



# Split-half reliability estimates of an online card sorting task in a community sample of young and elderly adults

Zhengkang Zhang<sup>1</sup> · Li-Zhuang Yang<sup>2</sup> · Teodóra Vékony<sup>3</sup> · Changqing Wang<sup>1</sup> · Hai Li<sup>2</sup>

Accepted: 3 March 2023 / Published online: 21 March 2023  
© The Psychonomic Society, Inc. 2023

## Abstract

Executive function is vital for normal social, cognitive, and motor functions. Executive function decline due to aging increases the risk of disability and falls in older adults, which has become an urgent public health issue. Fast and convenient neuropsychological tools are thus needed to identify high-risk groups as early as possible to conduct a timely intervention. Card sorting tasks, such as Wisconsin Card Sorting Task (WCST) and its variants, are popular tools for measuring executive function. This study investigated the reliability of an open-source, self-administered, online, short-version card sorting task with a sample of young ( $n = 107$ , 65 females, age:  $M = 30.1$  years,  $SD = 5.5$  years) and elderly Chinese ( $n = 113$ , 53 females, age:  $M = 64.0$  years,  $SD = 6.7$  years). We developed an automated scoring and visualization procedure following the recent recommendations on scoring perseverative responses to make the results comparable to the standardized WCST. Reliability estimates of commonly used measures were calculated using the split-half method. All task indices' reliabilities were reasonably good in both old and young groups except for "failure-to-maintain-set." Elderly Chinese adults showed compromised task performance on all measures compared with the young Chinese adults at the group level. The R script of automated scoring and estimation of reliability is publicly available.

**Keywords** Wisconsin card sorting task · Scoring · Split-half reliability · Cognitive flexibility · Executive function

Executive function is the top-down cognitive control process in initiating, maintaining, and flexibly updating goal-directed behaviors (Diamond, 2013). Executive function decline is a hallmark of cognitive aging (Lacreuse et al., 2020), which increases the risk of cognitive impairment and

falls in the elderly group (Montero-Odasso & Speechley, 2018). Thus, fast, accessible, convenient executive function tests to identify high-risk individuals are valuable to facing the increasingly aging population challenge. Wisconsin Card Sorting Task (WCST) and its variants (Barceló, 2003; Berg, 1948; Eling et al., 2008; Grant & Berg, 1948; Greve, 2001; Nelson, 1976; Nyhus & Barcelo, 2009) are popular tools to measure executive function or cognitive flexibility in specific. However, WCST is a standard neuropsychological test usually administered by professionals, and its complex scoring procedure causes inconsistency in the literature (Miles et al., 2021). The study thus evaluates the feasibility of an open-source, self-administered, short-version online card sorting task (OCST) with an automated scoring procedure using a sample of young and elderly Chinese participants.

Zhengkang Zhang and Li-Zhuang Yang contributed equally to this work.

- ✉ Li-Zhuang Yang  
lzyang@cmpt.ac.cn
- ✉ Changqing Wang  
wangchangqing@ahmu.edu.cn
- ✉ Hai Li  
hli@cmpt.ac.cn

<sup>1</sup> School of Biomedical Engineering, Anhui Medical University, Hefei 230032, China

<sup>2</sup> Anhui Province Key Laboratory of Medical Physics and Technology, Institute of Health and Medical Technology, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, China

<sup>3</sup> Université Claude Bernard Lyon 1, CNRS, INSERM, Centre de Recherche en Neurosciences de Lyon CRNL U1028 UMR5292, 95 Boulevard Pinel, 69500 Bron, France

## WCST and its variants

WCST originates from ideas of the German psychology of thinking and clinical practices of assessing prefrontal function in the early 20th century (Eling et al., 2008). In 1948, Grant and Berg (1948) formalized the design of the "University of Wisconsin Card Sorting Test," the predecessor of the well-known

WCST. The WCST quickly became a famous test of prefrontal lobe damage following Milner's pioneering work (Milner, 1963). The WCST was then refined and published as a standardized neuropsychological test with norms and an improved scoring procedure (Heaton et al., 1981; Heaton et al., 1993).

Card sorting enables quantitative evaluation of executive function (Milner, 1963). Typically, there are four *stimulus cards* and a pack of *response cards*. All cards show patterns composed of different *forms* (e.g., triangle, star, cross, circle), *colors* (e.g., red, green, yellow, blue), and *numbers* (1, 2, 3, 4). The pack of response cards is shuffled before the test begins. Participants are instructed to figure out the sorting rule and classify each response card into one of the four stimulus cards. The sorting rule changed without warning after the participant made ten consecutive correct choices. The sorting rule usually changes in a fixed order, such as color-form-number-color-form-number, unknown to participants. As the stimulus cards have orthogonal properties, participants' choice history can indicate whether they can quickly form and flexibly adjust their mindset.

The WCST task has been undergoing continual modifications or improvements. Milner (1963) used 128 cards in her pioneering work. However, a shorter version (WCST-64) with 64 response cards might fit the clinical settings better (Greve, 2001). Nelson (1976) developed the Modified Wisconsin Card Sorting Task (M-WCST), which removed all ambiguous response cards that shared more than one attribute from the stimulus cards. Barceló proposed an innovative design, namely the Madrid card sorting test (MCST), which incorporates the task-switch and WCST paradigm enabling both behavior and neurophysiologic recording (Barceló, 2003, 2021). A beta version of online MCST is also available in Spanish and English. Opensource software has also accelerated the widespread usage of WCST-like card sorting task, such as the *Berg Card Sorting Task* provided by the Psychology Experiment Building Language (PEBL) (Fox et al., 2013; Piper et al., 2012). Lange and colleagues designed a self-administered computerized variant of the WCST (cWCST) characterized by an unpredicted sorting rule change and removal of all ambiguous cards as MCST (Lange & Dewitte, 2019; Steinke et al., 2020; Steinke et al., 2021). An online open-source WCST-like card sorting task is also available with the assistance of the powerful *jsPsych* library (de Leeuw, 2015; Vékony, 2022). However, as far as we know, few studies have verified the online, self-administered WCST-like card sorting task with a community sample.

## Application of WCST

Card sorting is the most widely accepted task in assessing executive function deficits (Stuss & Benson, 1984). A survey of 747 North American psychologists revealed WCST

as one of the ten most frequently used neuropsychological assessment tools (Rabin et al., 2005). In a following survey study, the computerized version of WCST ranked as one of the two most commonly used automated test tools (Rabin et al., 2014). A recent systematic review suggests that WCST was one of the top five assessment tools of executive function with the most validations for children and adolescents in low- and middle-income countries (Kusi-Mensah et al., 2022). The classical form of WCST originating from the Milner version (Milner, 1963) is undoubtedly a famous neurocognitive task frequently used by clinicians (Miles et al., 2021).

WCST and its variants also rank as the seventh most frequently used neurocognitive tool to evaluate executive functions in aging (Faria et al., 2015). The age-related performance decline on WCST measures was expected and supported by behavior (Haaland et al., 1987; Lineweaver et al., 1999; Marquine et al., 2021; Perez-Enriquez et al., 2021) and neuroimaging studies (Esposito et al., 1999; Heckner et al., 2021). However, there was also evidence that WCST was not sensitive to aging in the Taiwanese population (Shan et al., 2008). In addition to age, education level is another factor affecting WCST performance (Lineweaver et al., 1999; Marquine et al., 2021). The WCST performance deficit might not solely stem from the deficits in cognitive flexibility but also the reduced working memory (Hartman et al., 2001; Lange et al., 2016).

Although WCST is a popular tool for assessing cognitive flexibility (Uddin, 2021), the construct validity of WCST as a pure measure has yet to be questioned (Nyhus & Barceló, 2009). Optimal WCST performance depends on multiple cognitive components, including set-shifting related to the frontoparietal network and rule inference related to the frontostriatal network (Lange et al., 2017). Recent brain imaging studies also imply that large-scale functional brain networks subserve the cognitive flexibility component measured by WCST, questioning the anatomical specificity of WCST (Nomi et al., 2017). Some researchers have thus advocated refining the WCST to make the measure "pure" or specific to a cognitive process (Barceló, 2021; Nyhus & Barceló, 2009). However, the refined WCST task, such as MCST, is more like a task-switching paradigm, which might not be comparable to the classical WCST measures (Miles et al., 2021).

