Advanced course in questionnaire technique and clinimetrics

Exercises in validity – suggested solutions

The questions from the exercise text are repeated in italics. The answers are suggestions; in particular the discussion points can be answered in more than one way.

EXERCISE 1.1 (criterion validity, continuous instrument and gold standard)

- *Go through step one, two and four in the process of criterion validation (see slide 15). Discuss whether the choice of gold standard and sample is appropriate for the purpose, and whether the scores from the instrument and the gold standard were obtained independently.*

Step 1 – criterion. The SF-12 is a short version of the SF-36, developed to reproduce the SF-36 summary scores with fewer items. Using SF-36 PCS as the criterion is therefore reasonable. It is not a true gold standard, however (slide 12): SF-36 PCS has its own measurement error, and agreement with it shows that the SF-12 reproduces the SF-36, not that either of them measures physical health perfectly. Since both scores are on the same continuous scale (same units), Bland-Altman's limits of agreement is the appropriate method (slide 16).

Step 2 – sample. Validity must be assessed in the population in which the instrument is to be used. If the SF-12 is to be used in patients with severe health anxiety, this sample is relevant. However, the patients were included in an RCT with its own inclusion and exclusion criteria, so they may not be representative of all patients with severe health anxiety. Using baseline data is an advantage, since the scores are not affected by the treatment. In general, a dedicated validation study is preferable to re-using data collected for another purpose.

Step 4 – independence. The SF-12 and the SF-36 were completed as separate questionnaires, so the two scores are not computed from the same answers. On the other hand, the 12 SF-12 items are also part of the SF-36, and both were completed at the same time. Patients may therefore remember and repeat their answers, which would make the agreement look better than it is. The order of the questionnaires and the time between them would be relevant information.

- *Calculate the limits of agreement using Table 1 (see slide 26).*

Limits of agreement = mean difference ± 1.96·SD = −2.54 ± 1.96·3.71 = −2.54 ± 7.27, i.e. **(–9.81 ; 4.73)**.

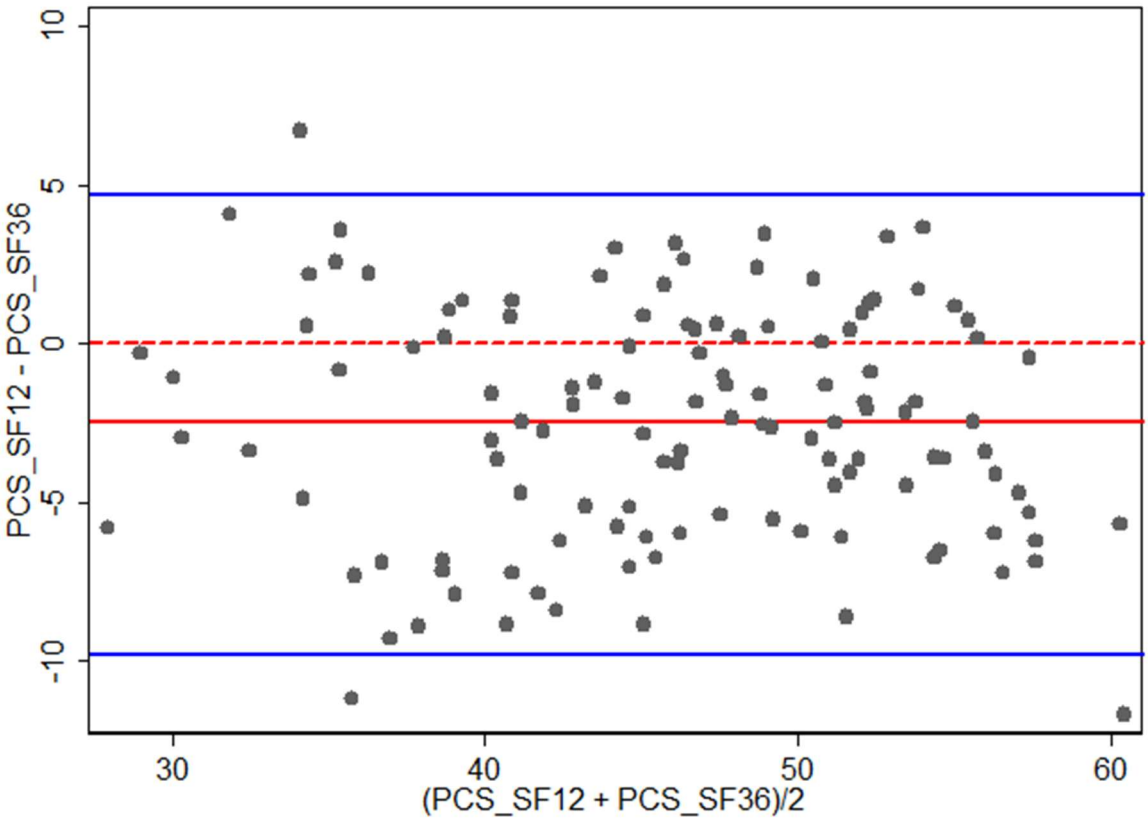
Interpretation: for 95% of patients from this population, the SF-12 PCS is expected to lie between 9.8 points below and 4.7 points above the SF-36 PCS.

