Scale separation reliability : what does it mean in the context of comparative judgment?

Reference:

Verhavert San, De Maeyer Sven, Donche Vincent, Coertjens Liesje.- Scale separation reliability : what does it mean in the context of comparative judgment?

Applied psychological measurement - ISSN 0146-6216 - (2017), p. 1-18

Full text (Publisher's DOI): http://dx.doi.org/doi:10.1177/0146621617748321

Scale Separation Reliability: What does it mean in the context of Comparative Judgement?

San Verhavert[1], Sven De Maeyer[1], Vincent Donche[1], and Liesje Coertjens[2]

Cite as :

Verhavert, S., De Maeyer, S., Donche, V., and Coertjens, L. (2017), Scale Separation Reliability: What does it mean in the context of comparative judgement. Applied Psychological Measurement (online first). https://doi.org/10.1177/0146621617748321

[1] University of Antwerp
[2] Université Catholique de Louvain, Louvain-la-Neuve

Corresponding Author:
San Verhavert (http://orcid.org/0000-0003-0633-9753), Department of Training and Education Sciences, Faculty of Social Sciences, University of Antwerp, Gratiekapelstraat 10, 2000 Antwerpen, Belgium, +32 32654628, san.verhavert@uantwerpen.be.

Abstract

Comparative Judgement (CJ) is an alternative method for assessing competences based on Thurstone's Law of Comparative Judgement. Assessors are asked to compare pairs of students work (representations) and judge which one is better on a certain competence. These judgements are analysed using the Bradly-Terry-Luce model resulting in logit estimates for the representations. In this context the Scale Separation Reliability (SSR), coming from Rasch modelling, is typically used as reliability measure. But, to our knowledge it has never been systematically investigated if the meaning of the SSR can be transferred from Rasch to CJ. As the meaning of the reliability is an important question for both assessment theory and practice, the current study looks into this. A meta-analysis is performed on 26 CJ assessments. For every assessment split-halves are performed based on assessor. The rank orders of the whole assessment and the halves are correlated and compared with SSR values using Bland-Altman plots. The correlation between the halves of an assessment was compared with the SSR of the whole assessment showing that the SSR is a good measure for split-half reliability. Comparing the SSR of one of the halves with the correlation between the two respective halves showed that the SSR can also be interpreted as an inter-rater correlation. Regarding SSR as expressing a correlation with the truth, the results are mixed.

Keywords

Comparative Judgement (CJ), Scale Separation Reliability (SSR), reliability theory, Rasch measurement, IRT

Introduction

There is a constant need for reliable assessments whether in everyday classroom assessment or high stakes selection procedures in professional contexts. In this context Comparative

Judgement (CJ) has been proposed as an assessment method providing reliable results (Pollitt, 2004, 2009). As the name states, the method of Comparative Judgement is based on comparisons in contrast to absolute judgements. Judges are presented with pairs of students work – further called representations – and are asked to judge which one is better with regard to the competence under assessment. Based on these judgements, done by several judges, a scale value can be estimated.

Already since the early days of CJ as an assessment method it has been considered inefficient, requiring a large number of comparisons to obtain estimates that have an acceptable/good reliability. Or as stated by Bramley, Bell and Pollitt:

> The most salient difficulty from a practical point of view is the monotony of the task and the time it takes to get a sufficient number of comparisons for reliable results. (1998, p. 14)

Therefore, one of the most important methodological questions in CJ to date is: how can the efficiency (in number of comparisons) of a CJ assessment be increased without affecting the reliability of the final estimates?

But how can this question ever be answered if it is not known what the reliability measure means? Because of the similarities between the models behind CJ, IRT and Rasch measurement, the reliability measure used in CJ has been adopted from Rasch measurement (Bramley, 2007, 2015; Pollitt, 2012). Although this is arguable, the differences between the CJ and the Rasch measurement method are substantial enough not to assume that this measure has the same meaning in both contexts (for further details see later). Therefore, the main focus of the current study will be how the reliability measure in CJ can be interpreted.

The first section of this article will be structured as follows. First, what CJ is will be discussed.

In doing so, the theoretical underpinnings leading up to the measurement model and its related reliability measure will be presented. Next, a theoretical framework on reliability and

elaboration on the ways reliability can be estimated will be presented. Finally, the reasoning behind the current study will be discussed.

What is CJ

CJ was introduced in educational assessment in 1995 by Pollitt and Murray who derived the method from Thurstone's Law of Comparative Judgement (Thurstone, 1927a, b). The starting point for this law was the psychophysical observation that an object in the environment or representation (e.g. an essay) has a *psychological impact* in an observer (e.g. assessor) and that this impact or impression can change over time even as the object remains constant. Consequently, any statement (e.g. judgement on the quality) based on this impression will change accordingly (Thurstone, 1927b). This was later formulated again, in educational assessment, by Laming (2003) who stated that an absolute judgement does not exist and that every judgement is a comparison. The latter can be found in Thurstone's derivation of the Law of Comparative Judgement. Thurstone assumed that the psychological impact cannot be observed directly and that if the objects producing these impacts can be ordered based on a certain characteristic, then the corresponding impacts must also follow the same ordering. Therefore, the only way that one can measure the impact is by asking the observer to compare two objects and state which one is better on a certain characteristic (Thurstone, 1927b). For example, assessors are asked to compare two essays on their quality regarding the competence argumentative writing.

If an observer is thus presented with several pairs of representations of which (s)he has to judge which one of the two possesses more of a specified quality, it is possible to estimate from these judgements a scaled location of the representations based on the normal function (Thurstone, 1927b). These estimates are commonly called ability estimates or ability values. Something similar can be found in paired comparison research. There, scale values

are estimated using the Bradley-Terry-Luce model (BTL model; Bradley & Terry, 1952; Luce, 1959) which can be obtained from the original formulation of Thurstone's law after some simplifying assumptions (Thurstone's case V; Thurstone, 1927a) and by substituting the normal function by a logit function (Andrich, 1978). Earlier, Thurstone's law was already identified as "a comparable method of analysis" (Bradley, 1953, p. 32) for paired comparison data. Andrich (2004) also pointed out that although Item Response Theory (IRT) with the two Parameter Logistic model (2PL model) and the Rasch model are conceptually different, Thurstone's Law of Comparative Judgement is considered as the forerunner of both paradigms (Andrich, 2004). All these analysis models are mathematically related, which becomes clear if the BTL model is formulated as follows:

$$p(x_{ij} = 1 | v_i, v_j) = \frac{e^{(v_j - v_i)}}{1 + e^{(v_j - v_i)}} \quad (1)$$

with $x_{ij} = 1$ if representation j is considered better than representation i, v_i and v_j are the estimated ability values, in logit scores, of the respective representations.

