**Scoring of Divergent Thinking Tests: A Review and Systematic Framework –**

**Online Supplemental Material**

**An Illustration of Various Scorings**

In this supplemental material we present Table S1, which includes many different scoring approaches for the exact same hypothetical responses that could have been given as responses to an Alternate Uses Task with *cardboard box* as the stimulus. Table S1 additionally includes scoring weights of the responses for each of the possible scoring dimensions above the horizontal thick line. Below this thick line, various methods for aggregating the response-level weights into a DT score are presented and calculated for each dimension.

The first scoring weights in Table S1 are for fluency scores and can be found in the second column. These scoring weights depend on the choice of criteria of response adequacy as outlined in the main article. In this example, repetitions were not allowed and, consequently, the second occurrence of *flower pot* which appeared as a synonymous idea, namely *growing vegetables in it*, received a weight of zero (implying exclusion). Furthermore, the response *feed the box to your car* received a weight of zero because inadequate ideas were not considered as scorable. All other responses received a weight of one and summing these weights yields the fluency score of 11 for these responses. Finally, the average of these fluency scoring weights yields a relative frequency of adequate ideas.

The third column of Table S1 includes the assigned semantic categories taken from Torrance (1966). Such categories build then the basis for all variants of flexibility scoring weights (fourth column of Table S1). For example, the classic flexibility scoring yields a weight of one for every response that is assigned to category that has not been assigned yet. Consequently, the first three responses receive a weight of one, because the responses were all assigned for different categories (growing, animal shelter, and furniture), but the fourth response was also assigned to the furniture category and, hence, receives a weight of zero. Other categories that were assigned more than once were games, transportation, and household appliances. For these categories, only the first occurrence receives a weight of one and all additional occurrences receive a weight of zero. The sum of these scoring weights yields the classic flexibility score and the average yields a flexibility ratio score.

Alternatively, flexibility can be calculated by means of category shift scores (Acar, Runco, & Ogurlo, 2018; Guilford, 1967; Nusbaum & Silvia, 2011). Here the scoring weights are derived slightly differently. In these types of scoring, the first response receives always a weight of zero, because one cannot start with a switch (see fifth and sixth column of Table S1). The scoring labelled *Switch I* (e.g., Acar et al., 2018) weighs every switch between categories by one and every stay in a category by zero (see fifth column in Table S1). *Switch II* (e.g., Nusbaum & Silvia, 2011) is different from Switch I because back-switches to categories that have already been used are not counted. That is the reason why classic flexibility and Switch II are always correlated perfectly, because they differ only in terms of the weight for the first response which is one for classic flexibility and zero for Switch II.

The seventh column in Table S1 includes relative frequency of occurrence estimates based on a sample size of *N* = 143. Thus, ideas with a relative frequency of 1/*N* = 1/143 = 0.007 are unique ideas. These frequencies could be directly used for originality scoring by taking the average. However, these relative frequencies are not intuitively interpretable (i.e., people with lower scores have higher levels of ability) and should be transformed. For example, the infrequency transformation (Mouchiroud & Lubart, 2001; Runco, Okuda, & Thurston, 1987) can be used which is simply 1 – relative frequency (see eighth column in Table S1). These infrequency weights could be averaged (Mouchiroud & Lubart, 2001) or aggregated by means of the .75-quantile (Forthmann, Holling, Çelik, Storme, & Lubart, 2017), for example, to yield an originality score (both are shown below the second horizontal thick line in the table). Summing up infrequency weights yields a score that has been labelled weighted fluency by Runco et al. (1987).

One of the most prominent, but also most criticized (Runco, 2008; Silvia et al., 2008), approaches to score originality is uniqueness scoring (e.g., Wallach & Kogan, 1965). The scoring weights for this type of scoring are presented in column nine in Table S1. The scoring weights are 1 when relative frequency of occurrence equals 1/*N* (which is 0.007 in the example; see column seven in Table S1) and zero for all other responses. This scoring is a special case of threshold scoring with 2/*N* as the threshold that needs to be surpassed. This makes also clear that the threshold is a function of sample size which is a rather undesired feature of this approach (see also Runco, 2008; Silvia et al., 2008).

