Some Reflections on the Practical Applicability of GIIU

Jakob Bergman

In Gunnemyr, Mattias & Jönsson, Martin L. (2023) *Post Hoc Interventions: Prospects and Problems.* Lund: Department of Philosophy, Lund University. https://doi.org/10.37852/oblu.184

ISBN: 978-91-89415-60-7 (print)
978-91-89415-61-4 (digital – pdf)
978-91-89415-62-1 (digital – html)

DOI: https://doi.org/10.37852/oblu.184.c503

Published by the Department of Philosophy, Lund University.
Edited by: Mattias Gunnemyr and Martin L. Jönsson
Cover layout by Cecilia von Arnold, Pufendorf Institute for Advanced Studies

3

Some Reflections on the Practical Applicability of GIIU

Jakob Bergman[1]

Abstract. This chapter discusses developments of GIIU in different ways. As the chapter tries to outline possible lines of future research, it is by nature exploratory, on the verge of being speculative. We discuss the issue when the bias depends on the score in a non-linear way and outline a test to detect different type biases of this kind. We also discuss issues where candidates are awarded multiple scores, when and how to apply GIIU.

Introduction

The post-hoc intervention General Informed Interval scale Update (GIIU) was originally suggested by Jönsson and Sjödahl (2017) as GIRU. It was further developed by Jönsson and Bergman (2022) and Jönsson (2022), and applied by Bergman and Jönsson (2022) on a grant application data.

In this paper we discuss some of the limitations of GIIU with focus on its practical applicability, and how these limitations may be mitigated.

Non-Linear Biases

Jönsson and Bergman (2022) state as a presupposition of GIIU that 'the prejudice operates in an approximately linear way' and also give examples of a non-linear bias where e.g. students of a certain ethnic group are always failed (regardless of their performance) or women are never awarded the highest grade(s). This type of bias would obviously be very hard to correct, as there is usually no way of knowing which of the students who should not have been

[1] Jakob Bergman, Senior Lecturer in Statistics at the Department of Statistics, Lund University.

failed or which women among those receiving the highest grade should have received an even higher grade. If auxiliary information is available, e.g. the opinion of a second evaluator or the scores of some other test, this could be used to indicate which individuals might be subject to a biased evaluation score. However, since such a situation with auxiliary information is not the one for which GIIU is designed and we furthermore also expect it to occur rarely in practice, we will not consider it further in this paper.

Another type of non-linear bias can arise when the bias is a function of the evaluation score. It has been shown in the literature that a biased evaluation is more common when there is greater uncertainty in evaluation. If the candidate is clearly very good or very bad, and the score is obvious, there is less room for subjectivity and hence biased assessments. However, if the candidate's performance is (partially) contradictory or ambiguous, it was shown by e.g. Dovidio and Gaertner (2000) and Hodson et al. (2002) that evaluators awarded lower scores to black candidates than white candidates, which was taken as indication of aversive racism. From our point of view, 'ambiguous' would in general mean a mid-range score. The bias is then a function of the score, where the bias is the greatest for mid-range values and smaller (or even non-existent) for the smallest and largest values.

As an example, we assume that scores are awarded as real numbers on scale from one to nine. We also assume that an evaluator is negatively biased towards one group adding a negative bias to the scores of members of that group. We can easily construct two such functions where the bias is greater for scores close to five and smaller for values close to one or nine, using the absolute and squared deviation from five, respectively:

$$bias(Score) = \frac{|Score - 5|}{2} - 2, \quad Score \in [1, 9] \tag{1}$$

$$bias(Score) = \frac{(Score - 5)^2}{2} - 2, \quad Score \in [1, 9] \tag{2}$$

In both cases, the bias functions will equal minus two when the score is five, and zero when the score is one or nine. The difference lies in how rapidly the bias decreases. Table 1 illustrates the two bias functions for integer values from one to nine. Note that for the quadratic function, the bias decreases more slowly than for the absolute deviation.

The type of non-linear biases introduced in (1) and (2), may easily be mitigated using GIIU, if we know the form of the bias, i.e. what the bias function looks like. Without any prior knowledge, this is an extremely hard, if

Table 1: Two examples of non-linear bias, where for each example the bias is the greatest for mid-range scores and smaller for the extreme scores. The two biases exemplified are calculated using (1) and (2).

Score	(1)		(2)	
	Bias	Biased score	Bias	Biased score
1	0	1.0	0	1.000
2	−0.5	1.5	−0.875	1.125
3	−1.0	2.0	−1.500	1.500
4	−1.5	2.5	−1.875	2.125
5	−2.0	3.0	−2.000	3.000
6	−1.5	4.5	−1.875	4.125
7	−1.0	6.0	−1.500	5.500
8	−0.5	7.5	−0.875	7.125
9	0	9.0	0	9.000

not impossible, task. Even if there is prior knowledge or auxiliary information, it is still a difficult task to estimate a bias function, unless one has very detailed knowledge of the form of the function or one knows exactly which individuals that have received biased scores. Otherwise one would need to make very strong assumptions about the distribution of scores within the groups, e.g. that all groups have the same distribution of scores. We find such assumptions to be not very realistic, and hence questionable.