If the 2PL model is formulated as follows:

$$p(x_{ij} = 1 | \beta_j, v_i) = \frac{e^{a_i(\beta_j - v_i)}}{1 + e^{a_i(\beta_j - v_i)}} \quad \text{(Birnbaum, 1968)}$$

with β_j the student ability and v_i the item difficulty and a_i the item discrimination parameter, then the Rasch model is merely the 2PL model with the discrimination parameter set to 1 (Birnbaum, 1968).

$$p(x_{ij} = 1 | \beta_j, v_i) = \frac{e^{(\beta_j - v_i)}}{1 + e^{(\beta_j - v_i)}} \quad \text{(Rasch, 1960)}$$

Thus, the person parameter of the 2PL model or the Rasch model is replaced by a second item parameter and the discrimination parameter of the 2PL model is fixed to 1. Despite the mathematical similarity, as proven by Andrich (1978), the BTL model and the IRT and Rasch models clearly have a different parametrization. Therefore, it seems not justifiable to just

copy measures of reliability from Rasch measurement or IRT, as carried out in previous research (Bramley, 2007; Heldsinger & Humphry, 2010; Pollitt, 2012), without checking if their meaning is generalizable to the context of CJ. To our knowledge, these checks have not been done up until now.

The reliability measure in CJ is called the Scale Separation Reliability (SSR), in analogy of the naming in Rasch literature from where the measure was taken over (see Bramley, 2015 for details), and is formulated as follows: (Bramley, 2015)

$$SSR = \frac{G^2}{(1+G^2)} \tag{2}$$

With

$$G = \frac{\sigma_\beta}{RMSE}$$

Where $\sigma_\beta$stands for the standard deviation of the true scores (β) and RMSE is the Root Mean Squared Error.

Reliability Theory

In Classical Test Theory (CTT) reliability is defined as the variance in observed scores that is attributable to true scores (Brennan, 2011; Webb, Shavelson, & Haertel, 2006) or what is assumed as the truth (Brennan, 2011). And although IRT does not entirely conform with CTT (see Brennan, 2011 for a brief discussion and further references) this perspective can also be recognised in IRT and Rasch measurement and thus in CJ.

As Shown in Appendix A the SSR, Equation (2), can be expressed as

$$SSR = \frac{\sigma_\beta^2}{\sigma_v^2} \tag{3}$$

where σ_β^2 and σ_v^2 are the variance of the true scores (β) and the observed/estimated scores (v) respectively. This is the mathematical expression of reliability in CTT. A similar formula

as equation (3) can be found in paired comparison research (Dunn-Rankin, Knezek, Wallace, & Zhang, 2004; Gulliksen & Tukey, 1958).

The variance of the true scores can be estimated from the variance of the observed scores using this formula:

$$\sigma_{\beta}^2 = \sigma_v^2 - MSE$$

where MSE stands for the Mean Squared Error (Andrich, 1982). The σ_v^2 and MSE can be calculated from the person parameters and their standard errors of estimation (resp.) like in Rasch analysis. Similar calculations can be made on estimates of proficiency and their standard errors as obtained in paired comparison research (Dunn-Rankin et al., 2004; Gulliksen & Tukey, 1958).

In practice, reliability can also be estimated from the correlation between two variations of the same assessment or parallel forms (Bramley, 2015; Webb et al., 2006). One way to create these parallel forms, in tests with multiple items, is to split this test in multiple halves on the items and then correlate the respective pairs. The mean of these correlations is then coefficient alpha (Cronbach's alpha; Cronbach, 1951) or the equivalent KR20 for dichotomous items (Webb et al., 2006).

In CJ however, it is impossible to do a split-half on the representations. Doing this could result in a reduction in the overlap between the pairs, which leads to incorrect or even missing ability estimates. As in CJ the assessor group can be seen as an integral part of the results – the judgements of all the assessors are pooled in the analysis – split-halves can be obtained by splitting the assessor group. This approach has already been taken in a few CJ

studies (e.g. Jones, Inglis, Gilmore, & Hodgen, 2013; Jones, Swan, & Pollitt, 2015). However, none of these studies have made the connection to the SSR.

The Current Study

Extending the idea of Jones and colleagues (2013, 2015) this study combines the idea of split-half correlations (on assessors) with the calculation of the SSR to check the interpretation and the validity of this reliability measure in the context of CJ assessments. This is done using an empirical approach.

The current study, investigates the value of three types of reliability in a CJ assessment context: the split-half reliability, inter-rater reliability and reliability as a correlation with the truth. Based on the idea of Jones and colleagues (e.g. 2013, 2015) of triangulating the SSR with split-half correlations as a way to support the reliability of CJ assessments, the meaning of the SSR measure is checked in the current study by directly comparing it with several correlations. Namely, assessments are split in halves and estimated logit scores of the respective halves were correlated. This correlation is then compared with the SSR of the whole assessment providing information on the SSR as split-half reliability. Further, as a CJ assessment can only be split in halves by judge, as argued by Bramley (2015) and demonstrated by Jones and colleagues, information can be obtained on the SSR as inter-rater reliability. This can be done by comparing the SSR measure of the estimates of one of the halves with the correlation between the estimates of the two respective groups. Finally, if one considers the whole assessment as the truth then correlating ability scores of the whole assessment with the scores of one of the halves can support the interpretation of SSR as a correlation with the truth when this correlation is compared to the SSR of the scores of the respective halve. This latter notion was extended in the following way. As the correlation of observed values with the truth is the main idea behind model fit, and the measure for model

fit R^2 is in essence the squared Pearson's r correlation, the squared correlation between the logit scores of the whole assessment and those of one half was compared with the SSR value of the logit scores of the respective half.

It should be remarked that the error variance in CTT is different from that in IRT, because the latter framework does not take item variance into account (Kim, 2012). This might have consequences for the comparability of reliability and Pearson's r correlation measures. Nevertheless, it does not pose a problem for split-half and inter-rater reliabilities as it was shown that the parallel forms reliability in IRT is equivalent to that in CTT (Kim, 2012) and thus with Pearson's r. Differences might arise when correlation with the truth, in other words squared-correlation reliability, is considered (Kim, 2012). In this study, this might lead to biased or inconclusive results.

As a correct interpretation of the reliability measure is methodologically important and practically relevant in assessments, the current study aims to question what the meaning/value is of the Scale Separation Reliability in contexts where Comparative Judgement is used. This is done using an empirical method.

Method

The Data

A meta-analysis is conducted on 15 CJ assessments, 26 assessor groups in total. This difference in numbers is due to how assessments are defined here. One CJ assessment can consist of multiple assessor groups resulting in multiple sets of estimates. In a CJ assessment, representations (products, e.g. essays) are compared regarding a specific competence. In one assessment two competences needed to be judged, resulting in two rank orders. This leads to 27 datasets being used.

Here follows a general description of the assessment characteristics to provide an idea on the range of assessments included in the analysis. For more specific details on the assessments see Appendix B. The majority of the assessments were conducted in higher education (n= 13), followed by secondary education (n= 6) and primary education (n= 1). The remaining assessments were conducted outside the context of education (n= 7). The assessments were conducted with 51 representations on average (min= 6, max= 201) and judged by an average of 28 assessors (min= 4, max= 93). The representations were compared 27 times on average (min= 13, max=105), leading to an average total of 548 comparisons (min= 60, max=2193) per assessment. The assessments resulted in an average SSR of 0.80 (min= 0.62, max= 0.93).

Procedure and Analyses

First the split-half procedure is discussed. Next, it is explained how the correlation coefficients were interpreted in the light of reliability. Afterward, detail are provided on the opportunity provided by some of the data, with regard to interpreting and confirming some results.

