SOCIAL SCIENCE

# Improving refugee integration through data-driven algorithmic assignment

Kirk Bansak,[1,2]* Jeremy Ferwerda,[2,3]* Jens Hainmueller,[1,2,4]*† Andrea Dillon,[2] Dominik Hangartner,[2,5,6] Duncan Lawrence,[2] Jeremy Weinstein[1,2]

**Developed democracies are settling an increased number of refugees, many of whom face challenges integrating into host societies. We developed a flexible data-driven algorithm that assigns refugees across resettlement locations to improve integration outcomes. The algorithm uses a combination of supervised machine learning and optimal matching to discover and leverage synergies between refugee characteristics and resettlement sites. The algorithm was tested on historical registry data from two countries with different assignment regimes and refugee populations, the United States and Switzerland. Our approach led to gains of roughly 40 to 70%, on average, in refugees' employment outcomes relative to current assignment practices. This approach can provide governments with a practical and cost-efficient policy tool that can be immediately implemented within existing institutional structures.**

Refugees are among the world's most vulnerable populations (*1*, *2*). After experiencing war, violence, and years of living in overcrowded refugee camps, refugees arrive in a new country with few resources and must acclimate to an unfamiliar local language, economy, and culture. Refugees frequently remain economically marginalized, with low levels of employment in the years following their arrival (*3*–*5*).

The assignment of refugees to different resettlement locations within a host country is one of the first policy decisions made during the resettlement process (*6*). It is also one of the most consequential in maximizing refugees' economic integration and self-sufficiency as a first step toward a more comprehensive integration into society (*7*–*9*). Three sets of factors affect refugee integration: geographical context, personal characteristics, and synergies between geography and personal characteristics (Fig. 1 and fig. S1). For instance, some resettlement locations in the United States offer better economic and social opportunities that can result in higher levels of refugee employment (Fig. 1A). In addition, refugees with certain characteristics, such as language and educational skills, are more likely to succeed economically regardless of the resettlement location to which they are sent (Fig. 1B). Finally, the expected employment returns associated with personal characteristics can vary across different resettlement locations (Fig. 1C). This indicates that there are synergies between places and people; certain characteristics will make a refugee a better match for a particular location. In Switzerland, for example, we find that the ability to speak French (i.e., among French-speaking African refugees) results in a larger payoff for refugees assigned to French-speaking cantons than for those assigned to German-speaking cantons (fig. S2).

Host countries' current procedures for determining how to allocate refugees across domestic resettlement sites do not fully leverage synergies between refugees and geographic locations. For instance, in the United States, refugees without existing U.S. ties are primarily assigned to resettlement locations according to the capacity of local resettlement offices at the time of arrival, without a systematic assessment of the local employment rate for refugees of similar profiles. In Switzerland, where most refugees initially enter as asylum seekers, the federal government attempts to reduce fiscal and social strain on individual localities by making assignment random and proportional across regions.

Prior research has proposed different schemes for refugee assignment both across countries (*10*, *11*) and within countries (*12*, *13*). These proposals include two-sided matching markets in which an optimized assignment is determined on the basis of match efficiency and/or the preferences of refugees and host locations (*14*). Although these approaches are theoretically appealing, there are practical barriers to their implementation, including a lack of systematic data on refugee preferences and the need for extensive political coordination.

We have developed a data-driven approach that, in contrast, can be immediately implemented by using existing data to optimize integration outcomes. Our algorithm has three stages: modeling, mapping, and matching. The modeling stage involves a supervised machine learning process that predicts the expected success for any quantifiable metric—for example, early employment—of new refugee arrivals across all possible resettlement locations. We designated historical resettlement data for model training, in which the unit of observation was a single refugee and which contained information on the refugees' background characteristics (e.g., country of origin, language skills, gender, age, etc.), time of arrival, assigned location, and measured employment success. These training data were then used to build a bundle of supervised learning models that predicted refugees' expected employment success as a function of their background characteristics. A separate model was fit for subgroups of refugees assigned to each location, thus yielding different models for each location and allowing for the discovery of refugee/location synergies. These fitted models were then applied to new, out-of-sample refugee arrival data to predict the expected employment success of each new arrival at each possible resettlement location.

