



Discuit – a tool for dividing items into equal sets

Dörte de Kok
d.a.de.kok@rug.nl

eScience Center Fellow 2022-2023

netherlands

eScience center

by SURF & NWO



Problem

- For experiments, treatment studies etc you need 2 (or more) sets of items that are balanced
 - For various factors, e.g. word frequency, concreteness, word class
- Solution so far
 - Make 1 long list of items
 - Split by hand



Problem

- › But what if this needs to be done for each participant?
 - Because 1 variable is performance on material in pretest
- › Also: often reduction of continuous variables into categories
 - High- vs low-frequent rather than original values



Proposed solution

- › Automatic splitting into sets
- › Python package: Discuit¹
 - Open source license
 - Available through [github](#)/[pypi](#)/[zenodo](#)



Discuit - functionality

- › Based on k-means/k-mode clustering
- › Division into sets based on
 - Categorical variables (e.g. word class, animacy)
 - Continuous variables (e.g. word frequency, AoA, concreteness)
- › Input list of items (.csv file) is split into desired number of sets based on variables (also in file)
- › Statistics to check item division



Example use case

- › Investigation of treatment efficacy (PhD-research Pauline Cuperus²)
- › For each participant: divide material in treated & untreated items
- › Example for participant DTR:
 - 107 experimental items (filling in verbs in past & future tense)
 - Variables: Log10 frequency³, AoA ratings⁴, concreteness ratings⁵, transitivity, instrumentality, regularity of past tense form, accuracy in 2 sessions before treatment (for both past and future)

Use case – input file

[illegible]

Discuit - Use

```
> pip install discuit
```

```
> python3 discuit/run_discuit.py example/dtr.csv 2  
--columns l n n n c c c c c a c --runs 3
```

Input file
with items

Number
of sets

Specify
type of
variable

Specify number
of runs

Use case – output

[illegible]



Use case - output

Variable	Set 1	Set 2	Test statistic	<i>p</i>
	Mean (SD)			
Frequency	1.947 (0.592)	1.998 (0.653)	$X^2(1) = 0.058$.810
Age of Acquisition	5.091 (1.132)	5.283 (1.314)	$X^2(1) = 0.845$.358
Concreteness	3.797 (0.591)	3.796 (0.564)	$X^2(1) = 0.049$.824

Note: Statistics produced with Discuit: Kruskal-Wallis Anova



Variable	Set 1	Set 2	Test statistic	<i>p</i>
	Number of items			
Transitivity				
- Intransitive	26	26	$X^2(1) = 0.000$	1.000
- Transitive	28	27		
Instrumentality*				
- Non-instr.	39	40	$X^2(2) = 0.157$.942
- Instrumental	14	12		
- Missing data	1	1		
Past tense form				
- Regular	35	31	$X^2(1) = 0.225$.636
- Irregular	19	22		
Pretest 1 – Past				
- Incorrect	32	32	$X^2(1) = 0.000$	1.000
- Correct	22	21		
Pretest 1 – Future				
- Incorrect	44	43	$X^2(1) = 0.000$	1.000
- Correct	10	10		
Pretest 2 – Past				
- Incorrect	15	14	$X^2(1) = 0.000$	1.000
- Correct	39	39		
Pretest 2 – Future				
- Incorrect	27	27	$X^2(1) = 0.000$	1.000
- Correct	27	26		

Categorical variables: Pearson Chi-square tests of independence. Performance on past tense items in pretest 2 was used as basis for absolute split.
*For 2 items, instrumentality could not be determined, 1 of these was added to each set



Summary/Discussion

- › With Discuit you can
 - Split your list of materials in multiple balanced sets
 - Based on clustering algorithms
 - Obtain automatic set comparisons (statistics)
- › This also works for actual/complex data
- › Contact me if you have data-sets I can run this on as further use-cases for an upcoming paper!



Credits



‣ Project was supported by eScience Center Fellowship 2022-23

References:

- ¹ De Kok, D. (2023). Discuit (v0.2.1). Zenodo. <https://doi.org/10.5281/zenodo.7839874>
- ² Cuperus, P. (submitted). *Aphasia therapy software: Research, development, and implementation* [Unpublished Doctoral Dissertation]. University of Groningen.
- ³ Van Heuven, W.J.B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, 67, 1176-1190.
- ⁴ Kuperman, V., Stadthagen-Gonzalez, H. & Brysbaert, M. Age-of-acquisition ratings for 30,000 English words. *Behav Res* 44, 978–990 (2012). <https://doi.org/10.3758/s13428-012-0210-4>
- ⁵ Brysbaert, M., Warriner, A.B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46, 904-911.



university of
groningen

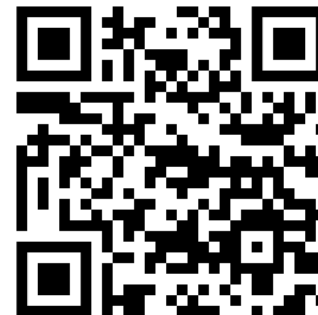
faculty of arts

neurolinguistics

Thank you for your attention!

Get in touch: d.a.de.kok@rug.nl

Download Discuit: pypi.org/project/discuit



Find this presentation on Zenodo: <https://doi.org/10.5281/zenodo.7985130>