Equivalence of standard and computerized versions of the Raven Progressive Matrices Test

John Eustis Williams a,*, David M. McCord b

a Department of Psychology, University of Northern Iowa, Cedar Falls, IA 50614, USA
b Department of Psychology, Western Carolina University, Cullowhee, NC 28723, USA

Available online 9 April 2004

Abstract

The present study examined the equivalence of the computer administered version of the Raven Standard Progressive Matrices (RSPM) with the standard paper-and-pencil administered version of the RSPM. In addition, the effects of state and trait anxiety as well as computer anxiety were investigated. Fifty undergraduate volunteers were administered the RSPM twice under one of four conditions: computer–computer, standard–standard, computer–standard, or standard–computer. No significant differences were found between mean scores and standard deviations across administrations or formats. Rank-order correlations revealed similar ranking across formats. Tentative support for the equivalence of the computerized version of the RSPM was found. Analyses revealed no significant differences in anxiety across formats and no significant correlations between anxiety and RSPM performance. Explanations and implications for further research are discussed.

Keywords: Computerised assessment; Non-verbal intelligence; Equivalence; RSPM

1. Introduction

Computers have been used in psychology for decades. Computerized testing is a growing field, and the availability of computers to the public is now greater than ever and will continue to grow. As the use of computers in psychological assessment continues, it is important to understand which factors interact with the format of computerized testing.
Computer anxiety is one of these factors. Most of the research on computer anxiety took place in the 1980s, when computer use increased dramatically. Though there are numerous advantages to using computer administration (e.g., control over the presentation of material, accurate response times), the most common question is whether or not traditional tests can be transferred to computer administration without the development of new norms.

One problem facing a test developer is the evidence of equivalence between formats of administration. The two forms of equivalence that are of concern are experiential equivalence and psychometric equivalence (Honaker, 1988). Experiential equivalence is important as an attempt to measure construct validity. Two versions of a test could be equivalent in terms of producing similar scores; however, both may be measuring different constructs. Psychometric equivalence is determined by showing no significant differences between group means and distributions. Showing that mean scores are statistically similar demonstrates that differing formats will not bias overall group scores.

In addition, it is important to examine variance in relation to format. Extreme variance differences indicate that range and standard deviation are not comparable to the norms set for the standard format, causing equivalence to be questioned. Rank order correlations provide reliability and similarity by directly comparing computer administered group scores to the standard format. Experiential equivalence is determined by showing no significant differences between attitude and perception of the test takers toward the test.

In general, tests that use presentation of verbal items and response have been shown to be psychometrically equivalent. However, nonverbal-type tests have had mixed results in terms of equivalence. None of the studies which investigated the equivalence of nonverbal-type tests considered the effects of individual perceptions of the test (experiential equivalence). Experiential equivalence, particularly the effects of computer anxiety, has been studied with verbal-type tests. However, nonverbal tests have not been studied with regard to computer anxiety. It could be that visual-spatial-type items are more susceptible to effects of computer anxiety than verbal-type items. Kubinger, Formann, and Farkas (1991) suggest that computer-related anxiety might explain the performance deficit found on a computerized form of the Raven Stanard Progressive Matrices (RSPM). Computer anxiety is assumed to cause greater randomness in responding and to interfere with the ability to learn techniques for solving the matrices and the ability to use those techniques on each additional item presented.

