Inside Categorization: Objectifying the Discriminations Produced by Policy Instruments

Adentrarse en la categorización: cómo objetivar las discriminaciones producidas por los instrumentos de gobierno

Pablo Cussac García*

Fecha de recepción: 10/11/2024 Fecha de aceptación: 16/01/2025

Policy instruments reproduce political and social inequalities. While different sociological objectification techniques allow us to approach the production of these discriminations, descriptive statistics, ethnographic observation, and semi-structured interviews also present significant limitations. This article addresses the methodological challenges posed by policy instruments and proposes an analytical framework based on the intersection of these three sources. Drawing from the experience of my doctoral thesis on teacher evaluation, I outline pathways for researchers to objectify the categorization process of policy instruments that rely to varying degrees on the quantification of individuals and their practices.

Keywords: categorization, instruments, observation, descriptive statistics, semi-structured interviews

Los instrumentos de gobierno son vectores de la reproducción de las desigualdades políticas y sociales. Si diferentes técnicas de objetivación sociológica permiten acercarse a la producción de dichas discriminaciones, la estadística descriptiva, la observación etnográfica o las entrevistas semi-estructuradas presentan también límites importantes. Este artículo aborda los desafíos metodológicos planteados por los instrumentos de gobierno y propone un marco de análisis basado en el cruce de estas tres fuentes. A partir de la experiencia de mi tesis doctoral sobre la evaluación docente, presento los caminos que el investigador/a puede tomar para objetivar los procesos de categorización de los dispositivos de gobierno que reposan en mayor o menor medida sobre la cuantificación de los individuos y sus prácticas.

***Palabras clave**: categorización, instrumentos, observación, estadística descriptiva, entrevistas semi-estructuradas*

* Centro de estudios y de investigaciones administrativas, políticas y sociales (CERAPS) de la Universidad de Lille

> I have a very clear memory of the portfolio of a teacher from the state of Guerrero. In the first lines, she explained, as best as she could, that she did not speak Spanish but an indigenous language and that she worked in a community with no teachers. That, indeed, she was no teacher but only covering that function. […] Even the writing… It was as if her keyboard had decomposed… You couldn't understand a thing; there were no separations between the words. It was as if a child had written it. (interview, teacher-evaluator)

Beneath this anecdotal excerpt from an interview with a Mexican teacher-evaluator appears to lie to a form of racism in the ranking of teachers' performance. Ultimately categorized as "insufficient" by the national evaluation instrument *Teacher's Professional Service*, this indigenous teacher from Guerrero is explicitly infantilized. Policy instruments such as this are often criticized—both publicly and academically—for being discriminatory, biased, and reproducing structural inequalities. But how can we precisely identify and characterize these discriminations? How do we generalize from usually anecdotal materials? This paper addresses these issues by advocating for the crossing of different qualitative and quantitative techniques. It draws on my doctoral dissertation, where I studied the workings of the Mexican *Teacher's Professional Service* through ethnographic observation and 78 interviews with different actors involved with the instrument, ranging from high-level bureaucrats to rural teachers.

The proliferation of actuarial and neo-managerial techniques in the public sector over the last five decades has been accompanied by abundant analyses of their social and political effects, as well as of the processes, sub-processes, and mechanisms through which these seemingly neutral tools become the vehicle of renewed forms of discrimination and inequality. Without delving into the specifics of these social processes, it is important to note that by relying on the quantification of individual performances rather than on the "nominal" differences between individuals (e.g., class, gender, race) as assessed by expert judgment, tools such as performance indicators, rankings, or benchmarks produce and legitimize new modes of classification, valuation, and hierarchization aligned with neoliberal ideas of individual merit and performance (Fourcade, 2016). Despite these myriad reflections, few studies have offered precise methodological frameworks for objectifying the (re)production of discriminations by policy instruments. One of the main challenges in such research is that contemporary discriminations often emerge from "infracategorical judgments"—stereotypes or cues that rarely align with the nominal categories used by public statistics and policy instruments (Monk, 2022). As a result, even if we can identify how a racial group, for example, is discriminated against by an evaluation or algorithm, it is difficult to pinpoint the precise micro- and meso-level processes that produce such discrimination. Indeed, quantification instruments often appear as "black boxes" built on complex architectures of data and categories that automate and blur the steps of judgment and the responsibility of the various actors and techniques involved in its production (Fourcade & Healy, 2017). Consequently, the question is not if policy instruments are discriminative, but rather *how and through whose actions* they become so.

