Latent structure of the Wisconsin Card Sorting Test: a confirmatory factor analytic study

Kevin W. Greve a,b,*, Timothy R. Stickle a, Jeffrey M. Love a,b, Kevin J. Bianchini a,b, Matthew S. Stanford a

a Department of Psychology, University of New Orleans, New Orleans, LA, USA
b Jefferson Neurobehavioral Group, Metairie, LA, USA

Accepted 14 September 2004

Abstract

The present study represents the first large scale confirmatory factor analysis of the Wisconsin Card Sorting Test (WCST). The results generally support the three factor solutions reported in the exploratory factor analysis literature. However, only the first factor, which reflects general executive functioning, is statistically sound. The secondary factors, while likely reflecting meaningful cognitive abilities, are less stable except when all subjects complete all 128 cards. It is likely that having two discontinuation rules for the WCST has contributed to the varied factor analytic solutions reported in the literature and early discontinuation may result in some loss of useful information. Continued multivariate research will be necessary to better clarify the processes underlying WCST performance and their relationships to one another.

© 2004 National Academy of Neuropsychology. Published by Elsevier Ltd. All rights reserved.

Keywords: Wisconsin Card Sorting Test; Confirmatory factor analysis; Assessment; Executive function

The Wisconsin Card Sorting Test (WCST; Heaton, Chelune, Talley, Kay, & Curtiss, 1993) is a well-established measure of executive function. Its value and popularity are illustrated by the ever-increasing number of studies incorporating the WCST. The WCST has been used in more than 600 papers, over 80% of which have been published in the last decade. Despite its frequent use as a clinical and research tool, debate continues concerning the nature of the cognitive...
processes which underlie WCST performance and the relationships among the various scores derived from the test.

To date, 17 factor analytic studies of the WCST in isolation have been published in peer-reviewed journals (all but one (Bowden et al., 1998) used exploratory factor analysis; EFA). These studies were reviewed and are noted by asterisks in the Reference section; a table summarizing these studies is available from the authors upon request. This review suggests that a 3-factor solution likely best accounts for the cognitive processes underlying performance on the WCST. These processes include the ability to shift set, problem solving/hypothesis testing, and response maintenance. However, within this body of literature as a whole, as well as the studies individually, there are a number of important limitations that make it difficult to know how to interpret this research and apply these findings to clinical practice.

First, sample sizes have generally been rather small, usually less than 200 and often less than 100. While technically the subject-to-variable ratios are generally within accepted limits (approximately 10 to 1; Tabachnick & Fidell, 1989), the solutions derived from the smaller samples are less stable and reliable than those derived from larger samples (Gorsuch, 1983).

Second, the use of orthogonal rotation has generally been the rule (except Goldman et al., 1996) though it appears that there is a high degree of correlation between the primary factor and the hypothesis testing factor. Greve et al. (2002) have argued that the major variables defining each factor are not independent. Therefore, the use of oblique rotation may be more appropriate for the WCST.

Third, Heaton et al. (1993) reported that Failure-to-Maintain-Set (FMS) was a relatively rare finding among the normal and clinical patients in the normative study. Therefore, the finding of a response maintenance factor represented almost solely by FMS may be artifactual. However, a number of clinical groups are characterized by FMS in the absence of significant perseveration (Greve et al., 2002; Stanford, Greve, & Gerstle, 1997; Stuss et al., 2000). Thus, the importance of FMS remains unclear.

Finally, when administered in standard fashion there are two termination criteria (i.e., completion of six categories or all 128 cards, whichever is first). This means that the number of trials completed by a given person may vary from as few as 70 to the full 128. This appears to introduce additional error given the relatively larger amount of variance accounted for in EFA’s in which all subjects complete the same number of trails. Thus, the variable termination criteria may contribute to instability or unreliability in the WCST factor structure.

The continuing disagreements about the number, composition, and orientation of the WCST’s factors highlight the drawbacks of EFA as a tool for the study of phenomena such as cognition. Confirmatory factor analysis (CFA) offers distinct advantages over EFA. First, CFA requires the investigator to specify a priori the number and orientation of factors that explain covariation in observed variables (Krueger, Caspi, Moffitt, & Silva, 1998; Stickle & Blechman, 2002). Additionally, the models are tested statistically for fit to observed patterns of behavioral covariation. CFA then, can be used to test plausible alternative hypotheses about the pattern of association of particular WCST scores for fit to observed patterns of behavior. Third, the data patterns obtained through CFA are more easily interpreted because they are constrained by a priori predictions which can be limited to those supported by theory, previous findings, or existing literature. Finally, a best solution is possible for any covariance matrix and the a priori model that best fits the observed data generally receives the strongest...
support. CFA, then, has the potential to clarify what parameters of cognition the WCST actually tests.

