelihood only secondarily, it is effectively treated as if it were an unconditional probability. In order to allocate jobs based on the

likelihood that the person would work in that job, we must use the unconditional probability that the person works in that occupation. This is given by the joint distribution of the conditional probability that they work in an occupation given that they are employed, together with the probability that they are employed. That is, the two probabilities must be multiplied. In the case where industry is used, the three probabilities must be multiplied.

This approach implies the abandonment of occupation and industry likelihood rankings. One could attempt to scupper the rankings approach by ranking each person's options, including unemployment. But this would introduce further problems, such as how to decide how much unemployment to allocate to those whose most likely state is unemployment. It should also be observed that, apart from the addition of "unemployed" to the occupation rankings, the rankings would remain exactly the same as in the original algorithm. This is because multiplying each occupation likelihood by the same non-negative employment likelihood will preserve their order. One may therefore entertain some skepticism that, after assigning all of those who are most likely to be unemployed as such, the distribution of jobs for the remaining occupations will be fixed. Recall that in the New York model, employment likelihood in fact refers specifically to employment in the childcare sector, and that almost all women have higher childcare employment probabilities than any man. The result of this adjustment to the algorithm would be that all eligible recipients would be assigned into unemployment, and all the jobs would be assigned to women. This would again yield an unrealistic distribution of jobs.

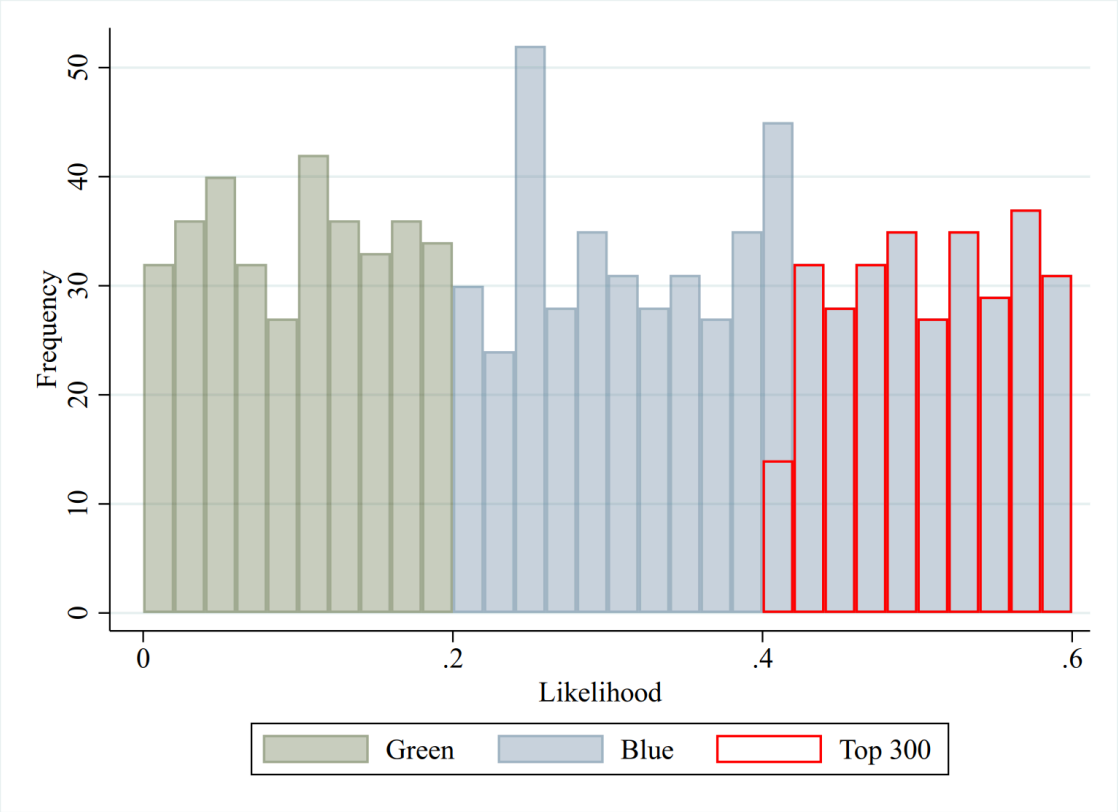
Another approach, in keeping with the spirit of the original algorithm, would be to assign jobs in decreasing order of the unconditional probability for that occupation. This is a dead end, as we will see. The method is applied iteratively through occupations. For a given occupation, job offers are generated for all eligible recipients by matching them with donors with that occupation, regardless of their occupation likelihood or employment likelihood. Job offers are then accepted or rejected as in the original algorithm. Those accepting their offer are then allocated jobs in order of their unconditional probability of working in that occupation, descending. If industry is being considered, then the probability used for ordering is the joint probability of working in that occupation and industry and of being employed.