A potential way forward could be to partition relevant social groups into three or more groups according to their scores. Assuming there are two salient social groups, A and B, we would thus create one group consisting of the third of the members of group A with lowest scores, one group of the third of the members of group A with the mid-scores, and one group of the third of the members of group A with the highest scores, and similarly partition the members of group B. We would then compare the mean scores for the A and B groups with the lowest scores, for the A and B groups with mid-scores, and for the A and B groups with the highest scores. If there is no bias, the difference in mean value would be close to zero for all three comparisons. If there is a constant bias, the mean difference would be positive (or negative) and about the same for all three comparisons. And if there is a bias of the type sketched above, we would expect the mean

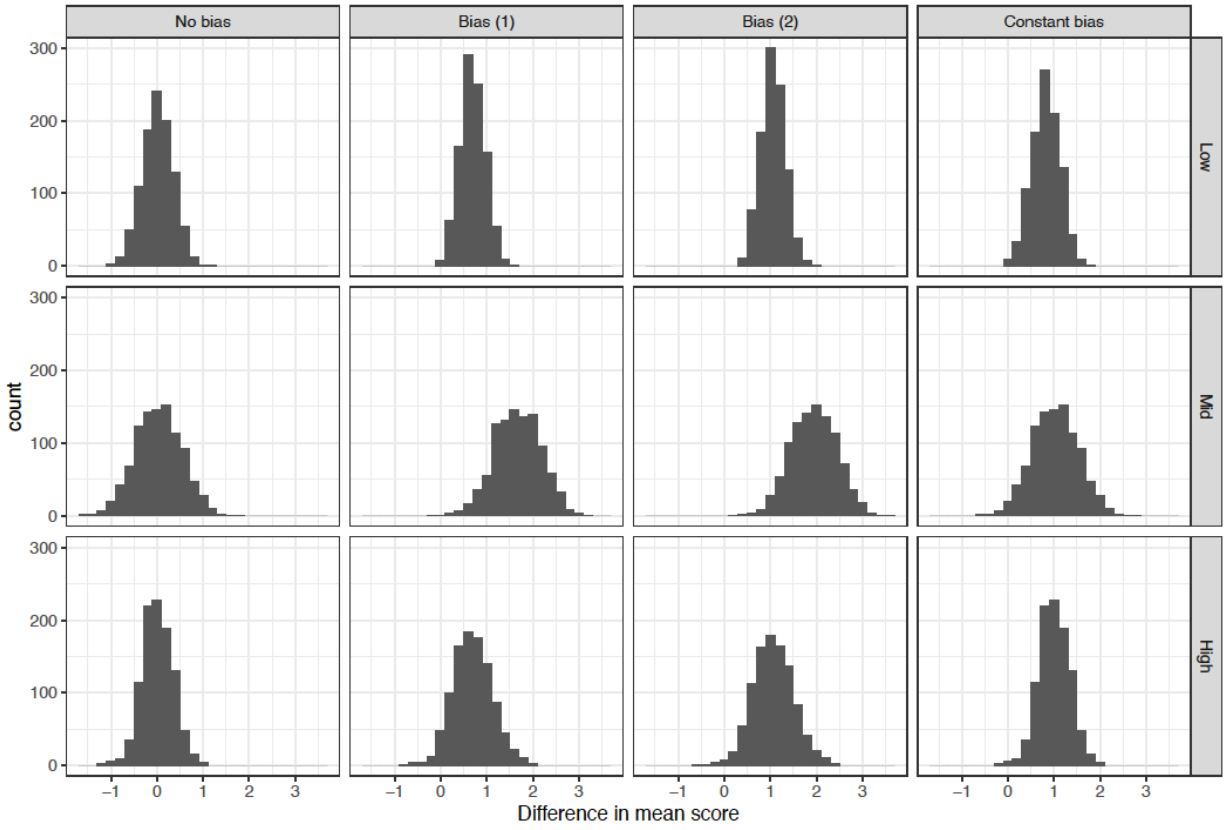

Figure 1: Histograms of the difference in mean scores from 1000 simulations of two groups with 99 candidates each. The top row shows the difference in mean score for the candidates with lowest scores, the middle row shows the difference in mean score for the candidates with the middle scores, and the bottom row shows the candidates with the highest scores.

difference to be greater in the mid-score group and smaller in the two other groups. One might also expect the variation to be smaller in the biased group with the lowest scores, as this group would consist of those candidates with genuinely low scores and those with low scores as a result of bias, thus pushing the scores towards the lower boundary, and conversely the variation to be slightly greater in the biased group with highest scores.

