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

(2) Information Literacy: Definition

- *"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

Information Literacy: Assessment

(for overviews, see McCulley, 2009; Walsh, 2009)

- **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 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; Kurbanoglou, Akkoyunlu & Umay, 2006)

(3) Experimental Study

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)

Medium of test administration	**Mode of test administration**	
	Unsupervised	*Supervised*
Computer	Group 1 (*n* = 34)	Group 2 (*n* = 32)
Paper-and-Pencil	Group 3 (*n* = 43)	Group 4 (*n* = 32)

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

Methods: Knowledge test

- **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?

- **Bibliographic databases usually have a thesaurus search.**
- Boolean operators can only be used with bibliographic databases.
- **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)

Measure	*Min*	*Max*	*M*	*SD*	Cronbach's Alpha	r_{it-t} (min-max)
Knowledge test	0.35	0.79	0.55	0.08	.70	.03 - .52
Self-efficacy scale	1.70	4.70	3.34	0.55	.72	.21 - .51

- Correlation (Knowledge Test – Self-Efficacy Scale): $r = .23$, $p < .01$

Means and Standard Deviations of Knowledge Test [*M*(*SD*)]

Group			
1 (PC, unsupervised)	**2** (PC, supervised)	**3** (P&P, unsupervised)	**4** (P&P, supervised)
0.52 (0.07)	0.55 (0.07)	0.56 (0.09)	0.57 (0.09)

Medium		Mode	
PC	P&P	unsupervised	supervised
0.54 (0.07)	0.57 (0.08)	0.54 (0.09)	0.56 (0.07)

- ANOVA:
  - significant effect of Medium: $F = 4.42$, df 1/137, $p < .05$, $\eta^2_{(part.)} = .031$ → **P&P > Computer**
  - marginally significant effect of Mode: $F = 2.99$, df 1/137, $p < .10$, $\eta^2_{(part.)} = .021$ → ***supervised >= unsupervised***
  - no interaction

Means and Standard Deviations Of Self-Efficacy Scale

Group			
1 (PC, unsupervised)	**2** (PC, supervised)	**3** (P&P, unsupervised)	**4** (P&P, supervised)
3.25 (0.55)	3.47 (0.59)	3.31 (0.51)	3.36 (0.55)

Medium		Mode	
PC	P&P	unsupervised	supervised
3.36 (0.58)	3.33 (0.52)	3.28 (0.52)	3.41 (0.57)

- ANOVA: all effects not significant

Psychometric Properties [Cronbach's α]

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

Scale	Group			
	1 (PC, unsupervised)	**2** (PC, supervised)	**3** (P&P, unsupervised)	**4** (P&P, supervised)
Knowledge test	.66	**.61**	**.79**	**.60**
Self-efficacy scale	.73	.76	.70	.72

Scale	Medium		Mode	
	PC	P&P	unsupervised	supervised
Knowledge test	.65	.73	**.75**	**.61**
Self-efficacy scale	. 75	.71	.72	.74

Correlations of Knowledge Test and Self-Efficacy Scale

Group			
1 (PC, unsupervised)	**2** (PC, supervised)	**3** (P&P, unsupervised)	**4** (P&P, supervised)
-.11	.22	**.31***	**.46****

Medium		Mode	
PC	P&P	unsupervised	supervised
.09	**.36****	.15	.32**

(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 noticable 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

Limitations / Future Perspectives

(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

Association of College and Research Libraries. (2000). *Information Literacy Competency Standards for Higher Education.* American Library Association.

Behm, T. (2015). Informationskompetenz und Selbstregulation: Zur Relevanz bereichsspezifischer Selbstwirksamkeitsüberzeugungen. In A.-K. Mayer (Hrsg.) *Informationskompetenz im Hochschulkontext – Interdisziplinäre Forschungsperspektiven* (S. 151-162). Lengerich: Pabst Science Publishers.

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.

Feldt, L. S. & Kim, S. (2006). Testing the difference between two Alpha coefficients with small samples of subjects and raters. *Educational and Psychological Measurement, 66*(4)*,* 589-600.

