Revising evaluation metrics for graduate admissions and faculty advancement to dismantle privilege

Academics are not immune to the biases contributing to persistent inequalities in society. We face an urgent need to overhaul and dismantle current evaluation practices that uphold inequities at multiple points along the academic pipeline. Graduate admissions and faculty advancement are two arenas of gatekeeping in which a reimagining and redistribution of weighting of commonly used evaluation metrics are warranted. We define and promote the use of dynamic, flexible holistic evaluation models that can be implemented by first recognizing and acknowledging the biases that contribute to racial and ethnic disparities in academia. Leaders of academic institutions must step up to drive adoption of these revised evaluation metrics.

Andres De Los Reyes and Lucina Q. Uddin

Academic evaluation practices and their shortcomings

The COVID-19 pandemic and the racial justice movement in 2020–2021 have placed societal power structures in stark relief against historical inequalities in access to institutions that facilitate upward mobility. Academic institutions dictate the career opportunities of key stakeholders (e.g., students, faculty, and staff). Thus, these institutions maintain a relatively outsized positioning within economic and social hierarchies that are central to day-to-day life and well-being. The time is ripe to examine practices by the academy that have long disadvantaged marginalized communities. The academy should not and cannot be immune to introspection about its own policies, practices, and power structures. Two sets of power structures require immediate attention, due to their current and continued impact on the academic pipeline: systems regarding admission of students into doctoral programs and criteria for the hiring, tenure, and promotion of faculty. Here we seek to contribute to an emerging dialogue on evaluation metrics for these high-stakes decisions in higher education.

The Graduate Record Examination (GRE), a standardized test historically widely used in graduate admissions decisions in the United States and Canada, is biased toward students from higher socioeconomic status and does not predict completion of the doctorate as effectively as presumed[1]. Issues of access have been at the forefront of discussions surrounding the value of the GRE as a tool for informing graduate admissions[2]. The #GRExit Twitter campaign reveals that graduate programs have been dropping the GRE as a requirement in recent years. Conversations about the equity of the online version of the test necessitated by the COVID-19 pandemic surround the concern that the online testing requirements may further disadvantage prospective students from rural and low-income backgrounds[3]. Even if GRE scores were found to reliably index academic career success, the vast majority of graduate students do not pursue careers in academia[4], further limiting the utility of this examination for predicting future employment and earning potential.

In addition to the GRE, factors traditionally considered in graduate admissions include grade point average (GPA), personal statements, resumes or curriculum vitae, letters of recommendation, and in-person or virtual interviews. Each of these forms of assessment is subject to its own issues surrounding bias and fairness, particularly those involving subjective evaluations. Empirical evidence suggests that undergraduate GPA, while correlated with graduate comprehensive exam scores and faculty ratings of graduate student performance, is not a very strong predictor of degree attainment or time to completion[5]. Personal statements and letters of recommendation are likewise limited in their ability to predict future academic performance, and they are vulnerable to a number of sociocognitive and rater biases[6]. Most surprisingly, previous research experience also appears to be unrelated to other predictors used in graduate admissions, as well as academic performance in graduate school, although this warrants further investigation, as only a small number of studies have been conducted on this topic[7]. It is important to also keep in mind that barriers to involvement in unpaid research factor into who has access to early research experiences[8]. This emerging literature suggests that many sources of information used in the graduate admission process are subject to bias and are not as objective and fair as widely assumed.

Faculty hiring, promotion, and tenure decisions also reflect policies and practices that create structural disadvantages for those who identify with historically marginalized groups. In particular, these decisions are often biased toward indicators of scholarly productivity with clear underlying power structures linked to them. For instance, hiring and promotion decisions are often biased toward individuals whose research can be framed within mainstream scholarly topics favored by senior faculty from majority backgrounds[9,10]. The recruitment of faculty of color who study topics related to diversity, equity, and inclusion (e.g., implicit race bias or health disparities in communities of color) is often relegated to specialty hiring initiatives. These initiatives often have little staying power or long-term funding to support repeated hiring efforts.

We argue that these policies and practices along the academic pipeline collectively create gaping holes, and thus we should not be surprised with the result: academia loses to other industries scores of talented individuals from historically marginalized backgrounds who could otherwise infuse

into our scholarly discourse vibrant, progressive, forward-thinking lines of work. Consider, for example, that among grant applicants at the National Institutes of Health (NIH), Black investigators achieved lower funding success rates than their White applicant counterparts; this disparity was in part accounted for by differences in the specific aims of work proposed by Black applicants, which were often focused on topics involving research at the community level[11].