Every assessment is split in halves by assessor group in every possible way. For instance, an assessment with 10 assessors results in 126 different possible split-halves of the data. In 54% of the assessments we limit the number of split-halves to 1000 because the number of assessors is too big to be manageable when all split-halves would be obtained. When there is an odd number of assessors one of both split-half groups contains one assessor more than the other.

For every assessment as a whole and every split-half group logit scores are estimated and the SSR is calculated. The logits of the corresponding split-half groups are correlated, as are the logits of each split-half group and those of the corresponding whole assessment. This

leads to three SSR's and three correlation coefficients per assessment per split-half or 55,662 correlations and as much SSR's in total. Per assessment and split-half group the mean of the SSR's and the mean of the correlations is calculated. It is then possible to compare each of these mean SSR's with each mean correlation coefficient as is shown in Figure 1. However, only five of the nine combinations (coloured black) are interpretable. These five combinations can be clustered into three interpretations. (a) If the correlation between the split-half groups is compared with the SSR of the whole assessment (bottom left plot), this provides information on the split-half reliability. (b) In the plots at the bottom middle and right, the correlation between the split-half groups is compared with the SSR of each group separately. Therefore, the correlation can be interpreted as an inter-rater reliability. (c) The reliability as a correlation between what is observed and what is considered as the truth can be found in the top row the middle plot and in the second row the right plot. In these two plots the correlation between one of the groups and the whole assessment is compared with the SSR of the respective group.

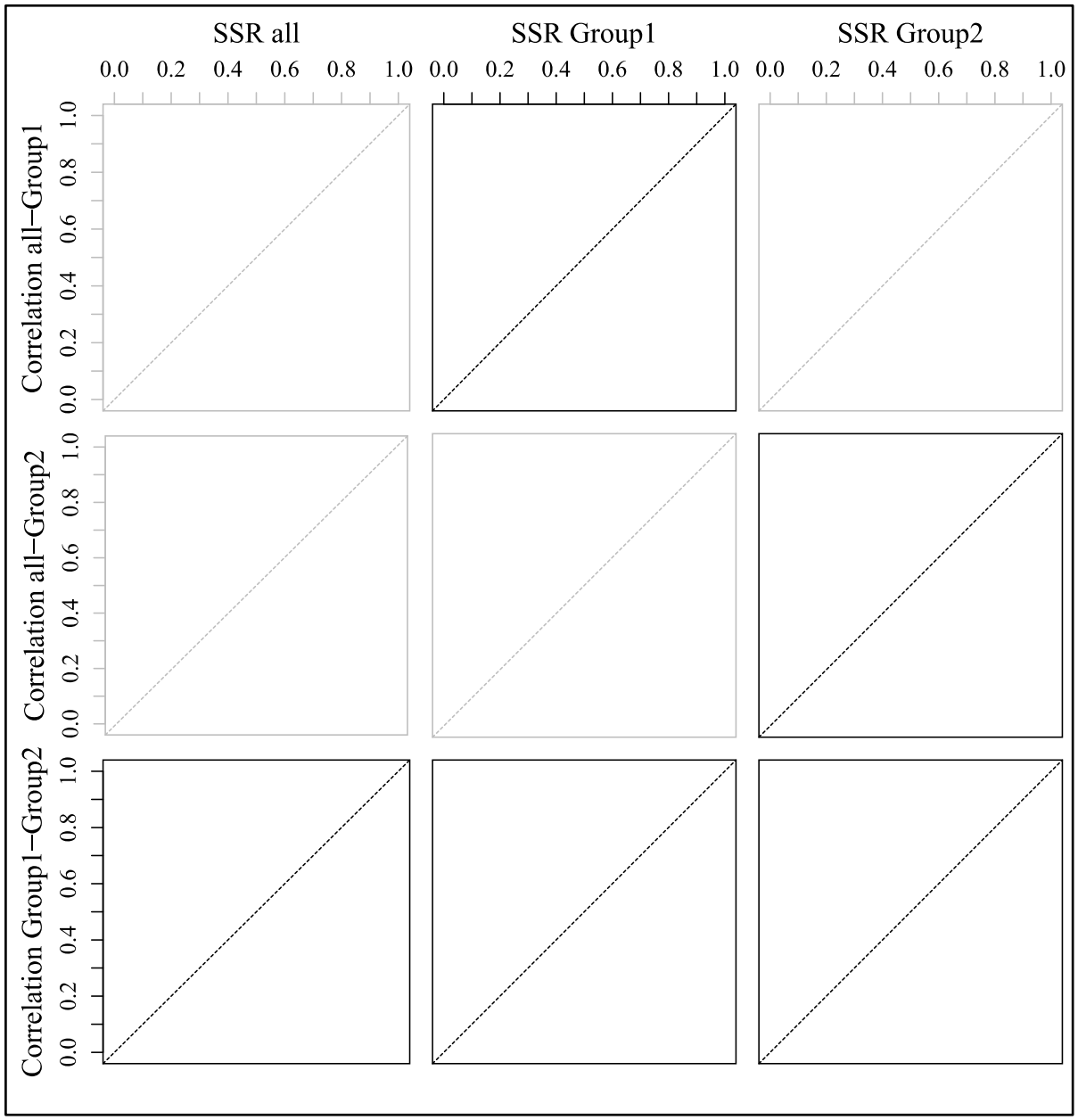

Figure 1. Example of SSR against correlation plot.
Note. SSR = Scale Separation Reliability

Plots of the mean SSR against the mean correlation are hard to interpret. The Bland-Altman plot (BA-plot), or Tukey's mean difference plot, provides more information (Bland & Altman, 1986; Kozak & Wnuk, 2014). In a BA-plot the mean of two values (measures) is plotted against the difference (d) between those values. In our study, this implies that for each of the 27 datasets we calculate the average of the SSR and the respective correlation and plot this against the difference between these two measures. The mean difference $(\bar{d})$ can

now be calculated and it can be assumed that 95% of the *real* difference values lie in the interval between $\bar{d} - 1.95 * \text{sd}$ and $\bar{d} + 1.95 * \text{sd}$. These borders are called the Limits of Agreement (LoA) and are a 95% confidence interval of the deviance between two measures. These LoA boundaries are however influenced by the sample size and thus require an error interval, the most optimistic and pessimistic estimate, based on the estimation error (Barnhart, Haber, & Lin, 2007; Bland & Altman, 1986). These plots can then be interpreted as follows. If zero is outside the LoA boundaries the measures do not agree. If zero is above the LoA then the SSR is an underestimation of the correlation. If zero is smaller, the SSR is an overestimation. When zero is inside de LoA boundaries the most extreme estimates of the LoA should be taken into account. If these are small enough, in absolute value, both measures agree, in that no large discrepancies are possible between the measures. It should only be defined what can be considered large. In terms of correlation, based on interpretations of correlations, an absolute difference $|d| \leq .3$ can be considered small, $.3 < |d| \leq .5$ is large but acceptable and $|d| > .5$ too large. The BlandAltmanLeh R package was used for the analyses (Lehnert, 2015).

