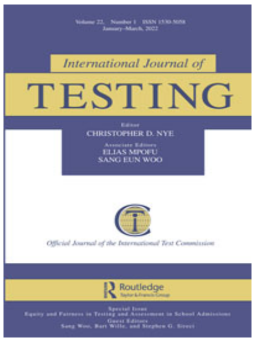

Routledge
Taylor & Francis Group

International Journal of Testing

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hijt20

Is there bias in alternatives to standardized tests? An investigation into letters of recommendation

Dev K. Dalal, Jason G. Randall, Ho Kwan Cheung, Brandon C. Gorman, Sylvia G. Roch & Kevin J. Williams

To cite this article: Dev K. Dalal, Jason G. Randall, Ho Kwan Cheung, Brandon C. Gorman, Sylvia G. Roch & Kevin J. Williams (2022) Is there bias in alternatives to standardized tests? An investigation into letters of recommendation, International Journal of Testing, 22:1, 21-42, DOI: 10.1080/15305058.2021.2019751

To link to this article: https://doi.org/10.1080/15305058.2021.2019751

View supplementary material

Published online: 27 Jan 2022.

Submit your article to this journal

View related articles

View Crossmark data

INTERNATIONAL JOURNAL OF TESTING
2022, VOL. 22, NO. 1, 21–42
https://doi.org/10.1080/15305058.2021.2019751

Routledge
Taylor & Francis Group

Check for updates

Is there bias in alternatives to standardized tests? An investigation into letters of recommendation

Dev K. Dalal, Jason G. Randall, Ho Kwan Cheung, Brandon C. Gorman, Sylvia G. Roch and Kevin J. Williams

University at Albany, State University of New York, Albany, New York

ABSTRACT

Individuals concerned with subgroup differences on standardized tests suggest replacing these tests with holistic evaluations of unstructured application materials, such as letters of recommendation (LORs), which they posit show less bias. We empirically investigate this proposition that LORs are bias-free, and argue that LORs might actually invite systematic, race and gender subgroup differences in the content *and* evaluation of LORs. We text analyzed over 37,000 LORs submitted on behalf of over 10,000 graduate school applicants. Results showed that LOR content does differ across applicants. Furthermore, we see some systematic gender, race, and gender-race intersection differences in LOR content. Content of LORs also systematically differed between degree programs (S.T.E.M. vs. non-S.T.E.M.) and degree sought (doctoral vs. masters). Finally, LOR content alone did not predict an appreciable amount of variance in offers of admission (the first barrier to increasing diversity and inclusion in graduate programs). Our results, combined with past research on LOR content bias, highlight concerns that LORs can be biased against marginalized groups. We conclude with suggestions for reducing potential bias in LOR and for increasing diversity in graduate programs.

KEYWORDS
Admissions decision making; diversity and inclusion; GRE; letters of recommendation; standardized tests

CONTACT Dev K. Dalal ddalal@albany.edu University at Albany, State University of New York, Albany, New York.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/15305058.2021.2019751

Deciding whom to admit into graduate programs is an important choice that carries ethical and legal imperatives necessitating fair and equitable approaches so that all applicants are reviewed impartially (Helms, 2006). Although standardized admissions tests, like the GRE, predict success during and after graduate school (Kuncel et al., 2001), some have expressed concerns that the GRE is biased against female applicants and applicants of color (AOCs), thus contributing to the limited diversity within graduate education (Helms, 2006). Traditionally, a test is considered biased if (1) two test takers, with the same underlying level of the construct being measured, have different observed scores on the same test (i.e., measurement bias), and/or (2) the test scores predict outcomes differently for different groups (i.e., prediction bias; Helms, 2006). Although extant research shows little evidence of prediction or measurement bias on the GRE (Helms, 2006; Kuncel & Hezlett, 2007), concerns remain about the possibility of bias and unfairness resulting from systematic variance attributable to subgroups influencing test scores (Helms, 2006, p. 849). One purported solution to increase representation of marginalized groups in graduate programs is using other, less standardized methods (e.g., letters of recommendation [LORs]; personal statements; unstructured interviews; Sackett & Kuncel, 2018) as part of a holistic evaluations of applicants (Buckley et al., 2018).

An assumption of these other methods, though, is that they are bias-free. For LORs, measurement bias refers to systematic differences in LORs content between groups of individuals, even though actual levels of the attribute are similar. Suggestions that LORs show no bias assume that the content of LORs does not differ systematically across subgroups. We caution, though, that such an assumption needs to be empirically evaluated. Furthermore, as part of a holistic assessment process, LORs may actually introduce bias into the admissions process that systematic, structured, and validated procedures are designed to minimize (Highhouse, 2008). Thus, using admissions systems with biased information may actually work against the goals of increasing diversity in graduate programs.

Here, we conduct a large-scale text analysis of LORs to examine whether there is systematic differences in LOR content (i.e., bias) based on the language used across gender, race, and the intersection of these groups. Furthermore, we explore if LOR content differs among applicants to master's and doctoral degrees, and across degree programs (i.e., S.T.E.M. and non-S.T.E.M.). Finally, we test if LOR content predicts offers of admission, including possible incremental prediction over GRE and undergraduate GPA. We believe this investigation is important not from a desire to discourage the use of LORs; rather, a desire to promote informed selection decisions protecting marginalized applicants from bias that can exist in LORs.

Removing standardization introduces avenues for bias

A growing number of people suggest focusing on information from interviews, personal statements, and LORs to make admissions decisions (Lucido, 2018). Decision makers use these unstructured components of an application to form holistic assessments of the candidates' likelihood of success in a graduate program (Posselt, 2016). However, these alternatives may not necessarily reduce bias in admission decisions, and therein not increasing diversity.

For instance, quality of LORs only modestly correlates with important graduate school outcomes (e.g., GPA; faculty ratings; Kuncel et al., 2014). Moreover, LORs suffer from construct and method confusion (Arthur & Villado, 2008); unstructured LORs purportedly "assess" a variety of constructs such as students' motivation, persistence (Kuncel et al., 2014), personality (e.g., Aamodt et al., 1993; Peres & Garcia, 1962), self-efficacy, and/or creativity (Kyllonen et al., 2005). Table 1 presents a selected list of constructs posited to relate to academic success, measurable via LORs, and highlights the variety of things LORs purport to measure. Such differences may preclude the systematic use of LORs. An additional concern, detailed later, is that the content of LORs may systematically differ across racial and gender groups (Madera et al., 2009, 2019) to describe students whose standing on underlying constructs may be similar.