This will not work because the selection algorithm is not actually treating probabilities as probabilities. An abstract example may help. Imagine throwing darts at a map of the world (with

a blindfold, if you are any good at darts). The probability of a dart landing on any country (or body of water) is equal to its area proportional to the map. Re-throw a dart if it lands in the water, or in a country that's already been hit. Simulating twenty dart throws with the "top probability first" approach, we would always hit exactly the largest twenty countries. But in reality, it is very likely that many smaller countries will be hit. If we are interested in a statistic that has some correlation with country size (such as total agricultural output), the "top probability first" simulation will consistently provide a biased estimate of that statistic.

This principle can be illustrated using simulated data. Figure 4 shows a histogram of uniformly distributed selection probabilities for 1,000 observations, where all observations with selection probability below 0.2 are colored green, and those with greater selection probability are colored blue. Selecting observations by sorting by selection probability and picking the top 300 yields the set of observations outlined in red.

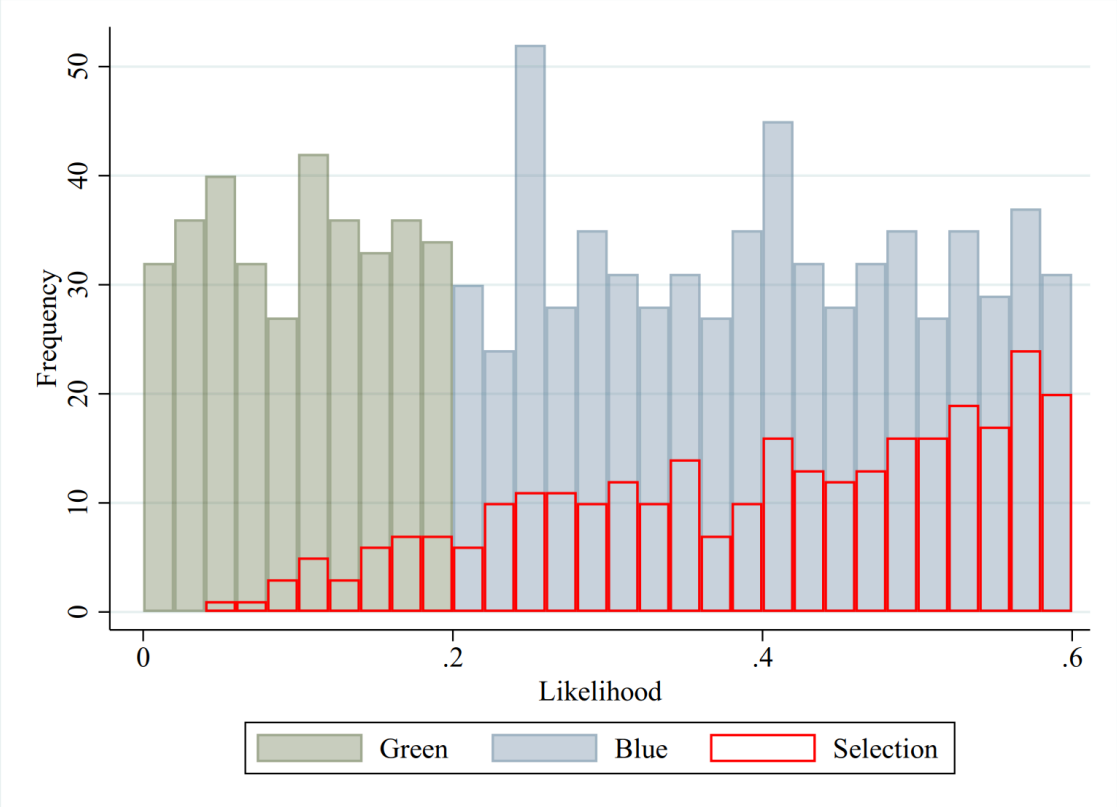
Figure 4: Selecting the 300 Most Likely Observations of Simulated Data



This selection algorithm will result in a selected group which is exclusively blue. If these likelihood values are interpreted as prior probabilities, our sampling would give us strong reason to doubt them. Post-hoc, it appears extremely unlikely that there is about 18 percent of the population that is green and had a 10-20 percent chance of being selected. Another way of viewing this is in terms of aggregate probabilities. The sum of all green likelihoods is 35, and the sum of all blue likelihoods is 295, which means that, in aggregate, we should expect about 10.6 percent of draws to be green.

