Equivalence of computerized versus paper-and-pencil testing of information literacy under controlled versus uncontrolled conditions: An experimental study

Anne-Kathrin Mayer & Günter Krampen

ZPID – Leibniz Institute for Psychology Information, Trier/Germany

13th European Conference on Psychological Assessment, Zurich/Switzerland, July 22-25, 2015
International Test Commission’s Guidelines on Quality Control in Scoring, Test Analysis, and Reporting of Test Scores:

“If a different test administration format is used (e.g., computerized administration of a paper and pencil test) it is necessary to compare the new test characteristics to the old, and sometimes to equate the new test to the old.” (ITC, 2014, p. 211)
(1) Equivalence of Assessment Tools
(e.g., Preckel & Thiemann, 2003; Van der Vijver & Harsveld, 1994)

(1) Quantitative equivalence → test norms
- Equality of test scores (mean levels, standard deviations, shape of the distribution)

(2) Qualitative equivalence → construct validity, e.g.
- Reliability / internal consistency (Cronbach’s Alpha)
- Factorial structure
- Correlations with other measures
- Group differences
Effects of Medium: P&P vs. Computer

• **Possible sources on nonequivalence**
  • *visual design*: perception of materials
  • *handling*: pencil vs. input devices (keyboard, mouse etc.)
  • *subjective evaluations of medium*: computer anxiety/aversion; „felt“ anonymity of computers

• **Previous findings** (numerous studies and meta-analyses, e.g. Mead & Drasgow, 1993, for cognitive ability tests; Kim, 1999, for achievement tests; see Gnambs, Batinic, & Hertel, 2011, for an overview):
  • **self-reports** (personality, clinical scales): structural/qualitative equivalence usually given; quantitative equivalence not guaranteed
  • **achievement tests**: more heterogenous results; overall: equivalence of power tests > speed tests / speed-power-tests
Effects of Mode: Uncontrolled vs. Controlled

• Possible sources on nonequivalence:
  • *researcher control* over sampling, time, place and environmental situation, person answering, technical devices used, internet speed etc. → standardization / internal validity
  • *presence of experimenter* → experimenter effects, test anxiety
  • *Objective/felt anonymity* → readiness to disclose information on sensitive or socially undesirable topics (e.g. Gnambs & Kaspar, 2014)

• Previous findings:
  • unsupervised testing has little impact on test scores or test validity (e.g., Wasko, Lawrence, & O’Connell, 2015)
  • no evidence that „cheating“ is a substantial problem (e.g., Ladyshewsky, 2014; Lievens & Burke, 2011) → often even higher test scores in proctored/controlled testing
• “to recognize when information is needed and ... to locate, evaluate, and use effectively the needed information.” (Association of College & Research Libraries ACRL, 2000, p. 2)

→ Essential for initiating and performing effective and efficient information searches in scholarly contexts as well as in everyday life
• **Objective assessments / Achievement tests:**
  - *Knowledge tests* (fixed choice format; multiple-choice format, e.g. Leichner, Peter, Mayer, & Krampen, 2013; scenario-based situational judgement test format, e.g. Rosman, Mayer, & Krampen, 2015)
  - *Standardized search tasks* (e.g. Leichner, Peter, Mayer, & Krampen, 2014)
  - *Bibliographies or portfolios* with scoring rubrics (e.g. Oakleaf, 2009)

• **Subjective assessments:**
  - Self-reports of *information behavior* (e.g. Heinström, 2005; Timmers & Glas, 2010)
  - *Self-efficacy scales* (e.g. Behm, 2015; Kurbanoglu, Akkoynlu & Umay, 2006)
Aims of Study:

Experimental examination of the equivalence of two information literacy measures (knowledge test/self-efficacy scale) under different administration conditions:

- **Medium** (paper-and-pencil vs. computerized testing)
- **Mode** (supervised vs. unsupervised testing)

with regard to

- means and standard deviations
- internal consistencies
- intercorrelations of objective and subjective assessments of IL
**Methods: Sample and Design**