In addition to unstandardized content of LORs, admissions decision makers advocate evaluating LORs using intuitive judgments. Such approaches allow the decision makers to combine information in an idiosyncratic way (Kuncel et al, 2013). Comparisons of standardized to intuition-based decision making across domains show using standardized methods for collecting and combining information outperforms the use of intuition (Highhouse, 2008; Kuncel et al., 2013). In sum, a greater emphasis on LORs can result in less valid inferences of graduate school performance, which can in turn impact admissions decisions. Perhaps more concerning, increased reliance on unstandardized information in LORs may introduce bias (Dalal et al., 2020) in such a way as to not be identifiable and therein irreducible (Highhouse, 2008).

Relying on intuitive evaluations of unstandardized application materials potentially introduces a trident of adverse outcomes. One adverse outcome is that intuitive evaluations are often less predictive than a standardized approach applied across decision makers and applicants. Research links this to reduced reliability of ratings stemming from intra-decision maker inconsistencies and inter-decision maker inconsistencies (Hastie & Dawes, 2001). A second issue faced when relying on holistic evaluations is that bias is not easily identifiable because intuitive decisions are

Table 1. Selected constructs identified as important for academic success and assessable via letters of recommendation.

Construct	Definition	Example content (if applicable)	Source(s)
Ability/ Creativity[1]	Terms used to convey someone's capacity to perform and identify unique solutions to problems.	Talent*; Intell*; Creat*	Kyllonen et al. (2005) Powers et al. (2020) Schmader et al. (2007)
Critical/Complex thinking[1]	The capacity to process, analyze, and solve complex problems.	Analy*; Answer*; Complex; Curio*	Kyllonen et al. (2005)
Maturity/ Emotional intelligence	The capacity to understand one's own emotions and the emotions of others, and to appropriately respond and/or regulate one's emotions.		Kyllonen et al. (2005)
Motivation/ Achievement[1]	A tendency to put in effort toward achieving one's goals and persist to task completion.	Able; Aspire; Drive; Persist*	Kuncel et al. (2014) Kyllonen et al. (2005)
Personality[1]	Behavioral tendencies in intellectance, conscientiousness, extraversion, agreeableness, and emotional stability	Social; Deep; Detail-Oriented; Complex*	Kuncel et al. (2014) Kyllonen et al. (2005)
Self-efficacy	A belief in one's ability to successfully perform actions/behaviors.		Kyllonen et al. (2005)

Note. *indicates capturing words that contain the letter string. Constructs without example content due to no empirical investigation of LOR content for that construct. 1—constructs included in the current investigation.

difficult to trace and decision makers often lack insight into how they decided (Dalal et al., 2020; Hastie & Dawes, 2001).

We focus on the third consequence of using unstandardized application materials; namely, the potential for decisions to be influenced by systematic differences in the presentation (i.e., LOR content) and evaluation (i.e., decision maker evaluations of LOR content) of information. Situational ambiguity promotes the manifestation of bias such that, in the absence of clear and specific information (a situation characteristic of unstructured data collection and combination procedures), decision makers may rely on automatic, schema-based processing, such as the use of heuristics, to facilitate information processing (Fiske & Neuberg, 1990; Kunda & Thagard, 1996). Because social identities (e.g., race, gender) are associated with collections of characteristics that can be readily applied, people may use gender and racial stereotypes to shape their thoughts and behaviors. Indeed, because stereotypes activate so easily, decision makers may only evaluate more deliberately when such stereotypes are robustly contradicted (Heilman & Haynes, 2005; Singletary & Hebl, 2009). Of particular concern for LORs is the potential for bias from both writers *and* evaluators of LORs. As such, the ambiguity of whether certain information should be included or not, and the lack of structure in evaluating LORs may invite both parties to rely on stereotypes when describing/evaluating applicants. In short, the lack of structure might invite subgroup differences in the content *and* evaluation of LORs.

Although limited with respect to LORs for graduate school admission (see Woo et al., 2020), research on LORs supports these concerns. First, in terms of evaluations of LOR content, Morgan et al. (2013) found that, holding the contents of the LOR constant, Black and male applicants were rated less positively than Caucasian and female applicants (the gender effect being attributed to a female-dominated program). Thus, even with the same content, evaluations of LOR content can be biased.

Second, there is evidence of bias in what writers include in LORs. Evidence of gender bias in LORs for male-dominated academic positions (e.g., Madera et al., 2009; 2019; Trix & Psenka, 2003) includes the use of more communal and less agentic language to describe females versus males, where communal language related negatively to hiring decisions (Madera et al., 2009). In addition, LORs written in support of female applicants, compared to those for males, are shorter and contain more recommendations for training and teaching versus research (Trix & Psenka, 2003). Finally, LORs for female candidates contained fewer assurances and more doubt-raising comments than letters for males (Trix & Psenka, 2003), with doubt-raisers negatively related to hiring recommendations (Madera et al., 2019).

Interestingly, the LORs described in these studies were all unstructured in nature. In contrast, Friedman et al. (2017) and Powers et al. (2020) compared content differences for unstandardized and standardized procedures for writing LORs, with standardization referring to the use of a shared template for letter writing. They showed that, in unstandardized LORs, the content differs between genders and races. The narrative content from LORs written using the standardized template, however, showed no differences across groups. Therefore, similar to findings in the interview literature (e.g., Levashina et al., 2014), standardizing LORs writing decreases systematic differences in LOR content across demographic groups.

Unstandardized LORs are more likely to result in biased evaluations because it opens the door to nondiagnostic information entering the decision environment (Dalal et al., 2020). Nondiagnostic information is a cue that is not related to the outcome being decided upon, therein reducing accuracy (Dalal et al., 2020). Reliance on unstandardized information in the form of LORs can introduce nondiagnostic information about an applicant into the decision process due to (1) the unstandardized approach to writing LORs (Friedman et al., 2017; Powers et al., 2020) where stereotypical language disadvantaging female applicants and AOCs is used disproportionately, and (2) from the decision makers reading LORs, whose intuitive assessments of candidates may rely on this stereotypical information (Morgan et al., 2013). In short, relying on holistic assessments of LORs to increase diversity may actually increase, not decrease, bias.