Figure 5 shows the same data set, with observations selected probabilistically. Each observation is chosen by simulating a draw from a discrete distribution with $N = 1000$ possibilities and each possibility having a draw probability equal to its likelihood value, normalized so they sum to 1.

Figure 5: Selecting Observations of Simulated Data Probabilistically



Selecting observations according to their probability of being selected yields a very different set of observations. It is visually apparent that "top probability first" selection is not an acceptable

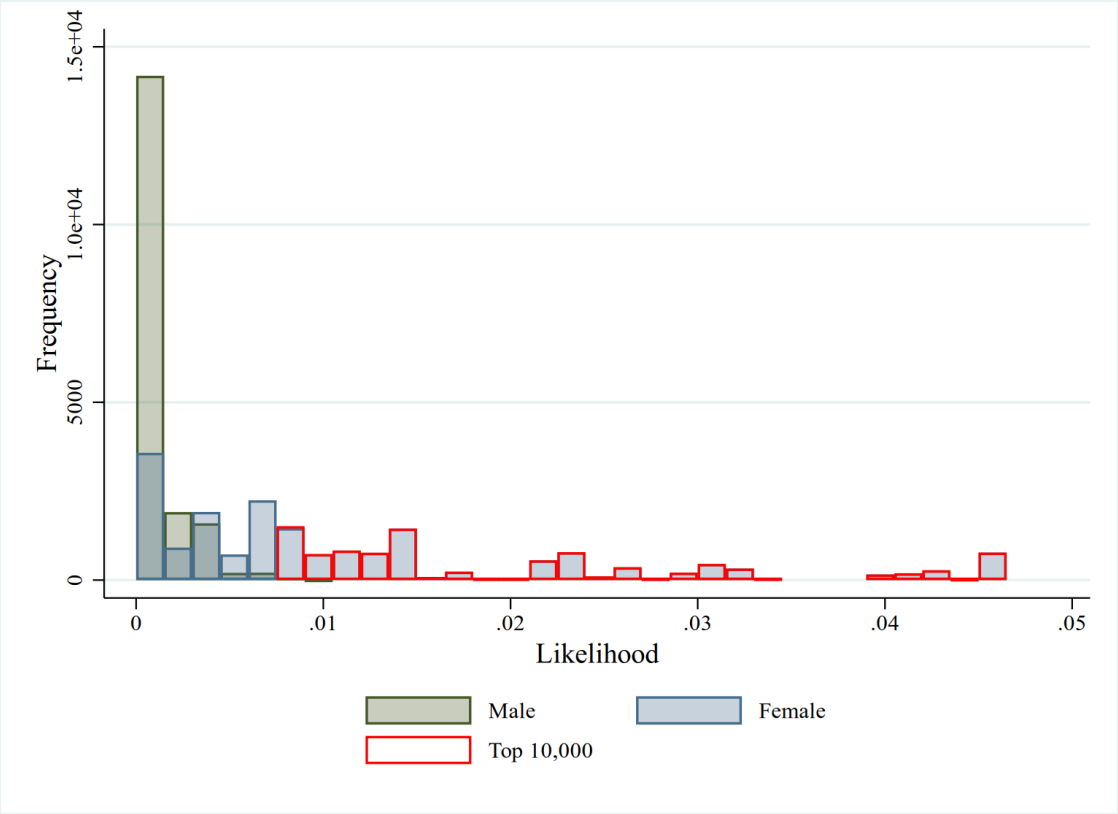
approximation for probabilistic selection. Whereas selecting the observations with maximum likelihood makes frequency a step function with respect to likelihood, selecting observations probabilistically results in selection frequency being a smooth linear function of likelihood. And whereas the first algorithm selects no green observations at all, probabilistic selection results in 33 of the 300 selected observations being green—that is, 11 percent. This is very close to the expectation calculated above that, on average, 10.6 percent of samples should be green.

