International Journal of Testing

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

Using personal statements in college admissions: An investigation of gender bias and the effects of increased structure

A. Susan M. Niessen & Marvin Neumann

To cite this article: A. Susan M. Niessen & Marvin Neumann (2022) Using personal statements in college admissions: An investigation of gender bias and the effects of increased structure, International Journal of Testing, 22:1, 5-20, DOI: 10.1080/15305058.2021.2019749

To link to this article: https://doi.org/10.1080/15305058.2021.2019749

INTERNATIONAL JOURNAL OF TESTING
2022, VOL. 22, NO. 1, 5–20
https://doi.org/10.1080/15305058.2021.2019749

Routledge
Taylor & Francis Group

Using personal statements in college admissions: An investigation of gender bias and the effects of increased structure

A. Susan M. Niessen and Marvin Neumann

Heymans Institute for Psychological Research, University of Groningen, Groningen, The Netherlands

ABSTRACT

Personal statements are among the most commonly used instruments in college admissions procedures. Yet, little research on their reliability, validity, and fairness exists. The first aim of this paper was to investigate hypotheses about adverse impact and underprediction for female applicants, which could result from lower tendencies to use agentic language compared to male applicants. Second, we examined if rating personal statements in a more structured manner would increase reliability and validity. Using personal statements (250 words) from a large cohort of applicants to an undergraduate psychology program at a Dutch University, we found no evidence for adverse impact for female applicants or more agentic language use by male applicants, and no relationship between agentic language use and personal statement ratings. In contrast, we found that personal statements of female applicants were rated slightly more positively than those of males. Exploratory analyses suggest that female applicants' better writing skills might explain this difference. A more structured approach to rating personal statements yielded higher, but still only 'moderate' inter-rater reliability, and virtually identical, negligible predictive validity for first year GPA and dropout.

KEYWORDS

College admissions; gender; personal statements; reliability; validity

Personal statements are among the most commonly used sources to gather information about applicants in college admissions procedures

CONTACT A. Susan M. Niessen a.s.m.niessen@rug.nl Heymans Institute for Psychological Research, University of Groningen, Groningen, The Netherlands.

for undergraduate and graduate programs (Clinedinst, 2019; Klieger et al., 2017; Woo et al., 2020). They are often used to collect information about motivation to study in a particular field, goals and interests, strengths and weaknesses, and writing skills (Kuncel et al., 2020; Kyllonen et al., 2005). A small meta-analysis ($k = 8–10$) by Murphy et al. (2009) showed that ratings of personal statements were poor predictors of academic achievement, with $r = .13$ for GPA and $r = .09$ for faculty performance ratings, and had no incremental validity ($\Delta R^2 = .002$) over admission test scores and prior grades. Plausible explanations are that personal statements are typically highly unstructured in nature (Kuncel et al., 2020; Woo et al., 2020), both in terms of instructions for applicants on what to include in their statement, and in how they are judged or rated, if formally rated at all. This lack of structure likely results in low reliability, construct validity, and hence, predictive validity.

Potential gender bias in personal statement ratings

Aside from low validity, there are concerns of possible biases resulting from using personal statements in admissions. One major concern is the possibility for gender bias (Woo et al., 2020). Some studies showed that male applicants tend to use more agentic and self-promotional language than female applicants (Babal et al., 2019; Osman et al., 2015; Ostapenko et al., 2018), which, while not previously tested, is expected to result in higher ratings on personal statements for male applicants. Moreover, assuming that the tendency to use agentic language is unrelated to academic achievement, ratings on personal statements could subsequently result in overpredicting male performance and underpredicting female performance.

The aim of this study was twofold. First, we aimed to investigate gender differences and predictive gender bias in ratings of personal statements. Since males have been found to use more self-promotional and agentic language, we expected that:

H1: Male applicants obtain higher personal statement ratings than female applicants.

H2: Male applicants use more agentic language in their personal statements than female applicants.

Since female students tend to perform better academically (Voyer & Voyer, 2014) and we expect higher personal statement ratings for male applicants, we also expected:

H3: Ratings on personal statements result in overprediction of male academic achievement and underprediction of female academic achievement.

