Jurnal Psikologi, 2021 (April), Vol. 20(1), 35-49

EQUIVALENCE OF TRADITIONAL AND INTERNET-DELIVERED TESTING OF WORD FLUENCY TASKS

Heni Gerda Pesau[1], Gilles van Luijtelaar [2]

[1]Psychology Faculty, Atma Jaya University, Makassar
Jl. Tanjung Alang No. 23, Makassar, Indonesia 90224
[2]Donders Centre for Cognition, Radboud University
PO Box 9104, 6500 HE Nijmegen, The Netherlands

heni_gerda@lecturer.uajm.ac.id

Abstract

Changes from traditional face-to-face to internet-delivered psychological assessment are urgently needed given the long-lasting pandemic, the general need for fast and efficient tests and test procedures, and easier availability and access for test-takers in remote settings. We used a quasi-experimental non-randomized group design for the comparison of two word fluency test procedures: one traditional that is face-to-face (n = 30) and one supervised via internet (n = 30). Participants were 17-31 years, education level high school and Bachelor. The letters S, K, T were used for the phonemic fluency test, for the emotion word fluency test subjects had to generate words related to subjective emotional feelings or the expression of emotions. The results showed that traditional administered and internet-delivered testing are equivalent (our hypothesis) as seen from the absence of significant differences between the two groups in the performances of all four word fluency tests ($p > .05$) and small effect sizes (Cohen's d range < .5). Significant correlations were found between the fluency tasks, irrespective of the way of test administration ($p < .05$). It can be concluded that the word fluency tasks can be assessed by supervised internet-delivered testing, but this is limited to a sample of young adults.

Keywords: equivalence; traditional testing; internet testing; word fluency tasks

INTRODUCTION

Advances in information technology are inevitable and affect all aspects of life including health care and health research, one of them is on neuropsychological assessment. Neuropsychological assessment is most often used for the diagnosis of brain-related cognitive problems, for clinical correlation with brain imaging findings, and the evaluation of the treatment response and the prediction of functioning potential (Harvey, 2012). The influence of technological advances on neuropsychological assessment is most clearly noticed in the implementation of tests that classically use traditional examiner-administered tests into Computerized Neuropsychological Assessment Devices (CNAD). CNAD can be defined as the use of various instruments such as computers, digital tablets, handheld devices, or digital devices that help testers or administrators to administer, decorate, or interpret tests related to brain function or factors related to neurologic health and illness that can be given by offline and also online way called internet-delivered testing (Bauer et al., 2012).

Internet-delivered testing can be defined as a subset of computer-based assessment where tests are conducted online or via internet in an unsupervised or controlled supervised and managed mode (Bartram & Coyne, 2005; Macqueen et al., 2018). The traditional examiner-administered test is one of the test administration methods in which participants interact directly with individuals who present stimuli, record verbal, movement, or written responses, and take notes on observing key behaviors (Bauer et al., 2012). Changes to the test method previously carried out through traditional examiner-administered to internet-delivered testing were not only inspired by technological developments but also by the need for faster and more efficient ways of test administration, the need for increased cost-effectiveness, global use of internet, and the need for faster and easier access, especially for test-takers in remote settings or areas

(Macqueen et al., 2018). Psychological and neuropsychological assessments through remote settings are urgently needed, especially given the current COVID-19 pandemic conditions which enforce physical distancing, although this need was present long before considering that not all test takers can conduct direct face-to-face tests (Velikonja et al., 2020).

Online computerized or internet-delivered testing is also used because it has several advantages when compared to traditional examiner-administered testing, among others, it can overcome the problem of distance by telephone or video call, participants can do the test at home or in the place where participants are and the time can be adjusted regarding procedures both with and without supervision (Feenstra et al., 2018), assist test takers who have difficulty going to the laboratory or clinic due to limited mobility and transportation which indirectly reduces the costs required (Bauer et al., 2012; Miller & Barr, 2017).

Actually, CNAD is not new in neuropsychological assessment, it has been widely used, e.g. to identify cognitive problems in children and adolescents with depression (Brooks et al., 2010), children with ADHD (Chamberlain et al., 2011), and in adults who experience mood disorders (Iverson et al., 2011). Some other tests were also adapted from traditional to digital or computerized versions, suited for healthy subjects. Examples are the computerized color Stroop Test which measures sensitivity for interference, selective attention, and cognitive flexibility (Din et al., 2019), and the Cambridge Computerized Cognitive Evaluation Adaptive Testing (CAMCOG-CAT) that has been used to measure various aspects of cognition among others in elderly suited to find indications of dementia and mild cognitive impairment (Zygouris & Tsolaki, 2015).

Research towards computerized psychological assessment has also been conducted in Indonesia (Setiatama & Kusrohmaniah, 2019). These authors investigated the effect of stress in a computerized version of the Stroop Test to obtain selective attention scores in young adults. Others have analyzed the comparison of psychometric properties between computer-based and paper-pencil versions of the Potential Academic Postgraduate Test (PAPs) and found that item difficulty, item discrimination index, and item fit index were equal in the computer-based and paper-pencil based version, so that PAPs can be administered in both forms (Marastuti et al., 2020). Other research was also carried out to investigate the effects of administrative time adaptation by using of Computer Answer Sheet on Culture Fair Intelligence Test (CFIT) 3A and 3B and it was found that adaptation administration time did not change the results of the intelligence test (Saptoto, 2018). It can be concluded that based on the limited number of research reports that the use of computerized tests in Indonesia, especially in the field of neuropsychology, has been successfully applied but is still limited.