We fear that a variety of institutional structures in academia have created the very environments academics often claim to be against; that is, where bases of knowledge fail to include sufficient diversity of thought. But there are pathways for developing new evaluative systems to reduce disparities in access to opportunities to flourish in the academy. In this commentary, we propose a rebalancing of key considerations across the academic pipeline, focusing on graduate admissions and on the faculty hiring, tenure, and promotion processes. By applying well-established principles of scholarly discourse to all our evaluative structures, we can fundamentally alter our evaluative processes. We close with concrete recommendations for academics to adopt more dynamic holistic approaches to evaluation to redress historical inequalities in academia.

Rationale and recommendations for implementing truly holistic evaluations

Academia is not a meritocracy. To reduce the influence of systemic inequalities on important decisions regarding graduate admissions and faculty advancement, we must first disabuse ourselves of the notion that academia is a pure meritocracy. Recent work suggests that above a certain threshold of 'applicant metrics', the benchmarks traditionally used to measure research success—including funding, number of publications, or the 'quality' of the journals in which a candidate's publications appear—are unable to completely differentiate applicants with and without faculty job offers[12]. This finding highlights the fact that the faculty hiring processes—and indeed all other types of admissions and gatekeeping practices—are necessarily subjective, as the market is flooded with more qualified candidates than there are available positions at every point along the academic pipeline. A survey of credentials that faculty hiring committees look for in candidates lists "scientific/programmatic and general fit" as one of the most valued attributes[13]. These 'fit' criteria are highly subjective and difficult, if not impossible, to quantify.

Need for the recognition and acknowledgement of biases. Indicators of scholarly productivity that are widely used by graduate admissions committees and committees for faculty hiring and promotion are biased, with clear underlying power structures linked to them[14]. These power structures confound our interpretation of scholarly work. The impact of this on graduate admissions is described above. Related to faculty evaluations, in Psychology, the power structures at mainstream academic journals often do not include board members with expertise on scholarly topics shared by faculty from historically marginalized backgrounds[9]. From an historical context, the epistemological perspectives that have governed much of the functioning of our academic spaces have their origins in White European Enlightenment. A key element of this perspective involves its presumed neutrality with regard to the racial and/or ethnic backgrounds of scholars and thus the lack of impact of these backgrounds on evaluations of scholarly merit. Yet, the very presumption that our colleagues' backgrounds have no impact on how we view their work, or that our backgrounds confer no impact on the conduct of our research, may further the status quo when it comes to faculty hiring and promotion. Thus, it has been argued that if universities truly wish to diversify their faculties, they need to move beyond conventional hiring criteria[15]. This will be increasingly imperative to address disparities in academia that will no doubt be further exacerbated by the current pandemic[16], which has further revealed stark disparities and inequities for individuals who identify as Black, Indigenous, or people of color (BIPOC)[17].

Much has been written about how citation counts and journal impact factors are limited in their ability to capture research quality and how producing and incentivizing research that is open, transparent, and reproducible should instead be the goal of science[18]. Still, the unspoken rules of tenure and promotion at most research-intensive universities emphasize the quality and quantity of research, teaching, and service as the only important factors under consideration for evaluation. We argue for a dramatic redistribution and reconsideration of these factors. Note that the evaluation of the quality and quantity of research, teaching, and service need not be abandoned entirely. Rather, they might be weighted differently, leaving room for consideration of additional factors not traditionally evaluated, including commitments to mentoring, community outreach, science communication, and contributions to departmental and institutional diversity, equity, and inclusivity initiatives. As key decision-makers in these contexts, we ought to strongly consider including criteria that might be related to or even embedded within research, teaching, and service, but nonetheless deserve a concrete place in decision-making. These criteria include specific expectations for mentoring, community outreach, science communication, and contributions to diversity and inclusivity. Arguably, these criteria are more subjective than other metrics. However, subjectivity should not preclude their consideration, as we often make important decisions based on subjective criteria (e.g., fit) in academia.

Ideally, unbiased evaluation rubrics would be put in place to replace existing biased ones. Originally (and ironically), GRE and journal impact factors were designed to provide unbiased rubrics, and key elements of these rubrics have the look and feel of indices that lack bias, including standardization and the applicability of the metrics across disciplines and fields. Yet, here too one can point to research indicating that the rubrics we thought were unbiased have turned out to contain biases after all. We previously cited evidence that calls into question such indices as the GRE and publication characteristics (e.g., journal impact factor and citation counts) as unbiased indicators. Consider also evidence of bias within the evaluation metrics regarding scholars' success at receiving grant funding. In US agencies like the NIH, policymakers have spent considerable attention developing seemingly unbiased rubrics to guide the evaluation of grant applications, including quantitative evaluations of proposal significance, study design, and the strength of the investigative team proposed to carry out the work. Nevertheless, recent work indicates that grant reviewers provide ratings based on these 'unbiased' rubrics that not only are unreliable, but are accounted for by factors other than the application's merit, namely reviewer characteristics[19]. Further, as mentioned previously, these same reviews produce racial disparities in funding success rates that appear to be accounted for, in part, by racial disparities among applicants' proposal aims and topics of investigation[11]. In sum, although we agree that producing a system predominated by unbiased metrics is a worthy goal, we have to also acknowledge that previous attempts to do so have encountered challenges, if not failures, in producing metrics that lack bias. Consequently, we contend that holistic approaches both acknowledge the fallibility of individual metrics and create