It should be remarked that the correlation as inter-rater reliability might be an overestimation. This might also, but to a lesser extent, be the case with split-half reliability. This is inherent to the CJ method. As was noted earlier, the judges are an integral part of the results. This is even more so because the algorithm constructing the pairs takes into account all previous, judged pairs, to not send out the same pair multiple times. Due to this part-dependence of pairs it is impossible to create complete independent halves.

This issue could in part be countered by the setup of some assessments. As can be seen in the table in Appendix B, some assessments (n= 5) were repeated by different assessor groups (2 to 3) thus providing assessment variations as in letting different groups of assessors compare the same representations with the same algorithm. If these variations are correlated,

a more correct estimate of inter-rater reliability can be obtained. This could then provide further support for this interpretation of the SSR.

Pearson's r is used as a correlation measure and the squared Pearson correlation is included as a further support for reliability as model fit. As remarked earlier, the latter should be interpreted with caution as there might be a difference in value between the squared Pearsons's r correlation and the squared-correlation reliability in IRT (Kim, 2012).

Results

In this section, only the results of the Bland-Altman plots are presented. Interested readers can find the plots of the SSR's against the correlations in Appendix C. The results are ordered according to the type of reliability they provide information for. We first focus on split-half reliability, then on inter-rater reliability and eventually on reliability as a correlation with the truth.

To investigate if the SSR could be interpreted as some form of split-half reliability the SSR measure of the whole assessment is compared with the mean of the split-half correlations for that assessment. In the BA-plot (figure 2), zero (black dotted line) is clearly within the LoA (dashed lines) and the most extreme estimates of the LoA (outermost grey dotted lines) are between -.1 and .4 which is just acceptable for correlations. Thus the SSR is a quite good estimate of the split-half correlation.

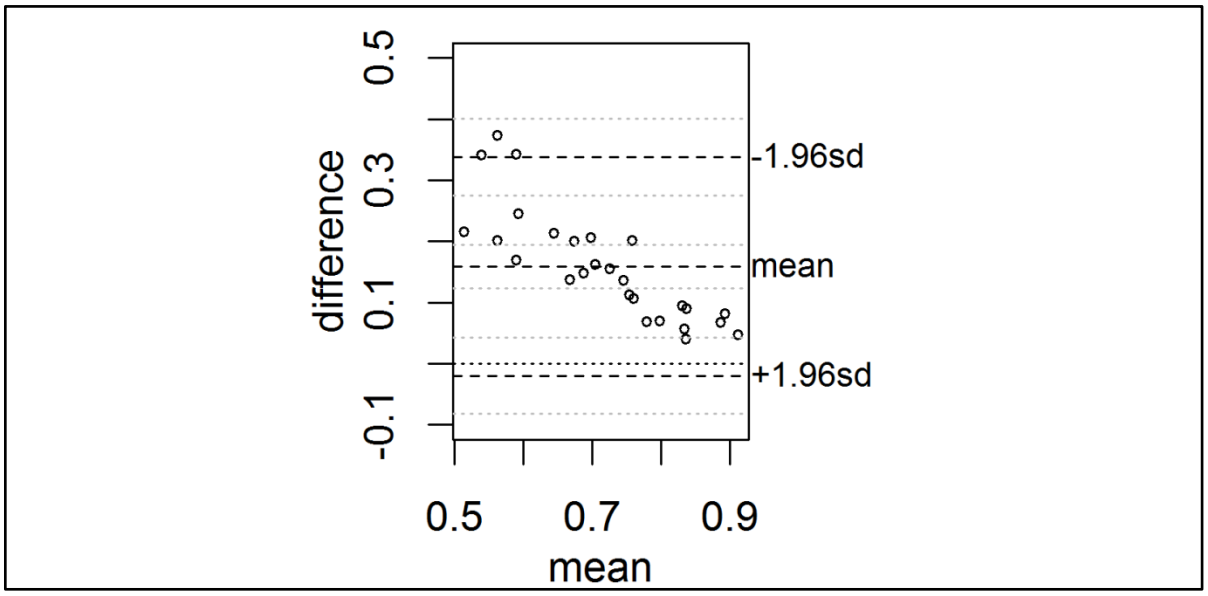

Figure 2. BA-plots for split-half reliability: comparison between the SSR of the whole assessment and Pearson's *r* correlation between both halves of the assessment

Comparing the mean of the correlations between two halves of an assessment and the SSR of one of these halves provides information on the SSR as inter-rater reliability (Figure 3.a., b.). Zero is within the LoA boundaries and the most extreme estimates for these boundaries are -.25 and .25 for the comparison with the group 1 SSR and between -.3 and .3 for the comparison with the group 2 SSR (Figure 3.a. and b.). This can be considered small. Therefore, the SSR could be considered as an inter-rater reliability.

As the split-half groups are not completely independent because of the CJ design, as stated earlier, these correlations might be an overestimate of the true inter-rater correlation. Therefore, within the assessments and if possible, SSR's of separate assessor groups are compared with real inter-rater correlations between these groups. These results confirm the results with the correlation between the split-half groups, as zero is inside the LoA and the extreme estimate boundaries around -.15 and .3 (Figure 3.c.). Again, the SSR appears a good measure for inter-rater reliability.

It should be remarked that the assessor groups are not completely comparable, in number and assessment expertise for all assessments, which could again result in an underestimate of the inter-rater reliability. Therefore it might be possible that these results are an overestimation of the agreement.