The only CFA of the WCST yet published was conducted by Bowden et al. (1998) in a sample of 75 college students. Six variables were examined: CAT, TC, TE, PE, PR, and NPE; FMS was not included. These authors reported that a 1-factor solution best fit the data which is not surprising in light of the discussion above. However, the sample size is inadequate for CFA. Given the ratio of sample size to degrees of freedom, the power is too low to detect adequacy of fit in a single model. In fact, a power analysis indicates that the minimum sample size for the tested models would need to be over 450 and preferably more in the range of 450–500 to adequately detect fit (MacCallum, Browne, & Sugawara, 1996). Thus, this analysis has insufficient power to determine what is good fit versus error and the finding that the WCST is best represented as a single factor is thus not reliable.

1. Purpose

The purpose of the present paper was to use CFA in a very large sample to provide a detailed investigation of the latent structure of the WCST by specifically comparing the 1-, 2-, and 3-factor solutions previously reported in the literature.

2. Method

2.1. Subjects

Data for this study were compiled retrospectively from a mixed sample of 1221 neurological (n = 620) and psychiatric (228) patients, and nonclinical control subjects (n = 373). Some of these participants have been the subject of previous exploratory factor analytic study (Greve, Ingram, & Bianchini, 1998; Greve, Bianchini, Hartley, & Adams, 1999; Greve et al., 2002). Subjects were included if they completed the WCST and had no external incentive to appear impaired. See Table 1 for subject characteristics.

2.2. Procedure

The WCST was administered in standard fashion (Heaton et al., 1993) and scored using commercially available software (Psychological Assessment Resources, 1990). Seven scores (Total Correct (TC), Perseverative Responses (PR), Perseverative Errors (PE), Non-Perseverative Errors (NPE), Percent Conceptual Level Responses (%CLR), Categories Completed (CAT), FMS) were submitted for CFA. Only raw scores were used. Several scores were excluded.

Total Trials and Total Errors (TE) were not included because these scores are a linear combination of two or more included scores (TC + PE + NPE, and PE + NPE, respectively) and were thus redundant. Trials-to-complete-the-first-category and Learning-to-learn
Table 1
Sample characteristics

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>1221</td>
<td>15</td>
<td>90</td>
<td>41.69</td>
<td>20.82</td>
</tr>
<tr>
<td>Education</td>
<td>1221</td>
<td>2</td>
<td>24</td>
<td>12.94</td>
<td>2.77</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>748</td>
</tr>
<tr>
<td>Female</td>
<td>473</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Handedness</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right</td>
<td>1067</td>
<td>90.96</td>
</tr>
<tr>
<td>Left</td>
<td>100</td>
<td>8.53</td>
</tr>
<tr>
<td>Mixed</td>
<td>6</td>
<td>0.51</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Diagnosis</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traumatic brain injury</td>
<td>193</td>
<td>15.81</td>
</tr>
<tr>
<td>Cerebrovascular accident</td>
<td>222</td>
<td>18.18</td>
</tr>
<tr>
<td>Dementia</td>
<td>104</td>
<td>8.52</td>
</tr>
<tr>
<td>Miscellaneous neurological*</td>
<td>101</td>
<td>8.27</td>
</tr>
<tr>
<td>Psychiatric</td>
<td>228</td>
<td>18.67</td>
</tr>
<tr>
<td>Healthy control</td>
<td>373</td>
<td>30.55</td>
</tr>
<tr>
<td>Total</td>
<td>1221</td>
<td>100.00</td>
</tr>
</tbody>
</table>

*Traumatic brain injury and other neurological patients with external incentive at the time of evaluation were excluded from these analyses to eliminate potential invalid data associated with malinger.

were not included because they required a degree of success many subjects did not achieve (the completion of one and three categories, respectively); those subjects would have been deleted automatically from the multivariate analyses, thus dramatically reducing and likely biasing the sample. Table 2 contains the descriptive statistics for the standard WCST scores.