- *Is there a systematic difference between the two scores (Table 1)?*

Yes. The mean PCS is 44.8 (95% CI 43.5 ; 46.2) on the SF-12 and 47.4 (46.0 ; 48.8) on the SF-36. The mean difference is **–2.54 (95% CI –3.21 ; –1.88), p < 0.001**. On average the SF-12 gives a PCS that is 2.5 points lower than the SF-36. The whole CI lies within ±5, so the systematic difference alone is smaller than what is considered clinically relevant.

- *Make a rough sketch of the Bland-Altman plot based on "Scatter plot.png".*

The sketch should have the average (PCS_SF12 + PCS_SF36)/2 on the x-axis (range approximately 28 to 61) and the difference PCS_SF12 − PCS_SF36 on the y-axis (range approximately −12 to 7), with horizontal lines at 0, at the mean difference −2.5 and at the limits of agreement −9.8 and 4.7. Most points lie between the two outer lines; only about three points lie outside (two below −9.8 and one above 4.7). With 123 patients about six (5%) would be expected outside, so this is in line with the limits of agreement.



- *Check the assumptions behind the analysis (see slide 27). In particular: is there any sign of a relationship between the differences and the averages?*
  - **Independent differences:** one pair of measurements per patient, all from baseline, and no clustering (e.g. families or clinics) → no reason to doubt independence.
  - **Same distribution of the differences (no relationship with the average and constant spread):** the scatter plot shows no clear trend and a roughly constant spread. At the highest averages (above about 55) all differences are negative, which could indicate that the SF-12 underestimates slightly more at high PCS values, but this is based on few points and is not convincing.
  - **Normally distributed differences:** the points in the QQ-plot (qq_diff_1_1.png) lie close to the straight line → the assumption is reasonable.
- *Make a summary of your findings, interpret and discuss the results. Would the required level of agreement be met if there were no systematic difference between the two scores?*

The assumptions are reasonably fulfilled. On average the SF-12 PCS is 2.5 (1.9 ; 3.2) points lower than the SF-36 PCS. The limits of agreement are (−9.8 ; 4.7), which is wider than the required level of agreement (−5 ; 5). Criterion validity of the SF-12 PCS at the individual level is therefore **not** established in this population.

Even without the systematic difference the requirement would not be met: the random error alone gives ±1.96·3.71 = ±7.3, which exceeds ±5. To meet the requirement, the SD of the differences would have to be at most 5/1.96 = 2.55. The lack of agreement is therefore mainly due to random error, not to the systematic difference.

Discussion: the conclusion depends on the purpose. For decisions about individual patients the SF-12 PCS cannot replace the SF-36 PCS. For comparing group means (e.g. in an RCT) the random error

averages out, and a systematic difference of 2.5 points may be acceptable, as long as the same instrument is used in all groups.

EXERCISE 1.2 (criterion validity, dichotomous instrument and gold standard)

- *Identify the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) in Table 2 (see slide 17).*

TP = 969, FP = 231, FN = 164 and TN = 14,696.

- *Calculate the sensitivity and specificity by "hand" (see slide 17).*

Sensitivity = TP/(TP + FN) = 969/(969 + 164) = 969/1,133 = **0.855**.

Specificity = TN/(TN + FP) = 14,696/(14,696 + 231) = 14,696/14,927 = **0.985**.

- *Calculate the negative and positive predictive value by "hand" (see slide 19).*

PPV = TP/(TP + FP) = 969/(969 + 231) = 969/1,200 = **0.808**.

NPV = TN/(TN + FN) = 14,696/(14,696 + 164) = 14,696/14,860 = **0.989**.