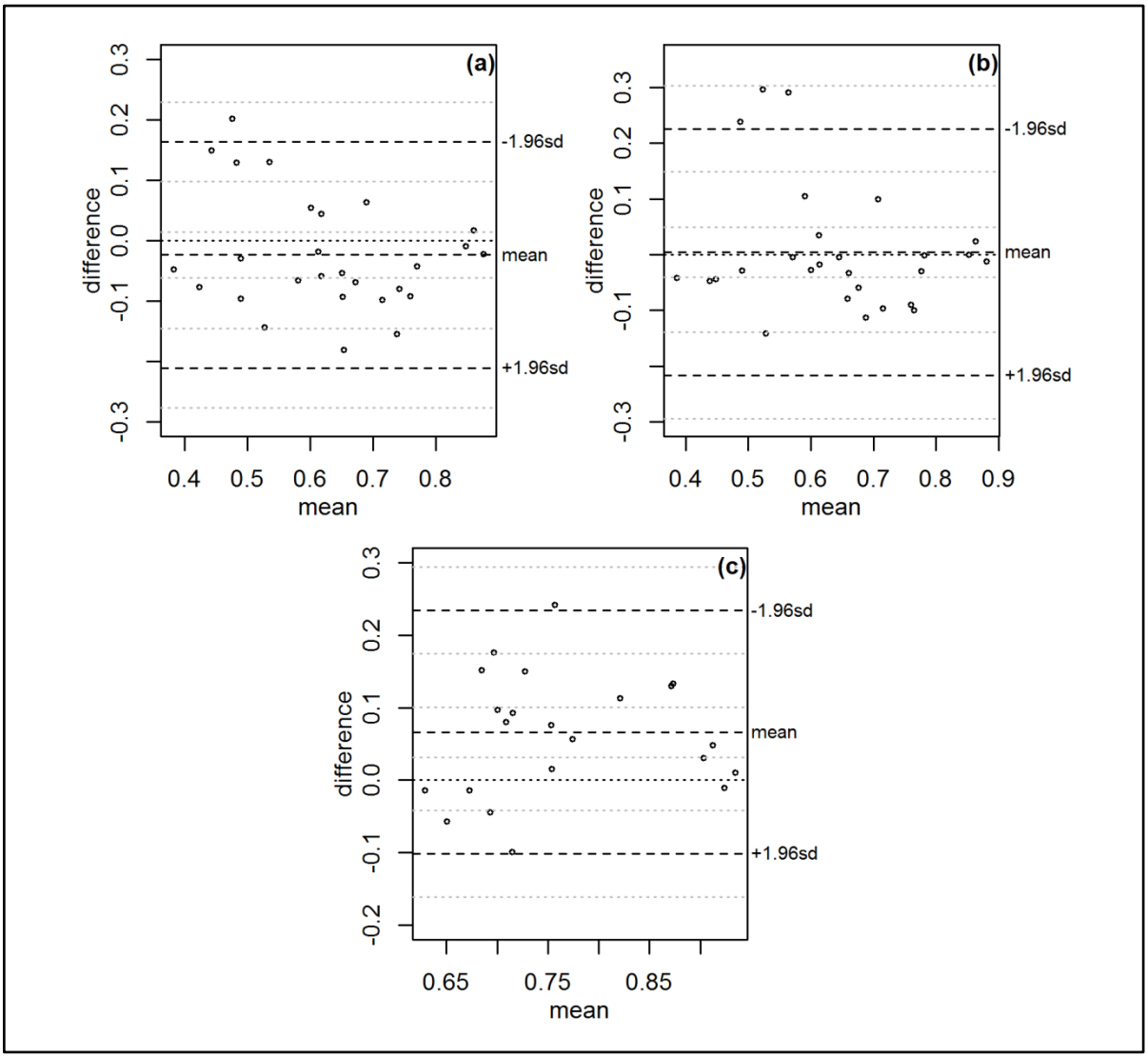

Figure 3. BA-plots for inter-rater reliability: comparisons between the SSR of group 1 and the correlation between the two split-half groups (a), between the SSR of group 2 and the correlation between the two split-half groups (b) and between the SSR of selected assessments and the true inter-rater correlation (c).

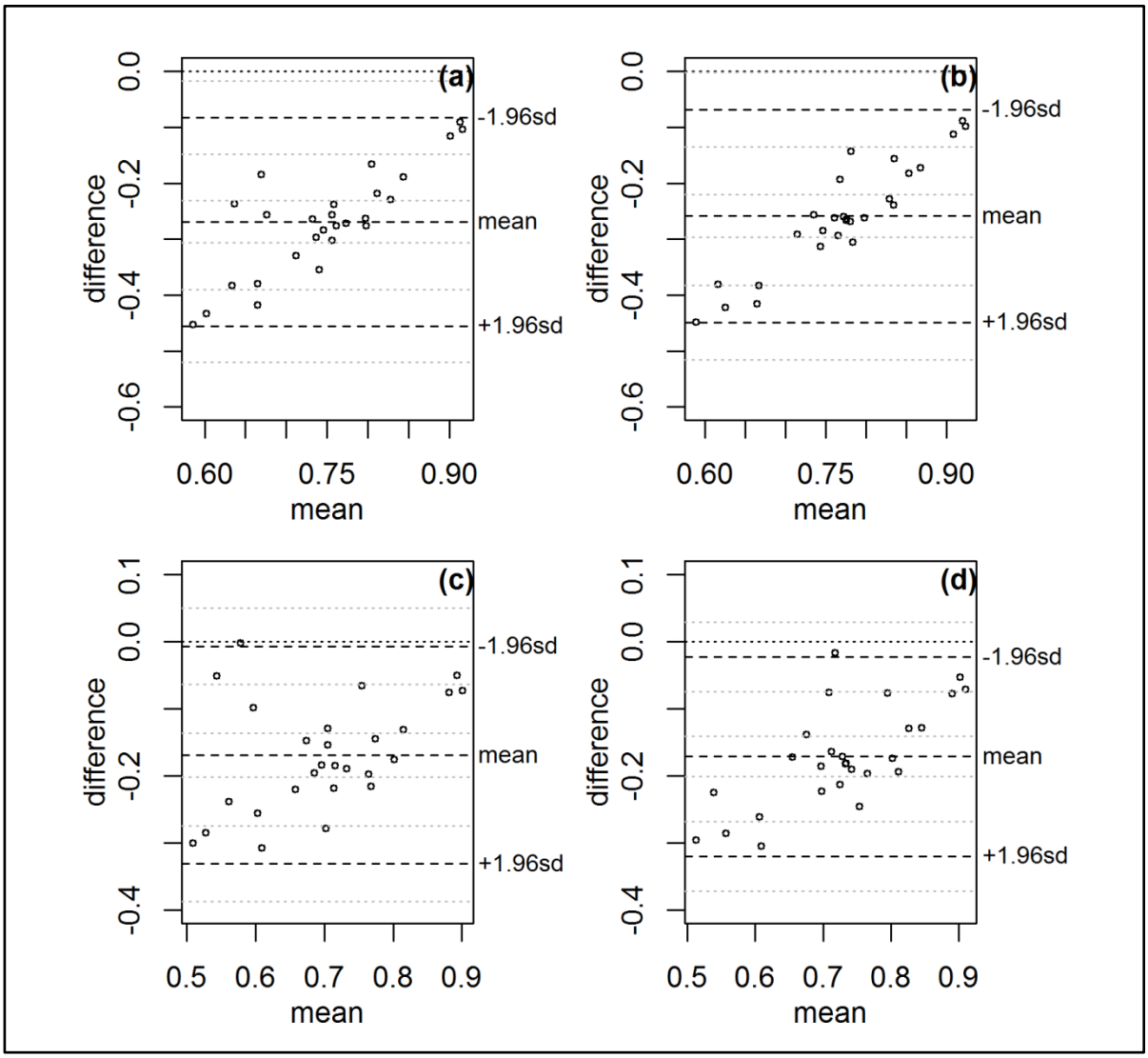

Figure 4. BA-plots for reliability as correlation with truth: Comparisons between the SSR of group 1 and the correlation between group 1 and the whole assessment (a) and (c) and the SSR of group 2 and the correlation between group 2 and the whole assessment (b) and (d). Plots (a) and (b) display Pearson's correlations and plots (c) and (d) the R^2.

For reliability as a correlation with the truth, the SSR of each half is compared with the correlation between the whole assessment and the respective half. If the difference between the SSR of either one of the groups and the correlation between the whole assessment and the respective group is considered (Figure 4.a. and b.) the LoA's concerning either group are below zero. This shows that these SSR's are underestimates of the respective

correlations. But, as argued, the correlation between observed values and the truth is better expressed by the measure of model fit (R^2). Hence looking at the difference between the SSR's of either group and the squared Pearson correlation between the whole assessment and the respective group (Figure 4.c. and d.), the results prove difficult to interpret. The zero lies above the LoA boundaries but still within the estimation error boundaries of the LoA's. The most extreme boundaries are still within the acceptable limits of around -.1 and around .3. It can be cautiously concluded that the SSR might be a good estimate for correlation with the truth but there is not enough data to be certain. On the other hand, one can expect these results as values of the SSR calculations might not completely correspond to the squared correlation values, as remarked earlier. This might also provide an explanation of the inconclusive results with the squared Pearson's r.

