

A peer-reviewed version of this preprint was published in PeerJ on 23 April 2015.

[View the peer-reviewed version](https://doi.org/10.7717/peerj.917) (peerj.com/articles/917), which is the preferred citable publication unless you specifically need to cite this preprint.

Kildebro N, Amirian I, Gögenur I, Rosenberg J. 2015. Test re-test reliability and construct validity of the star-track test of manual dexterity. PeerJ 3:e917 <https://doi.org/10.7717/peerj.917>

Title:

Test re-test reliability and construct validity of the star-track test of manual dexterity.

Authors:

Niels Kildebro¹, Ilda Amirian¹, Ismail Gögenur², Jacob Rosenberg¹

From:

¹ Center for Perioperative Optimization, Department Surgery, University of Copenhagen, Herlev Hospital, Herlev Ringvej 75, 2730 Herlev, Denmark

² Køge Hospital, Department of Surgery, Lykkebækvej 1, 4600 Køge, Denmark

Corresponding author:

Niels Kildebro, E-mail: nielskildebro@gmail.com, Phone: +45 31103626

Conflicts of Interest

None

17 **Abstract**

18 **Objectives:** To determine test re-test reliability and construct validity of the star-track test of
19 manual dexterity.

20 **Design:** Test re-test reliability was tested in a controlled study. Construct validity was tested in a
21 blinded randomized crossover study.

22 **Setting:** The study was performed at a university hospital in Denmark.

23 **Participants:** A total of 11 subjects for test re-test and 20 subjects for the construct validity study
24 were included. All were healthy volunteers.

25 **Intervention:** The test re-test trial had two measurements with 2 days pause in between. The
26 interventions in the construct validity study included baseline measurement, intervention 1: fatigue,
27 intervention 2: stress, and intervention 3: fatigue and stress. There was a 2 day pause between each
28 intervention.

29 **Main outcome measure:** Integrated measure of completion time and number of errors.

30 **Results:** All participants completed the study (test re-test n = 11; construct validity n=20). Test re-
31 test showed a strong Pearson product-moment correlation ($r = 0.90$, $n = 11$, $P < 0.01$) with no sign
32 of learning effect. The 20 subjects in the construct validity trial were randomized to the order of the
33 four interventions, so that all subjects completed each intervention once. A repeated measures
34 ANOVA determined that mean integrated measure differed between interventions ($p = 0.003$). Post
35 hoc tests using Bonferroni correction revealed that compared with baseline all interventions had
36 significantly higher integrated scores ranging from 47-59% difference in mean.

37 **Conclusion:** The star track test of manual dexterity had a strong test re-test reliability, and was able
38 to discriminate between a subject's normal manual dexterity and dexterity after exposure to fatigue
39 and/or stress.

40

41 Background

42 A surgeon's manual dexterity is often an outcome parameter in studies examining
43 environmental effects such as work environment or night shifts on surgeons (Amirian et al. 2014;
44 Dorion & Darveau 2013). Simulation tools are often used, but these are mostly time consuming
45 tests that are not readily available. Often a study needs a tool that is easy to administer and is
46 portable so that it can be used where the study calls for it. One such device was introduced in an
47 interventional study for measuring surgeons' accuracy (Dorion & Darveau 2013). The surgeons
48 were to follow a star shaped track with a pair of surgical scissors and each time the scissors touched
49 the border of the track, an error was counted. The track was to be completed 3 times and errors were
50 noted. The study stated that it was examined for test re-test reliability. They reported a Pearson's
51 correlation of $r = 0.955$ (Dorion & Darveau 2013; Savoie & Prince 2002). However, the method of
52 testing this was not described. The accuracy test's design allowed it to measure manual dexterity in
53 the dynamic phase and the subject had to use power grip, precision handling of a hand held object
54 and hand to eye coordination. These are all components of accuracy and are needed in instrument
55 handling and therefore important to surgeons' technical skills (Memon et al. 2010). Furthermore,
56 the test measures the subject's accuracy with a surgical tool. All these qualities make the test
57 appropriate for measuring the manual dexterity of surgeons.

58 The test needs further validation if it is to be used in further research (Fess 1995). The
59 psychometric qualities have not been tested thoroughly enough to state that the test is valid and
60 measures the intended characteristic (Fess 1995; Law 1987; Rudman & Hannah 1998). Furthermore
61 the test has no equipment constructions standards or instructions of use available, so the test lags
62 repeatability and reproducibility (Fess 1995; Law 1987; Rudman & Hannah 1998; Aaron & Jansen
63 2003). With further exploration of reliability and validity the test could be an excellent tool for

64 measuring manual dexterity, providing an assessment tool that requires short time to be
65 administered and is commercially available.

66 The purpose of this article was to provide construction standards, instructions for
67 application, test-retest reliability and construct validity for the star shaped test of manual dexterity.