## Scoring of the card sorting task

The standardized WCST can provide up to 16 main outcome measures (Chiu & Lee, 2021), several of which are redundant as they are linear combinations of other measures. Seven major indices are enough to validate the latent structure of the WCST, including *total correct*, *perseverative responses*, *perseverative errors*, *non-perseverative errors*,

*conceptual level responses, categories completed, failure-to-maintain-set* (Greve et al., 2005). Table 1 provides a brief explanation of the seven measures. Other indices, such as “*trials-to-complete-the-first-category*” and “*learning to learn*,” might not be suitable for factor analysis as many subjects might fail to learn the task structure making these indices zero. Some novel indices have also been validated, such as “Cognitive persistence” (Teubner-Rhodes et al., 2017).

*Perseverative errors* and *Perseverated responses*, which indicate the stubborn usage of outdated response rules, were widely used as cognitive flexibility indexes. Unfortunately, conceptual confusion and inconsistent scoring of perseveration are common in practices (Flashman et al., 1991; Miles et al., 2021). The seminal work of Flashman et al. (1991) spent about six pages elaborating on nine rules of scoring perseveration. After a decade, Miles and colleagues still have to use a lengthy tutorial paper to solve the inconsistency and recommend automated scoring to solve this issue (Miles et al., 2021). The critical concept of scoring perseveration behavior is the *perseverated-to-principle*, which refers to the repeatedly used incorrect sorting rule (Miles et al., 2021). Perseverated Responses thus refers to the number of responses that conform to the perseverated-to-principle, which can reveal the subject’s cognitive flexibility in mindset shifting. Even though the idea of perseverated responses seems intuitive initially, their calculation remains inconsistent in the literature. The status is due to the *ambiguous* trials, where the selected stimulus card shares two or more attributes with the current response card. Flashman and colleagues proposed the “*sandwich rule*” to deal with ambiguous trials, formalized later in the standardized WCST manual (Heaton et al., 1993). If the ambiguous response matches the “*perseverated-to-principle*” and is preceded and followed by an unambiguous perseverative error, it is scored as a perseverative response (Flashman et al., 1991, p191). However, several exceptional cases exist in practice, such as successive ambiguous responses being “*sandwiched*,” which might confuse the scorer. Moreover, a perseverative response can be correct as it might overlap with the right attribute (Miles et al., 2021). The scoring procedure is tedious and

error-prone without training. Figure 1A is a typical high-performance profile easy to score. However, Fig. 1B presents a typical low-performance profile with lots of scoring ambiguity, challenging an inefficient scorer. An automated scoring procedure is thus necessary to standardize clinical and research practices (Miles et al., 2021).

## Reliability of card sorting task measures

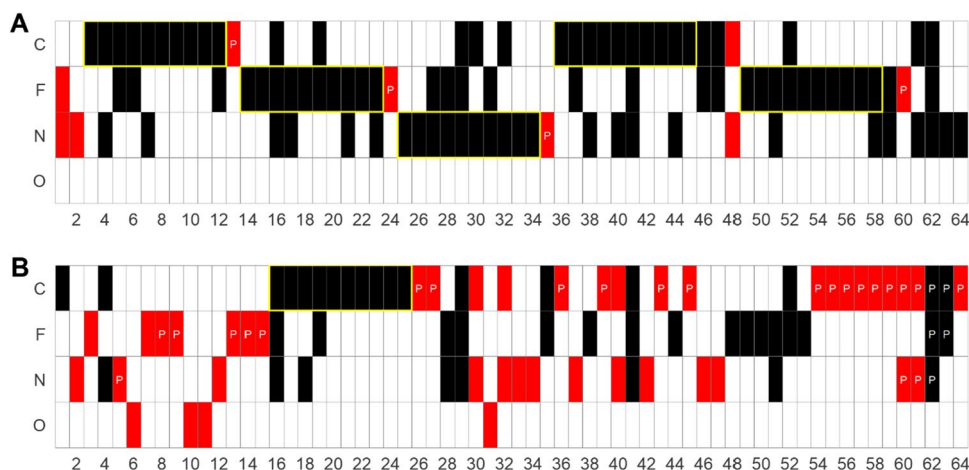
Reliability is fundamental in clinical settings and individual difference studies. First, the reliability estimates for a specific measure inform practitioners about the precision of the test scores. Second, low-reliability measurements reduce the statistical power to detect potential associations in individual difference studies (Hedge et al., 2018). However, the reliability of WCST and its variants received incomparable research attention concerning its widespread usage (Kopp et al., 2021).

A validation study indicates the test–retest reliability estimates of measures from M-WCST proposed by Nelson (1976) are only modest (ranging from .46 to .64) with a sample of 229 healthy community-dwelling old adults (Lineweaver et al., 1999). Even after one year, there was a practice effect on the measure of non-perseverative error (Lineweaver et al., 1999). However, Kopp and colleagues report that the M-WCST manifests desirable reliability estimates ( $> 0.9$ ) using the split-half estimation method in a sample of neurological inpatients ( $n = 146$ ) (Kopp et al., 2021). Chiu and Lee (2021) investigated the test–retest reliability of the classical WCST using a schizophrenia sample in Taiwan ( $n = 63$ ) with a 2-week interval. The study demonstrates that most WCST measures were acceptable except for non-perseverative errors and failure-to-maintain-set (Chiu & Lee, 2021). A recent study using a sample of healthy Argentinian adults (18–89 years old,  $n = 235$ ) indicates that the classical manual version of WCST has good reliability using Cronbach’s alpha coefficient (Miranda et al., 2020).

The inconsistency among the limited number of studies has two implications. First, many cognitive tasks,

**Table 1** Description of the seven OCST indices

Index	Meaning
Total correct (TC)	The number of correct trials
Perseverative response (PR)	The number of unambiguous trials matches the “perseverated-to-principle”
Perseverative errors (PE)	The number of incorrect perseverative responses
Non-perseverative errors (NPE)	The number of incorrect trials not belonging to perseverative responses
Conceptual level responses (CLR)	The number of three or more consecutively correct trials
Categories completed (CAT)	The number of categories completed
Failure-to-maintain-set (FMS)	The number of incorrect trials which occurred after five to nine consecutively correct trials



**Fig. 1** Two typical scoring cases. **A** High-performance response profile. **B** Low-performance response profile hard for an inexperienced scorer. *Note.* The *x*-axis depicts the 64 trials sequentially. The *y*-axis indicates the four dimensions: C (*Color*), F (*Form*), N (*Number*), and O (*Other*). The *colored tile* indicates the match between the response card and chosen stimulus card on each trial. O refers to a choice that did not match any of the C, F, or N dimensions. WCST is characterized by ambiguous trials where the chosen stimulus card might match the response card on more than one dimension. As shown by Fig. 1A, the examinee chose the stimulus card matching the form and number dimension at the first trial, which was incorrect, indicated by the *red color*. In the second trial, the examinee chose the stimulus card matching the num-

ber dimension, which was wrong. From the third trial, the examinee got the correct rule (*color*) and completed ten consecutive correct choices indicated by the *yellow borderline*. On trial 13, the correct rule shifted to form, and the examinee made a perseverative error (*indicated by a white “P”*). Figure 1B presents a hard-scoring scenario involving almost every specific scoring principle. For example, trials 60–63 are perseverative responses or perseverative errors as they were enclosed by unambiguous perseverative errors (*the sandwich rule*). Please refer to our Method section for a detailed description of the scoring method. The visualization method producing Fig. 1 is helpful for an informative visual check and clinical diagnosis, which is publicly available on our GitHub repository (see Code Availability section)

including WCST and its variants, might suffer from practice or learning effects. Thus, test–retest reliability might not be applicable in practice. Split-half reliability is a convenient estimate of internal consistency among different trials suitable for estimating the reliability of cognitive tasks (Parsons et al., 2019; Pronk et al., 2022; Steinke & Kopp, 2020). Second, the reliability estimates depend on specific measurements and the sample used, which cannot be generalized across task variants, different indices, populations, and test scenarios. Thus, more research efforts are needed to validate the usage of WCST and its variants in research and practice.