Current study

Here, we text analyzed over 31,000 LORs submitted on behalf of over 10,000 applicants to the graduate programs of a large public university to test for differences in LOR content across various groups. Given the limited theorizing on content differences in LORs for graduate applications, we do not offer hypotheses in this study. Instead, we explore six research questions. Our first research question seeks to establish the degree of variability in LOR content to determine whether LORs for admission to graduate programs do in fact differ in content:

RQ 1: Are there content differences across LORs?

To test this, we matched the content of LORs to dictionaries of psychological constructs that research suggests are important for graduate school performance (e.g., motivation; Kyllonen et al., 2005; Table 1). We also assess the overall sentimental tone of the LORs based on the language used to try to mirror a holistic evaluation of the valence of the LORs. Importantly, we setup our dictionaries before receiving LOR text limiting the potential for confirmation bias.

Next, we assess if these content categories predict admissions decisions. This allows us to assess if LORs, as a component of admissions portfolios, are related to being offered a spot in a program of study—the initial step in increasing diversity and inclusion in graduate programs:

RQ2: Does the content of LORs predict receiving an offer of admission?

The next three research questions address whether there are systematic differences in LOR content by race, gender, race-gender intersections, degree sought, and program of study. In this way, we index the measurement bias in LORs from the perspective of the letter writer:

RQ3: Does the content of LORs differ across applicant race, gender, and the intersections of race and gender?

RQ4: Does the content of LORs differ across masters and doctoral degrees?

RQ5: Does the content of LORs differ across S.T.E.M. versus Non-S.T.E.M. programs?

Although we might expect content differences based on applicants' degrees sought and program of study, if LOR writers are evaluating applicants on similar attributes independent of race or gender, we would not expect differences between races and/or genders in our analysis.[1]

Finally, we expand past work by testing if LOR content predicts offers of admission incrementally over GRE scores and undergraduate GPA:

> *RQ6: Does the content of LORs predict likelihood of offers of admissions over GRE and most recent GPA?*

The results of this study provides three main contributions to the discussions of intuitive evaluations of applications and how LORs may influence marginalized groups' likelihood of receiving a favorable admissions decision. First, by analyzing actual LORs for admissions, our study provides a realistic picture of LOR content and the degree of bias in LOR content for graduate school admission from the perspective of LOR writers. Second, our study uses LOR content to predict offers of admission, providing an objective criterion against which to test the predictive efficacy of LORs. Importantly, this sample includes the LORs for those offered and denied admission, helping avoid any range restriction (Kuncel et al., 1998). Finally, we undertake intersectionality analyses to provide nuance to the debate of racial and gender bias.

Method

Applicants and sample of letters

We analyzed 31,920 LORs submitted for 10,793 applicants for admissions to graduate programs for semesters between Spring 2018 and Fall 2020 (i.e., six admissions cycles).[2] Table 2 provides the demographic information about the sample of applicants. Where available, we recorded applicants' verified most recent undergraduate GPA (i.e., submitted one 4-point scale GPA), and verified GRE verbal, quantitative, and analytical writing percentiles. Admissions rates were 76.85% and 35.86% for masters and doctoral programs, respectively.

[1] This could also be said of the GRE; that is, if the GRE is evaluating applicants on similar attributes, independent of race or gender, there should be minimal subgroup difference on GRE scores (see Helms, 2006 for this perspective).

[2] A sample of programs, including S.T.E.M. designation, is available in the supplemental materials.

Table 2. Applicant demographics.

		N	% of reporting sample
Race			
	White	4,866	74.54
	Black	880	13.48
	Asian/Pacific Islander	572	8.76
	Indigenous	27	0.41
	Mixed-White[1]	166	2.54
	Mixed-Nonwhite[2]	17	0.26
Gender			
	Female	6,206	57.62
	Male	4,564	42.38
Degree			
	Doctoral	3,239	30.01
	Masters	7,554	69.99
Field of study			
	S.T.E.M.	6,019	55.77
	Non-S.T.E.M.	4,774	44.23
Admission decision			
	Admit	6,942	64.31
	Deny	3,852	35.69

Notes. 1—respondent indicated multiple races and White was one. 2—respondent indicated multiple races, none of which were White.

Analysis of LOR content

We used nine dictionaries to index the content and tone of LORs. Four dictionaries – personality, motivation, critical thinking, ability – represented psychological constructs relevant to performance in graduate school (Table 1). We limited the dictionaries to these four constructs as these represented popular constructs not assessed directly by the GRE (Kyllonen et al., 2005). The other five – standout words, positive emotional tone, negative emotional tone, tentativeness, certainty – index overall tone of the letter, approximate an overall intuitive assessment of a candidate, and investigate whether writers take different tones in LORs across groups.

Seven of the dictionaries came from LIWC2015 (Pennebaker et al., 2015), a popular text analysis software with 125 well-validated dictionaries (Tausczik & Pennebaker, 2010), and one dictionary came from a published study of LORs (Schmader et al., 2007). We developed the final personality dictionary by collecting personality adjectives from a comprehensive study of the Big-5 personality markers (Goldberg, 1992).[3] Rather than index individual Big-5 traits, potentially introducing false positives (i.e., indexing content incorrectly because of overlapping dictionary entries, Short et al., 2010), we created the dictionary with generally positive

[3] Dictionaries available in the supplemental materials.

personality adjectives. As such, higher scores represent greater positive personality content.

Consistent with text analysis best-practices (Short et al., 2010), we preprocessed the text of the LORs and removed non-English characters, English-language stopwords, and words less than three characters in length.[4] Following this, we used regular expression matching to count the number of times words from the respective dictionaries were included in the letter. We then normalized these counts with respect to the total word count of the post-processed LOR, such that final content scores represented the proportion of meaningful words in the letter that appear in the respective dictionaries. Thus, each letter analyzed received its own score for each of the nine dictionaries. Table 3 presents the means, standard deviations, and intercorrelations of these scores. Higher scores represented more content related to the construct and language tone.