69 **Method**

70 Equipment construction standards

71 The star-track test of manual dexterity consists of the following components:

- 72 ○ Replacement star Model 32532A from Lafayette instruments (lafayette instruments
73 2014)
- 74 ○ MakeyMakey from joylabz.com (JoyLabz 2014).
- 75 ○ Computer running the software Star Track 32bit.exe (Nørregaard 2014).
- 76 ○ Standard Metzenbaum surgical scissors

77
78 The replacement star Model 32532A from Lafayette instruments is a metal plate measuring 22 cm x
79 22 cm with a star shaped track in its center. The six-pointed star shaped track measures 15.3 cm
80 from point to opposite point. The track is 0.9 cm wide and is made of a non-conducting material. A
81 MakeyMakey is an inventor kit that can turn everyday objects into touchpads and combine them
82 with a computer. This is explained further at <http://makeymakey.com>. The MakeyMakey is used to
83 connect the metal plate and the Metzenbaum surgical scissors to a computer. Picture 1a shows a line
84 drawing of the test setup. In picture 1b detailed measurements of the dimensions of the metal star
85 shaped track are shown. A complete setup of the test with all its components is illustrated in picture
86 1c and the command window of Star Track 32bit and a sample test results window from the
87 program can be seen in picture 1d and 1e.

88

89 Instructions for administration of the test.

90 The test is setup in the following way. The program Star Track 32bit.exe needs to be installed on
91 the computer used for the test. It is recommended to use a laptop to increase the transportability of
92 the test. Star Track 32bit.exe is a freeware that can be downloaded and installed from
93 <http://bitbucket.org/lassebn/star-track>. MakeyMakey is used to connect the components of the star-
94 track test together using the following steps: 1: Connect the MakeyMakey to the computer via USB
95 cable. The MakeyMakey will autoinstall (JoyLabz 2014). 2: Connect the “space” part of the
96 MakeyMakey to the Metzenbaum surgical scissors using two alligator clips linked together. 3:
97 Connect the “ground” of the MakeyMakey to the side of the metal plate using an alligator clip.

98 The test should be performed in a quiet room without distractions. The examiner
99 places the metal plate about 10 cm from the edge of a table where the subject is sitting comfortably.
100 The examiner instructs the subject to use the scissor in the hand that he wishes to examine. Using
101 the scissors, the subject must follow the star shaped track ten times, five times clockwise and five
102 times counterclockwise. All ten rounds are completed continuously. The tip of the scissors must be
103 in contact with the star shaped track during the entire test. Each time the scissors come into contact
104 with the border of the track, an error is registered. Completion time and number of errors are
105 registered automatically by Star Track 32bit. The examiner should read the following instructions to
106 the subject: “To complete the test, you must follow the star shaped track with the surgical scissors.
107 You are to complete ten rounds, five rounds clockwise and five rounds counterclockwise. All ten
108 rounds are to be completed continuously. You are to complete the ten rounds as quickly as possible
109 with as few errors as possible. An error is counted every time the scissors touch the border of the
110 star shaped track. The scissors must touch the plate at all times during the test. “

111 With Star Track 32bit running on the computer, the examiner names the test result file. When
112 he/she presses the enter button the test will begin. When the subject completes the final round the
113 examiner presses “q” to stop the test.

115 Scoring

116 The Star Track 32bit program automatically records the time (in seconds) it takes to complete the
117 test. It also automatically records errors.

119 **Construct validity**

120 Design

121 We wished to study whether the star-track test would be able to distinguish between the base level
122 of manual dexterity of a person and when the person was fatigued and/or stressed. This was done by
123 conducting a randomized crossover study. Each subject was to complete the star-track test four
124 times. At each trial they were randomly assigned to different interventions. Each subject was to
125 complete all four interventions, and never the same intervention more than once. Each trial was
126 separated by two days pause. The interventions were: Baseline measurement, the subject completed
127 the star-track test without further intervention. Intervention 1: The subject was fatigued in his
128 dominant arm before completing the star-track test. Intervention 2: the subject was stressed while
129 performing the star-track test. Intervention 3: The subject was fatigued prior to the star-track test
130 and stressed while performing the star-track test. Completion time and number of errors in the star-
131 track test were measured at all four interventions using the Star Track 32bit software. The order in
132 which each subject received the four interventions was randomized
133 (<http://www.randomization.com>). Using the list, a research fellow not involved in the study packed
134 and sealed 4 opaque envelopes (labeled day 1, day 4, day 7 and day 10) for each subject. These

135 envelopes were opened on the respective days just before the test commenced, so that the subject
136 and examiner were blinded until that point.