## The present study

The classical form of WCST is widely used in clinical and research practices. In addition to the commercial and standardized WCST (Heaton et al., 1981; Heaton et al., 1993), open-source solutions are also widely used (Fox et al., 2013; Vékony, 2022). However, there were two unresolved issues. First, the scoring procedure of WCST-like tasks was inconsistent, which made the comparison among studies problematic (Flashman et al., 1991; Miles et al., 2021). Second, the reliability of WCST-like tasks received

insufficient research efforts. Third, online WCST-like tasks in cognitive aging studies should be validated. The present study operationalizes the expert consensus on WCST scoring (Flashman et al., 1991; Miles et al., 2021) into an automated scoring and visualization procedure using the *R* language. In addition, we investigate the split-half reliability of an online, WCST-like card sorting task (OCST) (Vékony, 2022) with a large community sample involving young and elderly participants. Our results suggest that most OCST measures manifest acceptable reliability and are sensitive to the age difference at the group level.

## Method

### Sample

We recruited 256 young (18–45 years) and old adults (55–81 years) from a community medical examination center of a subordinate county of Hefei City, Anhui province of China when they attended their regular annual health screen through advertisement. All participants had a corrected-to-normal vision. Six participants did not provide age, gender, or education information, and 30 participants who failed to complete the whole test were excluded from the analysis, leaving the final sample of 220

participants. The final young group included 65 females and 42 males (age,  $M = 30.1$  years,  $SD = 5.5$  years), and the old group included 53 females and 60 males (age,  $M = 64.0$  years,  $SD = 6.7$  years). The study was approved by the ethics committee of the Hefei Institutes of Physical Science and was conducted following the Declaration of Helsinki. All participants gave written informed consent and received monetary compensation.

### Procedure

Two nurses from the community medical examination center screened the volunteers. Eligible participants were taken to one of two testing rooms. The nurse registered participant information, introduced the task instructions, and familiarized the participants with the keyboard operation. Participants completed a short practice first. Then they completed the formal test by themselves. Some of them also completed a two-step decision-making task and an attention network task, which belongs to another parallel study. After the online test, the nurse measured the participants’ working memory using the digit-span test (Wechsler, 1997). The data for estimating split-half reliability is publicly available (see Data Availability Statement).

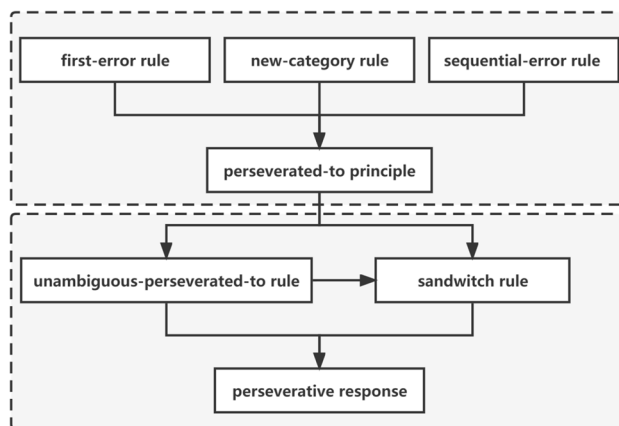
### Task design

The OCST (Vékony, 2022) follows the design of the short form Berg Card Sorting Task (Fox et al., 2013; Piper et al., 2012), which uses 64 response cards as the standardized WCST-64. We translated the task instruction into simplified Chinese and added a short practice to help the old participants who might be unfamiliar with keyboard operations. The task was deployed on the Tencent Cloud using the Python Django web framework. A typical trial began with a screen with four stimulus cards on the top row and a response card in the middle of the bottom row. The four stimulus cards had a red triangle, two green stars, three yellow diamonds, and four blue circles. The response card was drawn from 64 cards (4 colors  $\times$  4 forms  $\times$  4 numbers) without replacement. Participants needed to decide which of the four cards the response card belonged to by clicking on the corresponding keys and receiving feedback on whether they were correct. After the participant made ten consecutive right choices, the new sorting rule or category began. There was a maximum of six categories in the order of color-form-number-color-form-number unknown by the participants.

### Scoring procedure

We calculated the seven measures of Table 1. All the measures were straightforward except for perseverative responses and perseverative errors. We distilled the recommendations

in the literature (Flashman et al., 1991; Miles et al., 2021) into a logical flow and operationalized it into an R script publicly available on GitHub (see Code Available Statement). Figure 2 elaborates on the procedure and scoring principle of our scoring script. For each scoring principle, Supplementary Table 1 gave the source information in the quoted and cited form. We strongly recommend Flashman et al. (1991) and Miles et al. (2021) for interested readers. Please note that perseverative errors are a subset of perseverative responses (Miles et al., 2021). A perseverative response can be correct because of the large portion of ambiguous trials in the WCST. Perseverative errors can be easily identified if they are perseverative responses. The scoring procedure is a two-step decision. Firstly, the *perseverated-to-principle* was determined for each trial according to the *first-error rule*, *new-category rule*, and *sequential-error rule*. Then, an unambiguous error was classified as a perseverative error if the choice adhered to the perseverated-to-principle of that trial (*Unambiguous-perseverated-to rule*). For ambiguous trials enclosed by perseverative errors, we used the *sandwich rule* to judge whether it is a perseverative error (or perseverative response).



**Fig. 2** A schematic diagram of the automated scoring logic of perseverative response. *Note.* The first incorrect unambiguous choice dimension is considered the perseverated to principle (*first-error rule*). After achieving a category criterion (ten consecutive correct choices), the old rule became the perseverated-to principle (*new-category rule*). Suppose the participants make three sequential unambiguous incorrect choices on a particular dimension (ambiguous choices interlayered by the three trials do not influence the continuity). In that case, this dimension becomes the perseverated-to principle from the second unambiguous position of the sequence (*sequential error rule*). Trials that are unambiguous and cohere with this trial’s perseverated-to principle are tagged as a “perseverative response,” which are also “perseverative errors” (*unambiguous-perseverated-to rule*). Ambiguous responses match the perseverated-to principle, and enclosed by unambiguous perseverated responses are considered perseverative responses (*sandwich rule*). Please note that perseverative errors are a subset of perseverative responses. Perseverative errors are incorrect perseverative responses



## Split-half reliability calculation

Evaluating the split-half reliability involves splitting the trials into two halves, scoring each half for each subject, and calculating their Pearson correlation. The Spearman–Brown formula was then used to calculate the underestimation due to the usage of only half trials (Parsons et al., 2019). However, splitting a trial sequence into two halves is not that straightforward. Although first-second and odd-even splitting is common practice in evaluating questionnaires, recent simulation studies recommend the split method based on random permutation (Parsons et al., 2019), sampling-based (Steinke & Kopp, 2020), or Monte Carlo method (Pronk et al., 2022).

We conducted the split-half reliability estimation using *splithalfr* (Pronk et al., 2022) and *R* version 4.1.3 (R Core Team, 2022) using four split methods: the first-second half split, odd-even trial split, permuted split, and Monte Carlo-based split. For permuted and Monte Carlo splits, we made 5000 resamplings and calculated the mean Pearson correlation of each split. For first-second, odd-even, and permuted splits, Spearman–Brown correction was made. As the Monte Carlo splits method constructed a full-length trial sequence for each half, reliability was the median Pearson correlation. Because card sorting trials are not independent, it is unreasonable to split trials and then calculate task indexes for each half. Thus, we first tag each trial using our scoring procedure. Each trial is then labeled as a correct, perseverated response, perseverated error, non-perseverated error, conceptual level responses, or failure-to-maintain-set. We also assign a value of 1/10 to each trial belonging to an achieved category. The method, insight by Kopp et al. (2021), enables the split-half estimation of all seven measures. The reliability analysis script is publicly available (see Code Available Statement).

## Statistical analysis

The intercorrelations among the seven measures were evaluated using Spearman’s rank correlation coefficients and visualized using the *corrplot* package. We calculated the age group’s effect size for each measure using Cohen’s *d* and simulated the effect size distribution by performing bootstrap resampling ( $n = 5000$ ). We then used the Gardner–Altman estimation plot to visualize the result with the *dabestr* package in *R* (Ho et al., 2019). To examine whether card sorting performances declined with age, we performed multiple regressions with gender, education years, and digit span scores as covariates. As the linear trend might be driven by the group difference between the young and the elderly, we performed separate regression analyses for the elderly and young groups. Permutation tests of linear regression models

were conducted using the *permuco* package (Frossard & Renaud, 2021). The Bonferroni correction was used after repeated analysis of seven measures. Specifically, all *p* values were multiplied by seven to adjust the type I errors. The alpha threshold was .05 (two-tailed) for hypothesis testing. All statistical analyses were in the *R* version 4.1.3 environment (R Core Team, 2022), the script of which was publicly available.