The mapping stage involves transforming the refugee-level predictions from the modeling stage to a case-level metric. Mapping to a case-level metric is necessary because refugees are often not assigned to locations on an individual basis, but rather on a case-level basis, with cases most often being family units. Various mapping functions can be used. Our preferred case-level metric was the predicted probability that at least one refugee in the case would find employment at the location in question. This metric uses a simplifying assumption that the probabilities of employment for refugees within a case are independent, although we also tested alternative mapping functions—namely the mean, maximum, and minimum predicted probability of employment within each case—that do not require this assumption (*15*).

Finally, the matching stage involves assigning each case to a specific location to fulfill a chosen optimality criterion subject to constraints. Our algorithm is flexible and can accommodate multiple criteria and constraints. The optimality criterion we used in our applications was to maximize the average of the case-level metric (i.e., the global average of the probability that at least one refugee in each family gains employment). We also imposed constraints that represent real-world assignment restrictions, such as how many cases can be sent to different locations. To solve this constrained optimization problem, we used an optimal matching procedure with the RELAX-IV minimum cost flow solver (*16*, *17*); see supplementary materials and figs. S3 to S5 for details of the algorithm, data, measures, and statistical analysis (including out-of-sample classification accuracy and probability calibration).

For the algorithm to obtain reliable predictions, it is important that the historical assignment process not be determined by unobserved refugee characteristics. This criterion is currently met in many countries that assign refugees either randomly (according to burden-sharing constraints) or according to premeasured refugee characteristics that would serve as feature inputs into the algorithm. We assessed the performance of the

[1]Department of Political Science, Stanford University, Stanford, CA 94305, USA. [2]Immigration Policy Lab, Stanford University, Stanford, CA 94305, USA, and ETH Zurich, 8092 Zurich, Switzerland. [3]Department of Government, Dartmouth College, Hanover, NH 03755, USA. [4]Graduate School of Business, Stanford University, Stanford, CA 94305, USA. [5]Center for Comparative and International Studies, ETH Zurich, 8092 Zurich, Switzerland. [6]Department of Government, London School of Economics and Political Science, London WC2A 2AE, UK.
*These authors contributed equally to this work.
†Corresponding author. Email: jhain@stanford.edu

algorithm through applications in two such countries: the United States, where refugees are assigned primarily on the basis of capacity constraints, and Switzerland, where refugees are assigned randomly according to a proportional distribution key (see supplementary materials and tables S1 and S2 for details).

In the United States, reception and placement services (e.g., arranging location assignments, housing, etc.) for refugees are implemented by nine voluntary agencies in cooperation with the Department of State. After refugees are allocated to one of the agencies, placement officers centrally assign refugees to the agency's resettlement locations subject to local capacity constraints (*18*). Placement officers make assignment decisions prior to refugees' arrival and without interviewing the refugees. The premeasured characteristics of a case available to the placement officers can be viewed in the data, and hence can be used as feature inputs into the algorithm.

Refugees are granted work authorization upon arrival and encouraged to find employment as soon as possible. To track refugee resettlement success, the agencies are required to report the refugees' employment status at the end of the reception and placement period, 90 days after