- **Sample:**
  - \( N = 141 \) educational students (82.3\% BSc, 13.5\% MSc, 2.8\% Diploma)
  - Gender: 20.6\% male, 79.4\% female
  - Age: 19-43 years (\( M = 22.54, SD = 3.42 \))

<table>
<thead>
<tr>
<th>Medium of test administration</th>
<th>Mode of test administration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised</td>
<td>Supervised</td>
</tr>
<tr>
<td>Computer</td>
<td>Group 1 (( n = 34 ))</td>
</tr>
<tr>
<td>Paper-and-Pencil</td>
<td>Group 3 (( n = 43 ))</td>
</tr>
</tbody>
</table>

No differences between groups with regard to
- age, gender,
- final school grade, self-reported level of academic achievements
- participation in information literacy training
- amount of experiences with scholarly information searches
Methods: Procedure

• Registration of participants via E-Mail

• Randomized assignment to an experimental group:
  • **Group 1** (computer, unsupervised): mailed link to online-version of test battery → complete test battery and enter (individually chosen) personal code → come to lab to receive compensation (participation was checked on code list)
  • **Group 3** (P&P, unsupervised): come to lab to receive test battery and return the completed test battery later to receive compensation
  • **Group 2** (computer, supervised) / **Group 4** (P&P, supervised): individual appointment arranged via email → complete test battery in the lab while experimenter is present
Contents: searching and evaluating psychology information (Leichner, Peter, Mayer & Krampen, 2013)
• 35 multiple-choice items (3 response options, 0-3 correct)
• Scoring: 0-1 (p[correct])

Sample item:
“Which differences exist between Internet search engines (e.g. Google Scholar) and bibliographic databases?
  o Bibliographic databases usually have a thesaurus search.
  o Boolean operators can only be used with bibliographic databases.
  o The order of items on the results page is not affected by the number of clicks on each item.”
Methods: Self Efficacy Scale

- **Contents**: competencies regarding searching, accessing, and evaluating scholarly psychology information (Leichner, Mayer, Peter & Krampen, submitted)
- 10 items, 5-point Likert scale + „don’t know“-option
- Scoring: 1-5

**Sample items:**
- “When searching for literature on a certain topic, I know exactly in which order the available information resources should be used.”
- “When conducting a literature search on a certain topic, I am able to decide quickly whether a certain information resource is of relevance.”
### Descriptive Statistics (total sample)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Min</th>
<th>Max</th>
<th>M</th>
<th>SD</th>
<th>Cronbach’s Alpha</th>
<th>( r_{it-t} ) (min-max)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge test</td>
<td>0.35</td>
<td>0.79</td>
<td>0.55</td>
<td>0.08</td>
<td>.70</td>
<td>.03 - .52</td>
</tr>
<tr>
<td>Self-efficacy scale</td>
<td>1.70</td>
<td>4.70</td>
<td>3.34</td>
<td>0.55</td>
<td>.72</td>
<td>.21 - .51</td>
</tr>
</tbody>
</table>

- Correlation (Knowledge Test – Self-Efficacy Scale): \( r = .23, p < .01 \)
### Means and Standard Deviations of Knowledge Test [M(SD)]

<table>
<thead>
<tr>
<th>Group</th>
<th>1 (PC, unsupervised)</th>
<th>2 (PC, supervised)</th>
<th>3 (P&amp;P, unsupervised)</th>
<th>4 (P&amp;P, supervised)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.52 (0.07)</td>
<td>0.55 (0.07)</td>
<td>0.56 (0.09)</td>
<td>0.57 (0.09)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Medium</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC</td>
<td>P&amp;P</td>
</tr>
<tr>
<td></td>
<td>unsupervised</td>
</tr>
<tr>
<td></td>
<td>supervised</td>
</tr>
<tr>
<td>0.54 (0.07)</td>
<td>0.57 (0.08)</td>
</tr>
</tbody>
</table>