Internet-delivered testing in neuropsychological assessment apart from having several advantages over traditional examiners also presents certain challenges related to computer configurations and internet connections that can affect stimulus presentation (Parsons et al., 2017). Other things that need to be considered such as large variations in the test environment where the tester cannot control what is done as in traditional delivered testing, limited or interrupted internet network problems, and limited interactions between test takers and test administer (Macqueen et al., 2018). Another weakness encountered is the possibility that participants who take the test are not the actual participants, but it can be overcome by supervising during the implementation of the test called supervised internet-delivered testing (Parsons et al., 2017). Supervised or proctored internet-delivered testing can be defined as a mode that requires an administrator to log in a candidate

and confirm that the test had been properly administered and completed (Bartram & Coyne, 2005).

The administration of computerized tests, in general, are more standardized than traditional versions, but it can be a disadvantage because it is also less flexible. Due to the standard procedure on computerized tests, the tester could not adjust the instruction given to the client who may not understand and require further explanation, which is common practice in the traditional administered testing (Schmand, 2019). This might be even more the case in ethnic and cultural minorities, where individuals are not yet accustomed to using the tools needed in computerized tests (Schmand, 2019). In addition, other weaknesses related to user limitations include older individuals who will find it difficult to use computerized test devices (Tierney et al., 2014), as well as individuals who are constrained in using electronics devices due to a lack of resources and opportunities. Therefore, it was decided to include only participants which are familiar with electronic devices, so that the lack of experience will not hamper the administration process and test responses.

One important issue in the use of internet-delivered testing is that there is equivalence between traditional and internet-delivered testing. Psychometric properties might change when traditional forms are changed into online versions (International Test Commission [ITC], 2014; Macqueen et al., 2018). American Psychological Association within its Guidelines for Computer-Based Tests and Interpretations explained that one aspect of determining equivalence between internet-based and traditional versions is achieved "if the means, dispersions, and shapes of the score distributions are approximately the same" (Green, 1991). Equivalence is not defined as no difference at all (zero difference) because sampling error may persist (Counsell & Cribbie, 2015).

Previous studies were aimed to evaluate the equivalency of test administration: a comparison of traditional versus video-conferencing administration of a neurobehavioral screening test found that there was no significant difference between both ways of administration (Duffield, 2011). Another study examined the equivalence of a graphics tablet-based computer administration of the Rey Complex Figure Test (RCF) with a traditional administration of the RCF and found no significant performance differences between groups (Riordan et al., 2013). Also, experimental studies on the interaction of participant responses to health-related messages between the paper-pencil and internet survey methods found that the two methods were equivalent (Lewis et al., 2009). Equivalence was also found between an internet-based and a traditional version of a development test (Vosylys et al., 2012); between the traditional and the computer-based Halstead Category Test (Goette et al., 2019). Also, for the Montreal face-to-face cognitive assessment and video-based telehealth system (DeYoung & Shenal, 2018), and for video teleconference and face-to-face test conditions for the assessment of neuropsychological status the same results were obtained (Galusha-Glasscock et al., 2016).

However, other studies found non-equivalence in a prospective memory questionnaire comparison between an online and paper-pencil way of administration (Buchanan et al., 2005). A comparison between traditional administered and online face-to-face interviews screening tests for substance involvement showed clear differences (Khazaal et al., 2015), and also another study noted a lack of equivalence in surveys (Liao & Hsieh, 2017). In all, it can be concluded that traditional and internet-delivered testing can be equivalent or non-equivalent. Given the variety in the type of tests and the way they were administered, it remains obscure why this is the case.

One of the tests commonly used in neuropsychological assessment is the verbal word fluency test (VFT): it measures verbal

abilities in fluency in producing words, the underlying cognitive domains are executive function, self-monitoring, inhibition, working memory, and language skills (Lezak et al., 2012; Schmand, 2019; Shao et al., 2014). One type of verbal fluency is phonemic verbal fluency or also called letter fluency (Abeare et al., 2016), when tested, participants are asked to produce as many words beginning with a certain letter (for example A, B, C, etc.) within one minute (Kim et al., 2018). Neuroimaging data indicate that in phonemic fluency tests, activation in the left is greater than in the right prefrontal cortex (Robinson et al., 2012). Another verbal fluency task is the Emotion Word Fluency Task (EWFT). It establishes the respondent's fluency regarding words related to emotions. Emotion words implicate the involvement of affective brain regions including amygdala, anterior cingulate cortex, dorsolateral prefrontal cortex, and the right cerebral hemisphere (Abbassi et al., 2011). The EWFT belongs to the category of semantic word fluency tests but is quite unique compared to other verbal fluency tests considering its distinctive affective content (Abeare et al, 2016).

Digital or computerized version of the phonemic fluency tasks, such as in Brazilian children and teenagers (Dias & Seabra, 2014) has been used. Likewise in the computerized self-test (CST), a cognitive screening test for Alzheimer's disease and mild cognitive impairment, verbal fluency is one of the six cognitive domains. The CST uses internet and is administered through written and oral instructions (Dougherty et al., 2010). These two studies showed that the fluency test, which has a long tradition as a neuropsychological test, can be adapted to a computerized way of administration.

As mentioned above, the shift from traditional to internet-delivered testing cannot be avoided and the tendency to use internet for assessment will continue to develop. Internet tests must meet psychometric properties comparable to traditional tests including standardized assessment procedures and measurements. Therefore, in the current work, the equivalence of two ways of test administration of two verbal fluency tests will be established to see whether the data collection via internet yields the same results as the traditional way of administration. This study hypothesizes that there is equivalence between the two ways of test administration of the two word fluency tasks. This implies that there will be no differences in performance between the two groups on both fluency tests, and rather similar significant correlations among the different fluency tests in each group. It implies also that there will be similar relations between the performance of the two word fluency tests with the demographic factors education and age in both ways of test administration. Finally, the correlation coefficients between the two word fluency tests will also be used to estimate the validity of the two verbal tests.

METHOD

This study used a quasi-experimental method with no randomization of a preselected group, instead: the test assistants choose participants in their environment fulfilling the inclusion criteria. Participants belonged to the two nonrandomized groups: the first group took the traditional way of test administration; the second group took the tests online via internet. In the traditional group, the tester came to the participant's place and communicated face-to-face with the participants, gave instructions; and observed the participant directly. The internet-delivered testing group got the tests online via a video session, allowing to observe the participants and to communicate with them directly.