Several empirical studies have opened pathways for studying the production of discrimination by policy instruments. This literature takes inspiration from the "turn to

practice" initiated by science and technology scholars, which invites us to inquire about the micro-processes of interaction where instruments are applied and decisions are made (Camic, Gross & Lamont, 2011; Lamont, 2012). To study such practices, researchers have proposed different methodologies, including archival work (Hirschman & Bosk, 2020; Krippner & Hirschman, 2022), interviews and ethnographic observations (Brayne & Christin, 2021; Rosen et al., 2021), or quantitative techniques ranging from surveys to quasi-experimental methods (Tilcsik, 2021). Each of these approaches presents strengths and limitations. Quantitative research has precisely identified the causal mechanisms of discrimination, such as the components of the "architecture" of evaluation that can lead to discriminative outcomes (Rivera & Tilcsik, 2019). If these techniques can capture causality, it is at the expense of an analysis of the processes of classification and meaning-making that are key in producing inequalities—such as stigmatization, evaluation, or standardization (Lamont et al., 2014). Ethnographic work is better suited for capturing the practices, grammars, stereotypes, and representations that underlie these processes (Lamont, 2009). However, the exclusive observation of practices can make us neglect the instrument's broader institutional and organizational context, which can be better made explicit through interviews and archival materials. Qualitative scholars usually combine some or all of these materials to approach categorization and discrimination. But, if their respective merits are well-known, the trade-offs associated with building their complementarity have been under-theorized. Building on a "pluralistic" approach to methods (Lamont and Swidler, 2014), in this paper I propose a framework that combines three different sources—observations, interviews, and descriptive statistics—to identify the discriminations produced by the evaluation of teacher performance in Mexico.

The *Teacher's Professional Service* operated in Mexico from 2013 to 2019. This evaluation system, mandatory for all tenured teachers, defined national standards of "good teaching." Teachers who achieved performance levels categorized as "good" or "excellent" were eligible for salary bonuses, while those rated as "insufficient" faced the risk of losing tenure. The system also included mandatory competitive examinations for entry into the profession and promotion to middle-management positions, such as director or supervisor. All these evaluations threatened the institutionalized power of the corporatist teachers' union, the National Union of Education Workers, which had wielded significant influence over educational administration since the 1940s (Ornelas, 1995; 2019). As a result, the instrument was highly controversial and faced massive contestation, especially from the dissident sections of the union, the National Coordination of Education Workers (CNTE). These contested both the redefinition of corporatism by evaluation and the potentially discriminatory and decontextualizing effects of a standardized instrument. Beyond its political implications, the *Teacher's Professional Service* offers an opportunity to examine the social processes underpinning contemporary policy instruments. While the system included several instances of quantification of teacher practices—such as a standardized test of teacher knowledge (40% of the final result) or the quantification of the different "tasks" evaluated by a teaching portfolio (60% of the result)—it ultimately classified individual teachers into one of four performance categories: insufficient, sufficient, good, and excellent.

It did not rank teachers on a comparative table. As a result, two teachers could receive the same final result and face identical consequences, yet for entirely different reasons. Results stemmed not only from the quantification of teaching performance but also from the expert judgment of teacher-evaluators, the design of all evaluation procedures by the personnel of a private evaluation center *(Centro Nacional de Evaluación de la Educación Superior*, Ceneval), and the regulations of a public autonomous agency (*Instituto Nacional para la Evaluación de la Educación*, INEE). Who evaluates teachers in this complex *dispositif*? How are teachers evaluated? In what instances can we identify the eventual discrimination of teachers?