- *Find the sensitivity, specificity, the negative predictive value, the positive predictive value, and their 95% CI's in Table 3.*

	Estimate (%)	**95% CI (%)**
Sensitivity	85.5	83.3 ; 87.5
Specificity	98.5	98.2 ; 98.6
PPV	80.8	78.4 ; 82.9
NPV	98.9	98.7 ; 99.1

- *Why are the 95% CI's for the specificity and negative predictive value so narrow compared to the CI's for the sensitivity and positive predictive value?*

The standard error of a proportion is $\sqrt{p(1-p)/n}$. Two things make it small for the specificity and the NPV: the denominators are large (14,927 and 14,860, compared with 1,133 and 1,200 for the sensitivity and the PPV), and the proportions are close to 1, so $p(1-p)$ is small. For example:

- Sensitivity: $\sqrt{0.855 \cdot 0.145/1{,}133} \approx 0.010$, i.e. a CI of about ±2 percentage points.
- Specificity: $\sqrt{0.985 \cdot 0.015/14{,}927} \approx 0.001$, i.e. a CI of about ±0.2 percentage points.

- *Calculate the prevalence of cancer in the sample. Which PPV and NPV would you expect at this prevalence if the sensitivity and specificity were exactly as hypothesized? Are the researchers' four hypotheses consistent with each other (see slide 19)?*

Prevalence p = 1,133/16,060 = **0.071** (7.1%).

With sensitivity 0.85 and specificity 0.95:

- PPV = $0.85 \cdot 0.071 / (0.85 \cdot 0.071 + 0.05 \cdot 0.929) = 0.060/(0.060 + 0.046) \approx$ **0.56**
- NPV = $0.95 \cdot 0.929 / (0.95 \cdot 0.929 + 0.15 \cdot 0.071) = 0.883/(0.883 + 0.011) \approx$ **0.99**

The hypotheses are **not consistent**. If the sensitivity and specificity were exactly at their hypothesized minimum, the PPV would be about 56%, far below the hypothesized 80%. A PPV of 80% at this prevalence requires a specificity of about 98.4% (with sensitivity 85%). Conversely, the NPV hypothesis of 90% is so low that it is fulfilled almost automatically at this prevalence. Since the

PPV and NPV depend on the prevalence (slide 19), hypotheses about them must be formulated with the prevalence of the target population in mind.

- *Summarize and interpret your findings. Are the data consistent with the researchers' hypotheses? Consider both the point estimates and the 95% CI's: which hypotheses are demonstrated, and which are merely not contradicted by the data?*

The hypotheses are minimum requirements ("at least"). A requirement is *demonstrated* if the whole 95% CI lies above it, and *not contradicted* if the point estimate is above it but the CI includes it:

- **Specificity** 98.5% (98.2 ; 98.6): lower limit above 95% → demonstrated.
- **NPV** 98.9% (98.7 ; 99.1): lower limit above 90% → demonstrated.
- **Sensitivity** 85.5% (83.3 ; 87.5): estimate above 85%, but the CI includes values below → not contradicted, but not demonstrated.
- **PPV** 80.8% (78.4 ; 82.9): estimate above 80%, but the CI includes values below → not contradicted, but not demonstrated.

All four point estimates meet the requirements, but the sensitivity and the PPV are only just at the limit. Whether the instrument is "sufficiently valid for its purpose" also depends on the purpose: in cancer screening, 164 of 1,133 cancers (14.5%) are missed, which may or may not be acceptable. Moreover, the PPV of 80.8% only applies to populations with a prevalence of about 7%; in a population with lower prevalence the PPV will be lower.

EXERCISE 1.3 (Construct validity)

- *Evaluate the hypotheses regarding convergent and discriminant validity (H1–H4) using Table 4, and check the assumptions underlying the analyses using the graphs.*

Assumptions (slide 40). Independence: one observation per patient → fulfilled. Linearity: the scatterplot matrix (scatter_1_3.png) shows approximately linear relationships between WI-7 and SCL-8 and between WI-7 and MCS; the relationship with PCS is weak, but there is no sign of non-linearity. Approximate bivariate normality (needed for the CI of Pearson's r): the QQ-plots (qq_1_3.png) are approximately straight, and the scatterplots show no outliers. There are some floor values (0) for WI-7 and a ceiling for SCL-8 (several observations at or near 100), so the deviations are mild. Pearson's and Spearman's correlations are almost identical, so the conclusions do not depend on the choice of correlation coefficient.

Hypothesis	Expected	Pearson r (95% CI)	Result
H1: WI-7 vs SCL-8	r ≥ 0.50	0.68 (0.52 ; 0.79)	Confirmed
H2: WI-7 vs MCS	0.30 – 0.50	0.44 (0.21 ; 0.62)	Confirmed
H3: WI-7 vs PCS	r < 0.30	0.38 (0.14 ; 0.57)	Not confirmed
H4: ordering	SCL-8 > MCS > PCS	0.68 > 0.44 > 0.38	Confirmed

The CI's are wide (about ±0.2) with n = 63–64, so it matters whether a hypothesis is supported by the whole interval or only by the point estimate (slide 39). H1 is supported by the whole CI (0.52 ; 0.79), which lies entirely above 0.50. H2 is supported only by the point estimate: the CI (0.21 ; 0.62) extends below 0.30 and above 0.50. H3 is not confirmed, but the CI (0.14 ; 0.57) also includes values below 0.30. In H4 the ordering holds, but the difference between the correlations with MCS and PCS is only 0.06; the convention on slide 38 that correlations with related constructs should be at least 0.10 higher than with unrelated constructs would not be met.