Most of the computer administered psychological tests that are in use today are personality tests such as the MMPI and other tests that involve reading simple questions and answering with simple responses. The computerization of tests involving complex graphics and manipulation of materials on the computer screen, such as the Wechsler series of intelligence tests or the RSPM, has not been as widespread. One of the first attempts to automate the administration of IQ tests was the use of the Totally Automated Psychological Assessment Console (TAPAC) (Gilberstadt, Lushene, & Buegel, 1976). The TAPAC was a machine that used a reel-to-reel tape player, a slide projector, and a console with buttons to be pushed for responses. The TAPAC was used to administer a battery of tests designed to correlate with the Wechsler Adult Intelligence Scale (WAIS). This battery consisted of the Shipley–Hartford Vocabulary subtest, RSPM Standard 1938 edition, WAIS Digit Span, WAIS Digit Symbol, and the Halstead Category test. Unfortunately, this battery was not compared to the standard versions of the included tests. It was assumed that the automated versions were comparable to their standard versions.
Two other studies addressed the issue of automating psychological assessment (Calvert & Waterfall, 1982; Watts, Baddeley, & Williams, 1982). These two studies specifically addressed the issue of equivalence between the standard format of administration and the automated administration for the RSPM. Calvert and Waterfall used an apparatus similar to the TAPAC. They found no significant differences for group means or time to complete the tests between formats of administration. Watts et al., however, did find significant differences between formats of administration and time to complete the test. The time difference was also statistically significant, with the automated version being faster during both testing sessions. Watts et al. used an apparatus almost identical to the TAPAC and the apparatus used by Calvert and Waterfall.

A noticeable difference between the two studies was the use of a tailored and adaptive version of the RSPM for the automated administration by Watts et al. Watts et al. decreased the number of items from 60 to 29 for the automated version. They did not adapt the standard administration. Due to the decreased length of the automated test, it is not surprising that individuals completed the automated form significantly faster than the traditional version. Watts et al. also suggested that by making the automated version shorter, it would make the test easier. However, they found that individuals performed approximately five points higher on the traditional version. Their results are unclear because they were attempting to compare tests which not only differed in format of administration but also in content and process of administration. The latter issues were not addressed in the study.

The technology to transfer performance or nonverbal-oriented tests directly to computers has recently become available. However, the little research done on nonverbal tests has found mixed results. Most of the studies that did determine equivalence have serious methodological shortcomings. Rock and Nolen (1982) found no significant differences between a computerized version of the Raven Coloured Progressive Matrices and the traditional version of the test. This study did not address computer anxiety, which is assumed to be lower in subjects from younger generations (Loyd, Loyd, & Gressard, 1987; Rosen, Sears, & Weil, 1987).

The cognitive processing involved in such a task as drawing is continuous and requires feedback on responses (Levy & Barowsky, 1986). It may be that computerized tests that involve such tasks as spatial relations and visualization are affected by the format change more than tests that involve simple presentation of text on a screen. Kubinger et al. (1991) investigated the equivalence of a computerized form of the RSPM to the standard version. Effects of format difference were investigated as one of the questions of the study. The authors concluded that the computerized version underrated the IQ scores of the standard version, on average by 13 IQ points. They stated that one possible explanation of this discrepancy could be “the stress-evoking characteristic of a computer in general, as for instance it may induce a testee to precipitate item responses” (Kubinger et al., 1991, p. 300). They suggest that certain items may be more susceptible to basis or computer anxiety than other items.

1.1. Hypotheses

The purpose of the present study is to evaluate the equivalence between the standard form of the RSPM and a computer administered version, considering the possible participant response differences due to computer anxiety. In order to investigate equivalence, a
repeated measures, counterbalanced design was used. This allows for information about reliability, similarity in mean scores, distribution, and ranking across formats to be obtained. In addition, it is hypothesized that:

1. Computer anxiety will increase as scores on the computer administered RSPM decrease.
2. Participants with low anxiety and/or low computer anxiety will show similar results regardless of administration format.
3. Participants, regardless of level of computer anxiety, will complete the computerized version faster than the standard version.

2. Methods

2.1. Participants

The participants in this study were undergraduate volunteers who enrolled in psychology courses at a local state university. Participation in a study was a required part of their coursework. The total sample consisted of 30 females and 20 males, with a mean age of 18.3 years. All but 2 of the 50 participants were freshman. The students were randomly assigned to one of four groups, and tested on two occasions. Groups did not show a significant difference in terms of gender \( \chi^2(1) = 2.00 \), or age \( [F(3,46) = 0.376] \). The test–retest interval averaged 53 days with a range from 20 to 99 days.