To investigate the feasibility of such a test, we conducted a small simulation study. In the study two groups of 99 candidates were constructed. Each candidate received a random score from a uniform distribution between 1 and 9. We calculated biased scores for the candidates from one of the groups using both (1) and (2). For comparison we also calculated biased scores with a constant bias, by subtracting 1 from all scores for one of the groups. (Biased scores below 1 were set to 1, to stay within score range.) The difference in mean score was then

Table 2: Standard deviations of the difference of the means from simulation study.

Partition	No bias	Bias (1)	Bias (2)	Constant bias
Low	0.3300	0.2641	0.2557	0.2995
Mid	0.4982	0.5020	0.4987	0.4982
High	0.3317	0.4219	0.4379	0.3317

calculated for the 33 candidates with lowest, the middle, and the highest scores, respectively. Figure 1 shows the results of the simulation study. As expected, the mean differences are greater on average for the biased scores. One can see a clear difference between bias 1 and 2 on the one hand, and the constant bias on the other, in that for the latter, the mid-range values have a similar mean as the low and high values, while for the former there is a shift in mean value for the mid-range values. One may also note that the ordering of the values seems to impact the distribution of the mean of the mid-range values the most, for both the unbiased and the biased cases, as this partition has the greatest variation. The standard deviations of the differences in mean scores are presented in Table 2. These may be compared to the standard deviation of the difference of the means of two random samples of 33 observations each from a uniform distribution which is $\sqrt{2(9-1)^2/12/33} = 0.5685$. As may be expected, the ordering reduces the variation, especially for the low and high values. As anticipated, the variation is smaller for the low biased values and slightly increased for the high biased values.

We believe that a test created along these lines could be used to distinguish between different types of biases. However, several important issues remain to be studied. A fundamental task is to find an appropriate test statistic, and to determine its (approximate) distribution under the null and alternative hypotheses. A complicating factor is the fact that the samples are ordered, so the observations are conditioned on being the e.g. smallest third. Relating to this of course also the task to investigate the power of such tests. A more general question is to study the number of partitions. Is the number of partitions dependent on the shape of the bias or is there an optimal number of partitions? How does the number of partitions relate to the size of the history? From a power point of view, one could expect a practical minimum number of observations per partition, but is there a practical maximum number? Should the partitions increase as the size of the history increases?

Multivariate Grades

A situation where one could easily imagine there being ambiguity, is where candidates are assessed in several ways, or on several dimensions, or by several evaluators. In all these cases there will be several scores on which the final evaluation must be based. Jönsson (2019) discusses the need for a stringent and formalised weighting of scores when admitting students to PhD programmes in a Swedish context. It seems reasonable that this would also be the case in general. From a statistical point of view, several scores per candidate is a multivariate (or multidimensional) score.

An important question when applying GIIU, is at what stage one should apply GIIU. Typically, one would apply GIIU to the univariate final scores, but one could also imagine applying GIIU univariately to the underlying scores. The latter would be particularly relevant if the scores are evaluations by different evaluators, where some, but maybe not all, are prejudiced. An alternative to treating each evaluator separately, one could generalize GIIU to a multivariate setting, where the mean difference between the two social groups is assumed to be the null vector 0 (or some other specified vector d). This hypothesis could, assuming multivariate normal distributed scores, be tested using Hotelling's T^2 (a multivariate generalisation of Student's t). This would require finding a relevant set of multivariate functions for updating the scores. A fourth option would be the case, when the scores are (assumed to be) unbiased, but the weighting is biased, i.e. the evaluator uses different weighting functions for different groups. Depending on the circumstances, this could be a type of mixture problem (Aitchison, 1986).

Conclusion

As has been briefly outlined in this chapter, there are a number of potential developments for GIIU. Some of these are possibly application specific, and might even need to be tailored to specific situations. There are also developments which will require more research.

References

Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman and Hall. (Reprinted with additional material in 2003 by Blackburn press.).

Bergman, J. and Jönsson, M. L. (2022). Gender bias in grant applications: inquiry and the potential for a post-hoc remedy. Submitted.

Dovidio, J. F. and Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science*, *11*(4), 315–319.

Hodson, G., Dovidio, J. F., and Gaertner, S. L. (2002). Processes in racial discrimination: Differential weighting of conflicting information. *Personality and Social Psychology Bulletin*, *28*(4), 460–471.

Jönsson, M. (2019). Allt sammantaget den bästa kandidaten: Om möjligheten till en reglerad sammanvägning av meriter som ett sätt att undvika osaklig meritvärdering vid antagning till forskarutbildningen. *Högre utbildning*, *9*(2), 65–80.

Jönsson, M. (2022). On the prerequisites for improving prejudiced ranking(s) with individual and post hoc interventions. *Erkenntnis*, page in press.

Jönsson, M. L. and Bergman, J. (2022). Improving misrepresentations amid unwavering misrepresenters. *Synthese*, *200*.

Jönsson, M. L. and Sjödahl, J. (2017). Increasing the veracity of implicitly biased rankings. *Episteme*, *14*(4), 499 – 517.