Since we hypothesize that using agentic language explains the differences in personal statement ratings, we expected:

H4: Controlling for the use of agentic language reduces gender-based differential prediction.

Structure in personal statement ratings

Second, we investigated if evaluating personal statements in a more structured way improved reliability and predictive validity of personal statement ratings, as previously suggested (Kuncel et al., 2020; Murphy et al., 2009; Woo et al., 2020). To that end, we compared general, impressionistic ratings and ratings based on anchored rating scales on several dimensions, made based on the same personal statements. We have the following hypotheses:

H5: Single, impressionistic ratings of personal statements are unreliable and poor predictors of academic achievement.

H6: Using more structured ratings of personal statements will lead to improvements in reliability and predictive validity.

In previous studies, predictive validity was mainly examined for GPA. Since some hypothesized that personal statements may be more useful in predicting 'fit' related outcomes such as retention (Chiu, 2019; Murphy et al., 2009), we investigated predictive validity for first year GPA and dropout. Furthermore, we will explore if more structured ratings show different results regarding gender differences than general, impressionistic ratings.

Using personal statements in college admissions procedures

Finally, we present the effect of including personal statements in admissions procedures by demonstrating the validity of an admission test + personal statement composite under different hypothetical weighting schemes. Presenting composite validity under different possible weighting schemes adds to the traditional regression approach in terms of showing the practical implications of using personal statements in conjunction with other instruments, since optimal regression weights are seldom available or used in admissions procedures. We expect that using a suboptimally weighted composite of personal statement ratings and admission test scores will

result in lower predictive validity compared to just using the admission test alone, especially when general, impressionistic ratings are used.

Method

Sample

Archival personal statements of 806 applicants to an undergraduate Psychology program at a Dutch university were retrieved from the university records. Of all applicants, 627 were accepted and started the program, for whom information on gender, admission test scores, first year GPA, and dropout were retrieved as well. The mean age was $M = 19.62$ ($SD = 1.69$), 66% was female, 45% had the Dutch nationality, 46% had a German nationality, 7% had another EU-nationality, and 2% had a non-EU nationality. The program's courses were offered in Dutch and English, 42% took their courses in Dutch. A priori power analysis ($1 - \beta = .80$, $\alpha = .05$, one-tailed), showed that this sample size is sufficient to detect a predictive validity of $r = .10$ and small differences in personal statement ratings ($d = .30$).

Materials and procedure

All applicants had to submit a personal statement of approximately 250 words, indicating their motivation for the program. The mean number of words was $M = 232$ ($SD = 34$). Of all applicants, 38% wrote their statement in Dutch and 62% wrote their statement in English. Each personal statement was rated in two ways; by providing a single, impressionistic rating (second author and research assistant) and by using pre-defined dimensions with anchored rating scales (first author and research assistant). So, each statement was rated by two raters based on each approach. The scales were piloted to check calibration and clarity among the raters beforehand. The statements were presented to the raters in random order, and personally identifiable information such as names and addresses were omitted to ensure anonymity of the applicants. No other information, such as prior educational achievement, grades, test scores or applicant gender, were available when rating the statements. Rating the statements spanned about five weeks and was done in multiple sittings to limit fatigue effects.

Measures

To obtain a single, impressionistic rating, each statement was rated based on the following question: *How suitable is this applicant for studying*

Psychology at this university?, using a five-point scale (*not suitable at all* – *very suitable*). The mean rating across two raters was used in all analyses.

For the structured ratings, we consulted staff and the literature (Chiu, 2019; GlenMaye & Oakes, 2002; Max et al., 2010) to determine what kind of information is typically extracted from personal statements. The resulting five dimensions (*motivation for the discipline*, *expressing a future career perspective*, *motivation to study and learn*, *possessing relevant competencies and skills*, and *writing skills*) with anchored rating scales are shown in the Appendix. Again, mean ratings across the two raters were used in the analyses.