The deep politicization of teacher evaluation raised issues of accessibility to the field during my two stays in Mexico. I chose to approach the instrument through the bureaucracies implementing it. Taking this step back from political debates and actors allowed me to comprehensively question the *Teacher's Professional Service*. As they faced strong opposition, these organizations were eager to grant me (a white, European graduate student) access, as they considered me sympathetic to their rationalization attempts. Nevertheless, I soon realized that their eagerness to justify their policy contrasted with their reluctance to allow me to see the actual evaluation instruments and practices. I was only able to access them after one of my interviewees in the Mexican Education Department invited me to the grading process of the *Teacher's Professional Service*. This informal contact allowed me to bypass the leadership of Ceneval, even if I had to sign a non-disclosure agreement concerning the evaluation tools (see below). The observation of the grading of portfolios and standardized tests over a week in 2018 represents my first entry point into the workings of the instrument. Although this confidential process provided invaluable insights, it only partially illuminated the issues of categorization and discrimination. To address the questions that emerged from this first fieldwork, I conducted 13 in-depth interviews with teacher-evaluators in 2019. These interviews allowed me to examine the observed evaluation practices more closely and to better understand the representations underlying the judgment of evaluators. Finally, in 2023, during the writing of my dissertation, I analyzed the aggregated results of the evaluation to compare the qualitative evidence gathered with the larger picture of this national evaluation system. In the following three sections, I outline each of these steps, explain how they complement each other, and describe how each respectively contributed to my understanding of the categorization and discrimination of Mexican teachers under the *Teacher's Professional Service*.

The Construction of Categories: Observing Evaluation Practices

To investigate the workings of the Mexican evaluation system, I first observed the grading of the two main evaluation tools, the portfolio and the standardized test. This process involves teacher-evaluators—teachers who have already been evaluated with good results and have been certified to evaluate their colleagues—as well as tools such as an evaluation grid and statistical discrimination techniques operated by the personnel of Ceneval. This is a confidential process, and I was only granted access to observe it after signing a confidentiality agreement with the institution, which prevents me from disclosing the specific content of the evaluation grid and other evaluation methods. However, beyond the content of these tools,

my primary interest was in understanding how the various actors involved used them and how their evaluation practices contributed to the construction of the categories and groups that emerge from the *Teacher's Professional Service*. What should one observe in a process where hundreds of teacher-evaluators largely work on their computers in silence? How can one identify social practices and representations in an almost automatized process? How to observe this kind of process without disturbing it, and thus potentially affecting the work chances of teachers? To address these challenges, I adopted two strategies.

The first strategy involved focusing on moments of "trial," where the rules, criteria, and even the legitimacy of the evaluation process became explicit and open to discussion. One such instance of "trial" is the grading of the standardized test. Unlike the grading of portfolios, which includes hundreds of evaluators, this process is more intimate, lasting two days and involving only three teacher-evaluators, a coordinator, and a Ceneval statistician who intervenes intermittently. This setup facilitates the observation of evaluation practices. During a full day, the evaluators discuss the relative difficulty of each of the 120 questions on the test. To do this, they must imagine four "hypothetical candidates," one for each performance category, and assess the probability that each would answer each question correctly. Surprisingly, this technique does not lead to many discussions, and it seems as though the evaluators shared these avatars.[1] The moment of trial occurs on the second day, at the end of the grading process, when the evaluators and the statistician determine the cut-points between each performance category. These cut-points have to be derived from the difficulty of the exam, as assessed by the evaluators the previous day, rather than from a theoretical or *a priori* definition of how many teachers are "insufficient" or how many correct answers are necessary to be "excellent."

The final cut-points emerge from the collective work of evaluators and the statistician. While the statistician wants to produce a "normalized" distribution of results (one that places few teachers in the highest and lowest categories and most in the middle), evaluators seek to avoid a situation where many of their colleagues end up in the lowest category ("insufficient"), which involves potential sanctions. These positions are not contradictory. In the grading process I observed, the initial cut-points resulted in 14% of teachers classified as insufficient, 42% as sufficient, 34% as good, and 10% as excellent. This distribution was deemed too "spread out" by the statistician, and the evaluators agreed that 14% was an excessive number of "insufficients." They thus revised the difficulty estimates they had made the previous day, which led to a second distribution curve: 8% insufficient, 46% sufficient, 35% good, and 12% excellent. This new result "saved" an important portion of the previously insufficient teachers but was even more "spread out". The Ceneval statistician thus encouraged the evaluators to revise their expectations and increase the number of insufficient teachers, which should lower the ranks of the "sufficient" category—the highest category since the beginning. Following this advice, the evaluators adjusted their estimates again, arriving at a final, almost perfect normal distribution: 12% insufficient, 39% sufficient, 37% sufficient, and 12% excellent. The "normalization" goal of the *dispositif* and its embedding in different statistical and professional objectives thus became apparent. This

1 I questioned this observation during the interviews (see second section).

process reflects the dual nature of discrimination both as statistical description (the normal distribution) and as social prescription (sanctioning or "saving" those deemed insufficient).