This problem can also be understood more theoretically by examining the selection process. Assigning each person their most likely occupation and industry is a type of maximum a posteriori classification. For each individual, the chosen classification is the mode of the individual classification probabilities. Maximum a posteriori classification is often an effective way to classify individual data points. However, there is no reason to believe that the modes of the probabilities should, in aggregate, approximate the distribution of the probabilities themselves.

The problem cannot be adequately reduced to that of an invalid independence assumption. The problem is not merely that one person's chance of being a teacher depends on whether another person becomes a teacher. This is made clear by examining the blue/green simulation above. In the simulation, we may posit total independence of the selection probabilities of each observation, so that such an independence assumption is valid. The problem with maximum likelihood selection must therefore stem from deeper statistical problems, namely that the distribution of modal probabilities for each individual does not approximate the distribution of probabilities across the population.

Returning to the problem of job allocation and sex, we can make similar plots for the probability of employment in childcare administration. This is shown in Figure 6. The probability used is the unconditional probability (the joint distribution of employment probability and occupation probability given employment). The results of selecting 10,000 candidates using the "top probability first" approach are shown with a red outline.

Figure 6: Selecting the Top 10,000 Eligible Recipients by Likelihood of Being a Childcare Administrator



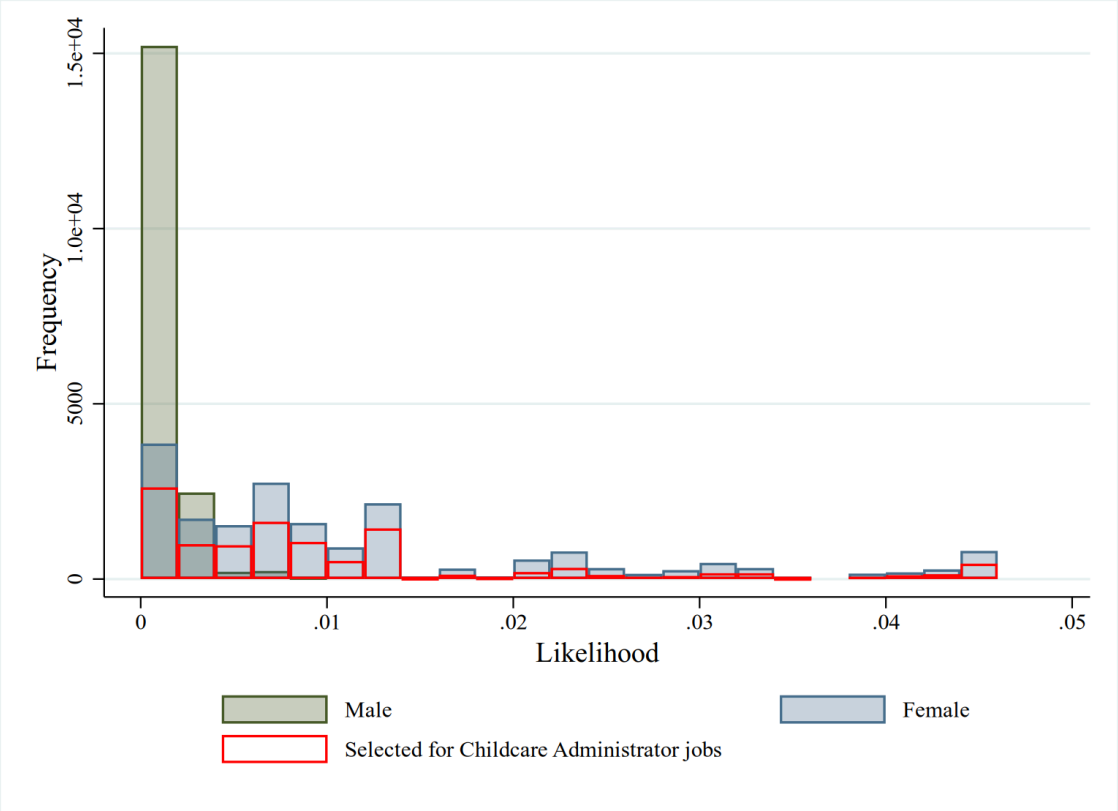
After adjusting the simulation to assign jobs to eligible recipients who accept their job offers in order of decreasing unconditional probability of employment in each occupation, the results are much as would be suggested by Figure 6. The bimodal nature of the employment likelihood distribution has resulted in unconditional childcare administrator probability distributions which are very different for each sex. Selecting the top 10,000 candidates exclusively selects women. This is the opposite problem we had originally, when men were being disproportionately selected for administrator jobs, and now the problem is (magnitudinally) worse. The results of administrator job allocation using this algorithm are shown in Table 7.