Participants

The participants were females and males living in Makassar, aged 17-31 years. Other inclusion criteria were a high school and bachelor level of education, being fluent in the Indonesian language (because the instruction and response were given in Indonesian) and without a neurological or psychiatric history. In the internet group, an additional criterion

was being familiar with and accustomed to using pc or tablets and internet tools. Participants of the traditional group were recruited and interviewed face to face regarding the inclusion criteria, while the participants of the internet delivered group were selected and registered by e-mail and Google Form. The study sample consisted of two times thirty participants (two-nonrandomized groups) who met the criteria set previously. All declared willing to participate, and that their data could be used for research purposes through informed consent. Demographic characteristics of the two groups of participants are presented in Table 1.

Table 1.
Demographic Characteristic of the Sample

Demographic Characteristic	Traditional Testing (n = 30)		Supervised-Internet Testing (n = 30)	
Age (years)	Male	Female	Male	Female
17-19	1	5	1	13
20-22	1	7	6	6
23-25	1	3	0	2
26-28	4	6	0	0
29-31	0	2	0	2
Mean (*SD*)	23.53 (3.83)		20.50 (3.13)	
Mean dif.	3.03			
t-value (*df* = 58)	3.36			
Sig. (2-tailed)	.001*			
Education level (in years)	Male	Female	Male	Female
Senior High School (12 years)	4	15	5	14
Diploma/Bachelor (≥ 15 years)	3	8	2	9
Mean (*SD*)	13.43 (1.92)		13.47 (1.96)	
Mean dif.	-.03			
t-value (*df* = 58)	-.07			
Sig. (2 tailed)	.95			

It was evaluated with Student's T-tests for independent groups and Chi-squared test whether there were differences in the demographic properties between the two groups. The results show that there is no difference between the distribution of the sexes in the two samples (χ^2 = .09, p > .05). There is a significant age difference between the two groups: the age of the traditional group (M = 23.53 years) and the internet group (M = 20.50 years). There is no significant difference in the length of education in years between the two groups, both have about 13 years of education.

Measurement

1. Phonemic Verbal Fluency Test (PVFT)
   The verbal fluency task is a test that is used to measure the fluency in producing words. This is done by asking participants to say as many words as possible within one minute in two forms of tasks, phonemic or semantic verbal fluency (Abeare et al., 2016); this study used phonemic verbal fluency (PVF). Generally, in PVFT, participants are asked to produce words that start with certain letters, for example, A, B, C, and others (Kim et al., 2018). Specifically for this study, the three letters S, K, T, were used. The letter categories were taken from Hendrawan and Hatta (2010).
2. Emotion Word Fluency Task (EWFT)
   EWFT is used to measure the fluency of respondents in producing words related to emotions, participants are instructed to say as many words related to emotions within one minute (Abeare et al., 2016).

After the instruction, the one-minute period started. All words spoken by participants were recorded. The number of correct words (excluding the wrong and repeated words) in each category were counted as correct. Correct words that are mentioned twice were counted once. Errors were words given not according to the provisions of letters on PVFT and words that do not include emotions on EWFT. Scored was the number of correct words per phonemic category as well as the total score of the three phonemic subtests, the latter variable is most often clinically used. For EWFT, responses are considered as correct if the word describes an emotional expression such as smiling or an emotional state or condition, such as feeling happy. Determination or classification of the correct responses was done through an inter-rater process.

Research procedure

The tests for the traditional examiner-administered group were administered by research assistants: third/fourth-year psychology students who have sufficient knowledge on how to administer psychological tests. They attended a two-day workshop aiming at providing them knowledge of ten neuropsychological tests, practicing the test administration procedures, and scoring system. Next, at the end of the two-day workshop, they proved their test administration proficiency by reporting the scores of all ten tests including the PVFT before they were allowed to start collecting the real data from the healthy subjects. In all, the PVFT and EWFT were given as part of the neuropsychological test battery where the tester gives direct face-to-face instructions to the participants and records the responses.

In internet-delivered testing, the tests were given by the researcher. It begins with giving instructions through a PowerPoint program and an audio recording of the researcher's voice so that the mode of the given instruction is the same for all participants. The instructions used were the same as on traditional examiner-administered instructions. In both settings, PVFT was given first with one minute each for the letters S, K, T then followed by one minute for the EWFT.

Prior to the implementation of the internet testing, a pilot study was done to explore the way to give instructions and which program to use. Based on these experiences, the following protocol was followed. Before starting the test, participants were asked to prepare their pc or tablet including a good video camera. The tester ensured that the participant understands and masters the use of media applications and has access to a working internet connection. Participants were also asked to be sure that they have allocated an appropriate time of the day with a quiet and comfortable place in their home situation so that no interference could occur. The tester made a video call with the participants by Zoom cloud meeting and made sure that there is a good quality video and audio signal. The test begins after the participants mentioned that they were ready. Next, the instructions were given in writing with a PowerPoint program and oral via the audio connection. During the test, the tester can observe the participant, his responses, and obstacles that might be faced by him, while the participant cannot see the tester and only sees a computer screen that contains the test instructions.

Data analysis

Parametric tests can be performed if the data meet two requirements: first, the distribution of the test scores (phonemic S, K, T, Total phonemic score, and emotion word fluency score) must be normal. The second prerequisite is that the variance of the data is homogenous across the two conditions. The normality and variance homogeneity test showed that the data from the two groups had a normal distribution ($p > .05$) and similar variances ($p > .05$). Therefore, scores on the tests were analyzed with parametric tests: independent sample *t*-test and Cohen's *d* were used to evaluate differences between the two ways of testing and to determine their effect size (Shaughnessy et al., 2012) for each of the

five fluency scores. Cohen's *d* was calculated as the difference between the mean scores on the two different administration formats, divided by the pooled standard deviation of scores. Effect size is defined as small when $d \leq .20$ (not zero) to .49, medium when $d > .49$ and $< .79$, and large when $d \geq .80$ (Cohen, 1988).