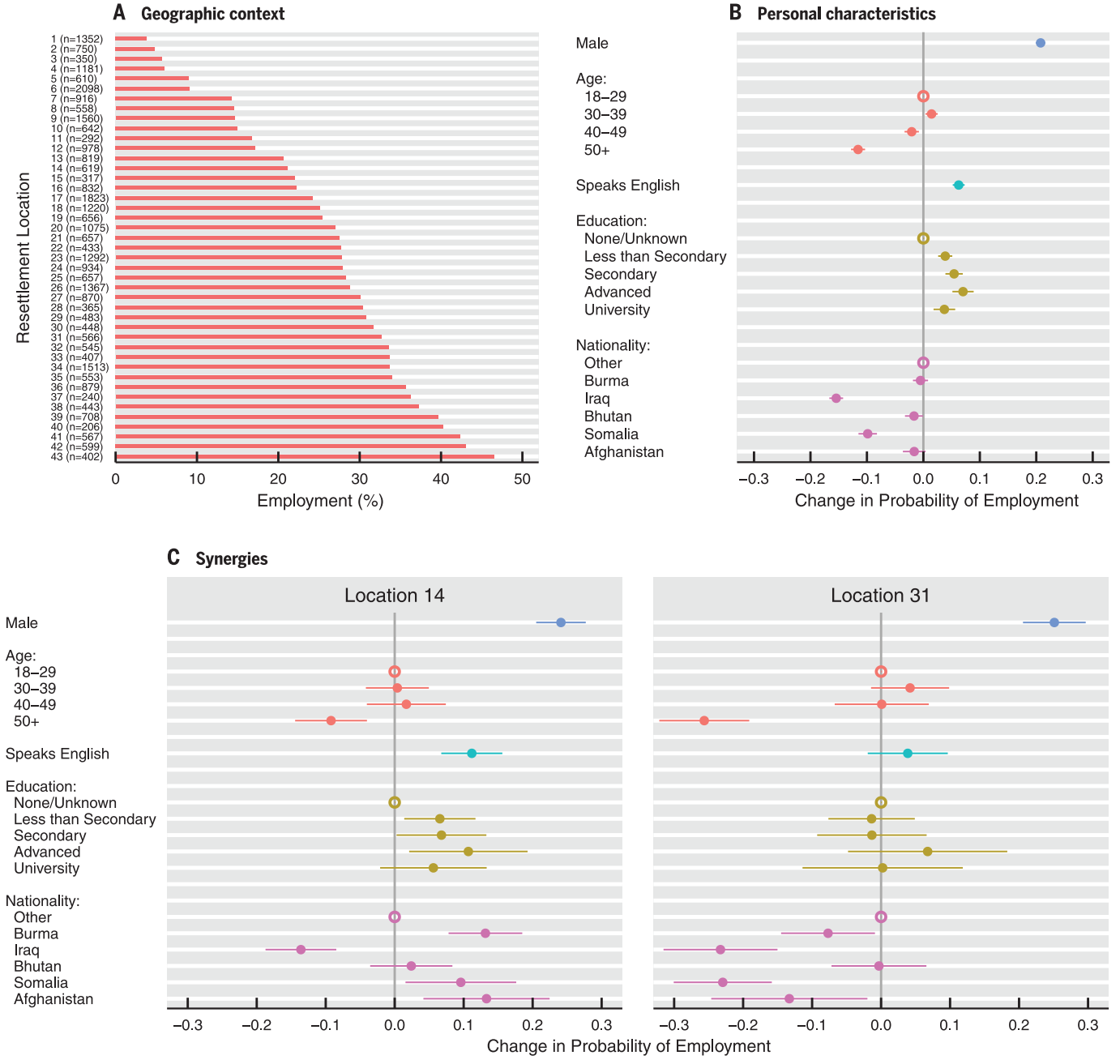


**Fig. 1. Variation in refugee employment in the United States.** (**A** to **C**) Refugee employment at 90 days after arrival varies as a function of refugees' assigned resettlement location (A), personal characteristics (pooled across refugees assigned to all locations) (B), and synergies between characteristics and locations (two example locations) (C). In (B) and (C), dots with horizontal lines indicate point estimates with robust 95% confidence intervals from ordinary least-squares regression. The open circles on the zero line denote reference categories. The data for all three panels include working-age refugees resettled by one of the largest U.S. resettlement agencies during the 2011–2016 period ($n$ = 33,782). These results are replicated for only working-age refugees without U.S. ties (i.e., "free cases") in fig. S1.