opportunities for strategically selecting metrics that reduce the likelihood that biases inherent in any one metric unduly influence decision-making.

Towards truly holistic approaches to academic evaluation. If we are to rely less on biased metrics such as the GRE or journal impact factors to make decisions regarding graduate admissions and tenure decisions, respectively, then what criteria should we use instead? One possible path forward would be to move away from over-reliance on these metrics in favor of a holistic approach that more adequately considers each individual scientist's personal experiences and contributions to society. In proposing a redistribution and redefinition of factors, we ought to also consider how we weigh these factors when making decisions regarding graduate school admissions and faculty hiring, tenure, and promotion. That is, should we continue to weigh one or more of these factors more heavily, relative to the others? We want to emphasize that some of the best work we can do in this space is to rid ourselves of the need for uniformity in weighting across graduate programs, departments, and institutions. That is, in all likelihood, any standardized shift we would make now might produce some changes in inequities, but would not be the 'correct' answer for all circumstances.

Towards implementing such a shift, some might argue that we must wait for an evidence base to accrue on proposed changes to the structures we highlighted previously. For instance, should some of our programs propose to pilot possible weighting strategies before beginning discussions like the one we raise in this commentary? We take a two-fold stance on evidence-gathering. First, waiting for evidence on possible alternative strategies, by construction, quashes discussion of these core issues. This is a particularly pressing concern when it comes to the graduate admission and faculty pipelines, because the current structures we rely on for decision-making stand on flimsy evidentiary grounds. Second, by beginning a discussion on alternative structures, we allow ourselves the ability to discuss methods of evidence-gathering for possible solutions to the pressing problems discussed here. Although the call for holistic evaluations is not new per se, it is important to reiterate this notion, as it has not yet been widely and systematically adopted by the academic community. We suggest that there is a need for both an increase in the number of factors considered during evaluation and a redistribution of the weighting of existing factors. In addition, we suggest that institutions should be given flexibility in implementation of this process, to facilitate empirical evaluations of decision-making processes and identification of processes that prove particularly effective in recruiting and retaining academics from diverse backgrounds.

Recent work demonstrates how decision-makers might implement holistic processes and, in doing so, may reduce key disparities. As with other post-undergraduate programs, medical residency programs frequently encounter racial and ethnic disparities in admissions, with a key bottleneck in the process being those invited to interview for residency placements[20]. To address this disparity in one medical residency admissions process, Barcelo and colleagues evaluated interview invite rates in a diverse pool of 547 applicants to a psychiatry residency program. The authors compared various approaches to evaluating applicants for an interview, including a traditional non-holistic model focused on such elements as applicants' standardized test scores and involvement in honors societies. The holistic model focused on a diverse set of criteria along a series of domains that reflected not only scholastic aptitude but also community service, clinical experience, and leadership experience, along with considerations of applicants' personal hardships or barriers they had overcome in the process of making their academic achievements. A third traditional modified model included elements from both the holistic and traditional models and was constructed by the authors to serve as an intermediary between these two models. The holistic model resulted in significant, large increases in invites to under-represented minority applicants (predicted probability of invite = 0.16) relative to the traditional model (predicted probability of invite = 0.08), with relatively little change in the predicted probability across models for non-under-represented applicants. In these analyses, factors such as the holistic model's increased emphasis on lived experiences and de-emphasis on standardized test scores predicted differences between models in invite rates[21].

Our goal in introducing this example is to emphasize that waiting for an evidence base to accrue before changes to academic evaluation practices can be implemented would simply result in a perpetuation of the systems currently in place. In fact, the residency study provides a window into how we might approach reconstructing our decision-making practices. Specifically, the holistic model the authors developed included emphasizing criteria that could traverse evaluation of applications across several disciplines and fields (e.g., leadership, lived experience, community involvement) as well as discipline-specific factors that the model emphasized (e.g., clinical experience) or de-emphasized (scores from a standardized medical exam). Further, the authors both tested the degree to which this model reduced disparities relative to a traditional model and examined factors that could account for any model differences they observed.