However, the observation of evaluation practices alone did not allow me to fully grasp the social and professional meaning of the evaluation categories and the statistical manipulations involved in their construction. As a result, I adopted a second strategy during my observations: engaging in informal situations and conversations. During lunch, coffee breaks, or even after-work drinks, I could ask evaluators to reflect on the work they were doing. These moments also provided space for anecdotes from previous years. For instance, some evaluators recalled how some of their colleagues had been excluded from the procedure after grading all portfolios as "excellent"—a practice identified and sanctioned by the internal auditing process. They also shared their strategies to avoid audits, such as grading the different components of portfolios as "sufficient" and "good" (audits are triggered when two different evaluators grade the same components of the same portfolio as "insufficient" and "excellent"). When having a drink with the evaluators involved in the grading of the standardized test described above, I could ask them if the normalization of results was a common practice. They acknowledged this practice while expressing their discomfort, as many of the test questions are discarded to construct the distribution (those questions answered correctly or incorrectly by more than 90% of teachers, and thus deemed ineffective for discriminating their performance). After one of the evaluators raised this issue, another answered, "Yes, yet you validated the procedure when you signed at the end and took your check." These informal discussions are thus an occasion to contextualize the observed practices, their meaning, the strategies of the actors involved, and the role they perceive themselves as playing in the evaluation process. Nevertheless, the social meaning of the categories remained elusive.

Grasping the Meaning of Categories Through Interviews

While the ethnographic observation of evaluation practices seems indispensable when studying policy instruments such as the *Teacher's Professional Service*, I found that understanding the processes of categorization required in-depth conversations with teacher-evaluators. These interviews provided an opportunity to reconstruct and compare their representations, their grammars of evaluation, and the ideas of merit and justice on which their work was based. To avoid common issues with semi-structured interviews—such as the forgetting of the precise practices that actors realized in the past or the *a posteriori* illusion of coherence in practices and representations—I could use my observations to resituate my interviewees (some of whom I had observed) in specific grading situations and ask them to walk me through their practices and thought-processes.

Returning to the construction of "hypothetical candidates" proved particularly useful in understanding the discriminations involved in the evaluators' practices. Evaluators would explain that they use various techniques to imagine these avatars, ranging from the personalization of the "hypothetical candidates," where relatives or friends are made the representatives of a particular category, to modes of abstraction and generalization where specific professional sub-groups are invoked as "insufficient." The insufficient category was

typically the easiest for them to imagine and talk about, as was the "excellent" category, where they always placed themselves. As one evaluator put it, "You'll say I have a great ego, but I imagined myself in level 4 [excellent]." This is not surprising, as reflecting on both the top and the bottom of the new evaluation system allowed them to construct its symbolic coherence and discriminate and distribute performances in it.