137

138 Subjects for construct validity

139 We aimed to include 20 subjects. The subjects were all volunteers and gave written informed
140 consent before inclusion. Subjects had to understand Danish (written and spoken). They were
141 excluded if they were diagnosed with heart, endocrine, neurological, autoimmune or psychological
142 disease, suffered from sleep disorders or had muscular-skeletal disorders of the upper extremities
143 (e.g. osteoarthritis, rotator cuff syndrome, hand injuries).

144

145 Method of achieving muscular fatigue

146 The fatigue was achieved by letting the subject hold a 2.5 kg weight in his dominant hand, and
147 holding the dominant arm to 90 degrees flexion. They were to hold this position without moving for
148 as long as possible. The subject then proceeded to complete the star-track test within 10 seconds.
149 This test has previously been used to measure muscular fatigue (Dorion & Darveau 2013) and is
150 described in occupational health literature as a way to achieve static muscular fatigue (Chaffin
151 1973).

152

153 Method of inducing stress

154 The brain can focus on performing a specific task at normal level, as long as the mental resources
155 exceed the demand of the task in progress. If multiple tasks are to be performed at the same time,
156 the demands of the tasks will at some point exceed the mental workload tolerance. This will cause
157 stress and subjects will begin making errors (Boles & Law 1998; Grier et al. 2008). According to
158 the theory of multiple resources, there are several mental resource pools, enabling several actions to

159 be performed simultaneously. However, if the actions performed require resources from the same
160 pool, they will cause stress more quickly (Wickens 2008). This allows for prediction of workload
161 overload by determining difficulty of the tasks undertaken and task interference. The star-track test
162 is visually perceived, requires spatial understanding and a manual response. The distraction was
163 designed to drain from these mental resources.

164 While the subject was performing the star-track test, the examiner would show the
165 subject 10 cards from a regular deck of cards. One card per round completed in the star-track test.
166 The card was placed near the metal plate of the star-track test, allowing the subject to have both the
167 star-track and the card in his field of vision. The subject had to identify the card by rank and suit
168 while performing the star-track test. According to the computational 3-D+1 model of multiple
169 resources (Wickens 2008), the difficulty of the tasks are both simple (following the star-track and
170 identifying cards). The tasks share demands of workloads at two levels (perception and cognition).
171 This gives a total interference of 4 (on a scale of interference from 0-8) (Wickens 2008). If the star-
172 track test is able to detect this workload overload, a higher integrated score (longer completion time
173 and/or more errors), compared with baseline should be scored while completing the test with
174 distraction.

175

176 **Test re-test reliability**

177 The reliability was tested with a controlled design. The purpose of this test was to determine the
178 test-retest effect and whether or not the test was consistent over time. The subjects completed the
179 star-track test with an interval of two days between tests. This design has been used previously to
180 perform test-retest trials of manual dexterity (Aaron & Jansen 2003). Completion time and errors
181 were measured at both tests using the Star Track 32bit software. To measure face-validity each
182 subject was asked if he understood the purpose of the star-track test, and what they believed it was

183 supposed to measure. We aimed to include 11 subjects. Inclusion and exclusion criteria were the
184 same as for the validity test.

185

186 Ethics and permissions

187 The study was registered at Clinicaltrials.gov (NCT02146443). The data collection was approved
188 by the Danish Data Protection Agency (journal no: HEH-2014-060, I-Suite no. 02972). The study
189 was exempt from approval by The Regional Danish Committee on Biomedical Research Ethics
190 (protocol no: H-6-2014-031). All subjects were volunteers who gave written informed consent, and
191 received no compensation for participating in the study.

192

193 Statistics

194 All statistics were calculated using IBM SPSS Statistics version 22 (SPSS, Chicago, IL, USA) and
195 Microsoft Office Excel 2007. To receive a complete estimate of a subject's manual dexterity, we
196 used an integrated measure for completion time and number of errors (Silverman et al. 1993). The
197 total number of ranks for conducted trials were found (80 for validity; 22 for the test-retest) and
198 mean rank was calculated. The difference of completion time and number of errors from respective
199 mean ranks was calculated as a % difference, and added on a per-subject basis to form an integrated
200 measure (Silverman et al. 1993). Since this was a pilot test, no sample size was calculated as no
201 data were available. Thus, sample size was determined by means of qualified estimate (Hertzog
202 2008). Study population age was described as median (range). We used the Shapiro-Wilk test of
203 normality to determine that data were normally distributed. Mauchly's Test was used to test for
204 Sphericity. We used repeated measures ANOVA with post hoc testing with Bonferoni correction for
205 intergroup measurements in the validation study and Pearson correlation coefficients for test re-test

206 reliability analysis. Test days were also compared with paired samples t-tests. $P \leq 0.05$ was
207 regarded as statistically significant.