Analysis of research questions[5]

To answer RQ 1 (i.e., are there content differences across LORs?), we conducted a within-letter ANOVA, controlling for applicant nesting, comparing differences in rates of content. We explored significant differences between all permutations of the nine content dimensions using post-hoc tests (Bonferroni-corrected family-wise (FW) Type I error $\alpha \leq .001$). For RQs 3 (i.e., LOR content differences across race, gender, and the intersection), 4 (i.e., LOR content differences between degrees), and 5 (i.e., LOR content differences between programs of study), we conducted linear regression, using robust standard errors to account for clustering and nonnormality (Wilcox, 2017; FW-Type I error corrected $\alpha \leq .006$). Finally, we used binary logistic regression, with robust standard errors, to address RQs 2 (i.e., does LOR content predict offers of admission?) and 6 (i.e., does LOR content predict offers of admission beyond GRE and GPA?). For RQ6, we used sequential logistic regression with most recent undergraduate GPA and GRE percentile included in model 1, and the nine LOR content dimensions added in model 2. We then assessed if any of the LOR content areas was a significant predictor of admissions offer (FW-Type I error corrected $\alpha \leq .004$). Furthermore, we control for S.T.E.M. designation when addressing RQ6 given extant research showing emphasis on different aspects of GRE between S.T.E.M. and non-S.T.E.M. fields (Bleske-Rechek & Browne, 2014). Table 4 shows

[4] See supplemental materials for full preprocessing details.

[5] The following analyses were preregistered: https://osf.io/xhtge and https://osf.io/dyw4z .

Table 3. Content analysis dictionaries, means, standard deviations, and intercorrelations.

		Content intercorrelations. Means (standard deviations) on diagonal								
Dictionary	Source of Dictionary	1	2	3	4	5	6	7	8	9
1. Personality	Authors	.03 (.02)*								
2. Motivation/Achievement	LIWC2015	.16	.08 (.03)*							
3. Critical thinking	LIWC2015	.16	.09	.05 (.02)*						
4. Ability/Creativity	LIWC2015	**.27**	**.36**	**.20**	.03 (.02)*					
5. Standout words	Schmader et al. (2007)	.08	.19	.06	.10	.01 (.01)*				
6. Positive emotional tone	LIWC2015	**.33**	**.36**	.16	**.21**	**.27**	.09 (.03)*			
7. Negative emotional tone	LIWC2015	.06	.02	.11	.04	−01	.02	.01 (.01)*		
8. Tentativeness	LIWC2015	.02	−.06	**.21**	−.05	.03	.04	**.23**	.01 (.01)*	
9. Certainty	LIWC2015	.14	.16	.08	.05	.09	**.32**	.06	.06	.02 (.02)*

Notes. N=31,920 LORs. Diagonal elements are the means proportion and standard deviations of the content dimension scores. All correlations, except the correlation between standout words and negative emotional tone are significantly different from 0, p < .0001. Bolded correlations: r > .20. *average proportion significantly different from 0, p < .0001, based on intercept-only random effects regression.

Table 4. Means, standard deviations and intercorrelations of applicant variables.

	1	2	3	4	5
1. Grade point average	3.42 (.39)				
2. GRE verbal %	**0.34**	49.96 (28.67)			
3. GRE quantitative %	0.11	0.16	51.83 (26.62)		
4. GRE analytical writing %	**0.36**	**0.67**	−0.04	46.08 (29.40)	
5. Admission decision[a]	0.14	0.14	0.07	0.12	64% Accepted

Notes. $N = 6{,}281 - 19{,}266$. Diagonal elements means and standard deviations. a—dichotomous variable coded 0-deny, 1-admit, point-biserial correlations in this row. All correlations significantly different than 0. Bolded correlations: $r > .20$.

the means, standard deviations, and intercorrelations of GPA, GRE, and admit decisions.

Results[6]

RQ1: are there content differences across LORs?

The categories of LOR content do differ across letters, from as low as 1% (tentativeness) to as high as 9% (positive emotional tone; Table 3). Results of the repeated measures ANOVA showed significant differences in the proportion of LOR content across letters, F (8, 86336) = 59,367.81, $p_{\text{G-G}}$[7] < .0001, $\eta^2 = .81$. Post-hoc tests showed that the average proportions were all significantly different from each other after correcting for FW-Type I error rates. In terms of psychological constructs, terms related to motivation (8%) and critical thinking (5%) featured most frequently. In terms of overall tone, as one would expect for LORs, positive emotional tone was most prevalent (9%). In sum, the answer to RQ1 is "yes;" there are content differences in LORs, and LOR content areas differ significantly among each other. LOR writers focus on different aspects of applicants' and use language that signals different evaluations. This was true even though dictionaries were determined *a priori* and independently from the LOR corpus.

[6] Given the large number of analyses and results, we focus our in-text discussion on only significant findings. Full results are available in the supplemental materials.

[7] Mauchly's test showed that sphericity was violated, $w = .06$, $p < .001$. Greenhouse-Geisser corrected p-value.

RQ2: does the content of LORs predict receiving an offer of admission?

Together, the nine content areas of LORs predicted about 0.4% of the variance in probability of an admissions offer. After controlling for FW-Type I error rates, four of the content areas were associated with increased chances of offers of admission, three psychological constructs: ability (*Odds Ratio* [*OR*] = 2.51), motivation (*OR* = 2.87), critical thinking (*OR* = 2.51), and one language tone: positive emotional tone (*OR* = 3.00). Although not passing the FW Type-I error corrected α-level, inclusions of standout terms did trend toward significant (*OR* = 2.64, p = .009). In sum, the answer to RQ2 is also "yes;" although a somewhat weak effect, LOR content does predict the chances of one receiving an offer of admission to a graduate program.

RQ3: does the content of LORs differ across applicant race, gender, and the intersections?