Discussion

The SSR measure from CJ has been adopted from Rasch measurement because of the algebraic similarity of the measurement models. However, the method of CJ and Rasch measurement are different enough not to assume that the reliability measures mean the same in both contexts. Therefore this study set out to answer the question how the SSR can be interpreted, more specific what the meaning is of the SSR in the context of CJ. Therefore a meta-analysis was conducted on 27 datasets (5 assessments or 26 assessor groups; see Appendix B). Using a split-half methodology SSR values were compared with several types of correlation using BA plots and corresponding LoA's. The assessments are diverse enough and the data set large enough that some generalising statements can be made. However, as this study set out to investigate the meaning of the SSR measure in CJ, it has to be remarked that these conclusions cannot be generalized to Rasch and IRT. Furthermore, it should be kept in mind that the analyses were conducted on the data from a set of 27 specific

assessments. Therefore it is necessary that the findings are replicated with more experimental studies.

The results strongly point in the direction that the SSR reflects the inter-rater reliability as the SSR of each split-half group shows congruency with Pearson's *r* correlation between both groups. However, it has been remarked that this correlation might be an overestimate of the inter-rater reliability because the groups are not independent. Therefore the confirmation was sought in assessments with different assessor groups. Here the SSR's were also close to Pearson's *r*. As these assessments were not set up to test inter-rater correlations however, the assessor groups were not constructed to be equivalent, so the correlations could be an underestimate of the potential inter-rater correlation. In sum, there are good and strong indications that the SSR reflects the inter-rater correlations but some results call for caution. These results should be further confirmed with a more experimental and controlled approach.

Regarding the most theoretical view on reliability, namely the correlations of the observed values with the truth, the SSR of each group differs from Pearson's r correlation. This could be due to the fact that reliability as a correlation with the truth is better reflected by the squared Pearson correlation or the measure of model fit (R^2). The squared Pearson correlation values indeed appear to lie closer to the corresponding SSR values. Cautiousness is however warranted. The results present a *borderline* case meaning there is not enough data to provide enough certainty over the results. Also, difference in conceptualization between CTT and IRT (Kim, 2012) might contribute to the fact that these values do not completely correspond. We can tentatively conclude that there is some evidence and a slight confirmation that the SSR might be interpreted as a theoretical reliability, namely a correlation with the truth.

Finally, there was also evidence that the SSR expresses split-half reliability. The SSR of the whole group appears not that different from Pearson's *r* correlation between both split-half groups. Evidence is thus pointing in the direction of the SSR as a split-half reliability but further research is needed.

It can be concluded that there are strong indications that the SSR provides an inter-rater reliability index which can be informative when using CJ. Some results also point in the direction of the SSR as a correlation with the truth and/or a split-half correlation. However, these indications are less strong and further research is recommended. Studies conducting assessments with a higher control on the equivalency of assessor groups are important to conduct as well as assessments where the rank order is known beforehand might provide some interesting findings.

The findings of this meta-analysis, based on a substantial yet specific sample of assessments, provide a first step toward a strong theoretical basis for the interpretation of CJ results. As this is study takes an empirical approach, these results need to be confirmed in more systematic studies.

These results provide initial information in the search toward adaptive algorithms to increase the efficiency of the CJ method. Even further, these results might give inspiration in the analyses of future simulation studies on these algorithms. Besides, this study reaches the assessment practice some handles to interpret the results of their CJ assessment.

Regarding the use of CJ in the assessment practice the efficiency question is an equally important methodological question with has important practical implications. This question also cannot be answered if it is unknown *how many* comparisons are actually needed to reach a certain level of reliability any way. This article focused on the basic

methodological and theoretical question of the meaning of the reliability, and future research is needed to question the numbers of comparisons actually needed.

Acknowledgment

The author thanks the two anonymous reviewers whose comments helped improve the manuscript.

The majority of the data was collected within and outside the University of Antwerp and with the cooperation of the following persons: Prof. dr. Kris Aerts, Cynthia De Bruycker, Benedicte De Winter, Ann-Kathrin Hennes, Stefan Martens Prof. dr. Nele Michels, Prof. dr. Jean-Michel Rigo, dr. Pierpaolo Settembri, dr. Joke Spildooren, Daniëlle Van Ast, Tine van Daal, Marie-Thérèse van de Kamp, Kristel Vandermolen, Kristof Vermeiren, and Ellen Volkaert. We would like to thank these people for their efforts.

Funding

This research is part of a larger project (D-PAC) funded by the Flanders Innovation & Entrepreneurship and the Research Foundation (grant number 130043).

References

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, *2*, 451–462. https://doi.org/10.1177/014662167800200319

Andrich, D. (1982). Index of Person Separation in Latent Trait Theory, the traditional KR-20 index, and the Guttman Scale response pattern. *Education Research and Perspectives*, *9*(1), 95–104.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, *42*(Suppl. 1), I7–I16. https://doi.org/10.1097/01.mlr.0000103528.48582.7c

Barnhart, H. X., Haber, M. J., & Lin, L. I. (2007). An overview on assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics*, *17*, 529–569. https://doi.org/10.1080/10543400701376480

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores (pp. 395-479). Reading, MA: Addison-Wesley.

Bland, M. J., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *327*, 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8

Bradley, R. A. (1953). Some statistical methods in taste testing and quality evaluation. *Biometrics*, *9*, 22–38. https://doi.org/10.2307/3001630

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*, 324–345. https://doi.org/10.1093/biomet/39.3-4.324

Bramley, T. (2007). Paired comparison methods. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 246–300). London, U.K.: Qualifications and Curriculum Authority.

Bramley, T. (2015). *Investigating the reliability of adaptive comparative judgement* (Cambridge assessment research report). Retrieved from www.cambridgeassessment.org.uk/Images/232694-investigating-the-reliability-of-adaptive-comparative-judgment.pdf

Bramley, T., Bell, J. F., & Pollitt, A. (1998). Assessing changes in standards over time using Thurstone Paired Comparisons. *Education Research and Perspectives*, *25*(2), 1–24.

Brennan, R. L. (2011). Generalizability Theory and Classical Test Theory. *Applied Measurement in Education*, *24*, 1–21. https://doi.org/10.1080/08957347.2011.532417

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. https://doi.org/10.1007/BF02310555

Dunn-Rankin, P., Knezek, G. A., Wallace, S. R., & Zhang, S. (2004). *Scaling Methods* (2 edition). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Gulliksen, H., & Tukey, J. W. (1958). Reliability for the Law of Comparative Judgment. *Psychometrika*, *23*, 95–110. https://doi.org/10.1007/BF02289008

Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, *37*(2), 1–19. https://doi.org/10.1007/BF03216919

Jones, I., Inglis, M., Gilmore, C., & Hodgen, J. (2013). Measuring conceptual understanding: The case of fractions. In A. M. Lindmeier & A. Heinze (Eds.), *Proceedings of the 37th Conference of the International Group for the Psychology of Mathematics Education (PME 37)*, 3, 113–120. Kiel, Germany: International Group for the Psychology of Mathematics Education.