Table 2
Descriptive statistics for the standard WCST scores

<table>
<thead>
<tr>
<th>N</th>
<th>Mean</th>
<th>S.D.</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Trials</td>
<td>1221</td>
<td>106.77</td>
<td>22.78</td>
<td>68</td>
</tr>
<tr>
<td>Total Correct</td>
<td>1221</td>
<td>68.45</td>
<td>14.15</td>
<td>27</td>
</tr>
<tr>
<td>Total Errors</td>
<td>1221</td>
<td>38.33</td>
<td>26.96</td>
<td>4</td>
</tr>
<tr>
<td>perseverative Errors</td>
<td>1221</td>
<td>27.26</td>
<td>27.43</td>
<td>3</td>
</tr>
<tr>
<td>Nonperseverative Errors</td>
<td>1221</td>
<td>23.20</td>
<td>21.09</td>
<td>3</td>
</tr>
<tr>
<td>% Concept Level Responses</td>
<td>1221</td>
<td>58.27</td>
<td>25.71</td>
<td>0</td>
</tr>
<tr>
<td>Categories Completed</td>
<td>1221</td>
<td>4.29</td>
<td>2.17</td>
<td>0</td>
</tr>
<tr>
<td>Trials to 1st Category</td>
<td>1109</td>
<td>18.01</td>
<td>17.94</td>
<td>10</td>
</tr>
<tr>
<td>Failure-to-Maintain-Set</td>
<td>1221</td>
<td>.95</td>
<td>1.23</td>
<td>0</td>
</tr>
</tbody>
</table>
3. Statistics

CFA was performed with the Mplus (Muthén & Muthén, 2001) statistical program. Three fit indices were computed for each model, with the overall pattern suggested by these indicators used to guide interpretation of model adequacy (see Bentler, 1990 and Loehlin, 1998, for a detailed discussion of these indices): (1) Comparative Fit Index (CFI); (2) Chi-square goodness-of-fit; (3) Root-Mean-Square-Error of Approximation (RMSEA). Evidence of good model fit implies that the data are consistent with the assumptions of the hypothesized model. That is, the pattern of observed covariances “fit” with or are reproduced by those generated from the hypothesized model such that the residual variance is within the limits predicted by sampling error (see Bollen, 1989).

Alternative models were compared on a combined basis of their parsimony, fit with existing theory, and the relative goodness of fit to the observed data as indicated by chi-square and practical fit indices. When alternative models were tested against each other for best fit to the data, a chi-square difference test was computed. This test examines the change in the chi-square statistic relative to the change in degrees of freedom. If the difference is statistically significant ($P < .05$), the model with the lower chi-square value provides a significantly better fit to the data. In all cases, overall fit of the model to the observed data was determined by examining at least two indices of fit. Given two alternative models with equivalent fit indices, the model that is more parsimonious (i.e., has fewer parameters) is preferred.

4. Results

The 1-factor model initially used all seven scores as indicators. In the 2-factor model, the indicators were (1) PE, PR, CLR, CAT and TC, and (2) CLR, CAT, TC, PE, NPE, and FMS. In the 3-factor model, indicators were: (1) PE, PR, CLR, CAT and TC, (2) CLR, CAT, TC, and NPE and (3) TC and FMS. Because PE is highly correlated with several other indicators, it rendered the models unreliable. PE was dropped from the final models, resulting in six indicators of the factors. Results of the analyses were mixed, with all models showing indications of lack of fit to the observed data. The 2-factor model provided poor fit, the 1-factor model showed somewhat better fit, and the 3-factor model producing fit indices just outside the lower range of those considered adequate. Table 3 shows fit indices for the comparative models.

Table 3

<table>
<thead>
<tr>
<th>Model</th>
<th>$X^2$</th>
<th>df</th>
<th>$P$</th>
<th>CFI</th>
<th>RMSEA (90% CI)</th>
<th>$\Delta X^2$</th>
<th>$\Delta df$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-Factor</td>
<td>146</td>
<td>5</td>
<td>&lt;.001</td>
<td>.96</td>
<td>.15 (.13–.17)</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>2-Factor</td>
<td>314</td>
<td>4</td>
<td>&lt;.001</td>
<td>.93</td>
<td>.22 (.20–.25)</td>
<td>n.a.</td>
<td>1</td>
<td>n.a.</td>
</tr>
<tr>
<td>3-Factor</td>
<td>59</td>
<td>3</td>
<td>&lt;.001</td>
<td>.97</td>
<td>.12 (.10–.15)</td>
<td>255</td>
<td>1</td>
<td>&lt;.001</td>
</tr>
</tbody>
</table>