arrival. To assess whether an optimized assignment could improve refugee outcomes, we analyzed de-identified data from one of the largest resettlement agencies for working-age refugees (ages 18 to 64; $n$ = 33,782) resettled during the 2011–2016 period. We split the data into training and test sets. For model training, we used data for the refugees who arrived from 2011 up to (but not including) the third quarter (Q3) of 2016, the most recent quarter with available data. We then applied the fitted models to predict the expected employment success at each location and determine the optimal assignment for the test set, refugees who arrived in 2016 Q3. For the test data, we focused on refugees who were free to be assigned to different resettlement locations ($n$ = 919), in contrast to refugees who are assigned according to the location of family or other ties. We also imposed constraints on the assignment such that each location could only receive as many cases under the optimized assignment as were received in actuality.

Our algorithmic assignment considerably increased expected refugee employment over the status quo assignment (Fig. 2). The median refugee's predicted probability of employment in the United States more than doubled, increasing from approximately 25% to 50%. Our optimized assignment increased the probability of finding employment across the entire distribution of refugees, including those who were least likely and most likely to find work.

In addition, the algorithmic assignment yielded higher employment rates in almost every location, including locations that had higher and lower baseline employment rates. On average, the employment rate was 34% under the actual assignment and 48% under the optimized assignment, which means that the optimized assignment would increase the employment rate above the baseline by roughly 41%.

We conducted a second test in the context of Switzerland, whose asylum process is similar to that of other European countries belonging to the Common European Asylum System. In Switzerland, asylum seekers who are not immediately rejected upon arrival are assigned to one of 26 cantons, where they wait for a decision on their asylum application. We focused on asylum seekers who received subsidiary protection status, which is Switzerland's largest refugee category (see supplementary materials). We drew upon data from the Swiss State Secretariat for Migration (SEM), which centrally manages the asylum process and assignment. In contrast to the U.S. case, the SEM uses a proportional random assignment of cases to locations, and tracks employment outcomes for several years after asylum seekers' arrival. This allowed us to benchmark our algorithm against a different status quo assignment mechanism and to optimize for a longer-term employment metric—specifically, refugees' employment at the end of their third calendar year in Switzerland. We focused on all working-age refugees who received subsidiary protection status and arrived from 1999 to 2013 ($n$ = 22,159), with refugees arriving in 2013 who were free to be assigned to any canton as the test set ($n$ = 888) and refugees arriving in all prior years as the training set. We also imposed the constraint that each canton gets assigned the same number of cases as in actuality, in which the number of cases, by law, is assigned in proportion to the population of the canton.

Our algorithmic assignment considerably increased expected refugee employment over the status quo assignment (Fig. 3). Similar to the U.S. context, our algorithm increased the predicted probability of finding employment across the entire distribution of refugees. On average, the third-year employment rate was 15% under the actual assignment and 26% under the optimized assignment. These results suggest that the data-driven assignment has the potential to increase third-year employment in the Swiss context by about 73%.

In the supplement, we present further results for both countries in which we applied alternative specifications for the algorithm. Specifically, we replicated testing with different time periods (figs. S6 and S7), alternative mapping functions (fig. S8), shorter- and longer-term outcomes (fig. S9), and varying lengths of the training data period (fig. S10). The results from these tests all show considerable gains.

Our analysis demonstrated large potential improvements, but we did not test the algorithm prospectively. Ideally, it should be tested in a randomized controlled trial design. In addition, further research is needed to determine whether it is more effective to optimize for short-term or long-term outcomes. In Switzerland, for example, we find considerable gains regardless of whether we optimize for second-, third-, or fourth-year employment (see supplementary materials). In the United States, however, longer-term employment outcomes are currently not tracked. Still, early employment is often highly predictive of long-term employment (*19*), and the use of shorter-term outcomes in the algorithm allows for faster learning of emerging and declining synergies based on more recent data, possibly resulting in a more effective assignment.