## Results

### Descriptive statistics

Table 2 provides descriptive statistics on the raw scores of the seven measures for elderly females ( $n = 53$ ), elderly males ( $n = 60$ ), young females ( $n = 65$ ), and young males ( $n = 42$ ), respectively. The intercorrelations among those measures were high except for failure-to-maintain-set. The intercorrelation matrix manifested itself into three clusters (Fig. 3). The first cluster was total correct, categories completed, and conceptual level responses with intercorrelation larger than .85. Another cluster comprised perseverative responses, perseverative errors, and non-perseverative errors with intercorrelations larger than .49. Failure-to-maintain-set is independent of the other measures except a small-to-moderate correlation with categories completed.

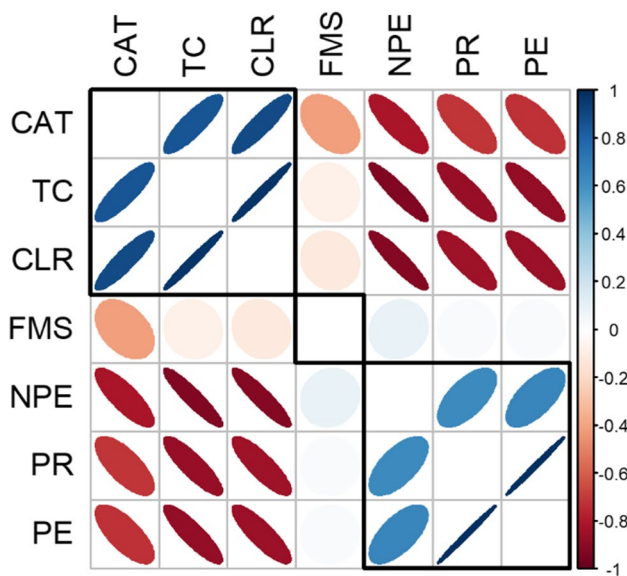
### Split-half reliability estimates

Table 3 describes the reliability estimates for the seven indices using the first-second, odd-even, permuted, and Monte Carlo split methods. In general, the first-second split method tended to underestimate, and the odd-even method tended to overestimate the reliability estimates. The permuted splits

**Table 2** Means and standard deviations for age and core task indices in old and young groups

	Old female <i>N</i> = 53	Old male <i>N</i> = 60	Young female <i>N</i> = 65	Young male <i>N</i> = 42
Age	62.2 (6.14)	65.6 (6.73)	30.9 (5.35)	28.9 (5.49)
TC	39.2 (9.78)	41.2 (10.4)	49.2 (8.48)	50.0 (9.52)
PR	13.5 (7.30)	11.8 (6.23)	7.71 (4.90)	7.74 (6.52)
PE	12.2 (6.07)	10.7 (5.50)	7.12 (3.92)	7.29 (5.65)
NPE	12.6 (6.24)	12.1 (6.41)	7.65 (5.94)	6.67 (4.89)
CLR	22.3 (11.0)	25.6 (12.0)	34.6 (10.8)	35.8 (12.2)
CAT	1.79 (1.23)	2.28 (1.56)	3.38 (1.50)	3.62 (1.53)
FMS	1.00 (1.18)	0.92 (1.03)	0.62 (0.76)	0.40 (0.59)

TC: Total Correct; PR: Perseverative Response; PE: Perseverative Errors; NPE: Non-perseverative Errors; CLR: Conceptual Level Responses; CAT: Categories Completed; FMS: Failure-to-Maintain-Set. The numbers are raw scores



**Fig. 3** The intercorrelations among the seven card sorting measures. Note: TC: Total Correct; PR: Perseverative Response; PE: Perseverative Errors; NPE: Non-perseverative Errors; CLR: Conceptual Level Responses; CAT: Categories Completed; FMS: Failure-to-Maintain-Set

and the Monte Carlo method yielded comparable estimates except for failure-to-maintain-set.

Figure 4 elaborates on the reliability difference between the young and elderly groups. Although there were slight differences, the reliability estimates for both groups were comparable. Category completed, conceptual level response and total correct were the top three measures with the highest reliability. The category completed was significantly larger than .9, followed by the conceptual level response and the total correct, around 0.9. The three measures fell into a desirable range suitable to be a clinical assessment tool. Reliability estimates of perseverative response, perseverative error, and non-perseverative error were around .8, acceptable in individual difference studies. However, the failure-to-maintain-set failed to manifest a reasonable reliability estimate.

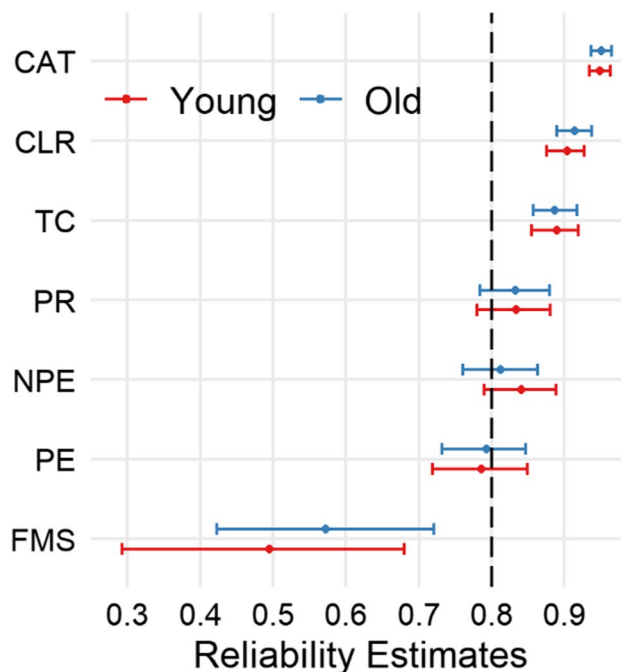
**Age effect**

Figure 5 illustrates the difference between the young and elderly groups. Elderly participants completed about two categories ( $M = 2.1, SD = 1.4$ ), which is lower than young participants ( $M = 3.5, SD = 1.5$ ) (categories completed: Cohen’s  $d = -.97, 95\% CI [-1.29, -.67]$ ). Consistently, they also made less correct responses ( $M = 40.3, SD = 10.1$ ) than the young participants ( $M = 49.6, SD = 8.9$ ) (total correct: Cohen’s  $d = -.97, 95\% CI [-1.27, -.67]$ ). Furthermore, the conceptual level responses were also lower in the elderly group ( $M = 24.0, SD = 11.6$ ) compared with the young

**Table 3** Split-half reliability estimates of the seven OCST measures in old and young groups

	First-second	Odd-even	Permutated	Monte Carlo
Young, $N = 107$				
TC	.77	.94	.87[.84,.91]	.89[.86,.92]
PR	.71	.89	.80[.73,.85]	.83[.78,.88]
PE	.72	.83	.72[.64,.79]	.79[.72,.85]
NPE	.69	.82	.81[.75,.86]	.84[.79,.89]
CLR	.80	.99	.89[.86,.92]	.90[.88,.93]
CAT	.75	1.0	.95[.93,.96]	.95[.93,.96]
FMS	-.27	-.23	-.04[-.35,.19]	.49[.29,.68]
Elderly, $N = 113$				
TC	.62	.89	.87[.83,.90]	.89[.86,.92]
PR	.53	.88	.80[.74,.85]	.83[.78,.88]
PE	.58	.82	.74[.66,.80]	.79[.73,.85]
NPE	.40	.67	.76[.70,.83]	.81[.76,.86]
CLR	.66	.98	.90[.88,.93]	.91[.89,.94]
CAT	.72	1.0	.95[.93,.96]	.95[.94,.96]
FMS	.25	-.05	.23[.01,.45]	.57[.42,.72]

Values in square brackets indicate the 95% high-density interval for permutated and Monte Carlo-based split. TC: Total Correct; PR: Perseverative Response; PE: Perseverative Errors; NPE: Non-perseverative Errors; CLR: Conceptual Level Responses; CAT: Categories Completed; FMS: Failure-to-Maintain-Set



**Fig. 4** Monte Carlo reliability estimates for the seven indices of OCST. Note: TC: Total Correct; PR: Perseverative Response; PE: Perseverative Errors; NPE: Non-perseverative Errors; CLR: Conceptual Level Responses; CAT: Categories Completed; FMS: Failure-to-Maintain-Set

group ( $M = 35.1$ ,  $SD = 11.3$ ) (Cohen's  $d = -.96$ , 95% CI [-1.27, -.65]).