The use of agentic language was operationalized as the proportion of the agentic words in each statement, using the agentic language dictionary developed by Pietraszkiewicz et al. (2019) for the LIWC linguistic analysis program. They report satisfactory convergent- and divergent validity and strong relationships with subjective ratings of agentic language use. Furthermore, we used a translate and back-translate procedure to develop a similar dictionary in Dutch (available via the first author). The analyses resulted in comparable, but slightly higher proportions of agentic language in statements written in Dutch than in statements written in English (see Table 1, Hedges' g = .29, 95% CI [.14, .43]). All personal statements were checked and corrected for spelling errors and typo's before linguistic analysis.

First year GPA, dropout, gender, and admission test scores were obtained from the university administration. First year GPA was the mean grade obtained in the first year (1–10, with 10 being the highest grade). Dropout was defined as unenrolling during the first year or not re-enrolling for the second year. The admission test score was the raw score on a 40-item psychology test that showed high predictive validity in prior studies (Niessen et al., 2018), comparable to other commonly used admission tests such as the SAT and ACT (Kuncel & Hezlett, 2010; Westrick et al., 2015).

Analytic approach

A proposal for this study was submitted to the editors before the data were collected and analyzed. The project proposal, R code used for analyses, and the Dutch LIWC dictionary file are available on OSF (https://osf.io/b5e2p/). All directional hypotheses were tested using one-tailed tests. For exploratory analyses or when the data indicated effects in the opposite direction than expected, no p-values were computed. Instead, we report descriptive statistics and effect sizes with confidence intervals. Hypotheses that followed from other hypothesized effects were not tested when the former hypothesized effects were not detected.[1]

Table 1. Descriptive statistics.

	All applicants			Enrolled applicants						
				All			Male		Female	
Variable	*M*	*SD*	Range	*M*	*SD*	Range	*M*	*SD*	*M*	*SD*
Personal statement rating – general	3.47	0.51	1.5 – 5.0	3.49	0.50	1.5 – 5.0	3.41	0.56	3.53	0.46
Personal statement rating – structured	2.96	0.52	1.3 – 4.6	2.97	0.52	1.3 – 4.60	2.89	0.55	3.01	0.49
Agentic language (%) – Dutch	4.87	1.73	0.0 – 9.71	4.87	1.73	0.0 – 9.71	4.70	1.73	4.94	1.73
Agentic language (%) – English	4.38	1.72	0.0 – 11.90	4.38	1.76	0.0 – 11.90	4.28	1.68	4.43	1.80
FYGPA				6.44	1.35	1.0 – 9.40	6.33	1.38	6.50	1.33

Note. Agentic language is the percentage of words defined as agentic in the personal statement, FYGPA = first year GPA.

Inter-rater reliability was estimated using intra-class correlations (ICCs, type 2), for single ratings and the average of two raters. Since we cannot verify whether the small differences in agentic language between statements written in Dutch or English reflect true differences, or are caused by factors such as differences in linguistic customs or using different dictionaries, all analyses including agentic language use were conducted separately for personal statements written in English and Dutch. Admission test + personal statement composite scores under different weighting schemes were computed by creating weighted sum scores using the standardized predictor scores.

The personal statements were submitted as part of the admissions procedure, but were not previously used in the admissions process. Therefore, range restriction in personal statement ratings was not expected. For verification, the ratio of standard deviation of enrolled applicants to the standard deviation in the entire applicant pool (u_x) was computed, resulting in $u_x = 0.97$ for general, impressionistic ratings and $u_x > 0.99$ for structured ratings. Hence, range restriction was negligible, so no corrections were applied.

Results

Descriptive statistics of all study variables are presented in Table 1.

Gender differences

To test the hypothesis that male applicants obtained higher personal statement ratings than female applicants when general, impressionistic ratings are made (H1), a one-tailed t-test was planned. However, contrary to this expectation, the descriptive statistics (Table 1) show that female applicants obtained slightly higher personal statement ratings than male applicants (Hedges' $g = .24$, 95% CI [.08, .41]). Since this difference is in the opposite direction than hypothesized, no statistical test was conducted.

To investigate whether male applicants used more agentic language than female applicants (H2), the proportion of agentic words in personal statements written by male and female applicants was compared. Descriptive statistics again show that the difference was in the opposite direction than expected; female applicants used more agentic language than male applicants in statements written in Dutch (Hedges' $g = .14$, 95% CI [–.12, .40]) and English (Hedges' $g = .08$, 95% CI [–.13, .30]), albeit slightly. Again, no statistical tests were conducted. Furthermore, we did not find an association between agentic language use and ratings

of personal statements written in Dutch (r = .04, 95% CI [–.08, .15]) or English (r = –.01, 95% CI [–.10, .08].