The validity of the instruments was determined by the Pearson's Product-Moment Correlation (*r*) (Odom & Morrow, 2006) between the three different PVFT scores and the correlations between the EWFT and the three PVFT test scores; the type of validity is criterion-related validity and refers to the relationship between scores obtained using the instrument and scores obtained using one or more other instruments or measures (Fraenkel et al., 2012). An $r > .6$ is considered as an excellent indication of convergent validity (Post, 2016).

This study also examined differences between traditional and internet-delivered testing of scores of fluency test and whether the demographic factors such as age, sex, and education level affected the score, and this was examined by multifactor (age, education, sex, way of test administration) MANOVA (multivariate analysis of variance since all dependent variables were tested together, the *F* were according to Pillai's trace, partial eta squared was used to establish the effect sizes). In addition, the correlation between scores of the different fluency tasks and between demographic factors were analyzed by Spearman's correlation coefficient. Data analysis was performed using SPSS 22.0 for Windows and an effect size calculator for *t*-test (Social Science Statistics, 2018). We used an alpha level of .05 for all statistical tests.

RESULT AND DISCUSSION

Data analysis was carried out to test whether the test scores obtained from the two forms of test administration, traditional and via internet showed equivalence. That is, among others, that there are no significant differences in test scores in the two forms of administration. The data, as presented in Table 2, indicate that there were no significant differences between traditional and internet-delivered testing on the scores of phonemic S, $t(58) = -1.31$, $p = .20$; phonemic K, $t(58) = -1.77$, $p = .08$; phonemic T, $t(58) = -.73$, $p = .47$; total score of VFT, $t(58) = -1.44$, $p = .15$; and of the EWFT, $t(58) = .97$, $p = .34$. The data in Table 2 indicate as well that the dispersion is rather similar for the two ways of test administration. These results indicate that the hypothesis is accepted: there is equivalence between traditional and internet-delivered testing since there are no significant differences in scores between the two test administration procedures. In addition, the effect sizes regarding the differences between the scores of the two ways of administration (Cohen's *d* ranged from .19 - .46) were small.

Table 2.
Comparison between the Two Ways of Test Administration

Test	TT	SIT	*Mean*	*Sig.*	*t-test* [a]	Effect size
	Mean (*SD*)	*Mean* (*SD*)	dif.	(2-tailed)		(*d*)[b]
PVFT (S)	12.20 (4.75)	13.90 (5.29)	-1.70	.20	-1.31	.34
PVFT (K)	13.60 (4.22)	16.00 (6.13)	-2.40	.08	-1.77	.46
PVFT (T)	12.90 (5.33)	13.83 (4.58)	-.93	.47	-.73	.19
Total VFT	38.70 (13.24)	43.73 (13.76)	-5.03	.15	-1.44	.37
EWFT	8.13 (3.66)	7.23 (3.56)	.90	.34	.97	.25

Notes. TT = Traditional Testing; SIT = Supervised-Internet Testing; PVFT = Phonemic Verbal Fluency Test; VFT = Verbal Fluency Test; EWFT = Emotion Word Fluency Test.
[a]*t*-value at $df = 58$.
[b]The values of Cohen's *d* effect sizes were calculated based on the *mean* and *standard deviation* scores.

The results obtained by comparing traditional groups and internet-delivered testing in completing fluency tasks indicate that there were no significant differences with only small effect sizes in the performance of the phonemic S, K, T, total of PVFT, and EWFT. Similar results, that is equivalence, were obtained in previous studies in which two test administration procedures were compared on a few neuropsychology tests (Duffield, 2011; Goette et al., 2019; Riordan et al., 2013). These results imply that at least some aspect of equivalence is reached as American Psychological Association within its Guidelines for Computer-Based Tests and Interpretations explained: "if the means, dispersions, and shapes of the score distributions are approximately the same" (Green, 1991).

In addition, the correlation coefficients in Table 3 showed significant and strong positive correlations from $r = .43$, $p < .05$ to r = .93, $p < .05$ for the phonemic subtests with the other phonemic test scores in both groups. The EWFT also has a significant positive correlation with the phonemic tests scores but much lower, namely ranging between $r = .38$, $p < .05$ to $r = .50$, $p < .05$ and this was again the case for both groups. The equal patterns of correlation coefficients in the two groups contribute to the equivalence of the two ways of test administration as well.

The high inter-correlated scores of the three phonemic tests reveal that they represent the same cognitive domains: executive function, self-monitoring, inhibition, working memory, and lexical access (Kim et al., 2018; Schmand, 2019), while the correlations of the phonemic tests with the scores of the EWFT are still significant, but much lower. The latter suggests that the EWFT differs from the phonemic fluency test. The reason for the lower correlation is that the EWFT score reflects a partially different cognitive domain, it is a semantic word fluency task that contains affective content involving different brain structures as the phonemic tasks. Most likely, affective regions including the amygdala, anterior cingulate cortex, dorsolateral prefrontal cortex, and the right cerebral hemisphere are additionally activated (Abbassi et al., 2011).

Table 3.
Inter-correlation of Phonemic and Emotion Word Fluency Tasks

		VFT (S)		VFT (K)		VFT (T)		Total VFT		EWFT	
		TT	SIT	TT	SIT	TT	SIT	TT	SIT	TT	SIT
VFT (S)	*r*	-	-	.83	.55	.75	.47	.92	.78	38	.43
	Sig.	-	-	.000*	.002*	.000*	.009*	.000*	.000*	.037*	.017*
VFT (K)	*r*	-	-	-	-	.79	.80	.93	.92	.42	.54
	Sig.	-	-	-	-	.000*	.000*	.000*	.000*	.021*	.002*
VFT (T)	*r*	-	-	-	-	-	-	.92	.87	.39	.30
	Sig.	-	-	-	-	-	-	.000*	.000*	.033*	.112
Total VFT	*r*	-	-	-	-	-	-	-	-	.43	.50
	Sig.	-	-	-	-	-	-	-	-	.018*	.005*
EWFT	*r*	-	-	-	-	-	-	-	-	-	-
	Sig.	-	-	-	-	-	-	-	-	-	-

Note. VFT = Verbal Fluency Test; EWFT = Emotion Word Fluency Test; TT = Traditional Testing; SIT = Supervised-Internet Testing.
*$p < .05$ (2-tailed).