Now, imagine if hundreds or thousands of programs, across myriad fields and disciplines, developed their own processes for evaluating applicants for interview invitations, with criteria both common across programs and unique to local program needs. Assuming regular or annual application cycles and thus thousands of 'data points' per cycle, within a short period academia would inherit a rather large 'database' of programs, evaluative models, and outcomes, along with variables to test for prediction of outcomes. With this structural change in how academia approaches this single decision (interview invites for graduate training programs), the variability among evaluative models would allow for scholarly work focused not only on examining whether particular approaches to interview decisions reduce long-standing disparities, but also on predicting important factors going forward, such as disparities in program admissions and the likelihood of positive trainee outcomes, including graduation rates and job placements. Over time, these data would prove critical in identifying evidence-based predictors of trainee and, eventually, faculty performance, thus allowing for refinement in institutional 'best practices' in decision-making as they pertain to evaluations of graduate student and faculty applicants.

In raising the potential for flexibility in holistic models, it is important to nonetheless consider the value of standardization, at least within specific models. For instance, with regards to graduate admissions, it has been suggested that the use of structured and standardized materials (personal statements, interviews, letters of recommendation) may be one way to guard against the kinds of cognitive biases that might contribute to racial disparities[6]. We would emphasize that standardization of any element of an application package along these lines need not necessitate that all programs adopt the same set of criteria. That is, different admission committees may vary in the relative weights they assign to personal statements, interviews, and letters of recommendation. Yet, they should endeavor to uniformly assess these elements in a standard way for each applicant in a given admissions cycle.

Barriers to progress. While we argue here that evaluation practices in academia should adopt more truly holistic approaches, we acknowledge that implementing systematic changes to policy will be a time-consuming and challenging endeavor. Faculty leading such efforts will need to seek and obtain approvals from the relevant governance bodies, and arriving at consensus surrounding the adoption of alternate metrics for evaluation will not be an easy task. As such, leaders of academic institutions, including senior faculty members and administrators, must step up to drive these changes and demonstrate through deeds their commitment to dismantling systems that favor the already-privileged.

Members of search and admissions committees must sincerely believe in the value of holistic evaluations for any proposed changes to hold weight. One way to facilitate this is to promote academic leaders with demonstrated commitments to mentoring, community outreach, science communication, and contributions to diversity and inclusivity. Academic leaders who already espouse these qualities will be in a good position to instigate and advocate for change.

Conducting holistic evaluations will no doubt require that committee members put extra time and effort into their evaluations. For graduate admissions, this may require phone calls and meetings with non-academic references for applicants. For faculty advancement, this will require time spent reading the candidate's publications and becoming familiar with the impact of the scholar's research by going beyond traditional metrics to solicit input from community stakeholders. As academic evaluation practices currently stand, committee members may over-rely on metrics like GRE scores and *h*-indices, not because they believe they are particularly valid, but rather for the simple reason that it is less work to focus on these quantitative indices than other more qualitative measures. The extra time and effort that will necessarily accompany holistic evaluations may need to be more adequately and creatively compensated for the individuals who are asked to complete them. That is, holistic review as a service commitment may need to be allocated a specific, valued space in the day-to-day work lives of key decision-makers in academia (e.g., faculty and administrators).

Concluding remarks

Academia's current decision-making practices linked to graduate admissions and to the recruitment and advancement of faculty perpetuate racial and ethnic disparities. We already have an evidence base indicating that our current practices created these disparities and may be making them worse. We require a new evidence base, one focused on revisions to our approaches and decision-making practices, that takes a holistic view of factors beyond traditional metrics. As disparities exist across disciplines and fields in academia, this should indicate to us that no one model of holistic evaluation will suffice for all decision-making contexts. Where some see challenges in this lack of standardization, we see opportunities. This very challenge may be the one element of structural change that becomes our saving grace.

This moment in history requires structural change focused on three principles. First, we should encourage flexibility in approaches to holistic review. That is, we should allow decision-makers the ability to create holistic models that include elements specific to particular disciplines or institutional programs, as well as elements that cut across disciplines and programs. This variability will facilitate short-term and long-term evaluations of new holistic review models. Second, after some period of time for data collection, we should create task forces or evaluative bodies focused on examining data from holistic review models and determining which factors in these models robustly predict key outcomes, including reductions in disparities at key decision-making moments (e.g., graduate admissions, faculty tenure and promotion). Third, as we develop holistic review models, we should also implement safeguards or explicit components into model development that focus on reviewing whether elements included in these models may introduce additional disparities or exacerbate existing disparities. For instance, our institutions might develop 'disparity review boards' for assessing admissions and hiring criteria, much like current institutional review boards (IRBs) for reviewing research protocols. Here, the emphasis would be on evaluating protocols for holistic review models, with a particular emphasis on whether the models include components for which research suggests potential disparities may arise. If a model includes such a component, the review board might make recommendations regarding monitoring this component for any disparities it creates, much like IRB protocols surrounding safety monitoring for clinical trials research.