The hypotheses that personal statements would result in overprediction of male performance and underprediction of female performance (H3), and that controlling for the use of agentic language would reduce differential prediction (H4), followed from the expectation that male applicants would receive higher personal statement ratings and would use more agentic language. Since we found no evidence for either of those expectations, potential differential prediction could not be explained by the hypothesized mechanism. Therefore, the analyses that were planned to investigate H3 and H4 were not conducted.

Rating structure

To investigate if rating personal statements in a more structured fashion increased reliability and validity compared to using general, impressionistic ratings, inter-rater reliability and predictive validity were computed for the two rating procedures. In agreement with H5, general, impressionistic ratings resulted in low inter-rater reliability (single-rater ICC = .32, 95% CI [.27, .37], average ratings ICC = .49, 95% CI [.42, .54]) and negligible to near-zero predictive validity for first year GPA (r = .09, 95% CI [.01, .16], p = .01, one-sided) and dropout (r = –.01, 95% CI [–.09, .07], p = .40, one-sided).

Ratings of personal statements made using a more structured approach resulted in somewhat higher reliability (H6, single-rater ICC = .50, 95% CI [.35, .61], average ratings ICC = .67, 95% CI [.52, .76]), although reliability was only 'moderate' according to most guidelines (LeBreton & Senter, 2008; Nunnally & Bernstein, 1994). Moreover, the increase in reliability was not accompanied by increased predictive validity for first year GPA (H6, r = .09, 95% CI [.02, .17], p = .01, one-sided) or dropout (r < .01, 95% CI [–.08, .08], p = .49, one-sided), as the correlations were virtually identical.

While gender differences in general ratings were not detected in the expected direction, we still explored gender differences in structured ratings. The results were very similar to those for general ratings, with statements of female applicants rated slightly higher than those of male applicants (Hedges' g = .23, 95% CI [.07, .40]).

Using personal statements in admissions

To demonstrate the effects of adding personal statements to an admission procedure that contains a commonly used valid predictor (i.e., an

Table 2. Admission test – personal statement composite validity under different weighting schemes.

Weighting scheme (%)		Composite *R* FYGPA	Composite *R* dropout
Admission test	Personal statement		
100	0	.45	–.28
90	10	.45	–.28
80	20	.44	–.27
70	30	.42	–.25
60	40	.39	–.22
50	50	.35	–.19
40	60	.30	–.15
30	70	.24	–.11
20	80	.19	–.08
10	90	.13	–.04
0	100	.09	–.01

Note. Dropout was coded 1, retention was coded 0.

admission test), Table 2 shows the validity of an admission test + personal statement composite under different hypothetical weighting schemes that could be used in practice. Since predictive validity and correlations between admission test scores and personal statement ratings were very similar, regardless of rating structure (general ratings: r = .17, 95% CI [.11, .24], structured ratings: r = .15, 95% CI [.08, .22]), the general ratings were used to generate the composite scores.

An optimal, regression-based composite of the admission test scores and personal statement ratings would result in R^2 = .20 for first year GPA, with relative weights analysis (Tonidandel & LeBreton, 2015) resulting in a weight of 98% for the admission test score and 2% for the personal statement rating. For dropout, an optimal composite would result in pseudo R^2 = .08, with relative weights of 99% for the admission test score and 1% for the personal statement. Additionally, weighting the statements 0% yielded identical results. These results and the hypothetical weighting schemes presented in Table 2 show that using the personal statement ratings in admissions procedures would not have improved predictive validity for GPA or dropout at best, and would have been detrimental to validity when receiving a weight above 10%.

Exploratory analyses

To shed some more light on the unexpected findings regarding gender differences, we exploratory investigated if writing skills (as perceived by the raters) could be a possible explanation for the higher ratings for female applicants. Women tend to score higher on writing skills than men (Kaufman et al., 2009; Petersen, 2018; Reilly et al., 2019), also when writing English as a foreign language (Keller et al., 2020). Furthermore, writing skills could have unintentionally influenced the general and

Table 3. Hierarchical regression analyses of personal statement ratings by gender and writing skills ratings.