Gnambs, T., Batinic, B. & Hertel, G. (2011). Internetbasierte psychologische Diagnostik. In L. F. Hornke, M. Amelang & M. Kersting (Hrsg.), *Verfahren zur Leistungs-, Intelligenz- und Verhaltensdiagnostik*, Enzyklopädie der Psychologie, Psychologische Diagnostik (Bd. II/3, S. 448-498). Göttingen: Hogrefe.

Gnambs, T., & Kaspar, K. (2014). Disclosure of sensitive behaviors across self-administered survey modes: a meta-analysis. *Behavior Research Methods*, 1-23.

Heinström, J. (2005). Fast surfing, broad scanning and deep diving. The influence of personality and study approach on students' information-seeking behavior. *Journal of Documentation, 61(2),*228-247.

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.

Kim, J. P. (1999, October). *Meta-analysis of equivalence of computerized and P&P tests on ability measures.* Annual Meeting of the Mid-Western Educational Research Association (Chicago, IL).

References (contd.)

Kurbanoglu, S., Akkoyunlu, B. & Umay, A. (2004). Developing the information literacy self-efficacy scale. *Journal of Documentation, 62*(6), 730-743.

Ladyshewsky, R. K. (2014). Post-graduate student performance in 'supervised in-class' vs. 'unsupervised online'multiple choice tests: implications for cheating and test security. *Assessment & Evaluation in Higher Education*, (ahead-of-print), 1-15.

Leichner, N., Mayer, A.-K., Peter, J., & Krampen, G. (under review). *On the relationship between self-assessed abilities and objective performance measures: The case of information literacy in the context of a randomized blended learning program evaluation.* Manuscript submitted for publication.

Leichner, N., Peter, J., Mayer, A.-K., & Krampen, G. (2013). Assessing information literacy among German psychology students. *Reference Services Review, 41*(4), 660–674. doi:10.1108/RSR-11-2012-0076

Leichner, N., Peter, J., Mayer, A.-K., & Krampen, G. (2014). Assessing information literacy using information search tasks. *Journal of Information Literacy*, 8(1), 3-20. doi:10.11645/8.1.1870

Lievens, F., & Burke, E. (2011). Dealing with the threats inherent in unproctored Internet testing of cognitive ability: Results from a large-scale operational test program. *Journal of Occupational and Organizational Psychology*, *84*(4), 817-824.

Mead, A. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*(3), 449-458.

Oakleaf, M. (2009). The information literacy instruction assessment cycle. *Journal of Documentation, 65*(4), 539-560.

Preckel, F., & Thiemann, H. (2003). Online-versus paper-pencil version of a high potential intelligence test. *Swiss Journal of Psychology/Schweizerische Zeitschrift für Psychologie/Revue Suisse de Psychologie*, *62*(2), 131-138.

References (contd.)

Rosman, T., Mayer, A.-K., & Krampen, G. (2015). Measuring psychology students' information-seeking skills in a situational judgment test format: Construction and validation of the PIKE-P Test. *European Journal of Psychological Assessment.* Advance online publication. http://dx.doi.org/10.1027/1015-5759/a000239

Timmers, C. & Glas, C. (2010). Developing scales for information-seeking behaviour. *Journal of Documentation, 66*(1), 46-69.

Wasko, L., Lawrence, A. D., & O'Connell, M. S. (2015, April). *What matters in the test environment?* 30th Annual Conference of the Society for Industrial and Organizational Psychology, Philadelphia, PA. Retrieved from: http://www.researchgate.net/profile/Matthew_Oconnell3/publication/275582078_What_matters_in_the_test_environment/links/553fb4d20cf2320416ebe940.pdf (July 23rd, 2015)

Weigold, A., Weigold, I. K., & Russell, E. J. (2013). Examination of the equivalence of self-report survey-based paper-and-pencil and internet data collection methods. *Psychological Methods, 18*(1), 53.