LORs for female applicants contained higher proportions of personality, ability, motivation, positive emotional tone, and certainty content compared to LORs for males. LORs for female applicants also contained lower proportions of content related to critical thinking and tentativeness when compared to males. Few racial differences emerged in LOR content after correcting for FW-Type I error rates. Specifically, compared to White applicants, Black applicants had lower proportions of standout words and tentative terms, and higher proportions of motivation terms.[8] Finally, intersectionality analyses revealed only two significant race-by-gender interactions after controlling for FW-Type I error rates. The negative emotional tone was significantly higher for Black females (M = .07, sd = .03) compared to White males (M = .03, sd = .01), White females (M = .01, sd = .01), and Black males (M[9] = –.09, sd = .03). The proportion of references to critical thinking was significantly lower for mixed-nonwhite males (M = –.22, sd = .11), compared to mixed-nonwhite females (M = .12, sd = .09) and White applicants (M_{males} = –.07, sd_{males} = .02; M_{females} = –.13, sd_{femles} = .01). In sum, the answer to RQ3 is "a little;" some LOR content does differ across gender, and between Black and White applicants, and there were few intersectional differences.

[8] See supplemental materials for detailed results including LOR content means by gender and race

[9] Values are standardized, so negative values are marginal means below the sample mean.

RQ4: does the content of LORs differ across degree sought?

With the exception of standout words, there were significant differences between LOR content for applicants to master's versus doctoral programs. Specifically, applicants to doctoral programs had a significantly higher proportion of critical thinking, and, surprisingly, tentativeness and negative emotional tone. Masters applicants had higher proportions of the other content areas.[10] The answer to RQ4 is "yes;" the content of LORs differs across degree sought.

RQ5: does the content of LORs differ across program of study?

Results comparing LOR content between S.T.E.M and non-S.T.E.M. disciplines likewise showed differences in content. Whereas LORs for applications to non-S.T.E.M. programs contained higher proportions of personality terms, positive emotional tone, motivation terms, and certainty tone, LORs for applications to S.T.E.M. programs contained higher proportions of ability terms, standout words, critical thinking terms, and tentative tone. In sum, the answer to RQ5 is also "yes;" the content of LORs differs between S.T.E.M. and non-S.T.E.M. programs.

RQ6: does LOR content predict likelihood of admissions offers over GRE and GPA?

Finally, results suggest that, after controlling for GRE scores, undergraduate GPA, and S.T.E.M. versus non-S.T.E.M. discipline, few LOR content categories predict chances of being offered admission. Specifically, for S.T.E.M disciplines, GRE quantitative percentile was a significant predictor of admissions offer ($OR = 3.56$); surprisingly, GPA was not, $pseudo\text{-}R^2 = .01$. Adding the nine content areas minimally increased the variance predicted, $\Delta pseudo\text{-}R^2 = .005$, and together, all the predictors predicted about 1% of the variance in offers of admission, $pseudo\text{-}R^2 = .01$. The only LOR content area that was a significant predictor of admissions offer was negative emotional tone—unexpectedly, a higher proportion of negative emotional tone was associated with an increase in the chances of being offered admission ($OR = 1.48$).

For non-S.T.E.M. disciplines, only GPA ($OR = 25.03$) was significantly related to offers of admission, $pseudo\text{-}R^2 = .09$. LOR content only minimally increased the variance predicted in admission offers, $\Delta pseudo\text{-}R^2$

[10] See supplemental materials for detailed results including LOR content means by degree sought and S.T.E.M. and non-S.T.E.M. disciplines.

= .01, and the predictors as a group predicted about 10% of the variance in admissions, *pseudo-R^2* = .10. GPA continued to relate to the chances of being offered admission, but none of the LOR content predicted offers of admission. In sum, the answer to RQ6 is "not really;" there is only minimal evidence that LOR content predicts admissions offers beyond GRE and GPA.

Exploratory analysis

In light of the growing concerns regarding gender representation in S.T.E.M. fields (Cheryan et al., 2017), we tested for systematic differences in the content of LORs at the intersection of gender and program of study.[11] Linear regression with robust standard errors showed no significant differences in LOR content at the intersection of gender and field of study after accounting for FW-Type 1 error.[12] In short, systematic differences in LOR content between male and female applicants were consistent across field of study.

Discussion

In an effort to increase diversity and inclusion in graduate education, some advocate for a more holistic assessment of applicants making use of unstructured information like LORs (Buckley et al., 2018). However, unstructured methods, like LORs, and holistic evaluations of their content, can be associated with more, not less, bias in decisions (Dalal et al., 2020). Here, we explored the potential for bias in the content of LORs by race, sex, and race-by-sex intersections, and if LOR content relates to admissions decisions. Although *ratings* of LORs are minimally related to graduate student *performance* (Kuncel et al., 2014), there is little data on the *content* of LORs for graduate school applicants, including how the content differs among groups, and if LOR content is related to graduate admission *offers*—the first step in increasing inclusion in graduate education. Four broad conclusions from our study fill this gap.

First, letter writers vary the content of their LORs, highlighting different applicant qualities and using language to signal different overall evaluations. Alone, however, LOR content predicts a small amount of variance in admissions offers (i.e., less than 1%). Second, there are meaningful differences in the content of LORs between male and female

[11] We thank an anonymous reviewer for suggesting this analysis.

[12] See supplemental materials for full results.

applicants, and some differences between Black and White applicants. When focused on only the four LOR content areas that predict admissions, compared to males, females have a higher proportion of terms connoting ability, motivation, and positive emotional tone, and a lower proportion of terms that connote critical thinking. Compared to White applicants, Black applicants have a higher proportion of terms signaling motivation. Investigation of intersectional race-by-gender and gender-by-area of study differences were, fortunately, less evident, but did suggest some disadvantages for Black female applicants being evaluated with more negative emotional tone and mixed race nonwhite male applicants receiving fewer mentions of critical thinking. Combining these results with other studies of bias in LOR content, there appears to be evidence of systematic differences in the content of LORs across race, gender, and the intersection of the two. Depending on what an LOR evaluator considers important for graduate student success, evaluating LORs for admissions may favor or disadvantage marginalized groups.