The preferred approach is to use probabilistic selection, where each person's probability of being selected for a job is the joint probability that they work in that occupation (and industry) and that they are employed. Figure 7 shows the distribution of joint likelihood scores selected in this manner.

Table 7: Sex Distribution of Childcare Administrator Jobs, Actual vs Simulated, Using "Top Probability First" Algorithm

Sex	Actual count	Percent of actual	Simulated count	Percent of simulated
Male	476	9.26	476	2.34
Female	4,665	90.74	19,842	97.66
Total	5,141	100	20,318	100

Figure 7: Selecting 10,000 Eligible Recipients Probabilistically, Using Likelihood of Being a Childcare Administrator



Notably, significant parts of the selection include likelihood bins that contain men. The distribution of childcare administrator jobs allocated to each sex, shown in Table 8, shows a small shift in the gender composition of childcare administrators, from 9.3 percent to 12.1 percent men.

Table 8: Sex Distribution of Childcare Administrators, Actual vs Simulated, Assigned Probabilistically

Sex	Actual count	Percent of actual	Simulated count	Percent of simulated
Male	476	9.26	2,467	12.13
Female	4,665	90.74	17,865	87.87
Total	5,141	100	20,332	100

Some of this, but not all, is explained by the gender composition of the pool of eligible recipients, which is 53 percent men. The shift is small enough not to warrant alarm—it likely captures other salient population-level differences among the unemployed of each gender group. It is in accordance with the idea that the characteristics of job holders in aggregate might change as the unemployed move into employment, which is the chief purpose of using the LIMM for job allocation, rather than arithmetic methods. Table 9 shows a similar shift for childcare worker jobs when using probabilistic allocation. Men are neither assigned a wildly disproportionate number of childcare worker jobs, nor excluded entirely.

Table 9: Sex Distribution of Childcare Workers, Actual vs Simulated, Assigned Probabilistically

Sex	Actual count	Percent of actual	Simulated count	Percent of simulated
Male	3,021	6.46	19,166	10.36
Female	43,720	93.54	165,777	89.64
Total	46,741	100	184,943	100

The recommendation of this paper can therefore be summarized as follows. For each job, candidates should be selected through random sampling without replacement. The probability of selecting any particular candidate should be their unconditional probability of working in that job, equal to the joint probability that they work in that occupation (and industry) and that they are employed—normalized to sum to one across all candidates. This ensures that the distributions of characteristics among job recipients respects the distributions of job probabilities

among members of those characteristic groups. It yields results that seem, subjectively, much more realistic than can be obtained using ranking, and resolves the theoretical problem of using the conditional probability of working in an occupation given being employed as if it is the unconditional probability of working in that occupation.

DISCUSSION

The present paper suggests methodological improvements to the Levy Institute Microsimulation Model based on problems encountered while developing the New York model in Istenes (2023). While these problems have some theoretical relevance to prior investment simulation models using the LIMM, the distortions seen in the New York model are attributable to the differences in how rankings and likelihoods were computed. While the New York model uses only three occupations and a single industry, other simulations use on the order of ten occupations and ten industries. One consequence of this is that, while in the New York model the probability used for the initial ranking is the probability that a person will work in an occupation given that they are employed in the childcare sector, in the other models it is the probability that the person will work in an occupation given that they are employed at all. Since the gender distribution of childcare sector workers is generally skewed strongly toward women, and the gender distribution of the employed is generally skewed toward men, the conditional probability that a man will work as a childcare administrator given that he is employed is much closer to the unconditional probability that he works as a childcare administrator than the conditional probability that he will work as a childcare administrator given that he is employed in the child care sector. The problems stemming from the misuse of the probability of working in an occupation or industry conditional on their working are therefore likely to be much less severe.