We see academia writ large at a crucial inflection point. Much of society is seemingly cognizant of and knowledgeable about long-standing racial and ethnic disparities across its constituent institutional structures. We should assume that much of the world will pay close attention to all of these structures and, in particular, to the actions its leaders take to reduce these disparities. With its sizable influence on economic advancement and life satisfaction, academia will see its fair share of this attention on the global stage, and the time to take bold action is now. ❐

Andres De Los Reyes [1] ✉ and Lucina Q. Uddin [2,3] ✉

[1]*Department of Psychology, University of Maryland, College Park, MD, USA.* [2]*Department of Psychology, University of Miami, Coral Gables, FL, USA.* [3]*Neuroscience Program, University of Miami Miller School of Medicine, Miami, FL, USA.*
✉*e-mail: adlr@umd.edu; l.uddin@miami.edu*

Published online: 30 March 2021
https://doi.org/10.1038/s41593-021-00836-2

References

1. Miller, C. W., Zwickl, B. M., Posselt, J. R., Silvestrini, R. T. & Hodapp, T. *Sci. Adv.* **5**, t7550 (2019).
2. Langin, K. *Science Careers* https://doi.org/10.1126/science.caredit.aay2093 (29 May 2019).
3. Hu, J.C. *Science Careers* https://doi.org/10.1126/science.caredit.abd4989 (24 June 2020).
4. Larson, R. C., Ghaffarzadegan, N. & Xue, Y. *Syst. Res. Behav. Sci.* **31**, 745–750 (2014).
5. Kuncel, N. R., Hezlett, S. A. & Ones, D. S. *Psychol. Bull.* **127**, 162–181 (2001).
6. Woo, S. E., LeBreton, J., Keith, M. & Tay, L. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/w5d7r (2020).
7. Miller, A., Crede, M. & Sotola, L.K. *Int. J. Sel. Assess.* https://doi.org/10.1111/ijsa.12312 (2020).
8. Kim, Y. K. & Sax, L. J. *Res. High. Educ.* **50**, 437–459 (2009).
9. Roberts, S. O., Bareket-Shavit, C., Dollins, F. A., Goldie, P. D. & Mortenson, E. *Perspect. Psychol. Sci.* **15**, 1295–1309 (2020).
10. Clauset, A., Arbesman, S. & Larremore, D. B. *Sci. Adv.* **1**, e1400005 (2015).
11. Hoppe, T. A. et al. *Sci. Adv.* **5**, w7238 (2019).
12. Fernandes, J. D. et al. *eLife* **9**, e54097 (2020).
13. Wright, C. B. & Vanderford, N. L. *Nat. Biotechnol.* **35**, 885–887 (2017).
14. Petersen, A. M. et al. *Proc. Natl. Acad. Sci. USA* **111**, 15316–15321 (2014).
15. Sensoy, Ö. & DiAngelo, R. *'Harv. Educ. Rev.* **87**, 557–580 (2017).
16. *Nature* **591**, 7 (2021).
17. Clark, U. S. & Hurd, Y. L. *Nat. Hum. Behav.* **4**, 774–777 (2020).
18. Dougherty, M. R. & Horne, Z. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/9g5wk (2019).
19. Pier, E. L. et al. *Proc. Natl. Acad. Sci. USA* **115**, 2952–2957 (2018).
20. Talamantes, E., Henderson, M. C., Fancher, T. L. & Mullan, F. *N. Engl. J. Med.* **380**, 803–805 (2019).
21. Barceló, N. E. et al. *Acad. Psychiatry* **45**, 34–42 (2020).

Acknowledgements

A.D.L.R. is supported by the Institute of Education Sciences, U.S. Department of Education, through grant no. R324A180032 to the University of Maryland at College Park. LQU is supported by the National Institutes of Health, U.S. Department of Health and Human Services, through grant no. R01MH107549, and by the Canadian Institute for Advanced Research (L'Institut Canadien de Recherches Avancées). The opinions expressed are those of the authors and do not represent views of these funding agencies.

Competing interests

The authors declare no competing interests.

Additional information

Peer review information *Nature Neuroscience* thanks Melanie Woodin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.