Next, the result (Table 4) showed there was no significant correlation between age, sex, and education with phonemic S, K, T, total VFT, and EWFT scores, both on traditional and internet-delivered testing. The correlation coefficients between the demographic factors and the scores of the tests within each of the groups also showed a similar pattern of only non-significant correlations for both groups.

Table 4.
Correlation Between Demographic Factors and Word Fluency Scores

		Age		Sex		Education	
		TT	SIT	TT	SIT	TT	SIT
VFT (S)	r_s	-.21	.11	-.16	-.10	.02	-.08
	Sig.	.27	.58	.39	.60	.93	.67
VFT (K)	r_s	-.22	.21	-.20	-.05	-.09	.13
	Sig.	.24	.28	.29	.81	.63	.50
VFT (T)	r_s	-.21	.13	-.25	-.04	-.04	.15
	Sig.	.27	.51	.19	.85	.85	.45
Total VFT	r_s	-.22	.17	-.22	-.07	-.05	.09
	Sig.	.24	.37	.24	.70	.80	.64
EWFT	r_s	-.04	-.11	-.01	.06	-.11	-.14
	Sig.	.83	.57	.96	.75	.55	.46

Note. VFT = Verbal Fluency Test; EWFT = Emotion Word Fluency Test; TT = Traditional Testing; SIT = Supervised-Internet Testing.

Finally, the effects of the demographic factors and the way of test administration on the phonemic fluency tests (S, K, T, and EVF) were explored together, the outcomes are presented in Table 5. The outcomes of the MANOVA, as reported in Table 5, are complementary to the outcomes of the *t*-tests as presented in Table 2. They answer the major hypothesis of whether there are quantitative differences between the word fluency test scores administered either in a classical way or via internet. Through this simultaneous analysis of all dependent variables (phonemic S, K, T, Total, and EVF), we found no significant difference between the dependent variables for the way the tests were administered $F(4, 44) = .49$, $p = .74$; $\eta_p^2 = .04$; for the factor age $F(4, 44) = 1.54$, $p = .21$; $\eta_p^2 = .12$; for the factor sex $F(4, 44) = .63$, $p = .64$; $\eta_p^2 = .05$; and level of education, $F(4, 44) = 1.28$, $p = .29$; $\eta_p^2 = .11$. All effect sizes were moderate to small. Also, the first and second-order interactions were non-significant with small effect sizes, demonstrating that the effects of age, sex, and education were equally absent and did not differ between the traditional and internet-tested groups.

The results showed equivalence of the two ways of test administration. The different forms of administration did not cause significant differences in the scores of the tests, next they showed similar correlation patterns for the two ways of administration between the five test scores as well as a similar lack of demographic effects in both groups.

Table 5.
Multivariate Analysis of Variance Outcomes

Independent Variables	*df*	*F*	*Sig.*	η_p^2
Testing administration	4, 44	.49	.74	.04
Sex	4, 44	.63	.64	.05
Age	4, 44	1.54	.21	.12
Education	4, 44	1.28	.29	.11
Testing administration & sex	4, 44	.62	.65	.05
Testing administration & age	4, 44	2.04	.11	.16
Testing administration & education	4, 44	1.18	.33	.09
Sex & age	4, 44	.19	.95	.02
Sex & education	4, 44	.83	.52	.07
Age & education	4, 44	2.08	.10	.16
Testing administration & sex & age	4, 44	1.18	.33	.10

Note. The missing interactions are due to an insufficient number of degrees of freedom

There are several factors that influence the equivalence of the two test administration procedures. Nonequivalence for the two test administration procedures has been ascribed to differences in test material (Liao & Hsieh, 2017) and an internet administration effect (Buchanan et al., 2005). Controlling some of the factors on the internet delivered testing can minimize the administration's negative impact of the internet testing and in this way, the attainment of test equivalency can be inhibited.

First, the limited interaction is considered as one of the weaknesses of digital or online testing (Schmand, 2019). The observations by the tester during the traditional examiner-administered test, e.g., the facial expressions and body language and whether the participant is ready and comfortable to take the test is part of a clinical assessment and evaluation routine. The face-to-face testing facilitates interactions and asking questions regarding whether the instructions given were clear or not. We tried to do the same on internet-delivered testing so that the test implementation conditions were close to the face-to-face test conditions, namely by implementing supervised internet-delivered testing with the webcam switched on. Also, others mentioned that in case tests are given via internet or videoconferencing, one must ensure that the videoconference interaction mimics the traditional face-to-face test administration as much as possible (Grosch, 2011). One of the reasons why equivalence was obtained could be that supervised internet testing was chosen (Macqueen et al., 2018) so that the testers and participants could interact, participants could ask if the instructions were not clear. The researcher in this study arranged the test situation by asking participants by activating the video so that the researcher could ask questions and observe the readiness of the test taker. Supervised internet testing has the additional advantage that it minimizes the possibility of faking, cheating, or replacing (Parsons et al., 2017).

Second, external factors related to disturbances or interferences from the surrounding environment where participants take the test were minimized, as well the quality of the internet connection, and the mastery of tools and applications were checked whether the quality of audio and video were good (Grosch, 2011; Tierney et al., 2014). Next, in order to minimize disturbances or obstacles from the environment, participants were asked to choose a suitable time and a quiet place and to check the internet network and applications before the test was administered. The selection of research participants in the current study is also considered to affect the obtained equivalence: we included participants aged 17-31 years with high school or bachelor education that master and often use the video meeting application.