Jones, I., Swan, M., & Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education*, *13*, 151–177. https://doi.org/10.1007/s10763-013-9497-6

Kim, S. (2012). A note on the reliability coefficients for Item Response Model-based ability estimates. *Psychometrika*, *77*, 153–162. https://doi.org/10.1007/s11336-011-9238-0

Kozak, M., & Wnuk, A. (2014). Including the Tukey mean-difference (Bland–Altman) plot in a statistics course. *Teaching Statistics*, *36*, 83–87. https://doi.org/10.1111/test.12032

Laming, D. (2003). *Human judgment: The eye of the beholder* (1st ed.). London: Cengage Learning EMEA.

Lehnert, B. (2015). BlandAltmanLeh: Plots (Slightly Extended) Bland-Altman Plots. (Version 0.3.1) [R package]. Retrieved from CRAN.R-project.org/package=BlandAltmanLeh

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York, NY: Wiley.

Pollitt, A. (June 2004). Let's stop marking exams. Presented at the IAEA Conference, Philadelphia, PA.

Pollitt, A. (September 2009). Abolishing marksism and rescuing validity. Presented at the IAEA Conference, Brisbane, Australia.

Pollitt, A. (2012). Comparative judgement for assessment. *International Journal of Technology and Design Education*, *22*, 157–170. https://doi.org/10.1007/s10798-011-9189-x

Pollitt, A., & Murray, N. L. (1995). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Studies in language testing 3: Performance testing, cognition and assessment* (pp. 74–91). Cambridge, U.K.: Cambridge University Press.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from www.R-project.org

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Thurstone, L. L. (1927a). A law of comparative judgment. *Psychological Review*, *34*, 273–286. https://doi.org/10.1037/h0070288

Thurstone, L. L. (1927b). Psychophysical analysis. *The American Journal of Psychology*, *38*, 368–389. https://doi.org/10.2307/1415006

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). 4. Reliability coefficients and Generalizability theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26, pp. 81–124). Amsterdam, The Netherlands: Elsevier.

Appendix A

Proof that the SSR $= \frac{trueSD^2}{obsSD^2}$

$$\text{SSR} = \frac{G^2}{(1+G^2)}$$

With

$$G = \frac{\sigma_\beta}{RMSE}$$

Fill out G in α

$$\text{SSR} = \frac{\left(\frac{\sigma_\beta}{RMSE}\right)^2}{\left(1+\left(\frac{trueSD}{RMSE}\right)^2\right)}$$

$$\text{SSR} = \frac{\sigma_\beta^2}{RMSE^2\left(1+\frac{\sigma_\beta^2}{RMSE^2}\right)}$$

If $RMSE = \sqrt{MSE} = \frac{\sum_i SE_i^2}{n}$ then $RMSE^2 = MSE$

$$\text{SSR} = \frac{\sigma_\beta^2}{\left(MSE + \cancel{MSE}\frac{\sigma_\beta^2}{\cancel{MSE}}\right)}$$

If $\sigma_\beta^2 = \sigma_v^2 - MSE$ then $\sigma_v^2 = \sigma_\beta^2 + MSE$

$$\text{SSR} = \frac{\sigma_\beta^2}{\sigma_v^2} \quad \blacksquare$$

Appendix B

Table with Assessment Details

Table B1. *Assessment details*

Assessment[a]	Domain	Assessor Group[b]	N_A[c]	N_R	N_{CR}	N_{CT}[d]	SSR
Argumentative Writing: Having Children	Secondary Education	Teachers	55	135	18	1224	0,81
Argumentative Writing: Organ Donation	Secondary Education	Teachers	52	136	13	890	0,74
Argumentative Writing: Stress of Students	Secondary Education	Teachers and Students Teacher Training	42	35	27	474	0,88
Visual Skills (visual arts)	Secondary Education	Arts Teachers	12	147	27	2193	0,86
Debriefing Notes Political Negotiation	Higher Education	Professors European Politics	4	84	15	622	0,72
Job Selection (CV Screening)	Job Application	HR Consultants	7	42	22	463	0,88
Job Selection (CV Screening)	Job Application	Students Industrial Psychology Group1[e]	51	42	15	308	0,62
Job Selection (CV Screening)	Job Application	Students Industrial Psychology Group2[e]	50	42	15	306	0,66
Narrative Writing	Primary Education	Students Teacher Training	40	201	20	2000	0,83

Note. $N_{A\,=}$ number of assessors; N_R = number of representations; N_{CR} = number of comparisons per representation; N_{CT} = number of comparisons in total.

[a]Assessments with the same name had the same representations but a different assessor group.

[b]We specifically distinguish between student groups and peers because not all assessments were peer assessments.

[c]Approximately; the number of comparisons for the majority of representations.

[d]Because of note [c] and missing data this might not be completely equal to (number of representation * comparisons per representation) / 2

[e]Official title of program: Personnel Management and Industrial Psychology.

[f]Official title of program: Physical Medicine and Rehabilitation.

[g]Official title of program: Training and Education Sciences.

Table B1. (Continued)

Assessment[a]	Domain	Assessor Group[b]	N_A[c]	N_R	N_{CR}	N_{CT}[d]	SSR
Entity Relationship Models	Higher Education	Professors Engineering Science	4	30	15	228	0,76
Entity Relationship Models	Higher Education	Peers (Students Engineering Science)	28	30	19	280	0,79
Paper Evidence Based Diagnostics	Higher Education	Peers (Students Rehabilitation[f])	93	93	21	969	0,81
Advanced Quantitative Methods	Higher Education	Peers (Students Education Sciences[f])	30	44	20	424	0,81
Qualitative Interview Techniques	Higher Education	Students Education Sciences Group1[g]	42	10	105	525	0,93
Qualitative Interview Techniques	Higher Education	Students Education Sciences Group2[g]	41	9	79	356	0,93
Qualitative Interview Techniques	Higher Education	Students Education Sciences Group3[g]	41	9	78	357	0,92
Mood Boards	Higher Education	Professors Interior Architecture	5	20	20	200	0,75
Mood Boards	Higher Education	Peers Students Interior Architecture Group1	16	20	18	180	0,8

Note. N_A = number of assessors; N_R = number of representations; N_{CR} = number of comparisons per representation; N_{CT} = number of comparisons in total.

[a]Assessments with the same name had the same representations but a different assessor group.

[b]We specifically distinguish between student groups and peers because not all assessments were peer assessments.

[c]Approximately; the number of comparisons for the majority of representations.

[d]Because of note [c] and missing data this might not be completely equal to (number of representation * comparisons per representation) / 2

[e]Official title of program: Personnel Management and Industrial Psychology.

[f]Official title of program: Physical Medicine and Rehabilitation.

[g]Official title of program: Training and Education Sciences.