However, if occupation and industry probabilities are conditioned on employment in general, this means it is likely that the distributions of conditional probabilities for childcare jobs are much more gender-differentiated than those encountered above. The probability that a man is a childcare worker, given that he works in childcare is not much different than the probability that a woman is a childcare worker, given that she works in childcare, as is shown in Figure 3. The probability that a man is a childcare worker given that he is employed is, in most contexts, likely

to be much lower than the same probability for a woman. It is likely that the first-stage ranking of eligible recipients by occupation (and industry) likelihood usually selects exclusively women as most likely to take jobs in childcare. Merely integrating employment probability by using the joint distribution will not help in this case. It is necessary to select candidates probabilistically.

One further note about implementing the recommended selection algorithm is warranted. This algorithm allocates jobs by iterating across job types and allocating available jobs to people in the eligible pool. If all jobs of a given type (that is, of a given occupation and industry) are allocated before proceeding to the next, then the pool of people randomly selected to fill that next job type has been distorted by the loss of those already assigned jobs. This would make the outcome of the algorithm dependent on the order of the occupations and industries, which is undesirable. This problem did not appear in the New York model because there were only three job types, and the first two in the list—administrator and assistant—had relatively few job openings available. When more occupations and industries are used, as is the case in all other LIMM implementations, this problem would distort the characteristic distributions of recipients of jobs which are reached later in the iterative allocation process. In order to mitigate this problem, jobs should be allocated in many passes over the job types, allocating only a small proportion of available jobs at a time. For example, in each step of allocation, 2 percent of available jobs of that type (or twenty jobs of that type if fewer than 1,000 remain) could be allocated using the statistical matching and probabilistic selection process, so that all jobs would be allocated after 50 passes through the set of job types. This will ensure that all job types have roughly equal access to the eligible candidates. The effectiveness of this approach should be verified by modifying the occupation and industry iteration order and seeing if the results change.

There is no question that the investment simulation models using the complete LIMM obtain reasonably realistic results, and that the correct conclusions are drawn from them. The results, when viewed in aggregate, align well with expectations about sectoral employment. The recommendations in the present paper should yield small adjustments in most LIMM implementations. The LIMM has been an effective tool for demonstrating the gendered benefits of social care investment. Specifically, it has provided greater precision to the argument that social care investment is a powerful tool for creating employment and income for women. The

methods shown here enhance that precision, honing the tool, and providing it with stronger theoretical foundations on which to advance its argument.

REFERENCES

Istenes, B. 2023. "Simulating Jobs Created by the New York Universal Child Care Act." *Theses - Graduate Programs in Economic Theory and Policy*, January. https://digitalcommons.bard.edu/levy_ms/48.

Kim, K., İ. İlkkaracan, and T. Kaya. 2019. "Public Investment in Care Services in Turkey: Promoting Employment & Gender Inclusive Growth." *Journal of Policy Modeling* 41 (6): 1210–29. https://ideas.repec.org//a/eee/jpolmo/v41y2019i6p1210-1229.html.

Masterson, T. 2013. "Quality of Statistical Match and Simulations Used in the Estimation of the Levy Institute Measure of Time and Consumption Poverty (LIMTCP) for Turkey in 2006." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2296203.

Masterson, T., R. Antonopoulous, L. Nassif Pires, F. Rios-Avila, and A. Zacharias. 2022. "Assessing the Impact of Childcare Expansion in Mexico: Time Use, Employment, and Poverty," August. https://www.levyinstitute.org/publications/assessing-the-impact-of-childcare-expansion-in-mexico-time-use-employment-and-poverty.

Zacharias, A., T. Masterson, and K. Kim. 2009. "Distributional Impact of the American Recovery and Reinvestment Act: A Microsimulation Approach." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1420776.

Zacharias, A., T. Masterson, F. Rios-Avilla, M. Nikiforos, K. Kim, and T. Khitarishvili. 2019. *Macroeconomic and Microeconomic Impacts of Improving Physical and Social Infrastructure: A Macro-Micro Policy Model for Ghana and Tanzania Final Project Report: Understanding the Interlocking of Income and Time Deficits for Men and Women in Ghana and Tanzania: Revisiting Poverty Measurement, Rethinking Policy Responses*. https://doi.org/10.13140/RG.2.2.31042.94409.