In contrast to more expensive interventions (such as language or job training programs) that are sometimes implemented long after refugees' arrival, our approach is cost-efficient and implemented before refugees' arrival, giving them the strongest foundation possible from which to integrate into host societies. Furthermore, our approach modifies an existing policy process, facilitating its immediate implementation, and it is dynamic in that it adapts to synergies over time. Because of the algorithm's data-driven learning capacity, policy-makers do not need to invest in identifying the precise sources of those synergies—local economic conditions, social environments, resettlement office efficacy, etc.—to harness their benefits.

Our approach also preserves the ability of policymakers to set their own parameters and priorities. Specifically, policy-makers can choose their



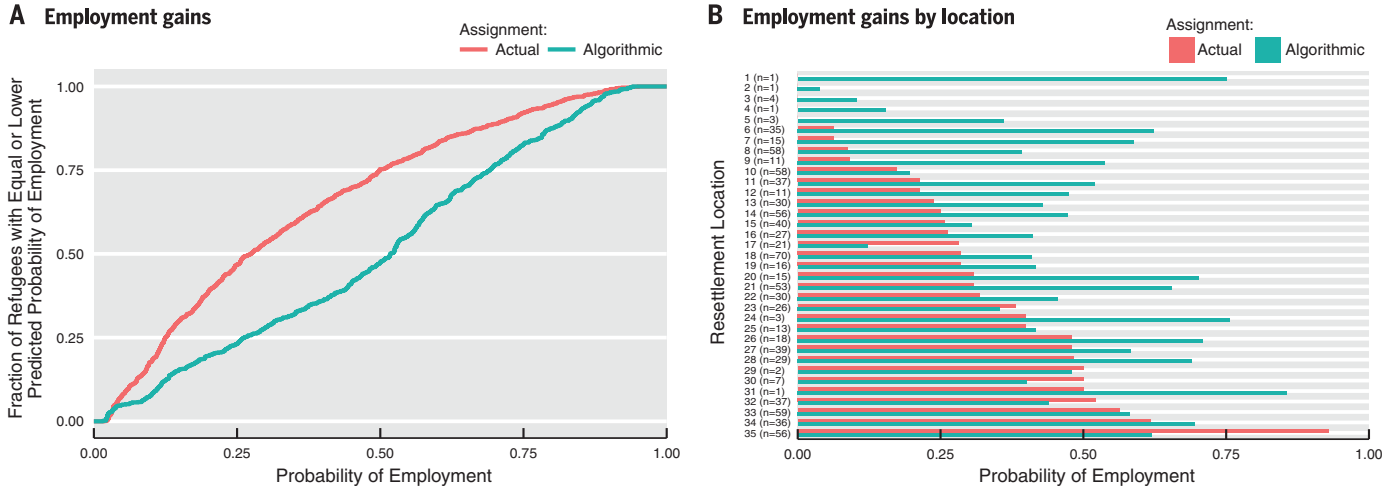


**Fig. 2. Employment gains from data-driven refugee assignment in the United States.** (**A**) Empirical cumulative distribution functions (ECDFs) of the refugees' predicted 90-day employment probabilities under their actual and algorithmic assignments. (**B**) Actual and algorithmic employment rates by resettlement location.