208

209 **Results**

210 Construct validity

211 A total of 20 subjects completed this study, 9 females and 11 males, with a median age 26 years
212 (range 22-29). Of the participants 3 were left-handed and 17 were right-handed. The integrated
213 measures scores for each of the four test arms of the crossover study can be found in table 1. We
214 tested for normality using the Shapiro-Wilk test. It showed that the data for all four test arms of the
215 crossover study did not violate the assumption of normality (baseline $p = 0.89$; intervention 1 $p =$
216 0.67 ; intervention 2 $p = 0.79$; intervention 3 $p = 0.44$). A repeated measures ANOVA was done to
217 determine if the integrated measures significantly differed from each other. Mauchly's Test of
218 Sphericity indicated that the assumption of sphericity of the data had not been violated ($\chi^2 (5) =$
219 6.90 , $p = 0.23$) and thus no correction was used in the repeated measures ANOVA. It was
220 determined that mean integrated measure differed significantly between interventions ($p = 0.003$).
221 Post hoc tests using Bonferroni correction revealed that compared with baseline all interventions
222 had significantly higher integrated scores, indicating that the test was able to differentiate between
223 the baseline and the interventions (see table 2). Furthermore; intervention 3 scored higher integrated
224 measure than intervention 1 and 2, with a mean difference in integrated measure of 0.12 and 0.10
225 respectively (see figure 1), although this difference was statistically insignificant ($P = 1$ for both).

226

227 Test re-test reliability:

228 A total of 11 subjects completed this study, hereof 5 females. The median age was 27 years (range
229 22-35). Two of the subjects were left-handed and nine were right-handed. The integrated measures
230 scores for each test day are presented in table 3. A Pearson product-moment correlation was run to
231 determine the relationship between the test days. The data showed no violation of normality
232 (Shapiro-Wilk test of test day 1 $p = 0.32$ and test day 2 $p = 0.25$), linearity or homoscedasticity.
233 There was a strong, positive correlation between the integrated measures of the two test days ($r =$
234 0.90 , $n = 11$, $P < 0.01$). Test day 1 and test day 2 were compared with paired samples t-tests to
235 ensure that the Pearson correlation coefficient was not high due to a consistent difference (e.g.
236 learning effect). There was no significant difference in integrated measure ($p = 0.21$).
237

238 Discussion

239 The star-track test was able to detect a difference between the baseline measurement and all three of
240 the interventions in the construct validity study. This indicates that the test was able to discriminate
241 between a person's baseline and impaired manual dexterity due to fatigue and/or stress. The test re-
242 test reliability showed that the star-track test had a strong test re-test reliability.

243 The purpose of the star-track test was to be an evaluative tool for measuring manual
244 dexterity, and to measure changes in individuals. More specifically, the target population of the test
245 was subjects with no impairment or disease in the upper extremities. The test involved a surgical
246 instrument, it was meant to be used in future research to evaluate the manual dexterity of surgeons.
247 If the test is to be used as a descriptive tool, more studies should be conducted where normative
248 data should be collected on different groups of subjects standardized for age, gender, surgical
249 experience and maybe various impairments of the upper extremities. The star-track test has no
250 predictive value yet. To gain this, studies where surgeon's integrated measures in the star-track test
251 are compared to patient outcomes would be needed.

252 The reliability of the star-track test of manual dexterity has been explored in previous
253 studies (Dorion & Darveau 2013; Savoie & Prince 2002). In this article, we confirmed the previous
254 findings of a strong test-retest reliability of the star track test. Also, the trials ruled out any
255 significant learning effects, which is especially important for the consistency of the test.
256 Furthermore, we described the method of obtaining the reliability results in detail, which had not
257 been done before. To make the star-track test accessible, equipment construction standards were
258 provided along with instructions for administration. We used components for construction of the
259 test that were commercially available, so that the test can be reconstructed and reproduced. This
260 further established the consistency and reliability of the test. As the data of the test were gained by
261 means of a computer program, we did not examine for inter-rater and intra-rater reliability, as the
262 standardized computer program minimized these factors.

263 All subjects easily understood that the test measured manual dexterity, which
264 indicated that the test had good face validity and was easy to understand. The content validity had
265 already been established, as the test was used to test the accuracy of surgeons in a previous study
266 (Dorion & Darveau 2013). However, manual dexterity is a more complete measure of a surgeon's
267 skill than accuracy. Dexterity is the ability to manipulate objects with your hands with a specific
268 purpose in mind (Dunn et al 1994; Baum & Edwards 1995). Dexterity can be subdivided into a
269 static phase, and a dynamic phase which involves powergrip (adapting hand strength) and precision
270 handling of handheld objects (Kamakura et al. 1980). The characteristics of manual dexterity are
271 accuracy and speed (Aaron & Jansen 2003). In this study we expanded the measurement to be a
272 more complete concept of manual dexterity by using an integrated measure of errors and
273 completion time. We believe that by doing this we have increased the content validity of the star-
274 track test.