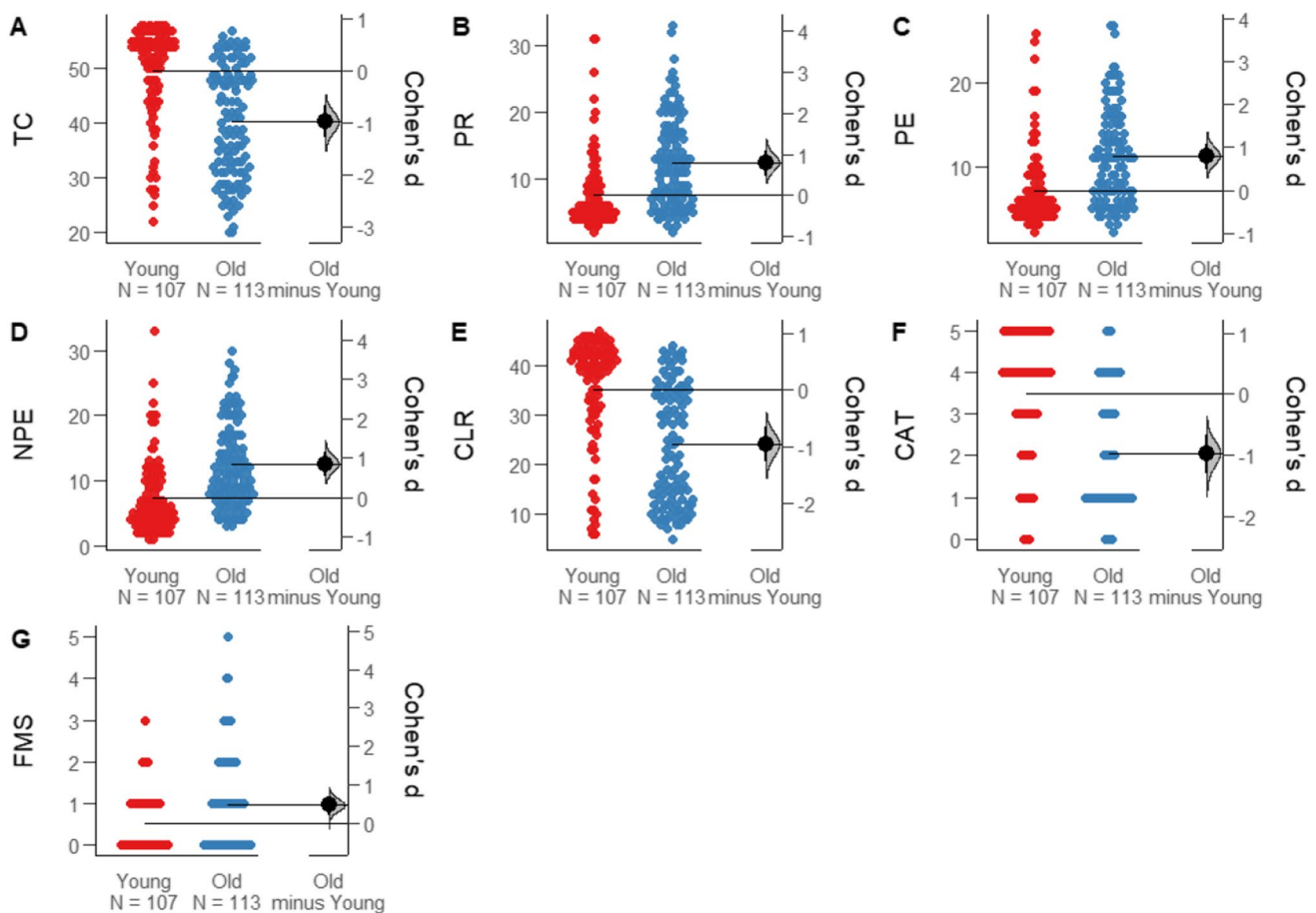
Elderly participants ( $M = 12.6$ ,  $SD = 6.8$ ) manifested more perseverative responses than the young participants ( $M = 7.7$ ,  $SD = 5.6$ ) (Cohen's  $d = .78$ , 95% CI [.46, 1.08]). The elderly ( $M = 11.4$ ,  $SD = 5.8$ ) also committed more perseverative errors than the young group ( $M = 7.2$ ,  $SD = 4.6$ ) (Cohen's  $d = .80$ , 95% CI [.48, 1.09]). On the measure of non-perseverative error, the elderly group ( $M = 12.3$ ,  $SD = 6.3$ ) got poor performance compared with the young group ( $M = 7.3$ ,  $SD = 5.5$ ) (Cohen's  $d = .85$ , 95% CI [.54, 1.14]). All participants made few failure-to-maintain set errors (range from 0 to 5); one-half of the young participants made zero errors, and one-half of the elderly made less than one error. The young group ( $M = .5$ ,  $SD = .7$ ) made fewer failure-to-maintain-set errors than the elderly group ( $M = 1.0$ ,  $SD = 1.1$ ) with medium effect size (Cohen's  $d = .46$ , 95% CI [.20, .69]).

We also examined whether the seven measures manifested age-specific decline in the elderly and young group

separately with education years, gender, and digit span score as covariates. However, none of the continuous age effects reached a significance level after Bonferroni correction (all adjusted  $p > .05$ ). In addition, our results did not yield any effect of gender (all adjusted  $p > .05$ ). For the young group, there was a positive association between education years and category completed (adjusted  $p = .021$ ). However, the effect of education years did not reach the significance level in the elderly group (all adjusted  $p > .05$ ).

## Discussion

WCST and its variants have become a popular clinical and research tool for assessing executive function since its origin. The classical form of WCST has many ambiguous trials, making scoring perseveration responses challenging. Although significant progress has been made in standardizing the scoring procedure (Flashman et al., 1991; Heaton et al., 1981; Heaton et al., 1993), there is still great



**Fig. 5** Gardner–Altman estimation plot depicting the age group difference on the seven measures. **A** Total Correct (TC); **B** Perseverative Response (PR); **C** Perseverative Errors (PE); **D** Non-perseverative

Errors (NPE); **E** Conceptual Level Responses (CLR); **F** Categories Completed (CAT); **G** Failure-to-Maintain-Set (FMS)



controversy and inconsistency (Miles et al., 2021). The present study contributes automated scoring and informative visualization procedure freely available as R scripts, which can benefit clinical psychologists and researchers. We also report split-half reliability estimates for the seven frequently used WCST measures for a publicly available online card sorting task (OCST) in a young and elderly community sample. Our results suggest that most WCST measures manifest acceptable reliability and are sensitive to the age difference at the group level.

### The automated scoring and visualization tool

The standardized WCST can provide up to sixteen measures (Chiu & Lee, 2021). Most of them are straightforward except for perseverative response and perseverative error (Flashman et al., 1991; Miles et al., 2021). The challenge is mainly due to the ambiguous trials where the response card shares more than one dimension with the chosen stimulus card. There are two solutions to this issue. First, remove all ambiguous response cards, as are the M-WCST (Nelson, 1976), MCST (Barceló, 2003), and cWCST (Steinke et al., 2021). However, these modified versions might not be comparable to the classical form of WCST. Second, providing systematic tutorials, as done by Flashman et al. (1991) and Miles et al. (2021). However, it is still hard to master all the scoring principles. An automated scoring tool is thus necessary to solve the scoring inconsistency of WCST-like tasks (Miles et al., 2021). The present study provides an open-source and transparent scoring procedure that strictly coheres to the expert consensus. The scoring procedure can facilitate the scoring of OCST (Vékony, 2022), a publicly available card sorting task that follows the typical design of WCST. It can be easily modified if a custom task script is used. Furthermore, we also contribute an informative visualization tool facilitating clinical diagnosis, scoring check, or self-education of the scoring principle. The automated scoring and visualization method is valuable for neuropsychological services in developing countries where trained professionals are lacking.