As with other forms of contemporary discrimination, evaluators did not rely on categorical differentiation between their colleagues but rather on "infracategorical discriminations" (Monk, 2023). This practice was facilitated by the fact that, aside from the written texts and the few photographs included in the portfolio, evaluators did not have access to any personal information (such as the names) of the teachers being evaluated. Moreover, the interview setting made them remember only portfolios that stood out, either for their quality or lack thereof—contrary to the observation of the continuous flow of grading, where most portfolios are "sufficient" or "good." This led them to reconstruct the evaluated person and ascribe her identity and professional situation. For instance, instead of explicitly referring to the racial or indigenous identity of an "insufficient" teacher, evaluators would say that the text was "as if written by a child" (as in the excerpt at the beginning of this paper) or that it was a teacher "from *la sierra*" (the mountains). Another common form of discrimination involving personal characteristics was the association between old age and "insufficiency." One evaluator recalls thinking about "that old teacher that knows a lot, that reads a lot, why doesn't she make an effort? […] But the fear, the tension, are stronger than her…" In other cases, reflecting on who qualifies as an "insufficient" teacher allowed evaluators to activate intra-professional cleavages. For example, one evaluator thought about the "insufficient" teacher as "teacher Óscar", an actual teacher from her school: "He was in an administrative position and when he came back, he was lost." I encountered similar remarks in other interviews, where evaluators referred to the "lazy teacher from the school" or "those that have been teaching for more than ten years and don't even have their degree." These cues associated insufficiency with union militancy, echoing the scandals of the (illegal) selling or the (legal) inheriting of teacher positions or the union's political patronage, including placing phantom employees in educational administrations. Insufficiency was also a way of discriminating against teachers with alternative educational backgrounds, such as those who graduated from universities rather than from the normal schools of teacher training, that many consider to "lack vocation," to teach "while they look for another job," and to have "insufficient interest in didactics and pedagogy." These snapshots bear witness to how interviews allowed me to access the categorization processes evaluators follow to assess and classify their colleagues. In this process, pre-existing social and professional identities, categories, and cleavages are activated. I could not have grasped this mechanism through observation alone, as reflecting and making explicit the thought-processes underlying the categorization of individuals requires an intimate and time-consuming setting that only the interview situation—conducted in the evaluators' homes or quiet coffeeshops—can provide. Interviews thus reinforce and complement the materials produced via observation, allowing to better objectify discriminations without overgeneralizing from anecdotal

evidence. These materials were especially crucial given that I was granted access to the grading process for only a week (40 hours) and could not be certain about the generalizability of my observations.

From Categories to Groups: Combining Qualitative and Quantitative Evidence

The problem of generalization is a recurrent challenge in qualitative studies. One potential solution is the combination of different sources (such as observations and interviews) to triangulate information and minimize biases. In this third and final section, I extend this combination of sources to include quantitative materials, specifically descriptive statistics. These are particularly important when studying the discriminations produced by policy instruments such as Mexico's *Teacher's Professional Service*. As is often the case, its results are public, which allowed me to test whether the discriminations identified during the categorization process translate into group differentiation and hierarchization within the profession. I did not excessively extrapolate from the reading of descriptive statistics, and I adopted a reflexive approach to these materials. Indeed, these statistics do not always rely on the categorical and infracategorical differences that would allow me to test the qualitative evidence gathered via observations and interviews. While some variables, such as age and gender, are well reflected, others, such as degree or unionization, are not (for personal confidentiality reasons), which led me to rely on complementary statistical sources. This cross-fertilization of qualitative and quantitative materials allowed me to infer correlations with confidence but not causality—a point developed below.

To test the correlation between age and the categorization as insufficient, I relied on the public presentation of the first performance evaluation results in the Mexican Senate (SEP, 2016). This analysis showed that among the 19,679 teachers categorized as "insufficient" in 2015, the oldest segment of the profession was overrepresented. On average, 15% of male teachers had insufficient results, a rate that doubled for the 55-60 age group (31%) and that went up to 35% for those over 60 years old. While women generally achieved better results (11% categorized as insufficient), the proportions remained similar. Among women aged 55-60, 22% were categorized as insufficient, rising to 28% for those over 60. In light of my previous qualitative analysis, I could understand the processes that led to this overrepresentation of old teachers as insufficient. Not only did they probably experience problems with an online evaluation that reflected on their portfolios, but they were also often considered less "effortful" and more "fearful" by evaluators.

By crossing evaluation results with measures of poverty (CONEVAL, 2020), I could indirectly interrogate the correlation between race, geographical location, and "insufficiency." Or, in other words, verify if teachers from "*la sierra*" were discriminated against by the evaluation standards. Indeed, teachers categorized as insufficient primarily come from the poorest regions of the country. Five out of the six states whose rate of insufficient teachers at least doubles the national average (12.43%) also present higher poverty rates than the national average (41.9%). These states are Zacatecas (29% insufficient teachers), Sonora

(31%), Oaxaca (31%), Guerrero (31%), Michoacan (35%), and Chiapas (44%).[2] Except for Sonora, these regions have high poverty rates, with up to 66% of the population in Guerrero and 76% in Chiapas living in poverty. These states also share an important characteristic: they are all rural, low-density population areas, which probably hinders their teachers' opportunities for professional socialization and preparation of the evaluations. Their situation contrasts with that of the states achieving the best evaluation results (half or less the national average of insufficient teachers). Despite their many differences, all these states present poverty rates below the national average, ranging from 40.8% (Yucatan) to 18.1% (Baja California Sur). Although less straightforward, these results allowed me to objectify the intuition stemming from my interviews and observations: many insufficient teachers come from poor, rural, and sometimes indigenous areas.