275 Construct validity was explored in this study, and the findings were, that the more
276 stressed and/or fatigued a subject was, the higher the integrated measure of manual dexterity. This
277 was in accordance with the hypothesis. It appears that the star-track test was responsive enough to
278 measure different levels of stress/and fatigue. It detected a mean difference of 10-12% between
279 intervention 3, where both muscular fatigue and stress was combined, and both intervention 1 and
280 intervention 2. It indicated that the test might be able to measure different intensities of stress and
281 fatigue's effect on the subject's manual dexterity.

282 The criterion validity of the test still needs to be established. This could be done in
283 future studies comparing data from the star-track test to other established accuracy tests and tests of
284 manual dexterity.

285 With the data presented in this article, we believe that the star-track test of manual
286 dexterity may be used in future research, when testing the accuracy and manual dexterity of
287 surgeons. The star-track test can be used to discriminate between a subject's normal manual
288 dexterity and after exposure to fatigue and/or stress.

289

290 **References**

- 291 Aaron DH, Jansen CW. 2003. Development of the Functional Dexterity Test (FDT):
292 construction, validity, reliability, and normative data. *Journal of Hand Therapy* 16:12-21.
- 293 Amirian I, Andersen LT, Rosenberg J, Gögenur I. 2014. Laparoscopic Skills and Cognitive
294 Function are not Affected in Surgeons During a Night Shift. *Journal of Surgical Education*
295 71:543-550.
- 296 Baum C, Edwards D. 1995. Position paper: occupational performance: occupational therapy's
297 definition of function. American Occupational Therapy Association. *American Journal of*
298 *Occupational Therapy* 49:1019-1020.
- 299 Boles DB, Law MB. 1998. A simultaneous task comparison of differentiated and
300 undifferentiated hemispheric resource theories. *Journal of Experimental Psychology:*
301 *Human Perception and Performance* 24:204-215.
- 302 Chaffin DB. 1973. Localized muscle fatigue--definition and measurement. *Journal of Occupational*
303 *Medicine* 15:346-354.
- 304 Dorion D, Darveau S. 2013. Do micropauses prevent surgeon's fatigue and loss of accuracy
305 associated with prolonged surgery? An experimental prospective study. *Annals of Surgery*
306 257:256- 259.
- 307 Dunn W, Hinojosa J, Schell B, Thumsun LK, henfeljer SD. 1994. Uniform terminology for
308 occupational therapy--third edition. American Occupational Therapy Association. *American*
309 *Journal of Occupational Therapy* 48:1047-1054.
- 310 Fess EE. 1995. Guidelines for evaluating assessment instruments. *Journal of Hand Therapy* 8:144-
311 148.
- 312 Grier R, Wickens C, Kaber D, Strayer D, Boehm-Davis D, Trafton JG, John MS. 2008. The
313 red-line of workload: Theory, research, and design. *Proceedings of the Human Factors and*

314 *Ergonomics Society Annual Meeting: Sage Publications*, p 1204-1208.

315 Hertzog MA. 2008. Considerations in determining sample size for pilot studies. *Research in*

316 *Nursing and Health* 31:180-191.

317 JoyLabz. 2014. Makeymakey. Available at <http://makeymakey.com> (accessed 19 March 2014).

318 Kamakura N, Matsuo M, Ishii H, Mitsuboshi F, Miura Y. 1980. Patterns of static prehension in

319 normal hands. *American Journal of Occupational Therapy* 34:437-445.

320 Law M. 1987. Measurement in occupational therapy: Scientific criteria for evaluation. *Canadian*

321 *Journal of Occupational Therapy* 54:133-138.

322 Lafayette Instrument Company. 2014. Replacement Star Model 32532A. Available at

323 http://www.lafayetteevaluation.com/product_detail.asp?itemid=216 (accessed 25 June

324 2014).

325 Memon MA, Brigden D, Subramanya MS, Memon B. 2010. Assessing the surgeon's technical

326 skills: analysis of the available tools. *Academic Medicine* 85:869-880.

327 Nørregaard L. 2014. Star Track 32bit.exe. Daybuilder Solutions. Available at

328 <http://bitbucket.org/lassebn/star-track> (accessed 5 December 2014).

329 Rudman D, Hannah S. 1998. An instrument evaluation framework: description and application

330 to assessments of hand function. *Journal of Hand Therapy* 11:266-277.

331 Savoie S TR, Prince F. 2002. Hauteur de la table d'operation et performance chirurgicale. Thesis.

332 Universite de Sherbrooke (Canada).

333 Silverman DG, O'Connor TZ, Brull SJ. 1993. Integrated assessment of pain scores and rescue

334 morphine use during studies of analgesic efficacy. *Anesthesia and Analgesia* 77:168-170.

335 Wickens CD. 2008. Multiple resources and mental workload. *Human Factors* 50:449-455.

336 **Legends**

337 **Picture 1a:** Line drawing of the test setup.