### Reliability of card sorting measures

The Monte Carlo results indicate that the reliability estimates of category completed, conceptual level response, and total correct are suitable to be used in clinical diagnosis usage ( $rel > 0.9$ ), and the perseveration response, non-perseverative errors, and perseverative errors are acceptable to be used in research ( $rel > 0.8$ ). An exception is failure-to-maintain-set, the reliability estimate of which is around 0.5. The low reliability of the failure-to-maintain-set measure was mainly due to the very few errors made.

Our estimates are generally superior to the standardized M-WCST reported by the manual, which quantifies a five-year interval test–retest reliability as 0.55 (Schretlen, 2010). However, as Schretlen (2010) estimated test–retest reliability and a different task version, readers should be cautious about comparing our findings directly with Schretlen (2010). Our reliability estimates are lower than the reliability of cWCST in Steinke et al. (2021), which reports that all measures achieved good reliability ( $rel > 0.9$ ). However, there are several fundamental differences between Steinke et al. (2021) and our study: First, the OCST in our study was a 64-trial, short-version task, while the cWCST had about 168 trials. An increase in trial length is beneficial for good reliability. Second, the OCST has ambiguous response cards. The exclusion of ambiguous trials in cWCST reduces the task difficulty, which might promote a consistent strategy in the task. Third, the cWCST only selected three kinds of error and three response time measures, while we reported seven widely used indices of WCST in the literature. Despite this, the reliability estimates are valuable as the OCST makes it the best to follow the task form and scoring of WCST-64, which is still one of the leading forces in clinical settings (Miles et al., 2021).

### Method of splitting

The present study reports four split-half reliability estimates using different splitting methods. The first-second splitting provides the worst, and the odd-even splitting provides the best reliability estimates in our research for most indices. The pattern is consistent with Kopp and colleagues' recent study, which investigates the split-half reliability of M-WCST with a clinical sample (Kopp et al., 2021). Their study systematically evaluates the impact of the trial grain size on reliability estimates, which ranges from the odd-even (grain size = 1) to the first-second approach (grain size = half of the trial length). There seems to be a decreasing trend as the trial grain size increases. The phenomenon might stem from confounding learning effects or strong dependence between trials (Pronk et al., 2022). Thus, the odd-even splitting might overestimate, while the first-second splitting might underestimate the reliability estimates of WCST and its variants.

Splitting by random permutation or Monte Carlo sampling might be the optimal choice. First, the random sampling averaged the bias due to arbitrary trial grain size. Second, the sampling approach can provide point estimates and a confidence interval to indicate the precision. Using random permutation splitting, Kopp and colleagues report encouraging reliability estimates ( $rel > 0.9$ ) for the M-WCST in the clinical setting (Kopp et al., 2021) and cWCST in a young sample in the lab setting (Steinke et al.,

2021). Our results yielded comparable reliability estimates between the random permutation and Monte Carlo sampling methods except for the measure of failure-to-maintain-set. The Monte Carlo method yields more reasonable estimates than the random permutation method on failure-to-maintain-set. As mentioned in the results section, about half of the young participants omitted zero errors, and half of the old participants omitted less than one error. That means the failure-to-maintain-set error was only in one split in many cases. The Monte Carlo method avoided this problem as it simulated a complete trial record.

### Sensitivity to cognitive aging

To deal with a world full of noise, people of our time need to filter interference, make forward-looking plans, inhibit useless or harmful behaviors, and change their mindset flexibly when the environment changes, the core of which is executive function. Despite its irreplaceability in human cognitive architecture, executive function is fragile and declines with age (Lacreuse et al., 2020). WCST and its variants are popular neuropsychological assessment tools for prefrontal or executive function. A recent meta-analysis suggests an association between prefrontal intactness and executive function. Moreover, compared with other executive function measures, the WCST indices have more robust correlations with the prefrontal volume size (Yuan & Raz, 2014).

The WCST and its variants have widespread usage in cognitive aging or normative studies (Esposito et al., 1999; Faria et al., 2015; Hartman et al., 2001; Heckner et al., 2021; Lineweaver et al., 1999; Marquine et al., 2021; Perez-Enriquez et al., 2021; Sanchez-Rodriguez et al., 2022; Shan et al., 2008). Most of those studies support age differences in the WCST measures, which is consistent with our findings. On seven major WCST indices, the old group (55~80 years) showed noticeable deterioration compared with the young group (18~45 years) in our study. To test whether there was continuous executive function decline with age in the old group, we also performed regression analyses controlling the confound of gender, education years, and digit span score. However, our results did not yield any significant linear aging trend on the seven measures. A study using Taiwan samples also reports age group differences but not a continuous linear decline with age (Shan et al., 2008). A recent study has revealed a complex aging pattern using a dataset of the Attention Network Test, which reveals that aging accompanies both improvement and decline (Verissimo et al., 2021). Thus, the executive function might have a non-linear dependence on aging. To verify this issue, a larger sample size with hierarchical sampling can benefit the linear and non-linear analysis of the aging effect.

### The collinearity among OCST measures

The standardized WCST test can provide up to 16 different indices. However, many indices are linear combinations of several other indices. For example, total error equals perseverative error plus non-perseverative error. Greve et al. (2005) conducted the first large-scale confirmatory factor analysis of the WCST (128-card version). Their study adopted seven major indices: total correct, perseverative error, perseverative response, non-perseverative error, conceptual level response, category completed, and failure-to-maintain-set. The perseverative error was removed from the final model due to the collinearity issue. Although the authors suggested a three-factor solution, the model fit was unsatisfactory, indicating the model might not reflect the data structure (Greve et al., 2005). Our correlation analysis reveals a serious collinearity problem among the seven measures used by Greve et al. (2005), with ten pairwise correlation coefficients larger than 0.8. An exception is failure-to-maintain-set, which only showed a moderate correlation with the category completed. Unfortunately, the reliability of the failure-to-maintain-set was unsatisfactory in our study. The collinearity issue questions the necessity of reporting many index scores in practice and research. As the current study used a 64-card, self-administered version in a Chinese community sample, whether the collinearity issues apply to other versions or populations should be checked in future studies. Moreover, the latent factor structure or the construct validity of WCST and its variants calls for additional research attention.

### Usage of OCST in online cognitive aging studies

The aging of the world is accelerating. Fast identification of individuals with abnormal aging risk can help the community and family take quick actions to weaken the negative consequences such as neurodegenerative diseases and falls. The worldwide epidemic, such as COVID-19, also raises the emergency of developing *Digital Neuropsychology* to provide online accessible neuropsychological test services (Germine et al., 2019; Steinke et al., 2021). Steinke et al. (2021) comprise the first valuable research evaluating the split-half reliability of a self-administered, computerized version of WCST (cWCST) in the lab by recruiting young volunteers. However, as far as we know, the present study is the first to explore an online version of the WCST-like task (OCST) in the community with both young and old volunteers. Unlike the cWCST, the OCST follows the original design and improved scoring scheme of standardized WCST-64. Thus, our results provide valuable information to researchers and practitioners planning to use the classical WCST version in community-based research.

## Limitations and future directions

The present study's automated scoring, informative visualization, and split-half reliability estimation procedure can significantly benefit clinicians and researchers. Furthermore, we provide reliability estimates for a publicly available card sorting task in the community-dwelling young and elderly sample. However, the sample size of the present study still needs to be increased to make a reliable norm. In addition, the validity of using the card sorting task as a fast cognitive screening tool should be justified with everyday function measures. Last but not least, it is worthwhile in the future to directly compare different WCST versions in the community sample.

## Conclusions

Executive function decline is a hallmark of cognitive aging, which increases the risk of cognitive impairment and falls in the elderly group. Fast executive function tests to identify high-risk individuals are necessary for the care service in the community. Card sorting tasks have been widely used as a measure of executive function. The study investigated the usability of an open-source, self-administered, online short-version card sorting task with a sample of young and old Chinese. We developed an automated scoring procedure following the recent recommendations on scoring perseverative responses to make the results comparable to the standardized WCST. Reliability estimates of commonly used measures were calculated using the split-half method. All task indices' reliabilities were reasonably good except for "failure to maintain-set." The R script of automated scoring and estimation of reliability was publicly available.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.3758/s13428-023-02104-6>.