A better way to use thresholds is to use fixed values such as 5% or 1% (Runco, 2008) or to use differential weights depending on the surpassed threshold (e.g., Cropley, 1967). Scoring weights for threshold scoring with only one fixed threshold (e.g., 5%) is not illustrated in Table S1, but Cropley’s (1967) approach to score points according to predetermined thresholds is illustrated in the tenth column. This approach yields zero weights for all responses with a relative frequency of occurrence greater than or equal to 0.15. In addition, responses occurring less often than 0.15 receive a weight of one, responses occurring less often than 0.05 receive a weight of 2, responses occurring less often than 0.03 receive a weight of 3, and responses occurring less often than 0.01 receive a weight of 4.

The eleventh column includes an objective scoring of remoteness based on latent semantic analysis as outlined by Forthmann, Oyebade, Ojo, Günther, and Holling (2018). This procedure involves calculation of the cosine between a word vector representing the task prompt and a word vector including the words of the response (stop words removed). This cosine is a measure of semantic similarity and will then be transformed by 1-cosine to yield a distance metric (Dumas & Dunbar, 2014). Importantly, cosine values derived from vector-based methods such as latent semantic analysis are artifactually influenced by the number of words (Forthmann, Oyebade, et al., 2018). According to Forthmann, Oyebade, et al. (2018) this confounding effect can be controlled for by simulation techniques and we applied the suggested technique here as well. Thus, the presented values are corrected for a potential bias due to the number of words. Responses with larger scores are semantically more distant as compared to the vector representing the prompt. For example, using a cardboard box as a *basketball hoop* is semantically more distant to the cardboard box vector than the response to make a *sailing boat* out of it.

In column twelve, responses are scored by one when the participant has marked it as being among the top-2 answers. In this hypothetical example, the participant has chosen *flowerpot* and *tool to measure a room* as top responses. All other responses receive a score of zero. This scoring indicates whether a response will enter the final scoring or not and this is the reason why no aggregate scores are presented in the table. This is reflected in the last column of Table S1. Here only the subjective ratings for these top-responses enter the aggregate scores below the thick line. Top scoring renders the number of responses entering the final aggregation equal across participants and, hence, the ratings could be averaged or summed. Finally, column fourteen includes hypothetical subjective ratings on a 5-point Likert-type scale based on the rating instructions provided by Silvia et al. (2008).

Final remarks on Table S1 should relate to the aggregation part below the horizontal thick line in the center of the table. The aggregation approaches are a selection from classic and more recently used methods. There are even more choices available (e.g., Plucker, Qian, & Wang, 2011; Plucker, Qian, & Schmalensee, 2014) and the reader might consult these references to apply other variants to the example data presented in Table S1. One must be cautious with some of the presented approaches because they are oversimplified for the purpose of illustration. For example, the residual score is based on the same regression equation for all of the presented scorings. In real applications, however, the prediction of a summative score by means of fluency will yield different regression coefficients across scoring dimensions and the intercept might be included in the regression equation. The presentation of the top-scoring method also could be misleading. Normally, only the responses marked as top-responses are scored by raters to reduce the rating-effort. Thus, the ratings of all the other responses are not necessarily available (as implied by the values in Table S1).

**The Issue of Measurement Precision and Sample Size in Frequency Estimates**

Many variants of originality scorings are based on frequencies of occurrence of responses. Responses are considered to be more original when less participants propose the same response. However, the precision of frequency estimates depends on sample size and one should at least keep that in mind when a new study is planned. The issue is illustrated in Figure S1. On the y-axis the estimated frequency of occurrence in percent is depicted along with error bars referring to one-sided Clopper-Pearson 95% confidence intervals. These intervals correspond with the null hypothesis that frequency of occurrence is significantly lower than one of the depicted originality thresholds of 5%, 7%, or 10%, respectively (a directed hypothesis corresponds with a one-sided statistical test or confidence interval). In case that an originality threshold of 5% is targeted for scoring, the confidence intervals depicted in Figure S1 are only narrow enough to refute the null hypothesis (frequency is equal to or larger than 5%) with a minimum sample size of *N* = 200. Hence, the shown estimates associated with smaller sample sizes than 200 are less trustworthy here, when we decide if the response with a frequency of occurrence of 2% surpassed the originality threshold of 5% or not (i.e., deciding if an estimate of 2% is significantly lower than a 5% threshold). Another observation that can be made from Figure S1, is that the minimum sample size for a CI that does not cover the targeted originality threshold depends also on the chosen threshold (higher thresholds require lower minimum sample sizes).