Predictors	General ratings				Structured ratings			
	B [95% CI]	r_p	R^2	ΔR^2	*B* [95% CI]	r_p	R^2	ΔR^2
Model 1			.01				.01	
Gender	0.12 [0.04, 0.20]	.11			0.09 [−0.01, 0.18]	.07		
Model 2			.11	.10			.07	.06
Gender	0.06 [−0.01, .014]	.06			0.04 [−0.05, 0.13]	.03		
Writing skills	0.22 [0.17, 0.28]	.32			0.21 [0.14, 0.27]	.26		

Note. r_p = partial correlation.

structured non-writing-skills-oriented ratings. Including writing skills as one aspect to be rated in the structured rating form (single-rater ICC = .39, 95% CI [.30, .47], average ratings ICC = .56, 95% CI [.46, .64] allowed an exploration of this possibility. The average writing skills rating across two raters was used in the analyses presented below.

Writing skills ratings were moderately related to general, impressionistic ratings (r = .36, 95% CI [.30, .41]) and structured ratings (with the writing skills item excluded, r = .26, 95% CI [.20, .32]). Furthermore, female applicants were rated slightly higher on writing skills than male applicants (Hedges' g = .34, 95% CI [.17, .50].[2] In addition, exploratory hierarchical regression analyses (Table 3) show that the (already small) relationship between gender and personal statement ratings was substantially reduced when writing skills ratings were included in the models. The regression coefficients for gender reduced by at least half and the partial correlations, showing the relationship between gender and personal statement ratings with writing skills ratings partialled out, were substantially smaller than their zero-order correlations.

Discussion

We found no evidence of adverse impact against female applicants in personal statement ratings, nor that male applicants use more agentic language in their personal statements. The latter is at odds with earlier studies that found that men used more agentic language in personal statements for medical residency programs (Babal et al., 2019; Osman et al., 2015). Furthermore, we found no evidence for the suggested relationship (Woo et al., 2020) between agentic language use and personal statement ratings. Contrary to our expectations, we found that personal statements of female applicants were rated slightly more positively than those of male applicants. Exploratory analyses suggest that these small differences might be explained by better writing skills of female applicants (Kaufman et al.,

2009; Reilly et al., 2019), and the relationship between (perceived) writing skills and personal statement ratings. However, since these analyses were conducted to explore possible reasons for unexpected results, these results should be interpreted very tentatively.

While no evidence for substantial gender bias was detected, the results concerning reliability and predictive validity once again paint a bleak picture for using personal statements in college admissions, at least when rated in the ways we investigated. In line with previous studies (Murphy et al., 2009), general, impressionistic ratings resulted in low inter-rater reliability and predictive validity for first year GPA. Additionally, the near-zero correlations with dropout provide a lack of evidence for the suggestion that personal statements may be more useful for predicting 'fit'-related outcomes than performance-related outcomes (Chiu, 2019; Murphy et al., 2009). Furthermore, rating personal statements in a more structured manner improved inter-rater reliability to some extent, but reliability was still only 'moderate' and we found no evidence for improvements in predictive validity. So, increasing rating structure, at least in the way adopted in this study, does not seem to solve the issues concerning the psychometric quality and utility of personal statement ratings in college admissions.

Finally, our findings align with earlier conclusions that using personal statement ratings derived using an unstructured or a more structured approach for the purpose of predicting academic performance (Murphy et al., 2009), either alone or in combination with a valid admission test, is ill advised. In terms of predictive validity, and hence, making admission decisions, spending time reading and rating personal statements in this way seems a waste of time at best. Moreover, if given substantial weight, which seems to be quite common in practice (Klieger et al., 2017), including personal statements in admissions procedures has a substantial detrimental effect on validity.

Limitations, strengths and future research

We investigated some previously posited hypotheses that had not been tested before. Furthermore, the data allowed estimations of validity and reliability of personal statements that were demonstrably unaffected by range restriction, which is rare when using data obtained in applied settings.