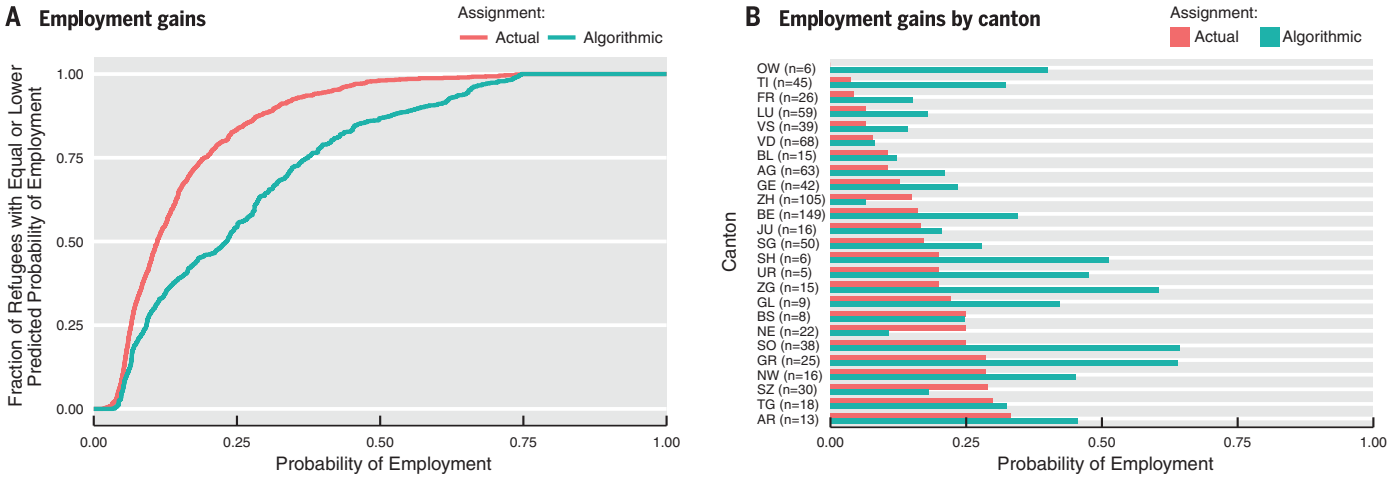


**Fig. 3. Employment gains from data-driven refugee assignment in Switzerland.** (**A**) ECDFs of the refugees' predicted third-year employment probabilities under their actual and algorithmic assignments. (**B**) Actual and algorithmic employment rates by canton. See table S3 for canton names.

own preferred integration success metrics, optimality criteria, and constraints for the assignment. For example, in our U.S. application, we find that the average employment gains achieved under our optimized assignment could be improved even further by slightly relaxing the constraints and allowing the algorithm to increase or reduce the number of cases assigned to different resettlement locations (fig. S11). In addition, if systematic data on refugee preferences regarding their geographic placement would become available, these data could also be incorporated into the algorithmic assignment by, for instance, optimizing for a weighted average of preferences and predicted integration success. Lastly, the algorithm could also be used to complement rather than replace placement officers' expertise. For instance, in a computer-assisted assignment process, the algorithm might provide several recommendations, and placement officers could use their own discretion to determine the final assignment or override any suggestions.