2.2. Test instruments

2.2.1. Raven Standard Progressive Matrices

Computerized and standard versions of the RSPM (Raven, 1958), were used in this study. The RSPM consists of 60 diagrammatic problems, each with eight possible choices for answers. Burke (1972) has estimated the test–retest reliability of the RSPM at 0.96. Concurrent validity between the Wechsler Adult Intelligence Scale Full Scale IQ and the RSPM has been estimated at 0.57 (McLaurin & Farrar, 1973). However, in the same study the RSPM was not found to be significantly correlated with grade point average. One study investigated the use of the original 1948 British norms with American participants. This study indicated that the use of these norms was valid with American participants. The RSPM scores were significantly correlated with Full Scale, Verbal Scale, and Performance Scale IQs, 0.85, 0.84, 0.75, respectively (Vincent & Cox, 1974).

The computerized version was obtained from J.C. Raven Ltd., and is designed for research purposes only at the time of the study. There is no current validity or reliability information available for the computerized version.

In the standard administration each diagrammatic problem was presented on a page in the test booklet. Responses were indicated by circling or marking the appropriate answer on an answer sheet. In the computerized version of the RSPM, each problem was presented on a computer screen and the mouse was used to indicate responses. The computerized version was designed to run on Macintosh computers and was written as a HyperCard program.
2.2.2. The Computer Anxiety Rating Scale (CARS-H)

The CARS-H (Heinssen, Glass, & Knight, 1987) was used to assess the computer anxiety levels among all participants. The CARS-H consists of 20 items formatted as a Likert type five-point scale. The five-point scale ranges from *strongly disagree* to *strongly agree*. A total computer anxiety score was obtained ranging from 20 to 100, with higher scores reflecting higher computer anxiety.

The validity of the CARS-H was estimated by comparison to other situational anxiety scales, math anxiety, and trait anxiety questionnaires. CARS-H scores were correlated with both situations ($r = 0.48$ and 0.37, respectively). The CARS-H also correlated with Loyd et al.'s (1987) computer attitude scale, specifically with its three subscales (Anxiety $r = -0.69$, Liking $r = -0.60$, Confidence $r = -0.74$). A correlation with computer experience was also obtained ($r = -0.33$). Test–retest reliability was found to be adequate ($r = 0.70$, $p < 0.0001$).

2.2.3. The State-Trait Anxiety Index (STAI)

The STAI (Spielberger, Gorsuch, & Lushene, 1970) was used to assess general anxiety among all participants. The STAI consists of 40 items and gives scores for both state (STAI-S) and trait (STAI-A) anxiety. Test–retest reliability is high for the state anxiety score ($r = 0.73–0.86$) and relatively low for the trait anxiety score ($r = 0.16–0.54$). Validity was obtained by comparison to the Taylor Manifest Anxiety Scale and the IPAT Anxiety Scale (Cattell & Scheier, 1976).

2.3. Procedure

Participants were randomly assigned to one of four groups as part of a counterbalanced, repeated measures design. Each participant was administered the RSPM twice with an average of 53 days between administrations. Participants in Group I were administered the computer version both times (Computer–Computer: C–C, $n = 12$) and in Group II they were administered the standard version both times (Standard–Standard: S–S, $n = 11$). Participants in Groups III and IV were administered the computer version and the standard version in counterbalanced order: Group III (Computer–Standard: C–S, $n = 12$) and Group IV (Standard–Computer: S–C, $n = 15$).

All participants signed a consent form which included a description of the experiment, rights to confidentiality, and their right to end participation. After consent was obtained seven participants decided not to continue with the experiment, leaving the 50 participants upon whom the data analysis is based. Before the first administration of the RSPM, all participants completed the STAI-A. Before the second administration of the RSPM, all participants completed the CARS-H.

3. Results

3.1. Outliers

Fig. 1 shows the difference scores between administrations 1 and 2 of the RSPM. Based on these difference scores, three participants’ data were excluded from further analyses. These outliers were greater than 2.5 standard deviations below the mean. In addition, 1 participant was excluded due to missing data.
### 3.2. RSPM mean scores

Mean scores and standard deviations for the RSPM were compared among the four group conditions and are summarized in Table 1. In order to compare similarity of scores across formats, a univariate repeated measures ANOVA was used. Group, Administration, and Gender main effects and interactions were examined. No significant main effects or interactions were found. The lack of significant main effects implies that the mode of administration did not produce differences in mean scores.