The results presented in this manuscript were based on data obtained from a single undergraduate program at one university, representing one specific personal statement format (a brief format of 250 words), and one level of selectivity (low, in this case), which limits the generalizability of our findings. Therefore, replications with larger and more diverse

samples are strongly encouraged. A question raised by an anonymous reviewer was whether agentic language use would be considered indicative of good 'fit' to a psychology program. While prior research in an organizational setting found that agentic descriptions were related to perceived competence and hireability (Rudman, 1998), we did not find relationships between agentic language use and personal statement ratings. It is possible that agentic language is perceived more favorably in some disciplines than in others. However, we think assuming that using language that demonstrates confidence, competence and ambition would be perceived positively in general, is plausible.

In this study, the raters were unaware of the gender of the applicants while rating the personal statements. Perhaps results would have been different if that had not been the case (Heilman, 2001; Rudman, 1998), and gender-blind rating may not be representative of real-world admissions procedures; names would usually provide quite accurate cues on applicant gender. Furthermore, our study was based on ratings made by a few raters without experience in rating personal statements. This limitation prevents investigating the effects of rater characteristics such as rater gender and experience. For example, it is possible that male and female raters would respond differently to agentic language use, or to statements written by male and female applicants in general. It is also possible that training raters, for example using frame-of-reference training (Roch et al., 2012), would result in higher inter-rater reliability and predictive validity, or would affect the influence of structure on reliability and validity. Future research should shed more light on these possibilities.

The degree of structure in the structured rating process may not have been sufficient to yield reliability and validity gains. The large body of research on the effect of structure on interview reliability and validity (Huffcutt et al., 2014; Levashina et al., 2014) shows that higher degrees of structure, both regarding what information is asked of the applicants and the response rating process, are more beneficial. The highest level of structure is reached when all applicants are asked to answer the exact same questions and each individual answer is rated using formal, anchored rating scales. We were confined to adjusting the structure of the rating process alone, and in the absence of detailed instructions to applicants on what to write about in their statements, the degree to which the rating process could be structured was limited as well (Huffcutt et al., 2013). Even in rating the quite broad dimensions we defined, we already noticed that some applicants did not write about some of them at all, complicating the rating process. Possibly, increasing the amount of structure analogous to the highest level of interview structure would result in more positive results. However, when adopting such an approach,

such personal statements would perhaps best be defined as open-ended biodata measures, or written structured interviews, which is considered a different type of instrument (Murphy et al., 2009).

Additional challenges in ensuring that personal statements would yield valid and fair assessments are the large amount of tips and tricks on writing personal statements available online and the apparent abundance of plagiarism (Shuker, 2014). A notable example is over 200 applicants to colleges in the U.K. in 2007 using the exact same opening sentence in their personal statement ("Ever since I accidentally burnt holes in my pyjamas after experimenting with a chemistry set on my 8th birthday, I have always had a passion for science", Shuker, 2014). Additionally, the difficulties in verifying and controlling the amount and type of support applicants receive in writing their statements is a potential source of bias. Wright and Bradley (2010) found that applicants to UK medical schools from state schools received lower scores on their personal statements than applicants from grammar and independent schools (often populated by students from higher SES backgrounds), but perform equally well in medical school. Applicants from state schools also received less support in writing their personal statements (Wright, 2015).

Furthermore, if (perceived) writing skills affect personal statement ratings, that could result in adverse impact for applicants with a migration or low SES background. While assessing writing skills is often mentioned as one of the attributes assessed in personal statements (Chiu, 2019; Kyllonen et al., 2005), they do not seem to be valid indicators of writing skills (Kuncel et al., 2020; Powers & Fowles, 1997). If assessing writing skills is considered relevant, other, more valid tools should be used.

Conclusion

In short, while our results do not provide support for earlier concerns of adverse impact and bias against female applicants when personal statements are used, the results in terms of validity and reliability were in line with earlier meta-analytic findings (Murphy et al., 2009). Given these findings, we echo earlier advice not to use personal statements in admissions procedures (Kuncel et al., 2020), at least when they are intended to contribute to predicting academic performance, and until evidence is presented that they can yield reliable and valid ratings. Whether the latter is possible remains an open question.