Note: CI: Confidence interval; CFI: Comparative Fit Index; RMSEA: Root-Mean-Square-Error Approximation; $\Delta X^2$ and $\Delta df$ changes in model relative to the preceding model in the table. Inter-factor correlations in the 3-factor model $F_1$, $F_2$ = .20, $F_1$, $F_3$ = .43, $F_2$, $F_3$ = −.13.
All three models yielded a significant Chi-square, suggesting that none fully described the data. The 1- and 3-factor models showed acceptable fit on the CFI (> .95), and none showed acceptable fit by RMSEA (≤ .10). Although evaluating the models by the CFI, the 3-factor model provides a good fit to the data, by RMSEA, all three models provide poor fit to the data. Only the 3-factor model provided a fit with low enough residual variance to warrant further investigation. Fit indices as a whole suggested that none of the models fit the observed data closely enough to be considered the “true” model, although the 3-factor model showed statistically significant improvement in fit over the 1- and 2-factor models.

The inadequacy of these models may be a result of a number of factors including: (1) significant overlap among variables; (2) different best-fit models for different clinical subgroups; (3) error due to differences in test length; (4) too few indicators per potential latent variable; (5) lack of multiple measures of the latent constructs; (6) lack of clear conceptualization of latent constructs; and, (7) potential effects of age and education. We conducted several follow-up analyses designed to examine these hypotheses.

Specifically, we tested the fit of the data to models generated from the above hypotheses using the standard variables suggested in EFA as indicators of 1-, 2-, and 3-factor models in each of the diagnostic subgroups. This set of analyses tested hypothesis 2. Additionally, we tested these models in the group with 128-card protocols. This set of analyses tested hypothesis 3. Finally, we tested 1-, 2-, and 3-factor models using a set of non-overlapping variables with at least three indicators per latent variable. These analyses examined hypotheses 1 and 4 above. Hypothesis 5 would require inclusion of additional neuropsychological tests, which is beyond the scope of this study.

Generally, the results of these follow-up analyses were unsatisfactory. A 3-factor model indicated in EFA studies using data only from participants who had the full 128 card protocol (n = 565) showed evidence of good fit (CFI = .999; \( \chi^2 \) [2] = 6.2, \( P = .04; \) RMSEA = .06 [90% CI: .01–.11]). In the 3-factor model, indicators were, (1) PR, CLR, CAT & TC, (2) CLR, CAT, TC, & NPE and (3) TC and FMS. One- and 2-factor models demonstrated very poor fit. Also, within the 128-card data set and within diagnostically homogeneous subgroups receiving the standard WCST, a 1-factor model containing only TC, CLR, PE, and CAT, also resulted in excellent fit (CFI = 1.0; \( \chi^2 \) [1] = 1.67, \( P = .20; \) RMSEA = .03 [90% CI: .00–.12]).

5. Discussion

A 3-factor model consistent with models derived from EFA provided best fit for the standard WCST variables derived from standard administration. However, neither it nor simpler models fit the observed data closely enough to be considered the “true” model. However, data from only persons who had completed all 128 cards resulted in an excellent 3-factor model. At the same time, regardless of the sample or subsample used and regardless of the quality of fit of multi-factor solutions, a 1-factor solution containing only the TC, CLR, PE, and CAT, the primary constituents of almost all first factors in any WCST EFA, produced excellent fit.

It has been argued that the Wisconsin Card Sorting Test measures general executive function (e.g., Heaton et al., 1993). The present study would suggest that it measures that construct very well. The second and third factors typically reported in EFA are arguably “error” factors in that
they are not good representatives of the associated latent variables except when all 128 cards are used. That is, the WCST is sensitive to other cognitive processes and performance on the WCST may serve as a gross indicator of these processes. However, it does not appear to measure them well. Greve et al. (2002) described the WCST factors in hierarchical terms in which greater low level deficits (e.g., cognitive inflexibility with associated perseveration) potentially obscure WCST evidence of other cognitive deficits (e.g., problem solving or distractability). These findings are consistent with Matarazzo’s (1990) conclusion that most neuropsychological batteries are really measures of \( g \) with much smaller subfactors measuring more specific cognitive abilities.

The perseveration scores on the WCST are generally the most sensitive to brain damage (Heaton et al., 1993). WCST perseveration has generally been associated with, though not exclusively localized to, dorsolateral pre-frontal cortex (e.g., Demakis, 2003; Milner, 1963). At the same time, Stuss et al. (2000) demonstrated that ventromedial/orbitofrontal lesions may produce higher levels of FMS. Similarly, Impulsive Aggressive individuals tend to have higher NPE scores (Stanford et al., 1997). Thus, this CFA supports three factors and there are certain patient subgroups who show differential impairment on the secondary factors when all 128 cards are given. So while it may be true that the WCST does not measure those underlying constructs well, the hierarchical relationship of the factors suggests that the absence abnormal of NPE or FMS may have little clinical meaning while their presence may indicate specific deficits worthy of follow-up with more specific tests.