- **ANOVA:**
  - significant effect of Medium: $F = 4.42$, $df = 1/137$, $p < .05$, $\eta^2_{(part.)} = .031$ $\Rightarrow$ P&P > Computer
  - marginally significant effect of Mode: $F = 2.99$, $df = 1/137$, $p < .10$, $\eta^2_{(part.)} = .021$ $\Rightarrow$ supervised $\geq$ unsupervised
  - no interaction
# Means and Standard Deviations Of Self-Efficacy Scale

<table>
<thead>
<tr>
<th>Group</th>
<th>1 (PC, unsupervised)</th>
<th>2 (PC, supervised)</th>
<th>3 (P&amp;P, unsupervised)</th>
<th>4 (P&amp;P, supervised)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.25 (0.55)</td>
<td>3.47 (0.59)</td>
<td>3.31 (0.51)</td>
<td>3.36 (0.55)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Medium</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC</td>
<td>P&amp;P</td>
</tr>
<tr>
<td>3.36 (0.58)</td>
<td>3.33 (0.52)</td>
</tr>
</tbody>
</table>

- ANOVA: all effects not significant
Psychometric Properties
[Cronbach’s α]

- Comparison of alphas: Feldt-Test (F-Test; Feldt, 1969; Feldt & Kim, 2009)

<table>
<thead>
<tr>
<th>Scale</th>
<th>Group</th>
<th>1 (PC, unsupervised)</th>
<th>2 (PC, supervised)</th>
<th>3 (P&amp;P, unsupervised)</th>
<th>4 (P&amp;P, supervised)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge test</td>
<td></td>
<td>.66</td>
<td>.61</td>
<td>.79</td>
<td>.60</td>
</tr>
<tr>
<td>Self-efficacy scale</td>
<td></td>
<td>.73</td>
<td>.76</td>
<td>.70</td>
<td>.72</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scale</th>
<th>Medium</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PC</td>
<td>P&amp;P</td>
</tr>
<tr>
<td>Knowledge test</td>
<td>.65</td>
<td>.73</td>
</tr>
<tr>
<td>Self-efficacy scale</td>
<td>.75</td>
<td>.71</td>
</tr>
</tbody>
</table>
## Correlations of Knowledge Test and Self-Efficacy Scale

<table>
<thead>
<tr>
<th>Group</th>
<th>1 (PC, unsupervised)</th>
<th>2 (PC, supervised)</th>
<th>3 (P&amp;P, unsupervised)</th>
<th>4 (P&amp;P, supervised)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-.11</td>
<td>.22</td>
<td>.31*</td>
<td>.46**</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Medium</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC</td>
<td>P&amp;P</td>
</tr>
<tr>
<td>.09</td>
<td>.36**</td>
</tr>
<tr>
<td></td>
<td>unsupervised</td>
</tr>
<tr>
<td></td>
<td>supervised</td>
</tr>
<tr>
<td></td>
<td>.15</td>
</tr>
<tr>
<td></td>
<td>.32**</td>
</tr>
</tbody>
</table>
(4) Summary and Discussion

- **Information literacy self-efficacy scale**
  - equivalent with regard to means, internal consistencies → robust measure, applicable independent of medium/mode of test administration

- **Information literacy knowledge test**
  - *small but noticeable effects of medium*: higher test scores, (numerically) higher reliability, and higher correspondence of test scores and self-assessments for P&P version compared to computer version → more careful completion of test?
  - *no consistent effects of mode*: equal test scores, equal level of correspondence between test scores and self-assessment, but (slightly) higher internal consistency under uncontrolled (vs. controlled) administration conditions
(1) Small sample sizes → robustness of findings?
(2) Homogenous sample of educational students with considerable test and computer experience, monetary compensation → generalizability?
(3) Limited anonymity in the „unsupervised“ conditions → better quality of data compared to „classical“ unproctored testing?
(4) „Low stakes“ test situation: more effects of mode of administration if test results are of high personal relevance?
Thank you for listening!

Contact:
Dr. Anne-Kathrin Mayer
ZPID – Leibniz Institute for Psychology Information
D-54286 Trier
mayer@zpid.de
References


Feldt, L. S. (1969). A test of the hypothesis that Cronbach’s alpha or Kuder-Richardson coefficient twenty is the same for two tests. Psychometrika, 34, 363–373.