A consequence that follows from these considerations is that measurement error in originality scores due to sampling error in frequency estimates is more likely with smaller samples. This issue associated with frequency-based scorings has been widely overlooked in the literature and has not yet been fully understood, but the illustration here provides a starting point for how planning of sample size needs to take the targeted originality threshold into account.

Finally, an example will be given for how the outlined logic that threshold scoring is a special case of testing a null-hypothesis (determined by the chosen threshold) against an alternative lower value of an estimated frequency. Hence, it is required to take further statistical power into account. For example, if a threshold of 5% is targeted for originality scoring, one can understand threshold scoring here as a test of an observed frequency (say 2% as in the example above) against the null-hypothesis that the frequency is equal or greater than 5%. Commonly, a power of 80% is a desired goal in sample size planning. With this information it is possible to use statistical functions to calculate the required sample size. For example, the power\_binom\_test() function from the R package MESS (Ekstrøm, 2018) can be used the following way:

power\_binom\_test(p0 = .05, pa = .02, power = .8, alternative="less")

The results of this function call indicate that for a test that a frequency estimate is below 5%, and in the case that the population frequency is 2%, a sample size of *N* = 259 is needed to achieve a power of 80%. Thus, beyond the considerations that were made in relation to Figure S1, additional power considerations demonstrate that required sample size should be even larger (i.e., *N* = 259 instead of *N* = 200). However, these early considerations of the issue should be treated as early thoughts into the right direction and highlight the importance of the issue for future research endeavors. The above considerations are all based on only one frequency estimate, but in the scoring process one is naturally interested in the frequencies of hundreds or thousands of different ideas which makes the situation even more complex. Currently, it is only safe to conclude that larger samples will yield more precise estimate of frequencies and that decisions close to a chosen threshold (e.g., is a response with frequency 4.9% lower than a 5% threshold) will be associated with more decision errors as compared to decisions in which the estimate is far away from a chosen threshold (e.g., is a response with frequency 40% really larger than a 5% threshold). While these issues have not yet been fully understood, we hope to encourage researchers and practitioners to pay a closer attention to these issues, in order to develop more definite guidelines when using frequency-based scoring.

**References**

Acar, S., Runco, M. A., & Ogurlu, U. (2018). The moderating influence of idea sequence: A re-analysis of the relationship between category switch and latency. *Personality and Individual Differences*. Advance online publication. doi: 10.1016/j.paid.2018.06.013

Cropley, A. J. (1972). Originality scores under timed and untimed conditions. *Australian Journal of Psychology*, *24*, 31-36. doi: 10.1080/00049537208255782

Dumas, D., & Dunbar, K. N. (2014). Understanding fluency and originality: A latent variable perspective. *Thinking Skills and Creativity*, *14*, 56–67. doi: [10.1016/j.tsc.2014.09.003](https://doi.org/10.1016/j.tsc.2014.09.003)

Ekstrøm, C. T. (2018). *MESS: Miscellaneous esoteric statistical scripts*. R package version 0.5.3. https://CRAN.R-project.org/package=MESS

Forthmann, B., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2017). Typing speed as a confounding variable and the measurement of quality in divergent thinking. *Creativity Research Journal*, *29*, 257-269. doi: 10.1080/10400419.2017.1360059

Forthmann, B., Oyebade, O., Ojo, A., Günther, F., & Holling, H. (2018). Application of latent semantic analysis is biased by elaboration. *Journal* *of Creative Behavior.* Advance online publication. doi: 10.1002/jocb.240

Guilford, J. P. (1967). *The nature of human intelligence*. New York, NY: McGraw-Hill.