As decision makers often lack insight into how they make their own decisions (Hastie & Dawes, 2001), it is unclear whether applicants from marginalized groups will be disadvantaged or not. This may partially explain why LOR content is not a stronger predictor of admissions decisions, as unstandardized LOR review allows for idiosyncratic emphasis given to different contact areas that vary between groups. Stated differently, inter-decision maker inconsistency (e.g., some decision makers weighing motivation more than others) as well as intra-decision maker inconsistency (e.g., the same decision maker weighing motivation importantly for one candidate, but not others) in the evaluation of LORs might explain why LOR content is not more predictive of offers of admission. These results suggest that evaluating LORs holistically may not necessarily increase diversity of graduate programs.

Results also showed that, after taking into account GRE and GPA, LOR content did not meaningfully improve the prediction of admissions offers. This could potentially be because graduate admissions committee members historically focus on the GRE first, leaving little variance left to be predicted by any other variable. Alternatively, this could reflect decision makers' reliance on more standardized information (GRE, GPA), which improves predictive validity, perhaps based on past experience wherein successful admissions decisions confirm the use of these standardized indicators. To test this possibility, we conducted an additional exploratory analysis investigating if LOR and undergraduate GPA predicted offers of admission for those students who applied to graduate programs but did not submit a GRE score (i.e., GRE optional programs; $N = 6{,}154$ applicants). Results from these analyses showed a similar pattern: Although undergraduate GPA was a significant predictor of admissions offers, none of the LOR content areas predicted offers of admission for GRE-optional programs.[13] This suggests that LOR content is unrelated to offers of admission even when GRE information is not available, and that overall, decision makers may be relying on things other than LORs when making admissions decisions. Another possibility is that admissions decision makers are not evaluating the LOR content similarly enough, therein reducing the reliability of LOR evaluations and in turn reducing the predictive capacity of LOR content—future research can disentangle these two.

Finally, our results also showed that LOR content differs among letters submitted to doctoral versus master's programs and between S.T.E.M. and non-S.T.E.M. disciplines. Regarding program of study, a general conclusion is, on the one hand, LORs associated with S.T.E.M. disciplines focus on concrete aspects of the candidate (i.e., standout terms, ability terms). On the other hand, LORs associated with non-S.T.E.M. disciplines focus on potential and behavioral patterns (i.e., personality, motivation). Unexpectedly, apart from critical thinking terminology, LORs on behalf of doctoral applicants contained lower proportions of many desirable content areas, and higher proportions of negative content areas. One speculation is that these content differences reflect differences in the standards letter writers hold for doctoral applicants compared to masters applicants, wherein letter writers express more qualifiers for the higher standards held for doctoral applicants. Another speculation is letter writers for doctoral applicants are likely to be academics whereas letter writers for masters applicants are likely a mix of academic and professional references. These groups might take different approaches to writing letters. These and other potential explanations will need further research.

Recommendations

Reiterating our commitment to increasing diversity and inclusion in graduate programs and helping to address the "leaky pipeline" that plagues many professions (Barr et al., 2008), we offer some recommendations for increasing diversity of graduate programs. First, combining our results with the research on structured LORs (e.g., Friedman et al., 2017; Powers et al., 2020), structured interviews (e.g., Levashina et al., 2014), and inter- and intra-decision maker inconsistencies (e.g., Hastie & Dawes, 2001), we recommend writing and evaluating LORs in a structured manner. Although the former is difficult, as letter writers are likely to use their own

[13] Full results available in the supplemental materials.

preferred styles, a growing area of research provides promising avenues for standardizing LORs by asking for ratings of applicants on important competencies (e.g., Liu et al., 2009; Powers et al., 2020), and prompting behavioral examples to support the narrative evaluations (e.g., Alweis et al., 2017; Walters et al., 2006). By comparison, standardizing admission decision makers' evaluations of LOR content is more tractable. Admissions committees can discuss what information and tone from LORs to consider in their admissions decisions, and each committee member can apply these rules consistently, across every applicant. The use of LOR rating forms can facilitate this. We expect that this form of standardization of LOR evaluation will improve the reliability and validity of LOR evaluations, though systematic future research is needed.

We offer two additional recommendations that, although not directly related to our investigation of LORs, are strategies that can assist with increasing diversity. First, universities can increase representation in their graduate programs by expanding the applicant pool to include more highly qualified applicants from all demographic groups through targeted recruitment (Newman & Lyon, 2009). This could include the use of minority and/or female recruiters, and inclusion of diversity-language and images in recruitment messages. These strategies attract more diverse talent (Avery et al., 2004; Avery & McKay, 2006) and can be adapted to the graduate admission context (Kilburn et al., 2019; Poock, 2007). Indeed, graduate programs that adopt pro-diversity recruitment practices (e.g., minority-support groups) are more successful in increasing diversity among their student bodies (Griffin et al., 2012; Slay et al., 2019). In addition, given the intrinsic role of faculty as mentors (Martinez et al., 2018), increasing faculty diversity may have trickle-down benefits on recruiting.

Second, as an anonymous reviewer noted, there is always room to consider avenues for improving the measurement of psychological variables. With respect to LORs, we advocate for the standardization of LOR writing and evaluation. With respect to the GRE and similar measures of cognitive abilities, increased standardization may also play a role in reducing subgroup differences. A productive avenue for increased test standardization efforts may be in the areas of test development by including a broader range of constructs assessed, and by increasing the equity of public education. Ackerman (2017) notes that content development in psychometrically-derived assessments lacks a broad view of intelligence, one that acknowledges the importance of relevant knowledge domains and intellectual skills (e.g., critical thinking, writing). This may reflect, in part, cultural differences about what constitutes "intelligence," and which cognitive abilities and skills are seen as most important to be targeted in standardized admissions testing (Flynn, 2018; Neisser et al., 1996). More broadly, increased investments in educational opportunities to marginalized communities may help address the "Education Debt" (Ladson-Billings, 2006) that has accumulated over the centuries of systemic mistreatment of communities of color and disproportionately harmed AOCs (Helms, 2006).

Study limitations and future directions

One limitation of our study is that we were unable to directly assess the potential bias from LOR evaluators. This is an important aspect to understanding the utility of LORs for admissions decision making, and, although some studies have looked at the evaluators' perspective (Morgan et al., 2013), more research is needed. This should include how the race and gender of the evaluator interacts with the race and gender of the applicant to influence the evaluation of LORs. Indeed, our data only had race and gender for the applicant. Future research should explore if LOR writers' and/or evaluators' race and gender influence LORs.