## REFERENCES AND NOTES


1. J. Smith, L. Daynes, *Lancet Glob. Health* **4**, e85–e86 (2016).
2. Médecins Sans Frontières, *The Illness of Migration: Ten Years of Medical Humanitarian Assistance to Migrants in Europe and in Transit Countries* (2013); www.aerzte-ohne-grenzen.de/sites/germany/files/attachments/msf-the-illness-of-migration-2013.pdf.
3. P. Connor, *J. Refug. Stud.* **23**, 377–397 (2010).
4. J. Hainmueller, D. Hangartner, D. Lawrence, *Sci. Adv.* **2**, e1600432 (2016).
5. M. Marbach, J. Hainmueller, D. Hangartner, *The Long-Term Impact of Employment Bans on the Economic Integration of Refugees* (Stanford-Zurich Immigration Policy Lab Working Paper 17-03, 2017); https://ssrn.com/abstract=3078172.
6. Although our focus is on refugee resettlement, the same issues also apply to the assignment of asylum seekers. Resettled refugees are persons who have officially been granted refugee status in advance of their arrival into the host country. In contrast, an asylum seeker is a person who has fled his or her home country and submitted a formal request for asylum in a host country. Authorities in that country then process the request and decide whether to grant official refugee status to the asylum seeker. This determination can take several years to conclude, and in the interim, the asylum seeker is typically placed within a specific location in the host country.
7. P.-A. Edin, P. Fredriksson, O. Åslund, *Q. J. Econ.* **118**, 329–357 (2003).
8. L. A. Beaman, *Rev. Econ. Stud.* **79**, 128–161 (2011).
9. A. P. Damm, *J. Labor Econ.* **27**, 281–314 (2009).
10. J. Fernández-Huertas Moraga, H. Rapoport, *J. Public Econ.* **115**, 94–108 (2014).
11. J. Fernández-Huertas Moraga, H. Rapoport, *CESifo Econ. Stud.* **61**, 638–672 (2015).
12. T. Andersson, L. Ehlers, *Assigning Refugees to Landlords in Sweden: Stable Maximum Matchings* (Lund University, 2016); http://project.nek.lu.se/publications/workpap/papers/wp16_18.pdf.
13. D. Delacrétaz, S. D. Kominers, A. Teytelboym, *Refugee Resettlement* (University of Melbourne, 2016); www.t8el.com/jmp.pdf.
14. Fernández-Huertas Moraga and Rapoport (*10*, *11*) couple an auction mechanism for tradeable refugee quotas with a preference-matching algorithm that optimizes over refugees' preferences for resettlement countries and countries' preferences over refugee types. Andersson and Ehlers (*12*) focus on within-country matching and develop an algorithm to find a stable maximum matching of refugees to landlords given induced preferences for landlords and refugee families. Delacrétaz *et al*. (*13*) provide algorithms that optimize match efficiency subject to multidimensional capacity constraints and incorporate refugee preferences and location priorities.
15. The formula we use is $y_{gij} = 1 - \prod_{i \in g}(1 - \alpha_{ij})$, where $\alpha_{ij}$ corresponds to the predicted probability of a positive employment outcome for refugee $i$ at location $j$, and $g$ denotes a particular case. Note that this formula uses a simplifying assumption that the probabilities of employment for refugees within a case are independent. See the supplement for additional details and alternative mapping functions—the mean, maximum, and minimum predicted probability of employment within each case—that do not require this assumption.
16. B. B. Hansen, S. O. Klopfer, *J. Comput. Graph. Stat.* **15**, 609–627 (2006).
17. D. P. Bertsekas, P. Tseng, *RELAX-IV: A Faster Version of the RELAX Code for Solving Minimum Cost Flow Problems* (MIT Laboratory for Information and Decision Systems, 1994).
18. Each location can only accommodate a limited number of cases each year. Other less common restrictions include the inability of certain locations to accept cases with severe medical conditions or particular languages. The assignment algorithm is designed to incorporate such constraints.
19. In the Swiss data, employment at the end of the first calendar year after arrival is associated with a 62–percentage point ($P < 0.0001$) increase in the probability of employment at the end of the second year and a 43–percentage point ($P < 0.0001$) increase at the end of the third calendar year; the $P$ values are from a linear regression of actual second- or third-year employment on first-year employment, respectively.


## ACKNOWLEDGMENTS

We acknowledge funding from the Ford Foundation for operational support of the Stanford Immigration Policy Lab, and from the Swiss National Science Foundation and the Carnegie Corporation of New York (J.H.). The funders had no role in the data collection, analysis, decision to publish, or preparation of the manuscript. The U.S. refugee registry data were provided to us under a collaboration research agreement with the Lutheran Immigration and Refugee Service (LIRS). This agreement requires that we do not transfer or disclose the data. Researchers can request access to the data from LIRS. The Swiss ZEMIS database is provided to us under a data use agreement with the Swiss State Secretariat for Migration (SEM), which requires that we do not disclose the individual-level data. Researchers can request access to the ZEMIS data from the SEM. We thank the SEM and LIRS for access to data and guidance, and G. Imbens, D. Laitin, R. Reich, L. Stanczyk, and S. Wager for helpful advice. For replication code, see Harvard Dataverse (doi:10.7910/DVN/MS8XES). The institutional review boards at Stanford University (IRB-40212) and Dartmouth College (00030198) approved this research. The U.S. resettlement agency that provided the data requested that we not disclose specific locations within the article.

## SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/359/6373/325/suppl/DC1
Supplementary Text
Figs. S1 to S11
Tables S1 to S3
References (*20*–*37*)

24 July 2017; accepted 13 December 2017
10.1126/science.aao4408