Measurement and Research Department Reports 2007-2

Assessing the Size of Halo-effects in Performance-Based Tests and a Practical Solution to Avoid Halo-effects

Timo M. Bechger
Gunter Maris
Ya Ping Hsiao

Measurement and Research Department Reports 2007-2

Assessing the Size of Halo-effects in Performance-Based Tests and a Practical Solution to Avoid Halo-effects

Timo M. Bechger

Gunter Maris

Ya Ping Hsiao

Cito
Arnhem, 2007

Abstract

The main purpose of this paper is to demonstrate how halo-effects may be detected and quantified using two independent ratings of the same persons. A practical illustration is given to show how halo-effects can be avoided.

Key words: rated data; halo-effect; performance-based testing; language testing; classical test theory.

1. Introduction

So-called *productive abilities* (e.g., speaking, writing) that require active behavior of examinees are usually measured via human judgement. That is, examinees demonstrate their ability on a number of assignments and experts are used to assess the quality of their performance. This simple fact gives rise to a myriad of complications. The most conspicuous one being that judges will usually disagree. In this note, we focus on the *halo-effect* which occurs when ratings are influenced by former ratings.

Halo effects can occur for many reasons. For example, judges may form a general impression after having seen a few performances and subsequent judgments may be heavily influenced by this first impression. Some raters may simply stop paying attention to the examinees's performances while others may (unconsciously) be tempted to make subsequent ratings consistent with earlier ratings. This is all speculation, however, for little is known about the complex cognitive processes of human scoring (e.g., Lumley, 2005).

Whatever its cause, when a halo-effect occurs, it implies a decrease in the number of independent opportunities for the candidate the demonstrate his or her proficiency. Thus, presenting a threat to the *reliability* of the examination. In the extreme case, only the first performance is rated and all subsequent ratings are equal to the first which leads to very high correlations between scores on different parts of the examination. However, since only one performance was rated we should not expect to make very precise statements about the examinees ability. In addition, halo-effects may increase the effect of the raters on the examination

marks and harm the *validity* of the exam.

The main purpose of this paper is to demonstrate how halo-effects may be detected and quantified using two independent ratings of the same persons. We start by explaining the basic principles of our approach followed by an application concerning a large-scale language exam in the Netherlands. Note that our approach is based on well-known findings from classical test theory, discussed in detail in general references such as Lord and Novick (1968) or Steyer and Eid (1993).

If a halo-effect can be detected, the next question is how to deal with it. The lack of a good understanding of the mechanisms causing halo-effects complicates solutions based on statistical (item response theory) modeling. In our view, halo-effects are therefore best avoided. The obvious way to do this is to have different raters judge different performances of the same candidate. As an illustration, the penultimate section describes an experiment where raters are assigned *at random* to different combinations of candidates and assignments. The paper ends with a discussion.

2. Theory

2.1. Two Reliabilities

Suppose that an examination has randomly been divided into two half-tests. Based upon a single rating, the two half-test are scored separately for each examinee and the correlation coeficient ρ_1 is computed between these two sets of scores. By construction, the half-tests are parallel or *exchangeable* and ρ_1 equals the reliability of each (Lord & Novick, 1968, Ch. 2). Specifically, if one repeatedly divides the item set in random halfs, the average correlation would equal the reliability.

Using *the Spearman-Brown prophecy formula* (Brown, 1910; Spearman, 1910), the reliability of the full-length examination is calculated as:

$$\rho_{XX'} = \frac{2\rho_1}{1+\rho_1} \tag{1}$$

When each examinee is rated by two independent raters there are two scores for each half-test. Schematically, the correlation matrix between these scores looks like Table 1. It is seen that, by adding a second rating, we now have to choose between *four* different split-half correlations that need not be the same: ρ_1, ρ_2, ρ_4, and ρ_5. The remaining correlation, ρ_3, between two ratings of the same test-half is a measure of rater reliability.

		R_1		R_2	
		T_1	T_2	T_1	T_2
R_1	T_1	1			
	T_2	ρ_1	1		
R_2	T_1	ρ_3	ρ_4	1	
	T_2	ρ_2	ρ_3	ρ_5	1

All these correlations would be affected when there are halo-effects.