Mouchiroud, C., & Lubart, T. (2001). Children’s original thinking: An empirical examination of alternative measures derived from divergent thinking tasks. *Journal of Genetic Psychology: Research and Theory on Human Development, 162,* 382– 401. doi: 10.1080/00221320109597491

Nusbaum, E. C., & Silvia, P. J. (2011). Are intelligence and creativity really so different? Fluid intelligence, executive processes, and strategy use in divergent thinking. *Intelligence*, *39*, 36-45. doi: 10.1016/j.intell.2010.11.002

Plucker, J. A., Qian, M., & Schmalensee, S. L. (2014). Is what you see what you really get? Comparison of scoring techniques in the assessment of real-world divergent thinking. *Creativity Research Journal, 26,* 135–143. [doi: 10.1080/10400419.2014.901023](http://dx.doi.org/10.1080/10400419.2014.901023)

Plucker, J. A., Qian, M., & Wang, S. (2011). Is originality in the eye of the beholder? Comparison of scoring techniques in the assessment of divergent thinking. *Journal of Creative Behavior*, *45*, 1-22. doi: 10.1002/j.2162-6057.2011.tb01081.x

Reiter-Palmon, R., Young Illies, M., Kobe Cross, L., Buboltz, C., & Nimps, T. (2009). Creativity and domain specificity: The effect of task type on multiple indexes of creative problem-solving. *Psychology of Aesthetics, Creativity, and the Arts*, *3*, 73-80. doi: 10.1037/a0013410

Runco, M. A. (2008). Commentary: Divergent thinking is not synonymous with creativity. *Psychology of Aesthetics, Creativity, and the Arts*, *2*, 93–96. doi: 10.1037/1931-3896.2.2.93

Runco, M. A., Okuda, S. M., & Thurston, B. J. (1987). The psychometric properties of four systems for scoring divergent thinking tests. *Journal of Psychoeducational Assessment, 5,* 149 –156. doi: 10.1177/073428298700500206

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., & Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, *2*, 68–85. doi: 10.1037/1931-3896.2.2.68

Torrance, E. P. (1966). *Torrance tests of creative thinking. Norms-technical* *manual (Research edition)*. Princeton, NJ: Personnel Press, Inc.

Wallach, M. A., & Kogan, N. (1965). *Modes of thinking in young children: A study of the creativity-intelligence distinction*. New York, NY: Holt,Rinehart, & Winston.