Notes

1. This was the case for differential prediction; the expected gender differences in personal statement ratings that could subsequently result in differential

prediction of academic achievement were not detected. Therefore, differential prediction analyses were not conducted, and the intended analytical approach for those analyses is not described. However, the hypotheses were retained in the paper for transparency.

2. Results were very similar for statements written in Dutch (Hedges' g = .31, 95% CI [.05, .57]) and English (Hedges' g = .35, 95% CI [.13, .56].

Acknowledgment

We thank Lotte Mensink for her help in organizing and collecting the data used for this study.

Disclosure statement

We have no conflicts of interest to disclose.

ORCID

A. Susan M. Niessen http://orcid.org/0000-0001-8249-9295
Marvin Neumann http://orcid.org/0000-0003-0193-8159

References

Babal, J. C., Gower, A. D., Frohna, J. G., & Moreno, M. A. (2019). Linguistic analysis of pediatric residency personal statements: Gender differences. *BMC Medical Education*, *19*(1), 392. https://doi.org/10.1186/s12909-019-1838-x

Chiu, Y. T. (2019). It's a match, but is it a good fit? Admissions tutors' evaluation of personal statements for PhD study. *Oxford Review of Education*, *45*(1), 136–150. https://doi.org/10.1080/03054985.2018.1502168

Clinedinst, M. (2019). *State of college admissions*. The National Association for College Admission Counseling. https://www.nacacnet.org/globalassets/documents/publications/research/2018_soca/soca2019_all.pdf

GlenMaye, L., & Oakes, M. (2002). Assessing suitability of MSW applicants through objective scoring of personal statements. *Journal of Social Work Education*, *38*(1), 67–82. https://doi.org/10.1080/10437797.2002.10779083

Heilman, M. E. (2001). Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. *Journal of Social Issues*, *57*(4), 657–674. https://doi.org/10.1111/0022-4537.00234

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2013). Employment interview reliability: New meta-analytic estimates by structure and format. *International Journal of Selection and Assessment*, *21*(3), 264–276. https://doi.org/10.1111/ijsa.12036

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2014). Moving forward indirectly: Reanalyzing the validity of employment interviews with indirect range restriction methodology. *International Journal of Selection and Assessment*, *22*(3), 297–309. https://doi.org/10.1111/ijsa.12078

Kaufman, A. S., Kaufman, J. C., Liu, X., & Johnson, C. K. (2009). How do educational attainment and gender relate to fluid intelligence, crystallized

intelligence, and academic skills at ages 22-90 years? *Archives of clinical neuropsychology: The official journal of the. Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *24*(2), 153–163. https://doi.org/10.1093/arclin/acp015

Keller, S. D., Fleckenstein, J., Krüger, M., Köller, O., & Rupp, A. A. (2020). English writing skills of students in upper secondary education: Results from an empirical study in Switzerland and Germany. *Journal of Second Language Writing*, *48*, 100700–100713. https://doi.org/10.1016/j.jslw.2019.100700

Klieger, D. M., Belur, V., & Kotloff, L. J. (2017). Perceptions and uses of GRE® scores after the launch of the GRE® revised General Test in August 2011. *ETS Research Report Series*, *2017*, 1–49. https://doi.org/10.1002/ets2.12130

Kuncel, N. R., & Hezlett, S. A. (2010). Fact and fiction in cognitive ability testing for admissions and hiring decisions. *Current Directions in Psychological Science*, *19*(6), 339–345. https://doi.org/10.1177/0963721410389459

Kuncel, N. R., Tran, K., & Zhang, S. H. (2020). Measuring student character: Modernizing predictors of academic success. In M. E. Oliveri & C. Wendler (Eds.), *Higher education admissions practices: A international perspective* (pp. 276–302). Cambridge University Press.

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

LeBreton, J. M., & Senter, J. L. (2008). Answers to twenty questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*(4), 815–852. https://doi.org/10.1177/1094428106296642

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

Max, B. A., Gelfand, B., Brooks, M. R., Beckerly, R., & Segal, S. (2010). Have personal statements become impersonal? An evaluation of personal statements in anesthesiology residency applications. *Journal of Clinical Anesthesia*, *22*(5), 346–351. https://doi.org/10.1016/j.jclinane.2009.10.007

Murphy, S. C., Klieger, D. M., Borneman, M. J., & Kuncel, N. R. (2009). The predictive power of personal statements in admissions: A meta-analysis and cautionary tale. *College and University*, *84*, 83–86.