Differences between the raters will give rise to differences between ρ_1 and ρ_5 or between ρ_2 and ρ_4. When ratings are exchangeable, there are no such differences and the correlations show the pattern in Table 2.

		R_1		R_2	
		T_1	T_2	T_1	T_2
R_1	T_1	1			
	T_2	ρ_1	1		
R_2	T_1	ρ_3	ρ_2	1	
	T_2	ρ_2	ρ_3	ρ_1	1

Exchangeability of the ratings simplifies matters but we are still left with *two*, possibly different, split-half correlations: ρ_1 and ρ_2. However, differences between ρ_1 and ρ_2 have a simple interpretation. Specifically, a halo-effect will increase the dependencies among ratings by the same rater so that ρ_1 will become larger than ρ_2. Hence,

$$\rho^*_{XX'} = \frac{2\rho_2}{1+\rho_2} \leq \rho_{XX'} \tag{2}$$

with equality when there are no halo-effects. Thus, compared to the $\rho^*_{XX'}$, $\rho_{XX'}$ may be inflated due to halo-effects.

2.2. An Effect-Size Measure

To quantify the size of a halo-effect we employ the general form of the Spearman-Brown formula:

$$\rho_{XX'} = \frac{k\rho^*_{XX'}}{1 + (k-1)\rho^*_{XX'}} \tag{3}$$

where k is the number of is the number of times the test would have to be lengthened to raise $\rho^*_{XX'}$ to the value of $\rho_{XX'}$. Solving for k gives:

$$k = \frac{\rho_{XX'}(1 - \rho^*_{XX'})}{\rho^*_{XX'}(1 - \rho_{XX'})} \tag{4}$$

Thus, the reliability of the test is estimated too large due to halo-effects and k expresses this effect in terms of examination length. We propose to use k as a measure for the size of the halo-effect.

2.3. Estimation and Testing

The correlations ρ_i are population quantities. In practice, we estimate $\rho_{XX'}$ and $\rho^*_{XX'}$ using sample correlations. Unless it is known that the ratings are exchangeable, we would first test whether ρ_1 equals ρ_5, and ρ_2 equals ρ_4. The corresponding sample correlations are based on independent ratings so that standard tests can be used (e.g., Steiger, 2005 and references therein).

When the hypothesis of equal correlations cannot be rejected, we simply average the sample correlations. Hence, ρ_1 is estimated as the average of the two *within-rating* correlations, and ρ_2 is estimated as the average of the two *between-rating, between-halfs* correlations. Formally, this gives ordinary least-squares estimates.

3. Practice

3.1. Halo-effects in the State Examination Dutch as a Second Language

The *State Examination Dutch as a Second Language (STEX)* measures the ability of non-native speakers of Dutch to use and understand Dutch as it is spoken, written, and heard in work and educational settings. The STEX includes separate exams for the productive abilities (speaking and writing), and the receptive abilities (reading and listening). Here, we consider the examination for speaking.

The examination consists of a number of assignments. Each assignment presents the examinee with a practical situation and he or she responds by speaking aloud. The utterances are recorded and send to two independent raters for judgement. The raters are chosen from a file of available raters such that no rater is assigned the same examinee twice. Raters are instructed to listen to the performance on each assignment and answer a set of questions concerning different aspects such as *tempo*, *content* or *vocabulary*.

Each rater passes judgement on all performances of an examinee and there is a real risk that halo-effects occur. To investigate the size of the halo-effect, we took data from the examination administered in July 2006. Parallel test-halfs where constructed by randomly assigning assignments to the two half test forms. The scores were simple sums of the ratings. The resulting correlations are in Table 3.

The pattern of the correlations strongly suggests that the ratings are exchangeable and there is no real need in this case for a statistical

		R_1		R_2	
		T_1	T_2	T_1	T_2
R_1	T_1	1			
	T_2	0.855178	1		
R_2	T_1	0.768225	0.714115	1	
	T_2	0.715133	0.769148	0.855767	1

test.[1] The difference between the estimated $\rho_1 = 0.85547$ and $\rho_2 = 0.71462$ suggests that a halo-effect has occured. It is easily calculated that $\rho_{XX'} = 0.9221$, and $\rho^*_{XX'} = 0.8335$. Using Equation 4, it follows that $k = 2.364$ which means that exam must be made about twice as long. Similar findings were found for examinations administered at other dates.