338 **Picture 1b:** Detailed measurements and dimensions of the metal plate with the star shaped track.

339 **Picture 1c:** A complete setup of the test with all its components.

340 **Picture 1d:** Command window of Star Track 32bit software.

341 **Picture 1e:** Sample test results window from Star Track 32bit software.

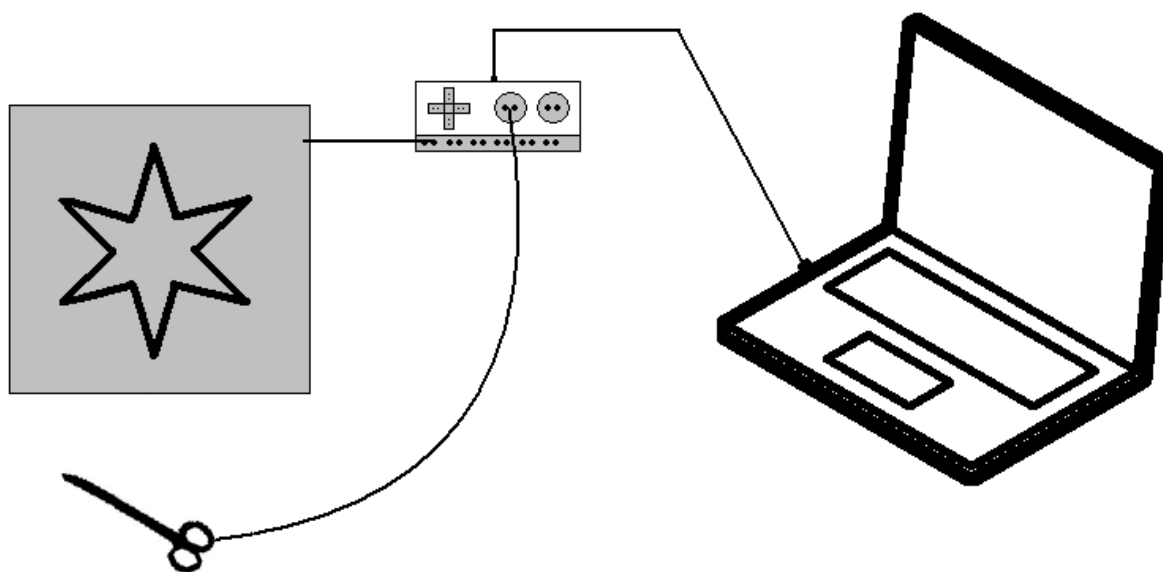
342 **Table 1:** Integrated measures of validity test. Integrated measure of time and error during
343 completion of the star-track test of manual dexterity during each of the four test arms.

344 **Table 2:** Post hoc tests of repeated measures ANOVA. Integrated measure of time and error during
345 completion of the star-track test of manual dexterity, baseline compared to the three interventions.

346 **Table 3:** Integrated measures of test re-test trial. Integrated measure of time and error during
347 completion of the star-track test of manual dexterity.

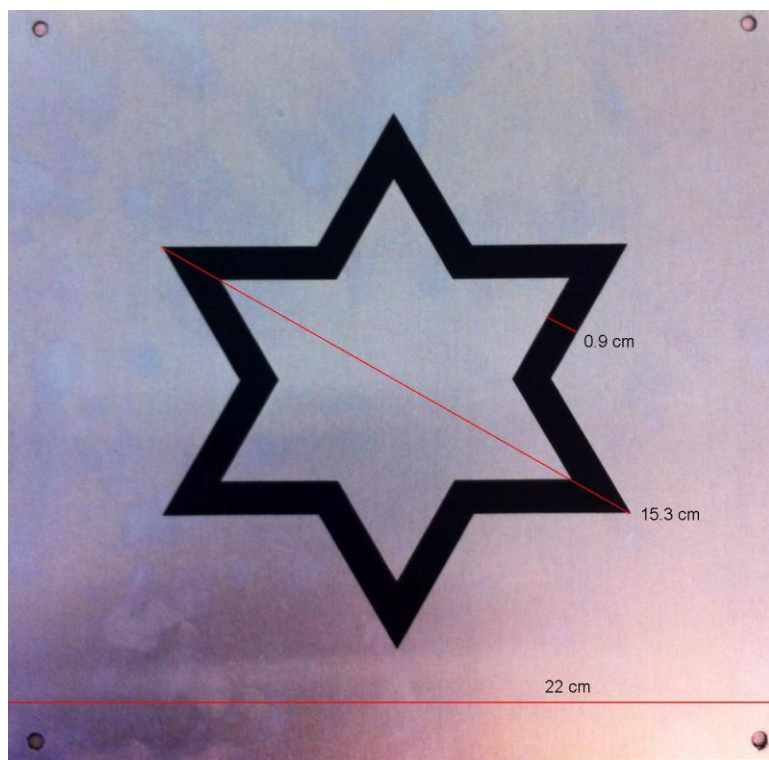
348 **Figure 1:** The mean integrated measures of time and error during completion of the star-track test
349 in the construct validation study.