Table B1 (Continued 2)

Assessment[a]	Domain	Assessor Group[b]	N_A[c]	N_R	N_{CR}	N_{CT}[d]	SSR
Mood Boards	Higher Education	Peers Students Interior Architecture Group2	19	20	22	224	0,76
Project Proposals Educational Innovation	Jury	Jury Educational Innovation	5	6	20	60	0,71
Selection Headmaster Training (Group1 Selection1)	Jury (Professional Development)	Selection Committee Group1	6	20	20	204	0,80
Selection Headmaster Training (Group1 Selection2)	Jury (Professional Development)	Selection Committee Group1	6	20	20	204	0,75
Selection Headmaster Training (Group2)	Jury (Professional Development)	Selection Committee Group2	14	16	16	130	0,81
Mathematical Problem Solving Task1	Secondary Education	Teachers Secondary Education Maths	14	58	20	588	0,86
Mathematical Problem Solving Task2	Secondary Education	Teachers Secondary Education Maths	14	58	18	518	0,86
Self-reflection Internship Medicine	Higher Education	Professors Medicine	9	22	19	206	0,77
Self-reflection Internship Medicine	Higher Education	Laymen	35	22	30	328	0,67

Note. $N_{A=}$ number of assessors; N_R = number of representations; N_{CR} = number of comparisons per representation; N_{CT} = number of comparisons in total.

[a]Assessments with the same name had the same representations but a different assessor group.

[b]We specifically distinguish between student groups and peers because not all assessments were peer assessments.

[c]Approximately; the number of comparisons for the majority of representations.

[d]Because of note [c] and missing data this might not be completely equal to (number of representation * comparisons per representation) / 2

[e]Official title of program: Personnel Management and Industrial Psychology.

[f]Official title of program: Physical Medicine and Rehabilitation.

[g]Official title of program: Training and Education Sciences.

Appendix C

Plots of SSR against correlation

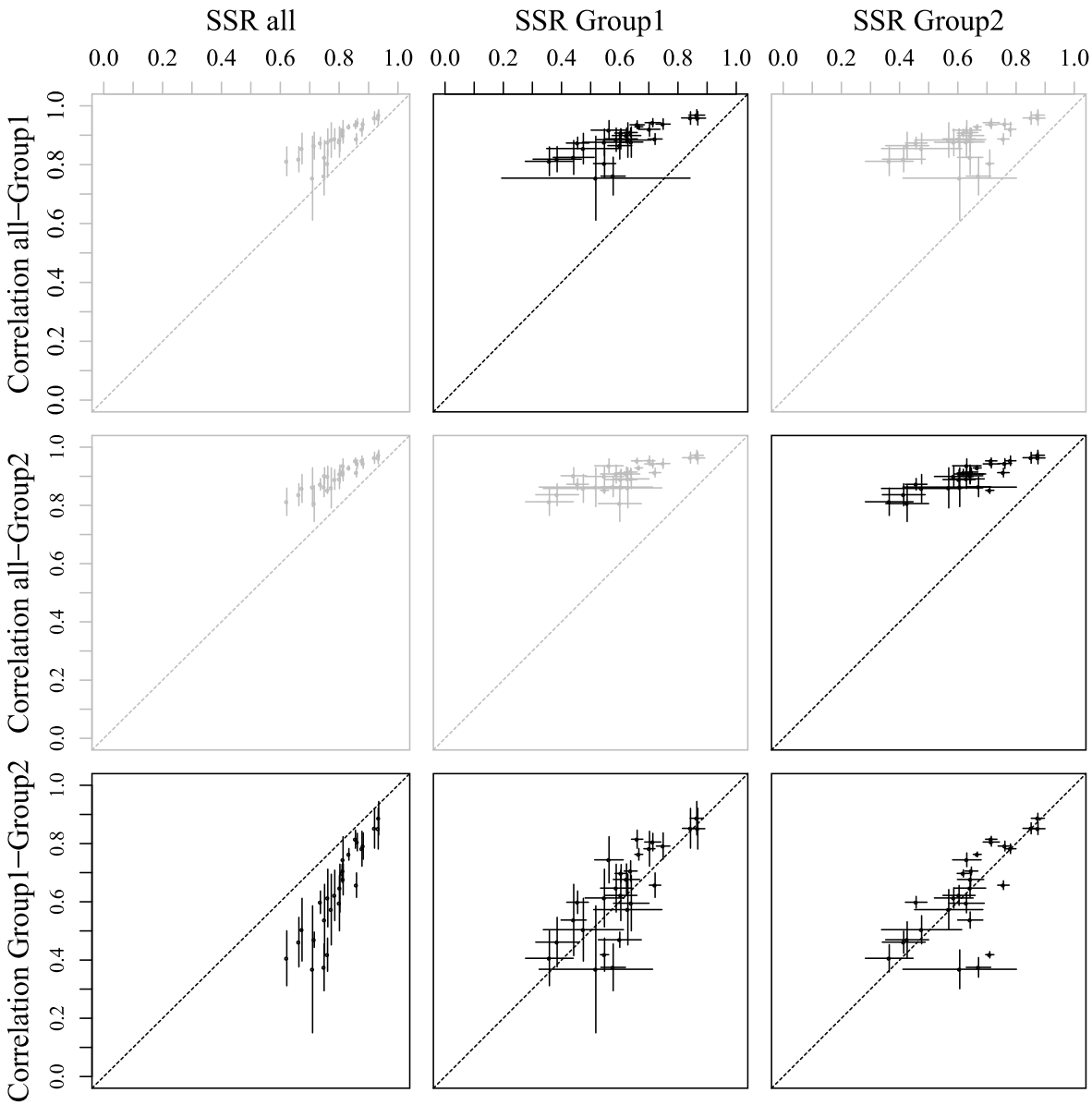

Figure C1. Plot of Scale Separation Reliabilities (SSR) against Pearson's r correlations.

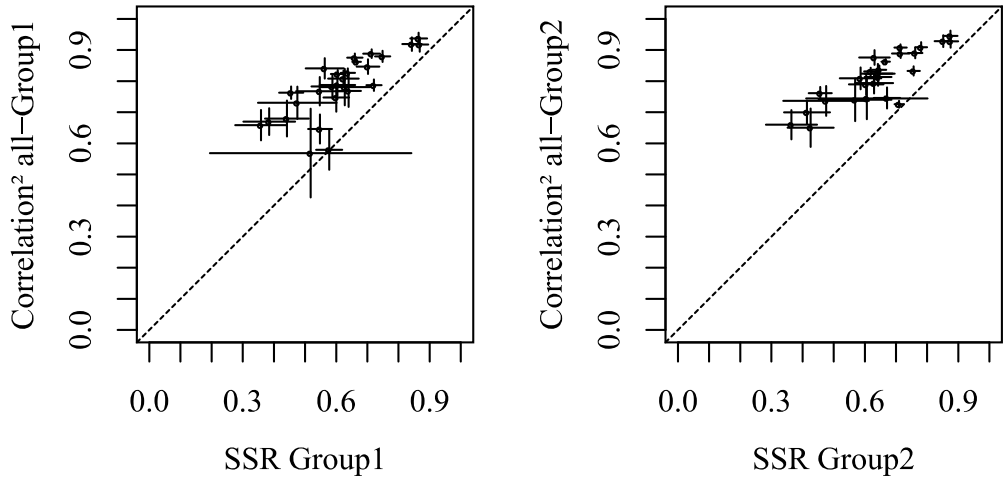

Figure C2. Plot of Scale Separation Reliabilities (SSR) against Pearson's r correlations squared.