Second, although we used past research to guide LOR content dictionaries, content we did not consider could relate to admissions decisions. Future research can explore other content domains. Related, as with all text analyses, our results are limited based on the content of the dictionaries (Short et al., 2010). Although most of the dictionaries are well-validated (Tausczik & Pennebaker, 2010), results should be replicated with different dictionaries and LORs.

Finally, the effect sizes in our study are not particularly large compared to other studies. On the one hand, the differences in effect sizes may be due to different criteria (i.e., admissions decisions versus performance). On the other hand, our effect sizes do leave open the question of what predicts offers of admission? As we have argued, unstructured evaluation of application material is less predictive than structured approaches. As such, it is possible that our small effect sizes are quantifying the limited validity of combining application information in an unstructured way. Replications and extension of this work will help disentangle these, and other explanations.

Conclusion

The results of our systematic analysis of over 37,000 LORs do not suggest that an admissions system emphasizing LOR content would be effective, as content categories signaling applicants' psychological attributes and the overall emotional tone and certainty of recommendations only weakly predicted admission offers, and this predictive value largely disappeared after considering GRE scores and undergraduate GPA. Our results also revealed differences in LOR content across the groups we studied.

Unfortunately, since evaluations of LORs are unstandardized, such differences in content categories can lead to differences in admissions offers, thereby affecting graduate program diversity. We recommend standardized evaluations of LOR content to encourage more equitable evaluation of marginalized applicants and to increase diversity in graduate education.

ORCID

Dev K. Dalal http://orcid.org/0000-0001-8968-7790
Ho Kwan Cheung http://orcid.org/0000-0002-8516-5466

References

Aamodt, M. G., Bryan, D. A., & Whitcomb, A. J. (1993). Predicting performance with letters of recommendation. *Public Personnel Management*, *22*, 81–90. https://doi.org/10.1177/009102609302200106

Ackerman, P. L. (2017). Adult intelligence: The construct and the criterion problem. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *12*(6), 987–998. https://doi.org/10.1177/1745691617703437

Alweis, R., Collichio, F., Milne, C. K., Dalal, B., Williams, C. M., Sulistio, M. S., Roth, T. K., & Muchmore, E. A. (2017). Guidelines for a standardized fellowship letter of recommendation. *The American Journal of Medicine*, *130*(5), 606–611.

Arthur, W., Jr., & Villado, A. J. (2008). The importance of distinguishing between constructs and methods when comparing predictors in personnel selection research and practice. *Journal of Applied Psychology*, *93*(2), 435–442. https://doi.org/10.1037/0021-9010.93.2.435

Avery, D. R., Hernandez, M., & Hebl, M. R. (2004). Who's watching the race? Racial salience in recruitment advertising 1. *Journal of Applied Social Psychology*, *34*(1), 146–161. https://doi.org/10.1111/j.1559-1816.2004.tb02541.x

Avery, D. R., & McKay, P. F. (2006). Target practice: An organizational impression management approach to attracting minority and female job applicants. *Personnel Psychology*, *59*(1), 157–187. https://doi.org/10.1111/j.1744-6570.2006.00807.x

Barr, D. A., Gonzalez, M. E., & Wanat, S. F. (2008). The leaky pipeline: Factors associated with early decline in interest in premedical studies among underrepresented minority undergraduate students. *Academic Medicine: Journal of the Association of American Medical Colleges*, *83*(5), 503–511.

Bleske-Rechek, A., & Browne, K. (2014). Trends in GRE scores and graduate enrollments by gender and ethnicity. *Intelligence*, *46*, 25–34. https://doi.org/10.1016/j.intell.2014.05.005

Buckley, L., Letukas, L., & Wildavsky, B., (2018). Introduction: The emergence of standardized testing and the rise of test-optional admissions. In L. Buckley, N. Letukas, & B. Wildavsky (Eds.), *Measuring success: Testing, grades, and the future of college admissions* (pp. 1–12). Johns Hopkins University Press.

Cheryan, S., Ziegler, S. A., Montoya, A. K., & Jiang, L. (2017). Why are some STEM fields more gender balanced than others? *Psychological Bulletin*, *143*(1), 1–35.

Dalal, D. K., Sassaman, L., & Zhu, X. (2020). The impact of nondiagnostic information on selection decision making: A cautionary note. *Personnel Assessment and Decisions*, *6*(2), 54–64. https://doi.org/10.25035/pad.2020.02.007

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in Experimental Social Psychology*, *23*, 1–74.

Flynn, J. R. (2018). Reflections about intelligence over 40 years. *Intelligence*, *70*, 73–83. https://doi.org/10.1016/j.intell.2018.06.007

Friedman, R., Fang, C. H., Hasbun, J., Han, H., Mady, L. J., Eloy, J. A., & Kalyoussef, E. (2017). Use of standardized letters of recommendation for otolaryngology head and neck surgery residency and the impact of gender. *The Laryngoscope*, *127*(12), 2738–2745. https://doi.org/10.1002/lary.26619

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

Griffin, K. A., Muñiz, M. M., & Espinosa, L. (2012). The influence of campus racial climate on diversity in graduate education. *The Review of Higher Education*, *35*(4), 535–566. https://doi.org/10.1353/rhe.2012.0031

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world* (2nd ed.). SAGE.

Heilman, M. E., & Haynes, M. C. (2005). No credit where credit is due: Attributional rationalization of women's success in male-female teams. *The Journal of Applied Psychology*, *90*(5), 905–916.

Helms, J. E. (2006). Fairness is not validity or cultural bias in racial-group assessment: A quantitative perspective. *The American Psychologist*, *61*(8), 845–859. https://doi.org/10.1037/0003-066X.61.8.845

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology*, *1*(3), 333–342. https://doi.org/10.1111/j.1754-9434.2008.00058.x

Kilburn, F., Hill, L., Porter, M. D., & Pell, C. (2019). Inclusive recruitment and admissions strategies increase diversity in CRNA educational programs. *AANA Journal*, *87*, 379–389.