### 3.3. RSPM mean score distribution

RSPM score distributions were visually examined for possible differences. In particular, comparing distributions for Groups III and IV allowed examination of variance within

![Histogram](image)

**Fig. 1.** Histogram of difference scores between administration 1 and administration 2 of the RSPM.

### Table 1
Mean scores and standard deviations for the RSPM

<table>
<thead>
<tr>
<th>Administration Groups</th>
<th>I (Computer–Computer)</th>
<th>II (Standard–Standard)</th>
<th>III (Computer–Standard)</th>
<th>IV (Standard–Computer)</th>
</tr>
</thead>
<tbody>
<tr>
<td>n = 10</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>RSPM Score</td>
<td>49.0 (4.9)</td>
<td>46.9 (6.7)</td>
<td>49.1 (3.1)</td>
<td>48.8 (4.4)</td>
</tr>
</tbody>
</table>

*Note:* Standard deviations are listed in parentheses.
participants under both administration modes. No significant differences were found using Hartley’s test of variance homogeneity \[F_{\text{max}}(2,10) = 0\] for the CS group, \[F_{\text{max}}(2,13) = 2.61\] for the SC group. Visual inspection of the distribution revealed an approximation of normality.

3.4. Test–retest of RSPM scores

Pearson test–retest correlation coefficients were computed to compare individual differences across both administration modes. Because no mean or variance differences were found for administration, Groups III and IV were combined for this analysis. Significant correlations were found in all groups. Group I (C-C) had a test–retest correlation of 0.952 \((p < 0.01)\); Group II (S–S) had a test–retest correlation of 0.826 \((p < 0.01)\); and the combined groups of III and IV yielded a test–retest correlation of 0.594 \((p < 0.01)\).

Fischer \(r\) to \(z\) transformations revealed a significant difference between the correlations for the C–C group and the combined C–S and S–C groups \((z = 2.93, p < 0.05)\). This implies that when administering the RSPM one format is preferable to retesting with another format. No other significant differences between the correlations of each group were found. No significant test–retest reliability differences were found between Groups I and II. This result implies that both modes of administration are equally reliable when used for both test administrations.

3.5. Anxiety and RSPM performance

STAI-T, STAI-S, and CARS-H means and standard deviations are shown in Table 2. Results of an ANOVA demonstrated no significant differences across all three anxiety scales for Group, Gender, or Group by Gender. In addition, none of the anxiety scales demonstrated significant correlations with mode of administration. However, it is noted that scores on the anxiety scales had restricted ranges: STAI-T (scores from 22 to 60; maximum range is 20–80); STAI-S (scores from 20 to 57, maximum range is 20–80); CARS-H (scores from 24 to 77; maximum range is 20–100). For all groups the mean, median, and mode were within two points of each other.

3.6. RSPM completion time

Results of an ANOVA on the time needed to complete the RSPM demonstrated a significant main effect for Administration, \(F(1,27) = 34.66; p < 0.001\). Examination of the

<table>
<thead>
<tr>
<th>Groups</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n)</td>
<td>10</td>
<td>11</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>STAI-T</td>
<td>36.50 (5.95)</td>
<td>40.45 (12.18)</td>
<td>41.00 (9.81)</td>
<td>35.00 (7.32)</td>
</tr>
<tr>
<td>STAI-S</td>
<td>33.00 (7.29)</td>
<td>34.82 (9.2)</td>
<td>36.08 (11.11)</td>
<td>33.07 (9.51)</td>
</tr>
<tr>
<td>CARS-H</td>
<td>43.75 (15.04)</td>
<td>47.45 (10.49)</td>
<td>41.75 (9.22)</td>
<td>38.80 (7.84)</td>
</tr>
</tbody>
</table>

Note: Standard deviations are listed in parentheses.
means and standard deviation (see Table 3) revealed that the second administration required significantly less time to complete than the first. Across all four groups the average completion time decreased by 4.94 min on the second administration. The format of the administration had no effect on completion time.