Other factors can contribute to equivalence, such as the design and presentation of the test materials, the way the instructions are given, and the input devices such as mouse and keyboard (Bartram & Coyne, 2005; Mayer & Krampen, 2015). Some of these issues can be overcome by a pilot study to ensure that written and audio instructions will be received clearly. Another issue is that the participants in our protocol did not have to use input devices such as a mouse or keyboard, instead, they responded in the form of spoken words. Previous research also showed that VFT scores were not influenced if the data were collected through the videoconference method because participants gave verbal responses directly, and this contrasts the outcomes of tests that require interaction with physical objects such as the Mini-Mental Status Exam (MMSE) test and clock drawing test (Brearly et al., 2017).

One of the external factors that was not controlled is that the traditionally administered fluency tasks were given as part of a neuropsychology test battery, while via internet only two tests were administered. Although equivalence of the two test procedures was achieved, this aspect is a limitation of our study, and it can be considered for further research. Further

research can also be carried out by the evaluation and comparison of other cognitive and neuropsychological tests via internet with a cross-over design. Important is that the verbal fluency tests do not require written answers, only verbal responses and are therefore well suited for being used via internet. Perhaps other tests requiring verbal responses, such as the digit span, the Stroop test, the Boston Naming Test, or Raven's auditory verbal learning and memory test can be adapted for internet usage and compared with traditional face-to-face administration. Other possibilities for further research are comparisons between paper and pencil tests versus online and real-time combined with surveys and questionnaires with written replies. It is however not expected that people with no or little experience with computer screens and usage of a mouse or touchscreens can be tested in this modern way. They might need supervision in doing this test through internet. The same might be the case for elderly people, as well as for different categories of neuropsychological patients.

However, assessment in those who have experience with internet tools is reliable, and is equivalent to the traditional way of assessment, although the clinical validity of the Indonesian versions of both verbal fluency tests still needs to be determined. The fact that subjects do not have to give written responses and that they can be supervised makes the online versions of the fluency tests easy and trustworthy instruments for subjects who are familiar with internet tools.

CONCLUSION

Traditional and internet-delivered testing are equivalent in that there is no significant difference in the results of fluency tasks both in phonemic (S, K, T) and emotion word fluency tasks in a sample of young adults. Next, there are no differences between the two ways of test administration in the contribution of the demographic factors age, sex, and level of education, also suggesting equivalence. Equivalence is also inferred from the inter-correlation patterns, both within the various test scores and between the demographic factors and test scores. Therefore, both word fluency tasks are suited to be used online. This study shows some weaknesses that can be considered by subsequent researchers such as small sample sizes, the small differences in age between the traditional and internet groups. Also, a larger range of levels of education and age might have been preferred, while still taking experience with electronic devices under control, and differences in the provision of different fluency tasks as part of a series of tests or only two tests. There is, however, no evidence that these not controlled factors had any influence on the scores of the word fluency tasks and jeopardize the conclusions that supervised internet testing of verbal fluency task yields equivalent test scores in the investigated sample, but that could be considered for further research. The results of this study are also limited to a young sample where the results obtained may be different in older age participants.

ACKNOWLEDGMENT

This study was sponsored by Atma Jaya University, Makassar, Indonesia. The authors would like to thank all participants and raters (students of Psychology Faculty of Atma Jaya University, Makassar). The authors would like to also thank Dr. Augustina Sulastri, Psikolog, Faculty of Psychology, Soegijapranata Catholic University, Semarang, who allowed the use of the data of Makassar participants that were collected for a larger research project regarding obtaining normative scores for relevant clinical neuropsychological tests in Indonesia

REFERENCES

Abbassi, E., Kahlaoui, K., Wilson, M.A., & Joanette, Y. (2011). Processing the emotions in words: The complementary contributions of the left and right hemispheres. *Cognitive, Affective, & Behavioral Neuroscience*, *11*(3), 372–385. https://doi.org/10.3758/s13415-011-0034-1

Abeare, C. A., Freund, S., Kaploun, K., McAuley, T., & Dumitrescu, C. (2016). The emotion word fluency test (EWFT): Initial psychometric, validation, and physiological evidence in young adults. *Journal of Clinical and Experimental Neuropsychology*, *39*(8), 738–752. https://doi.org/10.1080/13803395.2016.1259396

Bauer, R. M., Iverson, G. L., Cernich, A. N., Binder, L. M., Ruff, R. M., & Naugle, R. I. (2012). Computerized neuropsychological assessment devices: Joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. *Archives of Clinical Neuropsychology*, *27*(3), 362–373. https://doi.org/10.1093/arclin/acs027

Bartram, D., & Coyne, I. (2005). *International guidelines on computer-based and internet-delivered testing*. International Test Commission (ITC). https://ptc.bps.org.uk/sites/ptc.bps.org.uk/files/guidance_documents/international_guidelines_on_computer-based_and_internet_delivered_tests.pdf

Brearly, T. W., Shura, R. D., Martindale, S. L., Lazowski, R. A., Luxton, D. D., Shenal, B. V., & Rowland, J. A. (2017). Neuropsychological test administration by videoconference: A systematic review and meta-analysis. *Neuropsychology Review*, *27*(2), 174-186. https://doi.org/10.1007/s11065-017-9349-1

Brooks, B. L., Iverson, G. L., Sherman, E. M. S., & Roberge, M. C. (2010). Identifying cognitive problems in children and adolescents with depression using computerized neuropsychological testing. *Applied Neuropsychology*, *17*(1), 37–43. https://doi.org/10.1080/09084280903526083