**Author's contribution** LZ, HL, and ZZ designed the study. HL, TV, and CW provided research tools. LZ and ZZ analyzed the data and drafted the manuscript. All authors contributed to the paper revision and approved the final version.

**Funding** Data collection and preliminary analysis were sponsored by the Natural Science Fund of Anhui Province (2008085MC69), the Natural Science Fund of Hefei City (2021033), the General scientific research project of Anhui Provincial Health Commission (AHWJ2021b150), Collaborative Innovation Program of Hefei Science Center, CAS (2021HSC-CIP013), Anhui Province Key Laboratory of Medical Physics and Technology (LMPT201904).

**Data availability** The raw card sorting data is made available in an online depository (10.57760/sciencedb.06197), which can be used along with the source code.

**Code availability** The source code for all the analysis in the manuscript is available on GitHub under the MIT license (<https://github.com/ZhengKang-Zhang/WCST-index-calculation>). We also demonstrate an exemplary analysis using publically available WCST card sorting data.

## References

- Barceló, F. (2003). The Madrid card sorting test (MCST): a task switching paradigm to study executive attention with event-related potentials. *Brain Research Protocols*, 11(1), 27–37. [https://doi.org/10.1016/s1385-299x\(03\)00013-8](https://doi.org/10.1016/s1385-299x(03)00013-8)
- Barceló, F. (2021). A Predictive Processing Account of Card Sorting: Fast Proactive and Reactive Frontoparietal Cortical Dynamics during Inference and Learning of Perceptual Categories. *Journal of Cognitive Neuroscience*, 33(9), 1636–1656. [https://doi.org/10.1162/jocn\\_a\\_01662](https://doi.org/10.1162/jocn_a_01662)
- Berg, E. A. (1948). A simple objective technique for measuring flexibility in thinking. *The Journal of General Psychology*, 39, 15–22. <https://doi.org/10.1080/00221309.1948.9918159>
- Chiu, E. C., & Lee, S. C. (2021). Test–retest reliability of the Wisconsin Card Sorting Test in people with schizophrenia. *Disability and Rehabilitation*, 43(7), 996–1000. <https://doi.org/10.1080/09638288.2019.1647295>
- de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, 47(1), 1–12. <https://doi.org/10.3758/s13428-014-0458-y>
- Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, 64, 135–168. <https://doi.org/10.1146/annurev-psych-113011-143750>
- Eling, P., Derckx, K., & Maes, R. (2008). On the historical and conceptual background of the Wisconsin Card Sorting Test. *Brain and Cognition*, 67(3), 247–253. <https://doi.org/10.1016/j.bandc.2008.01.006>
- Esposito, G., Kirkby, B. S., Van Horn, J. D., Ellmore, T. M., & Berman, K. F. (1999). Context-dependent, neural system-specific neurophysiological concomitants of ageing: Mapping PET correlates during cognitive activation. *Brain*, 122(Pt 5), 963–979. <https://doi.org/10.1093/brain/122.5.963>
- Faria, C. A., Alves, H. V. D., & Charchat-Fichman, H. (2015). The most frequently used tests for assessing executive functions in aging. *Dement Neuropsychol*, 9(2), 149–155. <https://doi.org/10.1590/1980-57642015DN92000009>
- Flashman, L. A., Homer, M. D., & Freides, D. (1991). Note on scoring perseveration on the Wisconsin card sorting test. *The Clinical Neuropsychologist*, 5(2), 190–194. <https://doi.org/10.1080/13854049108403303>
- Fox, C. J., Mueller, S. T., Gray, H. M., Raber, J., & Piper, B. J. (2013). Evaluation of a short-form of the Berg Card Sorting Test. *PLoS One*, 8(5), e63885. <https://doi.org/10.1371/journal.pone.0063885>
- Frossard, J., & Renaud, O. (2021). Permutation Tests for Regression, ANOVA, and Comparison of Signals: The permuco Package. *Journal of Statistical Software*, 99(15), 1–32. <https://doi.org/10.18637/jss.v099.i15>
- Germine, L., Reinecke, K., & Chaytor, N. S. (2019). Digital neuropsychology: Challenges and opportunities at the intersection of science and software. *The Clinical Neuropsychologist*, 33(2), 271–286. <https://doi.org/10.1080/13854046.2018.1535662>
- Grant, D. A., & Berg, E. A. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *Journal of Experimental Psychology*, 38(4), 404–411. <https://doi.org/10.1037/h0059831>
- Greve, K. W. (2001). The WCST-64: A standardized short-form of the Wisconsin Card Sorting Test. *The Clinical Neuropsychologist*, 15(2), 228–234. <https://doi.org/10.1076/clin.15.2.228.1901>
- Greve, K. W., Stickler, T. R., Love, J. M., Bianchini, K. J., & Stanford, M. S. (2005). Latent structure of the Wisconsin Card Sorting Test: A confirmatory factor analytic study. *Archives of Clinical Neuropsychology*, 20(3), 355–364. <https://doi.org/10.1016/j.acn.2004.09.004>
- Haaland, K. Y., Vranes, L. F., Goodwin, J. S., & Garry, P. J. (1987). Wisconsin Card Sort Test performance in a healthy elderly