Intra-professional cleavages were harder to test. As with the race and ethnicity of the teachers and their students, variables such as unionization or academic background were not included in the individual (and anonymous) results made public by the administration. I thus used two sources to interrogate the new forms of intra-professional competition: results in the new competitive examinations to access the teaching profession (Cordero Arroyo & Jimenez Moreno, 2018) and enrollment data in normal schools for teacher training (*normales*) (INEE, 2017). If students from *normales* achieved better results in three of the four competitive exams held between 2014 and 2018, the proportion reversed in 2016-2017, when the percentage of university graduates eligible for a teaching position reached 68%, compared to 50% for students from *normales*. Furthermore, enrollment in these schools has plummeted since the introduction of the *Teacher's Professional Service*. Private *normales* have lost 45% of their students, becoming almost irrelevant. The much more historically important public *normales* have also lost around 21% of their students. These points to a recomposing of professional values and valorization criteria, with university graduates considered as having "less vocation" by *normalistas*, and probably less tied to the teachers' union, becoming increasingly important in the ranks of the profession.

This quantitative evidence allows me to reinforce the objectivation of the modes of discrimination produced and reproduced by the *Teacher's Professional Service*. To be sure, I do not identify causal mechanisms but rather point at strong correlations to confirm the generalization of the qualitative evidence that represented the heart of my doctoral research. If multivariate analysis could be conducted on this data, my goal was not to identify the exact effect of each variable but rather to show the multiple meanings and the categorization processes that lay behind the "insufficient" result. Echoing recent results in the sociology of instruments (Krippner & Hirschman, 2022), I thus show that it is not a single group or category that is considered insufficient, but rather that different individuals, reduced to a series of attributes, can be categorized as such for various reasons (but with the same implications). This, in turn, hinders the

2 One should note that the figures for Oaxaca, Guerrero, Michoacan, and Chiapas are magnified by the boycott of evaluation in these states by the dissident sections of the union. Teachers who did not present their evaluation materials were automatically categorized as "insufficient". When counting those "not presenting evaluation" as a distinct category, the "insufficient" results of these States lower substantially (below the national average in Oaxaca and Michoacán, with only a third to a quarter of summoned teachers present on the evaluation day), although they remain high in Guerrero (16%) and Chiapas (28%), the two poorest states.

potential identification of these individuals with the category and, thus, their potential political mobilization.

Discussion

In this paper, I have proposed a methodological framework to study the production of discrimination by policy instruments. The *Teacher's Professional Service* raises specific questions regarding the conceptualization of categorization, hierarchization, and discrimination processes, as well as about data availability and limitations. I adopted a strategy combining observations, in-depth interviews, and descriptive statistics to study this instrument. These materials complement and extend each other. Focusing on "moments of trial" during the observation of practices of evaluation allowed me to identify the rationale of the instrument and the role that various actors and tools play during the grading process. These elements enabled me, during a second fieldwork, to refer to specific moments and practices in my interviews and to objectify the representations and discriminations (both categorical and infracategorical) used by evaluators while assessing their colleagues. Finally, I scaled up the analysis through descriptive statistics to test if these group discriminations could be observed at the aggregate level. In doing so, I could escape the anxiety often felt by qualitative scholars when attempting to generalize their findings, and I was able to confirm both the existence of evaluation-induced discriminations and the social processes leading to their production.

Of course, this proposal has limitations and should be further developed and adapted to address the specificities of other policy instruments. One important limitation is the absence of causal identification in my quantitative analysis, which could be explored in further studies. Another limitation is that I did not include the "phenomenological dimension" (Fourcade & Healy, 2017) of evaluation in my inquiry into the production of discrimination by the *Teacher's Professional Service*. What does it mean to be insufficient or excellent? How does it affect teacher subjectivities and daily work? Only more systematic ethnographic work in schools could address these questions. Beyond the materials presented in this paper, like most studies of instruments, I also analyzed official documentation in my dissertation (e.g., technical documents, guidelines for evaluators, and examples of portfolios and evaluation grids), which may deserve a methodological reflection regarding its status as archival material and its role in the production of discriminations.