Buchanan, T., Ali, T., Heffernan, T. M., Ling, J., Parrott, A. C., Rodgers, J., & Scholey, A. B. (2005). Nonequivalence of on-line and paper-and-pencil psychological tests: The case of the prospective memory questionnaire. *Behavior Research Methods*, *37*(1), 148–154. https://doi.org/10.3758/BF03206409

Chamberlain, S. R., Robbins, T. W., Winder-Rhodes, S., Mller, U., Sahakian, B. J., Blackwell, A. D., & Barnett, J. H. (2011). Translational approaches to frontostriatal dysfunction in attention-deficit/hyperactivity disorder using a computerized neuropsychological battery. *Biological Psychiatry*, *69*(12), 1192–1203. https://doi.org/10.1016/j.biopsych.2010.08.019

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203771587

Counsell, A., & Cribbie, R.A. (2015). Equivalence tests for comparing correlation and regression coefficients. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 292–309. https://doi.org/10.1111/bmsp.12045

DeYoung, N., & Shenal, B. V. (2018). The reliability of the Montreal Cognitive Assessment using telehealth in a rural setting with veterans. *Journal of Telemedicine and Telecare*, *25*(4), 197-203. https://doi.org/10.1177/1357633X17752030

Dias, N. M., & Seabra, A. G. (2014). Teste de fluência FAS em crianças e adolescentes brasileiros: demandas executivas e efeitos de idade e gênero [The FAS fluency test in Brazilian children and teenagers: executive demands and the effects of age and gender]. *Arquivos de Neuro-Psiquiatria*, *72*(1), 55–62. https://doi.org/10.1590/0004-282X20130213.

Din, N. C., Chia, E., & Meng, T. (2019).

Computerized stroop tests: A review. *Journal of Psychology & Psychotherapy*, *9*(1), 2161–0487. https://doi.org/10.4172/2161-0487.1000353

Dougherty, J. H., Cannon, R. L., Nicholas, C. R., Hall, L., Hare, F., Carr, E., Arunthamakun, J. (2010). The computerized self test (CST): An interactive, internet accessible cognitive screening test for dementia. *Journal of Alzheimer's Disease*, *20*(1), 185–195. https://doi.org/10.3233/JAD-2010-1354

Duffield, T. C. (2011). *A Comparison of paper-pencil versus video-conferencing administration of a neurobehavioral screening test* [Master's thesis, University of Central Florida]. https://stars.library.ucf.edu/etd/1919

Feenstra, H. E. M., Murre, J. M. J., Vermeulen, I. E., Kieffer, J. M., & Schagen, S. B. (2018). Reliability and validity of a self-administered tool for online neuropsychological testing: The Amsterdam cognition scan. *Journal of Clinical and Experimental Neuropsychology*, *40*(3), 253–273. https://doi.org/10.1080/13803395.2017.1339017

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.). McGraw Hill. https://saochhengpheng.files.wordpress.com/2017/03/jack_fraenkel_norman_wallen_helen_hyun-how_to_design_and_evaluate_research_in_education_8th_edition_-mcgraw-hill_humanities_social_sciences_languages2011.pdf

Galusha-Glasscock, J. M., Horton, D. K., Weiner, M. F., & Cullum, C. M. (2015). Video teleconference administration of the Repeatable Battery for the assessment of neuropsychological status. *Archives of Clinical Neuropsychology*, *31*(1), 8-11. https://doi.org/10.1093/arclin/acv058

Goette, W.F., Schmitt, A., & Nici, J. (2019). Psychometric equivalence of the computerized and original halstead category test using a matched archival sample. *Assessment*, 00(0), 1-13. https://doi.org/10.1177/1073191119887444

Green, B. F. (1991). Guidelines for computer testing. In B. G. Terry & S. L., Wise (Eds.), *The computer and the decision-making process* (pp. 245–273). Lawrence Erlbaum Associates, Inc. https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1011&context=burosc omputerdecision

Grosch, M. C., Gottlieb, M. C., & Cullum, C. M. (2011). Initial practice recommendations for teleneuropsychology. *The Clinical Neuropsychologist*, *25*(7), 1119-1133. https://dx.doi.org/10.1080/13854046.2011.609840

Harvey P. D. (2012). Clinical applications of neuropsychological assessment. *Dialogues in clinical neuroscience*, *14*(1), 91–99. https://doi.org/10.31887/DCNS.2012.14.1/pharvey

Hendrawan, D., & Hatta, T. (2010). Evaluation of stimuli for development of the Indonesian version of verbal fluency task using ranking method. *Psychologia*, *53*(1), 14–26. https://doi.org/10.2117/psysoc.2010.14

International Test Commission. (2014). ITC guidelines on quality control in scoring, test analysis, and reporting of test scores. *International Journal of Testing*, *14*(3), 195-217. https://doi.org/10.1080/15305058.2014.918040

Iverson, G. L., Brooks, B. L., Langenecker, S. A., & Young, A. H. (2011). Identifying a

cognitive impairment subgroup in adults with mood disorders. *Journal of Affective Disorders*, *132*(3), 360–367. https://doi.org/10.1016/j.jad.2011.03.001

Khazaal, Y., Chatton, A., Monney, G., Nallet, A., Khan, R., Zullino, D., & Etter, J. F. (2015). Internal consistency and measurement equivalence of the cannabis screening questions on the paper-and-pencil face-to-face ASSIST versus the online instrument. *Substance Abuse: Treatment, Prevention, and Policy*, *10*(1), 2–11. https://doi.org/10.1186/s13011-015-0002-9

Kim, I., Millin, N. J., & Hwang, J. (2018). Word retrieval by verbal fluency tasks for young and old people: An fNIR study. *Clinical Archives of Communication Disorders*, *3*(1), 52–58. https://doi.org/10.21849/cacd.2018.00318

Lewis, I., Watson, B., & White, K. M. (2009). Internet versus paper-and-pencil survey methods in psychological experiments: Equivalence testing of participant responses to health-related messages. *Australian Journal of Psychology*, *61*(2), 107–116. https://doi.org/10.1080/00049530802105865