- *Evaluate the hypotheses regarding discriminative validity (H5–H6) using Table 5, and check the assumptions underlying the analyses using Table 5 and the graphs.*

Assumptions. Independence: one observation per patient → fulfilled. Normality within groups: the QQ-plots by comorbidity and by gender are approximately straight, with some floor values (0), mainly in the group with comorbidity. Equal standard deviations: 19.8 vs 23.6 (no vs yes comorbidity) and 25.5 vs 24.0 (women vs men) are of similar size, so the unpaired t-test assuming equal SDs is reasonable.

Hypothesis	Expected	Difference (95% CI)	Result
H5: no − yes comorbidity	≥ 15	24.0 (12.9 ; 35.1)	Confirmed
H6: women − men	≥ 10	5.0 (−8.3 ; 18.4)	Not confirmed

For H5 the difference is in the expected direction and larger than expected. For H6 the CI includes both 0 and 10, so the data are inconclusive regarding a gender difference. Expressed as standardised mean differences (slide 38), using the pooled SD (about 22 and 25 points, respectively): d ≈ 24.0/22 ≈ 1.1 (large) for comorbidity and d ≈ 5.0/25 ≈ 0.2 (small) for gender.

- *Discuss the rationale behind each hypothesis. Is the sample size adequate for each of the analyses (see slide 43), and how does it affect your conclusions?*

Rationale. SCL-8 measures emotional distress (anxiety and depression), a construct similar to health anxiety, so a correlation of at least 0.50 is plausible; a very high correlation would, however, question whether the WI-7 measures something distinct from general distress. MCS covers mental health in general, a related but dissimilar construct, and a correlation between 0.30 and 0.50 is expected. PCS covers physical health and was regarded as unrelated; but patients with health anxiety often report bodily symptoms and reduced physical functioning, so a correlation of about 0.4 is not implausible, and H3 may have been too strict. Psychiatric comorbidity is expected to be associated with more severe health anxiety, which gives a good rationale for H5. The rationale for H6 is weaker: there is no strong theoretical basis for a gender difference in health anxiety of 10 points.

Sample size. COSMIN requires a total n of at least 100 for both correlations and group comparisons, with 50–99 rated as acceptable (slide 43). With n = 63–64 all analyses are acceptable, but not very good. The consequence is imprecision: the CI's for the correlations are about ±0.2 wide, and the comparison of 21 men with 43 women gives a CI for the difference that is 27 points wide. Hypotheses may therefore fail to be confirmed because of imprecision rather than lack of validity.

- *Summarize and interpret your findings. How many of the hypotheses were confirmed, and what do you conclude about the construct validity of the WI-7?*

Four of six hypotheses were confirmed (H1, H2, H4 and H5), i.e. 67%. This is below the COSMIN criterion of at least 75% (slide 39), so construct validity of the WI-7 is **not sufficiently supported** by these data. The confirmed hypotheses are, however, the most central ones: the WI-7 correlates most strongly with emotional distress and least with physical health, and it clearly discriminates between patients with and without psychiatric comorbidity. The two non-confirmed hypotheses may reflect unrealistic expectations rather than lack of validity: physical health may not be unrelated to health anxiety (H3), and there is little theoretical basis for a gender difference (H6, which is moreover inconclusive). Note also that H2 is confirmed only by its point estimate. The hypotheses cannot be revised after seeing the data; a new study with a larger sample and well-founded hypotheses specified in advance would be needed.

EXERCISE 1.4

- *Think of your own data and choose a specific variable/scale. Specify hypotheses regarding direction and strength of correlations between your chosen scale and other variables/scales (see slide 38).*

No single answer. Good hypotheses:

- are specified before the data are analysed and are based on theory or earlier studies about the construct;
- state both the direction and the expected magnitude (e.g. $r \geq 0.50$, or a difference of at least 10 points);
- include instruments expected to be strongly related (convergent), weakly related (discriminant) and groups expected to differ (discriminative);
- preferably include hypotheses about the relative magnitude of correlations (e.g. r with A > r with B), which are less sensitive to the exact choice of thresholds.