References

Brayne, S., & Christin, A. (2021). Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts. *Social Problems*, *68*(3), 608–624. https://doi.org/10.1093/socpro/spaa004

Camic, C., Gross, N., & Lamont, M. (2011), *Social Knowledge in the Making*. University of Chicago Press

CONEVAL. (2020). *Informes de pobreza y evaluación de las entidades federativas*. CONEVAL.

Cordero Arroyo, G., & Jimenez Moreno, J. A. (2018). La política de ingreso a la carrera docente en México: Resultados de una supuesta idoneidad. *Education Policy Analysis Archives*, *26*(5), 1–27. https://doi.org/10.14507/epaa.26.3463

Fourcade, M. (2016). Ordinalization: Lewis A. Coser Memorial Award for Theoretical Agenda Setting 2014. *Sociological Theory*, *34*(3), 175–195. https://doi.org/10.1177/0735275116665876

Fourcade, M., & Healy, K. (2017). Categories All the Way Down. *Historical Social Research / Historische Sozialforschung*, *42*(1 (159)), 286–296. https://www.jstor.org/stable/44176033

Hirschman, D., & Bosk, E. A. (2020). Standardizing Biases: Selection Devices and the Quantification of Race. *Sociology of Race and Ethnicity*, *6*(3), 348–364. https://doi.org/10.1177/2332649219844797

INEE. (2017). *La educación normal en México. Elementos para su análisis*. INEE.

Krippner, G. R., & Hirschman, D. (2022). The Person of the Category: The Pricing of Risk and the Politics of Classification in Insurance and Credit. *Theory and Society*, *51*, 685–727. https://doi.org/10.1007/s11186-022-09500-5

Lamont, M. (2009). *How Professors Think: Inside the Curious World of Academic Judgment*. Harvard University Press.

Lamont, M. (2012). Toward a Comparative Sociology of Valuation and Evaluation. *Annual Review of Sociology*, *38*(1), 201–221. https://doi.org/10.1146/annurev-soc-070308-120022

Lamont, M., Beljean, S., & Clair, M. (2014). What is Missing? Cultural Processes and Causal Pathways to Inequality. *Socio-Economic Review*, *12*(3), 573–608. https://doi.org/10.1093/ser/mwu011

Lamont, M. & Swidler, A. (2014), Methodological Pluralism and the Possibilities and Limits of Interviewing. *Qualitative Sociology*, *37(2)*, p. 153-171. https://doi.org/10.1007/s11133-014-9274-z

Monk, E. P. (2022). Inequality without Groups: Contemporary Theories of Categories, Intersectional Typicality, and the Disaggregation of Difference. *Sociological Theory*, *40*(1), 3–27. https://doi.org/10.1177/07352751221076863

Ornelas, C. (1995). *El sistema educativo mexicano: La transición de fin de siglo*. Fondo de Cultura Económica.

Ornelas, C. (2019). *La contienda por la educación: Globalización, neocorporativismo y democracia*. Fondo de Cultura Economica.

Rivera, L. A., & Tilcsik, A. (2019). Scaling Down Inequality: Rating Scales, Gender Bias, and the Architecture of Evaluation. *American Sociological Review*, *84*(2), 248–274. https://doi.org/10.1177/0003122419833601

Rosen, E., Garboden, P. M. E., & Cossyleon, J. E. (2021). Racial Discrimination in Housing: How Landlords Use Algorithms and Home Visits to Screen Tenants. *American Sociological Review*, *86*(5), 787–822. https://doi.org/10.1177/00031224211029618

SEP. (2016). *Evaluación del Desempeño, 2015-2016: Resultados Finales y Hallazgos Preliminares*. Subsecretaría de Planeación, Evaluación y Coordinación. Presentation at the Senate's education commission.

Tilcsik, A. (2021). Statistical Discrimination and the Rationalization of Stereotypes. *American Sociological Review*, *86*(1), 93–122. https://doi.org/10.1177/0003122420969399