Kuncel, N. R., Campbell, J. P., & Ones, D. S. (1998). Validity of the Graduate Record Examination: Estimated or tacitly known? *American Psychologist*, *53*(5), 567–568. https://doi.org/10.1037/0003-066X.53.5.567

Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science (New York, N.Y.)*, *315*(5815), 1080–1081.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin*, *127*(1), 162–181. https://doi.org/10.1037/0033-2909.127.1.162

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology*, *98*(6), 1060–1072. https://doi.org/10.1037/a0034156

Kuncel, N. R., Kochevar, R. J., & Ones, D. S. (2014). Letters of recommendation in college and graduate admission: Reason for hope. *International Journal of Selection and Assessment*, *22*(1), 101–107. https://doi.org/10.1111/ijsa.12060

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, *103*(2), 284–308. https://doi.org/10.1037/0033-295X.103.2.284

Kyllonen, P., Walters, A. M., & Kaufman, J. C. (2005). Noncognitive constructs and their assessment in graduate education: A review. *Educational Assessment*, *10*(3), 153–184. https://doi.org/10.1207/s15326977ea1003_2

Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in US schools. *Educational Researcher*, *35*, 3–12. https://doi.org/10.3102/0013189X035007003

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology*, *67*(1), 241–293. https://doi.org/10.1111/peps.12052

Liu, O. L., Minsky, J., Ling, G., & Kyllonen, P. (2009). Using the standardized letters of recommendation in selection: Results From a multidimensional Rasch Model. *Educational and Psychological Measurement*, *69*(3), 475–492. https://doi.org/10.1177/0013164408322031

Lucido, J. A. (2018). Understanding the test-optional movement. In L. Buckley & N. Wildavsky (Eds.), *Measuring success: Testing, grades, and the future of college admissions* (pp. 145–170). Johns Hopkins University Press.

Madera, J. M., Hebl, M. R., Dial, H., Martin, R., & Valian, V. (2019). Raising doubt in letters of recommendation for academia: Gender differences and their impact. *Journal of Business and Psychology*, *34*(3), 287–303. https://doi.org/10.1007/s10869-018-9541-1

Madera, J. M., Hebl, M. R., & Martin, R. C. (2009). Gender and letters of recommendation for academia: Agentic and communal differences. *Journal of Applied Psychology*, *94*(6), 1591–1599. [Database] https://doi.org/10.1037/a0016539

Martinez, L. R., Boucaud, D. W., Casadevall, A., & August, A. (2018). Factors contributing to the success of NIH-designated underrepresented minorities in academic and nonacademic research positions. *CBE—Life Sciences Education*, *17*(2), ar32. https://doi.org/10.1187/cbe.16-09-0287

Morgan, W. B., Elder, K. B., & King, E. B. (2013). The emergence and reduction of bias in letters of recommendation. *Journal of Applied Social Psychology*, *43*(11), 2297–2306. https://doi.org/10.1111/jasp.12179

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*(2), 77–101. https://doi.org/10.1037/0003-066X.51.2.77

Newman, D. A., & Lyon, J. S. (2009). Recruitment efforts to reduce adverse impact: Targeted recruiting for personality, cognitive ability, and diversity. *Journal of Applied Psychology*, *94*, 298–317. https://doi.org/10.1037/a0013472

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *Linguistic inquiry and word count: LIWC2015*. Pennebaker Conglomerates.

Peres, S. H., & Garcia, J. R. (1962). Validity and dimensions of descriptive adjectives used in reference letters for engineering applicants. *Personnel Psychology*, *15*, 279–286.

Poock, M. C. (2007). A shifting paradigm in the recruitment and retention of underrepresented graduate students. *Journal of College Student Retention: Research, Theory & Practice*, *9*(2), 169–181. https://doi.org/10.2190/CS.9.2.c

Posselt, J. R. (2016). *Inside graduate admission: Merit, diversity, and faculty gatekeeping*. Harvard University Press.

Powers, A., Gerull, K. M., Rothman, R., Klein, S. A., Wright, R. W., & Dy, C. J. (2020). Race- and gender-based differences in descriptions of applicants in the letters of recommendation for orthopaedic surgery residency. *JBJS Open Access*, *5*(3), e20.00023. https://doi.org/10.2106/JBJS.OA.20.00023

Sackett, P. R., & Kuncel, N. R. (2018). Eight myths about standardized admissions testing. In J. Buckley, L. Letukas, & N. Wildavsky (Eds.), *Measuring success: Testing, grades, and the future of college admissions* (pp. 13–39). Johns Hopkins University Press.

Schmader, T., Whitehead, J., & Wysocki, V. H. (2007). A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex Roles*, *57*(7–8), 509–514.

Short, J. C., Broberg, J. C., Cogliser, C. C., & Brigham, K. H. (2010). Construct validation using computer-aided text analysis (CATA) an illustration using entrepreneurial orientation. *Organizational Research Methods*, *13*(2), 320–347. https://doi.org/10.1177/1094428109335949

Singletary, S. L., & Hebl, M. R. (2009). Compensatory strategies for reducing interpersonal discrimination: The effectiveness of acknowledgments, increased positivity, and individuating information. *Journal of Applied Psychology*, *94*(3), 797–805. https://doi.org/10.1037/a0014185

Slay, K. E., Reyes, K. A., & Posselt, J. R. (2019). Bait and switch: Representation, climate, and tensions of diversity work in graduate education. *The Review of Higher Education*, *42*(5), 255–286. https://doi.org/10.1353/rhe.2019.0052

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. [Database] https://doi.org/10.1177/0261927X09351676

Trix, F., & Psenka, C. (2003). Exploring the color of glass: Letters of recommendation for female and male medical faculty. *Discourse & Society*, *14*(2), 191–220. https://doi.org/10.1177/0957926503014002277

Walters, A. M., Kyllonen, P. C., & Plante, J. W. (2006). Developing a standardized letter of recommendation. *Journal of College Admission*, *191*, 8–17.

Wilcox, R. R. (2017). *Introduction to robust estimation and hypothesis testing* (4th ed.). Elsevier.

Woo, S. E., LeBreton, J., Keith, M., & Tay, L. (2020, August 18). Bias, fairness, and validity in graduate admissions: A psychometric perspective.