Table S1

*Illustration of idea-scoring weights and aggregation for various DT scorings by means of generated alternate uses for a cardboard box.*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Response | Fluency  | Assigned category (Torrance, 1966) | Flexibility (Torrance, 1966) | Switch I (Acar et al., 2018) | Switch II (Nusbaum, Silvia, 2011) | Frequency (Mouchiroud & Lubart, 2001) | Infrequency (Mouchiroud & Lubart, 2001) | Uniqueness (Wallach & Kogan, 1965) | Unusualness thresholds (Cropley, 1967) | LSA-distance (Forthmann, Oyebade, et al., 2018)  | Participants‘ Top-2 (Silvia et al., 2008) | Ratings on a 5-point scale (Silvia et al., 2008) | Subjective Top-2 weights (Silvia et al., 2008) |
| Flowerpot | 1.000 | Growing | 1.000 | 0.000 | 0.000 | 0.105 | 0.895 | 0.000 | 1.000 | 0.815 | 1.000 | 3.000 | 3.000 |
| Shelter for cavies | 1.000 | Animal shelter | 1.000 | 1.000 | 1.000 | 0.315 | 0.685 | 0.000 | 0.000 | 0.755 | 0.000 | 2.000 | - |
| Lampshade | 1.000 | Furniture | 1.000 | 1.000 | 1.000 | 0.049 | 0.951 | 0.000 | 2.000 | 0.866 | 0.000 | 4.000 | - |
| Chair | 1.000 | Furniture | 0.000 | 0.000 | 0.000 | 0.203 | 0.797 | 0.000 | 0.000 | 0.910 | 0.000 | 2.000 | - |
| Sled | 1.000 | Transportation (surface) | 1.000 | 1.000 | 1.000 | 0.028 | 0.972 | 0.000 | 3.000 | 0.861 | 0.000 | 4.000 | - |
| Plate | 1.000 | Household appliances | 1.000 | 1.000 | 1.000 | 0.049 | 0.951 | 0.000 | 2.000 | 0.882 | 0.000 | 2.000 | - |
| Sieve, when drilling holes in it | 1.000 | Household appliances | 0.000 | 0.000 | 0.000 | 0.007 | 0.993 | 1.000 | 4.000 | 1.197 | 0.000 | 4.000 | - |
| Growing vegetables in it | 0.000 | - | - | - | - | - | - | - | - | - | - | - | - |
| Basketball hoop | 1.000 | Games | 1.000 | 1.000 | 1.000 | 0.007 | 0.993 | 1.000 | 4.000 | 0.924 | 0.000 | 3.000 | - |
| Sailing boat | 1.000 | Transportation (surface) | 0.000 | 1.000 | 0.000 | 0.028 | 0.972 | 0.000 | 3.000 | 0.809 | 0.000 | 2.000 | - |
| Play hide-and-seek | 1.000 | Games | 0.000 | 1.000 | 0.000 | 0.238 | 0.762 | 0.000 | 0.000 | 0.870 | 0.000 | 1.000 | - |
| Feed the box to your car | 0.000 | - | - | - | - | - | - | - | - | - | - | - | - |
| Tool to measure a room | 1.000 | Education | 1.000 | 1.000 | 1.000 | 0.007 | 0.993 | 1.000 | 4.000 | 0.747 | 1.000 | 5.000 | 5.000 |
| Method of aggregation |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Average (ratio) | 0.846 | - | 0.636 | 0.727 | 0.545 | 0.094 | 0.906 | 0.273 | 2.091 | 0.876 | - | 2.909 | 4.000 |
| Summation | 11.000 | - | 7.000 | 8.000 | 6.000 | 1.036 | 9.964 | 3.000 | 23.000 | 9.636 | - | 32.000 | 8.000 |
| Residuala | - | - | 1.500 | 2.500 | 0.500 | -4.464 | 4.464 | -2.500 | 17.500 | 4.136 | - | 26.500 | - |
| *q*.50 (median; Mouchiroud & Lubart, 2001) | - | - | 1.000 | 1.000 | 1.000 | 0.049 | 0.951 | 0.000 | 2.000 | 0.866 | - | 3.000 | 4.000 |
| *q*.75 (Forthmann, Holling, Çelik et al., 2017) | - | - | 1.000 | 1.000 | 1.000 | 0.154 | 0.983 | 0.500 | 3.500 | 0.896 | - | 4.000 | 4.500 |
| Maximum (Reiter-Palmon et al., 2009) | - | - | 1.000 | 1.000 | 1.000 | 0.315 | 0.993 | 1.000 | 4.000 | 1.197 | - | 5.000 | 5.000 |
| Minimum | - | - | 0.000 | 0.000 | 0.000 | 0.007 | 0.685 | 0.000 | 0.000 | 0.747 | - | 1.000 | 3.000 |
| Top-2 (statistically determined; Silvia et al., 2008) | - | - | 1.000 | 1.000 | 1.000 | 0.277 | 0.993 | 1.000 | 4.000 | 1.061 | - | 4.500 | 4.000 |

*Note*. Assumed sample size: *N* = 143. aResidual scores for all scoring dimensions (in practice the regression equations will be different for different scorings) were calculated based on the following regression equation: Summative score = 0.5\*Fluency (the intercept was omitted for simplicity).



*Figure S1.* On the y-axis the frequency of occurrence in percent is depicted. On the x-axis sample size is depicted for all possible sample sizes between 1 and 1000 that can yield a result of 2% in a sample. Error bars are one-sided Clopper-Pearson 95% confidence intervals. It is noteworthy that the lower bound of a one-sided CI of a proportion is always zero when the alternative hypothesis is *less* than the proportion under the null hypothesis (i.e., the targeted originality threshold). These intervals correspond with the null hypothesis that the observed frequency is significantly lower as compared to the scoring thresholds of 5%, 7%, and 10%, respectively (red horizontal lines).