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). Neuropsychological assessment (5th ed.). Oxford University Press

Liao, P. W., & Hsieh, J. Y. (2017). Does internet-based survey have more stable and unbiased results than paper-and-pencil survey? *Open Journal of Social Sciences*, *05*(01), 69–86. https://doi.org/10.4236/jss.2017.51006

Macqueen, P., Howe, W., & Power, M. (2018). *Online psychological testing*. Australian Psychological Society. https://www.psychology.org.au/APS/media/Resource-Finder/Testing/Online-psychological-testing.pdf

Marastuti, A., Anggoro, W. J., Marvianto, R. D., & Al Afghani, A. A. (2020). Perbandingan properti psikometri antara tes PAPs berbentuk computer-based dan paper and pencil test. *Gadjah Mada Journal of Psychology*, *6*(1), 12-28. https://doi.org/10.22146/gamajop.51852

Mayer, A. K., & Krampen, G. (2015, July 22 - 25). Equivalence of computerized versus paper - and - pencil testing of information literacy under controlled versus uncontrolled conditions: An experimental study [Conference session]. *13th European Conference on Psychological Assessment*, Zurich/Switzerland. https://docplayer.net/10229346-Anne-kathrin-mayer-gunter-krampen-zpid-leibniz-institute-for-psychology-information-trier-germany.html

Miller, J. B., & Barr, W. B. (2017). The technology crisis in neuropsychology. *Archives of Clinical Neuropsychology*, *32*(5), 541-554. https://doi.org/10.1093/arclin%2Facx050

Odom, L.R. & Morrow, J.R, Jr. (2006). What's this r? A correlational approach to explaining validity, reliability and objectivity coefficients. *Measurement in Physical Education and Exercise Science*, *10*(2), 137-145. https://doi.org/10.1207/s15327841mpee1002_5

Parsons, T. D., Mcmahan, T., & Kane, R. (2017). Practice parameters facilitating adoption of advanced technologies for enhancing neuropsychological assessment paradigms. *The Clinical Neuropsychologist*, *32*(1), 16 - 41. https://doi.org/10.1080/13854046.2017.1337932

Post, M. W. (2016). What to do with "moderate" reliability and validity coefficients? *Archives of Physical Medicine and Rehabilitation*, *97*(7), 1051-1052. http://dx.doi.org/10.1016/j.apmr.2016.04.

001

Riordan, P., Lombardo, T., & Schulenberg, S. E. (2013). Evaluation of a computer-based administration of the rey complex figure test. *Applied Neuropsychology*, *20*(3), 169–178. https://doi.org/10.1080/09084282.2012.670171

Robinson, G., Shallice, T., Bozzali, M., Cipolotti, L. (2012). The differing roles of the frontal cortex in fluency tests. *Brain - A Journal of Neurology, 135*(7), 2202–2214. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3381725/pdf/aws142.pdf

Saptoto, R. (2018). Pengaruh Adaptasi Waktu Administrasi yang disebabkan Penggunaan Lembar Jawaban Komputer terhadap Hasil CFIT 3 A dan 3 B. *Jurnal Psikologi*, *45*(1), 52–65. https://doi.org/10.22146/jpsi.30853

Schmand, B. (2019). Why are neuropsychologists so reluctant to embrace modern assessment techniques? *Clinical Neuropsychologist*, *33*(2), 209–219. https://doi.org/10.1080/13854046.2018.1523468

Setiatama, T., & Kusrohmaniah, S. (2019). Pengaruh stres melalui sing-a-song stress test terhadap selective attention pada dewasa awal. *Gadjah Mada Journal of Professional Psychology*, *3*(1), 55. https://doi.org/10.22146/gamajpp.42780

Shao, Z., Janse, E., Visser, K., & Meyer, A. S. (2014). What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Frontiers in Psychology*, *5*(72), 1-10. https://doi.org/10.3389/fpsyg.2014.00772

Shaughnessy, J. J., Zechmeister, E. R., & Zechmesiter, J. S. (2012). *Metode penelitian dalam psikologi edisi 9* (E. Tjo, Trans). Salemba Humanika.

Social Science Statistics. (2018). *Effect size calculator for t-test*. [Computer software]. Social Science Statistics. https://www.socscistatistics.com/effectsize/default3.aspx

Tierney, M. C., Naglie, G., Upshur, R., Moineddin, R., Charles, J., & Liisa Jaakkimainen, R. (2014). Feasibility and validity of the self-administered computerized assessment of mild cognitive impairment with older primary care patients. *Alzheimer Disease and Associated Disorders*, *28*(4), 311–319. https://doi.org/10.1097/WAD.0000000000000036

Velikonja, D., Hamilton, J., Ball, S., Belfry, S., Cunningham, T., Frank, J.B., Goldelson, J., Lennox, C., Levitt, B., Kaplan, R., Kaplan, F.K., Kurzman, D., McKay, C., Williams, T., & Zakzanis, K. (2020). Guidelines for best practices in psychological remote/ tele assessments version I. *Ontario Psychological Association (OPA) & Canadian Academy of Psychologist in Disability Assessment (CAPDA) Working Group.* https://www.psych.on.ca/getattachment/Policy-Public-Affairs/OPA-Guidelines/Guidelines-for-Best-Practices-in-Psychological-Rem/OPACAPDA-RemoteTele-Assessment-V9.pdf.aspx?ext=.pdf

Vosylis R., Žukauskienė R., & Malinauskienė O. (2012). Comparison of internet-based versus paper-and-pencil administered assessment of positive development indicators in adolescents' sample. *Psichologija*, *45*(1), 7-21. https://doi.org/10.15388/Psichol.2012.45.1

Zygouris, S., & Tsolaki, M. (2015). Computerized cognitive testing for older adults: A review. *American Journal of Alzheimer's Disease and Other Dementias*, *30*(1), 13–28. https://doi.org/10.1177/1533317514522852