3.2. Random Assignment of Raters to Examinees

An obvious way to eliminate the halo-effect in the STEX is to have different raters judge the performances of an examinee on different assignments. The findings presented in the previous section served to convince the leadership of the STEX of the need to bring this into practice and a small pilot study was organized to see whether this was feasible.

For the pilot, 50 examinees where drawn from those that took the Juli 2006 examination and their performances were re-rated. On this occasion, rater pairs where randomly assigned to *combinations of exam-*

[1]For completeness, we used Steiger's MULTICORR program (Steiger, 1979) to test the pattern hypothesis of Table 2. This produced a Chi-square statistic of 0.0257 with 3 degrees of freedom.

inees and assignments. Hence, a halo effect cannot occur because different assignments are rated by different raters. The randomness of the assignment ensures that the first and second rating are exchangeable. The correlation matrix is given in Table 4.

		R_1		R_2	
		T_1	T_2	T_1	T_2
R_1	T_1	1			
	T_2	0.757038	1		
R_2	T_1	0.851374	0.796855	1	
	T_2	0.82865	0.878345	0.856656	1

Averaging the relevant correlations, we find that $\rho_1 = 0.8068$ and $\rho_2 = 0.8127$. Hence, $\rho_{XX'} = 0.8931$, and $\rho^*_{XX'} = 0.8967$. In this case, $k = 0.9623$ suggesting a small, negative halo-effect. As an aside, we note that the rater reliability ρ_3 is higher than the rater reliability for the regular examination (Chi-square = 8.7583, d.f.= 2, p= 0.0125). This could be due to the stimulating effect of participating in the pilot.

In this case, we know that a halo-effect cannot have occurred and k differes from 1 due to sampling error. To gain an impression of the sampling variation of k we used a resampling scheme. Specifically, we randomly switched the first and second raters in each of the rater pairs a number of times and each time calculated k. The mean estimated k was equal to 1.022 and the variance was 0.1513. [2]

[2]MS-excel was used to do the calculations.

4. Discussion

We have discussed how a halo-effect can be detected and how its size can be expressed in terms of examination length. It was the following observation that led to this result. On the one hand, one could argue that halo-effects decrease reliablity. On the other hand, it increases an estimate of reliability calculated using a split-half method with single ratings. The contradiction is solved if one considers that Equation 1 is based on the assumption that measurement errors are independent and that this assumption is violated when a halo-effect occurs. When the observed scores are the result of human judgement, measurement errors include variation in the quality of an examinee's performances as well as errors of judgement. When a halo-effect occurs, the latter are positively correlated across different ratings and (1) should no longer be interpreted as a reliability. Note that the problem persists if we use an estimation method based on item covariances such as *Cronbach's alpha* (Cronbach, 1951).

Note that the halo-effect is not synonymous to rater bias. *Rater bias* refers to under or over estimation of the quality of a performance while the halo-effect refers to dependencies between ratings that would not disappear if the ability of the examinee were known. Halo-effects do however increase the probability that raters biases affect the examination marks. In general, we believe that halo-effect should be avoided when possible. In the application this was done by random assignment of raters to examinee-assignment combinations. As an aside, we note that random assignment of raters has many more advantages. In the STEX, for

example, the quality of the ratings given by an individual rater is evaluated by comparison with the second ratings of the same candidates. Random assignment ensures that the contra-ratings are quite literally done by an *average of the available raters* which enhances the validity of the comparison. Further advantages of random assignment of raters are discussed in Maris and Bechger (2007).

References

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*, 296-322.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297-334.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lumley, P. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt: Peter Lang.

Maris, G., & Bechger, T. M. (2007). Scoring open ended questions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, p. 663-680). Amsterdam: Elsevier.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*, 271-295.

Steiger, J. H. (1979). Multicorr: a computer program for fast, accurate, small-sample tests of correlational pattern hypotheses. *Educational and Psychological Measurement*, *39*, 677-680.

Steiger, J. H. (2005). Comparing correlations. In A. Maydeu-Olivares (Ed.), *Contemporary psychometrics. a festschrift for Roderick P. Mcdonald.* Mahwah NJ: Lawrence Erlbaum Associates.

Steyer, R., & Eid, M. (1993). *Messen und Testen*. Berlin: Springer-Verlag.