The finding of a better defined factor structure when all subjects complete the same number of trials is consistent with the EFA results of Greve et al. (2002) for 128 cards and Kongs, Thompson, Iverson, and Heaton (2000) for the 64-card version (WCST-64). In both studies the three factor solution accounted for 99% of the variance. Given that those persons who completed all 128 trials can be considered to have “failed” the WCST and so are not necessarily representative of all persons taking the WCST, it is important to note that the Kongs et al. and Greve et al. studies required completion of all trials (64 and 128, respectively) regardless of the number of categories completed. Thus it is not unreasonable to expect that the results of the present analysis would generalize to less impaired persons. In effect, having two discontinuation rules (six categories or 128 cards) means that not everyone is administered the same test. This may help explain some of the inconsistencies in the WCST factor analytic literature. Not administering all 128 cards may also result in a failure to detect some extant deficits in certain impaired populations. For example, Stanford et al. (1997) reported that their impulsive aggressive sample completed an average of 5.3 categories. If testing had been terminated with six categories it is unlikely they would have differed from the control group whose average would have approached six categories. However, because all 128 cards were administered, the IA mean clearly differed from the control group mean of 8.6 categories completed.

The finding that having a uniform discontinuation rule results in a better defined factor structure enhances the potential value of the WCST-64, particularly since comprehensive norms have recently been published. However, tests with fewer items are generally less reliable and administering only 64 cards may result in lost information (see discussion of Stanford et al., 1997, above). Love, Greve, Sherwin, and Mathias (2003), in comparing the agreement between the WCST-64 and the standard version (with the two discontinuation rules), have
suggested that something different happens in the second half of the test than in the first. This is consistent with Axelrod, Paolo, and Abraham (1997) and Merrick, Donders, and Wiersum (2003). It is therefore important to determine if the standard WCST has incremental validity over the WCST-64. This is a question that is very amenable to CFA. Similarly, CFA can be used to address the question of whether or how performance changes over the course of 128 cards. Both of these issues have conceptual and practical implications.

CFA's of the WCST that include indicators from additional tests would also be useful. Numerous such EFA's have been reported (e.g., Boone, Ponton, Gorsuch, Gonzalez, & Miller, 1998; Greve et al., 1998; Greve, Farrell, Besson, & Crouch, 1995; Paolo, Tröster, Axelrod, & Koller, 1995) and can be used to guide CFA model development. In general, these analyses have supported the WCST (particularly Factor 1 variables) as a measure of executive function. At the same time the WCST factor(s) have often loaded independently of other neuropsychological tests. Such analyses using CFA, while potentially enlightening, were beyond the scope of this study. However, it seems likely that these CFA's would still be affected by the same factors that adversely affected the CFA of the WCST alone. Therefore, it may be valuable to use person-centered (e.g., Latent Class Analysis; Goodman, 1974; Lazarsfeld & Henry, 1968) analyses to realize the full value of the WCST.

6. Summary

The present study represents the first large scale CFA of the Wisconsin Card Sorting Test. The results generally support the three factor solutions reported in the EFA literature. However, only the first factor, which reflects general executive functioning, is statistically sound. The secondary factors, while likely reflecting meaningful cognitive abilities, are less stable except when all subjects complete all 128 cards. It is likely that having two discontinuation rules for the WCST has contributed to the varied factor analytic solutions reported in the literature and early discontinuation may result in some loss of useful information. It may therefore be useful for researchers to routinely use all 128 cards. Clinicians are somewhat more limited given that the extensive and excellent norms are based on the two discontinuation rule procedure. More generally, continued multivariate research will be necessary to better clarify the processes underlying WCST performance and their relationships to one another.

Acknowledgments

The authors would like to thank Adrienne Brennan, Matthew Heinly, Nora Vitone, and Cora Joffe for their assistance with data collection. Kelly Curtis’ review of multiple drafts of this paper is also appreciated. Some of the data upon which this paper is based were presented at the 30th and 31st Annual Meetings of the International Neuropsychological Society, Toronto, Ont., Canada, and Honolulu, Hawaii. Timothy R. Stickle is now at University of Vermont, Matthew S. Stanford is now at Baylor University, and Jeffrey M. Love is now at Pennsylvania State University.
References

Factor analytic studies reviewed.