- population. *Journal of Gerontology*, 42(3), 345–346. <https://doi.org/10.1093/geronj/42.3.345>
- Hartman, M., Bolton, E., & Fehnel, S. E. (2001). Accounting for age differences on the Wisconsin Card Sorting Test: Decreased working memory, not inflexibility. *Psychology and Aging*, 16(3), 385–399. <https://doi.org/10.1037/0882-7974.16.3.385>
- Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtiss, G. (1981). *Wisconsin Card Sorting Test: Manual* (1st ed.). Psychological Assessment Resources.
- Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtiss, G. (1993). *Wisconsin Card Sorting Test manual: Revised and expanded*. Psychological Assessment Resources.
- Heckner, M. K., Cieslik, E. C., Eickhoff, S. B., Camilleri, J. A., Hoffstaedter, F., & Langner, R. (2021). The Aging Brain and Executive Functions Revisited: Implications from Meta-analytic and Functional-Connectivity Evidence. *Journal of Cognitive Neuroscience*, 33(9), 1716–1752. [https://doi.org/10.1162/jocn\\_a\\_01616](https://doi.org/10.1162/jocn_a_01616)
- Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, 50(3), 1166–1186. <https://doi.org/10.3758/s13428-017-0935-1>
- Ho, J., Tumkaya, T., Aryal, S., Choi, H., & Claridge-Chang, A. (2019). Moving beyond P values: Data analysis with estimation graphics. *Nature Methods*, 16(7), 565–566. <https://doi.org/10.1038/s41592-019-0470-3>
- Kopp, B., Lange, F., & Steinke, A. (2021). The reliability of the Wisconsin card sorting test in clinical practice. *Assessment*, 28(1), 248–263. <https://doi.org/10.1177/1073191119866257>
- Kusi-Mensah, K., Nuamah, N. D., Wemakor, S., Agorinya, J., Seidu, R., Martyn-Dickens, C., & Bateman, A. (2022). Assessment tools for executive function and adaptive function following brain pathology among children in developing country contexts: A scoping review of current tools. *Neuropsychology Review*, 32(3), 459–482. <https://doi.org/10.1007/s11065-021-09529-w>
- Lacreuse, A., Raz, N., Schmidtke, D., Hopkins, W. D., & Herndon, J. G. (2020). Age-related decline in executive function as a hallmark of cognitive ageing in primates: An overview of cognitive and neurobiological studies. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 375(1811), 20190618. <https://doi.org/10.1098/rstb.2019.0618>
- Lange, F., & Dewitte, S. (2019). Cognitive flexibility and pro-environmental behaviour: A multimethod approach. *European Journal of Personality*, 33(4), 488–505. <https://doi.org/10.1002/per.2204>
- Lange, F., Kroger, B., Steinke, A., Seer, C., Dengler, R., & Kopp, B. (2016). Decomposing card-sorting performance: Effects of working memory load and age-related changes. *Neuropsychology*, 30(5), 579–590. <https://doi.org/10.1037/neu0000271>
- Lange, F., Seer, C., & Kopp, B. (2017). Cognitive flexibility in neurological disorders: Cognitive components and event-related potentials. *Neuroscience and Biobehavioral Reviews*, 83, 496–507. <https://doi.org/10.1016/j.neubiorev.2017.09.011>
- Lineweaver, T. T., Bond, M. W., Thomas, R. G., & Salmon, D. P. (1999). A normative study of Nelson's (1976) modified version of the Wisconsin Card Sorting Test in healthy older adults. *The Clinical Neuropsychologist*, 13(3), 328–347. <https://doi.org/10.1076/clin.13.3.328.1745>
- Marquine, M. J., Yassai-Gonzalez, D., Perez-Tejada, A., Umlauf, A., Kamalyan, L., Morlett Paredes, A., ... Heaton, R. K. (2021). Demographically adjusted normative data for the Wisconsin card sorting test-64 item: Results from the neuropsychological norms for the U.S.–Mexico Border Region in Spanish (NP-NUMBRs) project. *The Clinical Neuropsychologist*, 35(2), 339–355. <https://doi.org/10.1080/13854046.2019.1703042>
- Miles, S., Howlett, C. A., Berryman, C., Nedeljkovic, M., Moseley, G. L., & Phillipou, A. (2021). Considerations for using the Wisconsin card sorting test to assess cognitive flexibility. *Behavior Research Methods*, 53(5), 2083–2091. <https://doi.org/10.3758/s13428-021-01551-3>
- Milner, B. (1963). Effects of different brain lesions on card sorting: the role of the frontal lobes. *Archives of Neurology*, 9, 100–110. <https://doi.org/10.1001/archneur.1963.00460070100010>
- Miranda, A. R., Franchetto Sierra, J., Martinez Roulet, A., Rivadero, L., Serra, S. V., & Soria, E. A. (2020). Age, education and gender effects on Wisconsin card sorting test: Standardization, reliability and validity in healthy Argentinian adults. *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition*, 27(6), 807–825. <https://doi.org/10.1080/13825585.2019.1693491>
- Montero-Odasso, M., & Speechley, M. (2018). Falls in cognitively impaired older adults: Implications for risk assessment and prevention. *Journal of the American Geriatrics Society*, 66(2), 367–375. <https://doi.org/10.1111/jgs.15219>
- Nelson, H. E. (1976). A modified card sorting test sensitive to frontal lobe defects. *Cortex*, 12(4), 313–324. [https://doi.org/10.1016/s0010-9452\(76\)80035-4](https://doi.org/10.1016/s0010-9452(76)80035-4)
- Nomi, J. S., Vij, S. G., Dajani, D. R., Steimke, R., Damaraju, E., Rachakonda, S., ... Uddin, L. Q. (2017). Chronnectomic patterns and neural flexibility underlie executive function. *Neuroimage*, 147, 861–871. <https://doi.org/10.1016/j.neuroimage.2016.10.026>
- Nyhus, E., & Barcelo, F. (2009). The Wisconsin card sorting test and the cognitive assessment of prefrontal executive functions: A critical update. *Brain and Cognition*, 71(3), 437–451. <https://doi.org/10.1016/j.bandc.2009.03.005>
- Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science*, 2(4), 378–395. <https://doi.org/10.1177/2515245919879695>
- Perez-Enriquez, C., Garcia-Escobar, G., Florido-Santiago, M., Pique-Candini, J., Arrondo-Elizaran, C., Grau-Guinea, L., ... Sanchez-Benavides, G. (2021). Spanish normative studies (NEURONORMA-Plus project): Norms for the Wisconsin card sorting test, the modified Taylor complex figure, and the ruff-light trail learning test. *Neurología (English Edition)*. <https://doi.org/10.1016/j.nrl.2021.05.017>
- Piper, B. J., Li, V., Eiwaz, M. A., Kobel, Y. V., Benice, T. S., Chu, A. M., ... Raber, J. (2012). Executive function on the psychology experiment building language tests. *Behavior Research Methods*, 44(1), 110–123. <https://doi.org/10.3758/s13428-011-0096-6>
- Pronk, T., Molenaar, D., Wiers, R. W., & Murre, J. (2022). Methods to split cognitive task data for estimating split-half reliability: A comprehensive review and systematic assessment. *Psychonomic Bulletin & Review*, 29(1), 44–54. <https://doi.org/10.3758/s13423-021-01948-3>
- R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>. Accessed 2022-03-10.
- Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology*, 20(1), 33–65. <https://doi.org/10.1016/j.acn.2004.02.005>
- Rabin, L. A., Spadaccini, A. T., Brodare, D. L., Grant, K. S., Elbulok-Charcape, M. M., & Barr, W. B. (2014). Utilization rates of computerized tests and test batteries among clinical neuropsychologists in the United States and Canada. *Professional Psychology: Research and Practice*, 45(5), 368–377. <https://doi.org/10.1037/a0037987>
- Sanchez-Rodriguez, J. L., Juarez-Vela, R., Santolalla-Arnedo, I., Ruiz de Vinaspre-Hernandez, R., & Sanchez-Gonzalez, J. L. (2022). An item response theory analysis of the Wisconsin card sorting



- test in normal aging, Alzheimer's disease and Parkinson's disease: Neurophysiological approach. *Journal of Personalized Medicine*, 12(4), 539. <https://doi.org/10.3390/jpm12040539>
- Schretlen, D. J. (2010). *Modified Wisconsin Card Sorting Test (M-WCST): Professional manual*. Psychological Assessment Resources.
- Shan, I. K., Chen, Y. S., Lee, Y. C., & Su, T. P. (2008). Adult normative data of the Wisconsin Card Sorting Test in Taiwan. *Journal of the Chinese Medical Association*, 71(10), 517–522. [https://doi.org/10.1016/S1726-4901\(08\)70160-6](https://doi.org/10.1016/S1726-4901(08)70160-6)
- Steinke, A., & Kopp, B. (2020). RELEX: An excel-based software tool for sampling split-half reliability coefficients. *Methods in Psychology*, 2, 100023. <https://doi.org/10.1016/j.metip.2020.100023>
- Steinke, A., Lange, F., & Kopp, B. (2020). Parallel model-based and model-free reinforcement learning for card sorting performance. *Scientific Reports*, 10(1), 15464. <https://doi.org/10.1038/s41598-020-72407-7>
- Steinke, A., Kopp, B., & Lange, F. (2021). The Wisconsin card sorting test: Split-half reliability estimates for a self-administered computerized variant. *Brain Sciences*, 11(5). <https://doi.org/10.3390/brainsci11050529>
- Stuss, D. T., & Benson, D. F. (1984). Neuropsychological studies of the frontal lobes. *Psychological Bulletin*, 95(1), 3–28. <https://doi.org/10.1037/0033-2909.95.1.3>
- Teubner-Rhodes, S., Vaden, K. I., Jr., Dubno, J. R., & Eckert, M. A. (2017). Cognitive persistence: Development and validation of a novel measure from the Wisconsin Card Sorting Test. *Neuropsychologia*, 102, 95–108. <https://doi.org/10.1016/j.neuropsychologia.2017.05.027>
- Uddin, L. Q. (2021). Cognitive and behavioural flexibility: neural mechanisms and clinical considerations. *Nature Reviews Neuroscience*, 22(3), 167–179. <https://doi.org/10.1038/s41583-021-00428-w>
- Vékony, T. (2022). *Card sorting task created with jsPsych (version 1.0.1) [computer software]*. <https://doi.org/10.5281/zenodo.7093470>
- Verissimo, J., Verhaeghen, P., Goldman, N., Weinstein, M., & Ullman, M. T. (2021). Evidence that ageing yields improvements as well as declines across attention and executive functions. *Nature Human Behaviour*. <https://doi.org/10.1038/s41562-021-01169-7>
- Wechsler, D. (1997). *Wechsler Adult Intelligence Scale* (3rd ed.). Psychological Corporation.
- Yuan, P., & Raz, N. (2014). Prefrontal cortex and executive functions in healthy adults: a meta-analysis of structural neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, 42, 180–192. <https://doi.org/10.1016/j.neubiorev.2014.02.005>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.