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2018). Admission testing for higher education: A multi-cohort study on the validity of high-fidelity curriculum-sampling tests. *PLoS One*, *13*(6), e0198746. https://doi.org/10.1371/journal.pone.0198746

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Osman, N. Y., Schonhardt-Bailey, C., Walling, J. L., Katz, J. T., & Alexander, E. K. (2015). Textual analysis of internal medicine residency personal statements: Themes and gender differences. *Medical Education*, *49*(1), 93–102. https://doi.org/10.1111/medu.12487

Ostapenko, L., Schonhardt-Bailey, C., Walling Sublette, J., Smink, D. S., & Osman, N. Y. (2018). Textual analysis of general surgery residency personal statements: Topics and gender differences. *Journal of Surgical Education*, *75*(3), 573–581. https://doi.org/10.1016/j.jsurg.2017.09.021

Petersen, J. (2018). Gender difference in verbal performance: A meta-analysis of United States state performance assessments. *Educational Psychology Review*, *30*(4), 1269–1281. https://doi.org/10.1007/s10648-018-9450-x

Pietraszkiewicz, A., Formanowicz, M., Gustafsson Sendén, M., Boyd, R. L., Sikström, S., & Sczesny, S. (2019). The big two dictionaries: Capturing agency and communion in natural language. *European Journal of Social Psychology*, *49*(5), 871–887. https://doi.org/10.1002/ejsp.2561

Powers, D. E., & Fowles, M. E. (1997). The personal statement as an indicator of writing skill: A cautionary note. *Educational Assessment*, *4*(1), 75–87. https://doi.org/10.1207/s15326977ea0401_3

Reilly, D., Neumann, D. L., & Andrews, G. (2019 May-Jun). Gender differences in reading and writing achievement: Evidence from the National Assessment of Educational Progress (NAEP). *The American Psychologist*, *74*(4), 445–458. https://doi.org/10.1037/amp0000356

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology*, *85*(2), 370–395. https://doi.org/10.1111/j.2044-8325.2011.02045.x

Rudman, L. A. (1998). Self-promotion as a risk factor for women: The costs and benefits of counterstereotypical impression management. *Journal of Personality and Social Psychology*, *74*(3), 629–645. https://doi.org/10.1037/0022-3514.74.3.629

Shuker, L. (2014). It'll look good on your personal statement': Self-marketing amongst university applicants in the United Kingdom. *British Journal of Sociology of Education*, *35*(2), 224–243. https://doi.org/10.1080/01425692.2012.740804

Tonidandel, S., & LeBreton, J. M. (2015). RWA web: A free, comprehensive, web-based, and user-friendly tool for relative weight analyses. *Journal of Business and Psychology*, *30*(2), 207–216. https://doi.org/10.1007/s10869-014-9351-z

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, *140*(4), 1174–1204. https://doi.org/10.1037/a0036620

Westrick, P. A., Le, H., Robbins, S. B., Radunzel, J. M. R., & Schmidt, F. L. (2015). College performance and retention: A meta-analysis of the predictive validities of ACT Scores, high school grades, and SES. *Educational Assessment*, *20*(1), 23–45. https://doi.org/10.1080/10627197.2015.997614

Woo, S. E., LeBreton, J., Keith, M., & Tay, L. (2020). Bias, fairness, and validity in graduate admissions: A psychometric perspective. *Perspectives on Psychological Science*. https://doi.org/10.31234/osf.io/w5d7r

Wright, S. (2015). Medical school personal statements: A measure of motivation or proxy for cultural privilege? *Advances in Health Science Education*, *20*, 627–643. https://doi.org/10.1007/s10459-014-9550-4

Wright, S. R., & Bradley, P. M. (2010). Has the UK Clinical Aptitude Test improved medical student selection? *Medical Education*, *44*(11), 1069–1076. https://doi.org/10.1111/j.1365-2923.2010.03792.x