4. Discussion

Results of the current study support the view that the RSPM standard version and the RSPM computerized version are compatible and equivalent. No significant differences were found between individuals taking the RSPM on the standard version or the computerized version. The lack of significant difference in mean scores corresponds with the idea that different groups of random individuals will obtain relatively similar scores regardless of the format. The finding that distribution and variances were not significantly different allows for the supposition that the computerized version will produce scores within a normal curve similar to the pattern obtained by the standard version. Finally the rank order correlations suggest that an individual will not obtain a significantly different score when taking the RSPM on a computer. This suggests that the computerized version of the RSPM produces similar scores across individuals and a distribution similar to that of the standard version. In addition, participants who completed both formats obtained high correlations for RSPM scores. This correlation between individuals’ scores on both formats, along with the lack of difference in distribution and mean scores, allows for a direct comparison of either score with the normative data used in the RSPM manuals.

Results do not support the hypotheses that anxiety related to computer use, or state or trait anxiety, affects the equivalence between the two versions of the RSPM. Anxiety appeared to have no affect on RSPM scores. No significant mean score differences where found between groups or administration. Anxiety scores did not correlate significantly with RSPM scores regardless of test format. Individuals with high scores on the CARS-H or the STAI did equally as well as individuals with low scores. Of concern with this finding is the restricted range obtained in the anxiety scores. It is unclear whether the range was broad enough to accurately judge a correlation between anxiety and RSPM scores.

The participants in this study were all undergraduates from a regional university. It is interesting to note that computer exposure is assumed to be higher for this group of individuals (undergraduates) than for the general population. The relative lack of impact of computer anxiety on the computerized version of the RSPM is contrary to previous

<table>
<thead>
<tr>
<th>Administration</th>
<th>Groups</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>III</td>
<td>IV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 10</td>
<td>n = 11</td>
<td>n = 11</td>
<td>n = 14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Completion time</td>
<td>17.00</td>
<td>11.45</td>
<td>21.00</td>
<td>16.80</td>
<td>15.88</td>
<td>14.25</td>
<td>17.82</td>
</tr>
<tr>
<td>(2.68)</td>
<td>(2.91)</td>
<td>(4.06)</td>
<td>(5.36)</td>
<td>(4.85)</td>
<td>(4.17)</td>
<td>(3.37)</td>
<td>(3.61)</td>
</tr>
</tbody>
</table>

*Note: Standard deviations are listed in parentheses.*
research on the use of computerized administration of tests (Lushene, O’Neil, & Dunn, 1974). However, since Lushene et al. completed their study, computer use has increased dramatically. Therefore, it is not surprising that computer anxiety would be more difficult to observe today (Loyd et al., 1987). As computer use becomes even more common, the computer may become less of a stress provoking stimulus than it has in the past.

However, a significant main effect of administration on time to complete the RSPM was found. Because the effect is not related to the format of administration or the group, it is supposed that it is due to practice effects and familiarity with the test. The decrease in completion time did not affect overall RSPM scores. Similar equivalency studies, (Harrell, Honaker, Hetu, & Oberwager, 1987; Honaker, Harrell, & Buffaloe, 1989) which used a repeated measures counterbalanced design, also revealed a significant decrease in completion time on the second administration of the MAB-V and the MMPI.

The criteria for equivalence for computer administered tests, as noted in Honaker (1988) and the APA guidelines, are tentatively met by the present study, with the exception of comparison to external criteria. The computerized version of the RSPM has met the criteria for similar mean scores, distribution and variance, and test–retest reliability. It is not known whether the scores from each version of the RSPM correlate equally with other variables. Future studies can address this issue and use external variables to test for validity.

In summary, the current study offers a first look at the equivalence of the standard RSPM and the computerized version of the RSPM. Tentative evidence is found for equivalence between the test administration formats. Computer anxiety did not play a role in the determination of equivalence. In order to fully evaluate the equivalence of these two formats, further research is needed. If a study is conducted using the four group repeated measures design, the issues of power and sample size will need to be considered. However, important results could be obtained by the simpler two group repeated measures design. Rank order comparisons could be made and analysis of external criteria could be included.

References