350 **Picture 1a**



351

352 **Picture 1b**



353

354

355 **Picture 1c**

356

357

358

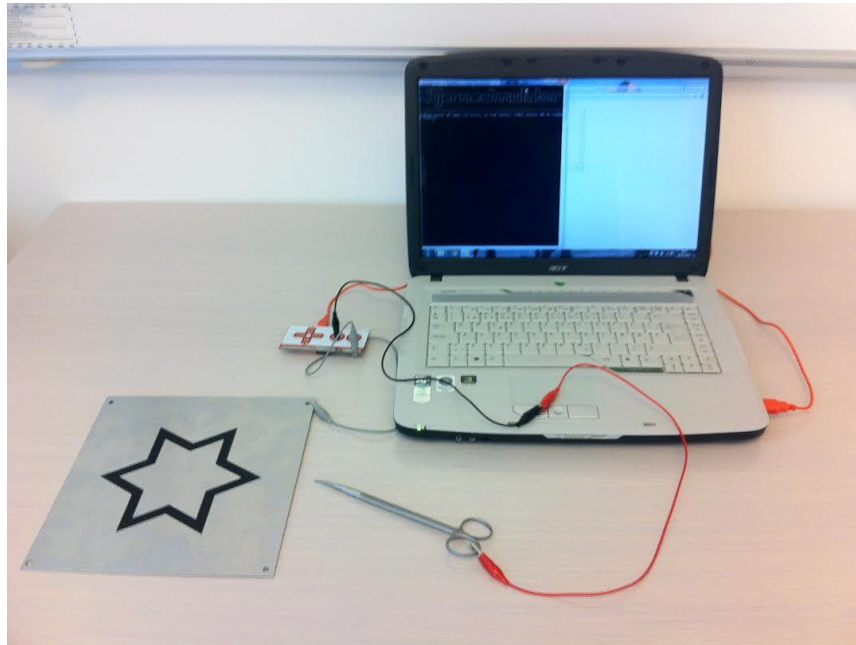
359

360

361

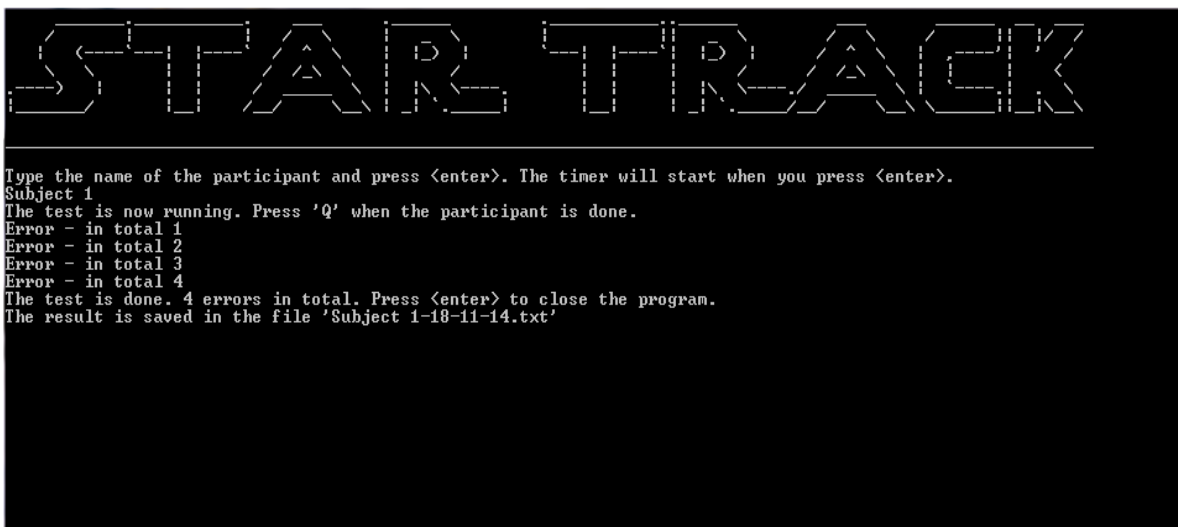
362

363



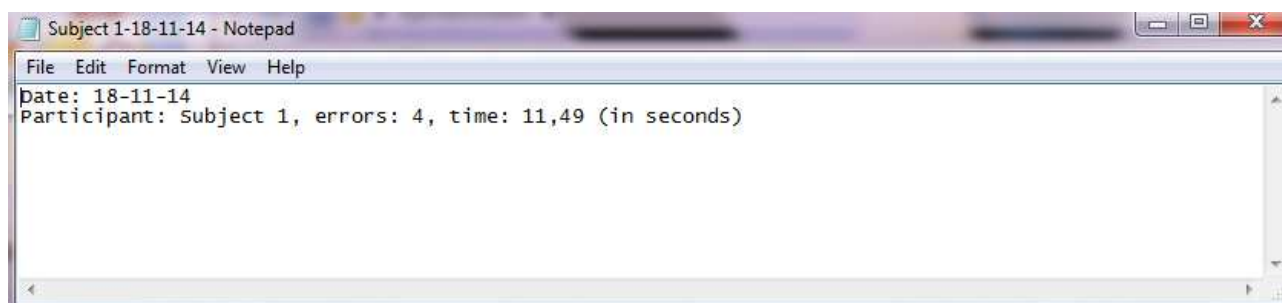
364 **Picture 1d**

365



366

367 **Picture 1e**



368
369

370 **Table 1**

Test arm	N	Integrated measure	SD
Baseline	20	-0.39	0.51
Intervention 1	20	0.08	0.61
Intervention 2	20	0.10	0.39
Intervention 3	20	0.20	0.55

373 **Table 2**

374

Comparison	Mean difference	p - values
Baseline - intervention 1	- 0.47 (-0.93; -0.01)	0.05
Baseline - intervention 2	- 0.49 (-0.92; -0.06)	0.02
Baseline - Intervention 3	- 0.59 (-1.09; -0.10)	0.01

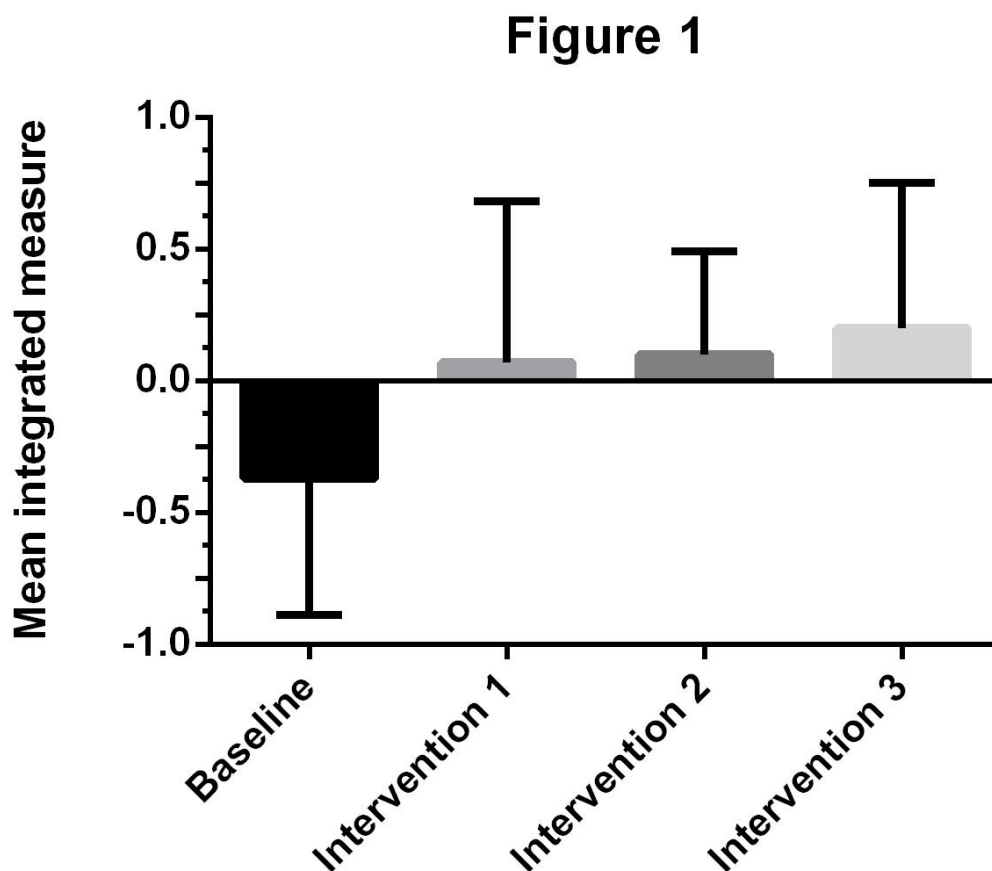
375
376
377
378

379 Values are presented as mean difference in integrated measure with 95% confidence interval. p -
380 values calculated with post hoc tests using the Bonferroni correction.

381 **Table 3**

Test re-test day	N	Integrated measure	SD
Test day 1	11	0.07	0.62
Test day 2	11	-0.05	0.72

382



384

385 Integrated measure is percent from mean integrated measure of study population. Whiskers
 386 represent standard deviation. A positive score is a poorer than average performance (e.g. longer
 387 completion time and/or more errors) when compared to the mean score, while a negative score is
 388 better than average.