Measurement of Change

Dr. Armin Zareiyan

Responsiveness

- The ultimate goal of medicine is to cure patients
  - Therefore, assessing whether the disease status of patients has changed over time is often the most important objective of measurements in clinical practice and clinical and health research
- we need measurement instruments with an evaluative purpose or application to detect changes in health status over time. These instruments should be responsive.
- **Responsiveness** is defined by the COSMIN panel as ***'the ability of an instrument to detect change over time in the construct to be measured'***

Responsiveness…

In essence, when assessing responsiveness the hypothesis is tested that **if patients change on the construct of interest, their scores on the measurement instrument assessing this construct change accordingly**

Change Scores & Reliability

- Three Questions related to Change in scores Over Time:

1.Reliabilty of Changes:

- Does a change in score truly represent change , or it merely reflect random fluctuation in measurement?

2. Responsiveness:

Does a person's change in scores on a measure correspond to a commensurate improvement (or deterioration) in the construct?

3. Interpretation of a change score:

What does a change score mean? Is the change large enough to be consider clinically significant?

Measuring Change (Stability)

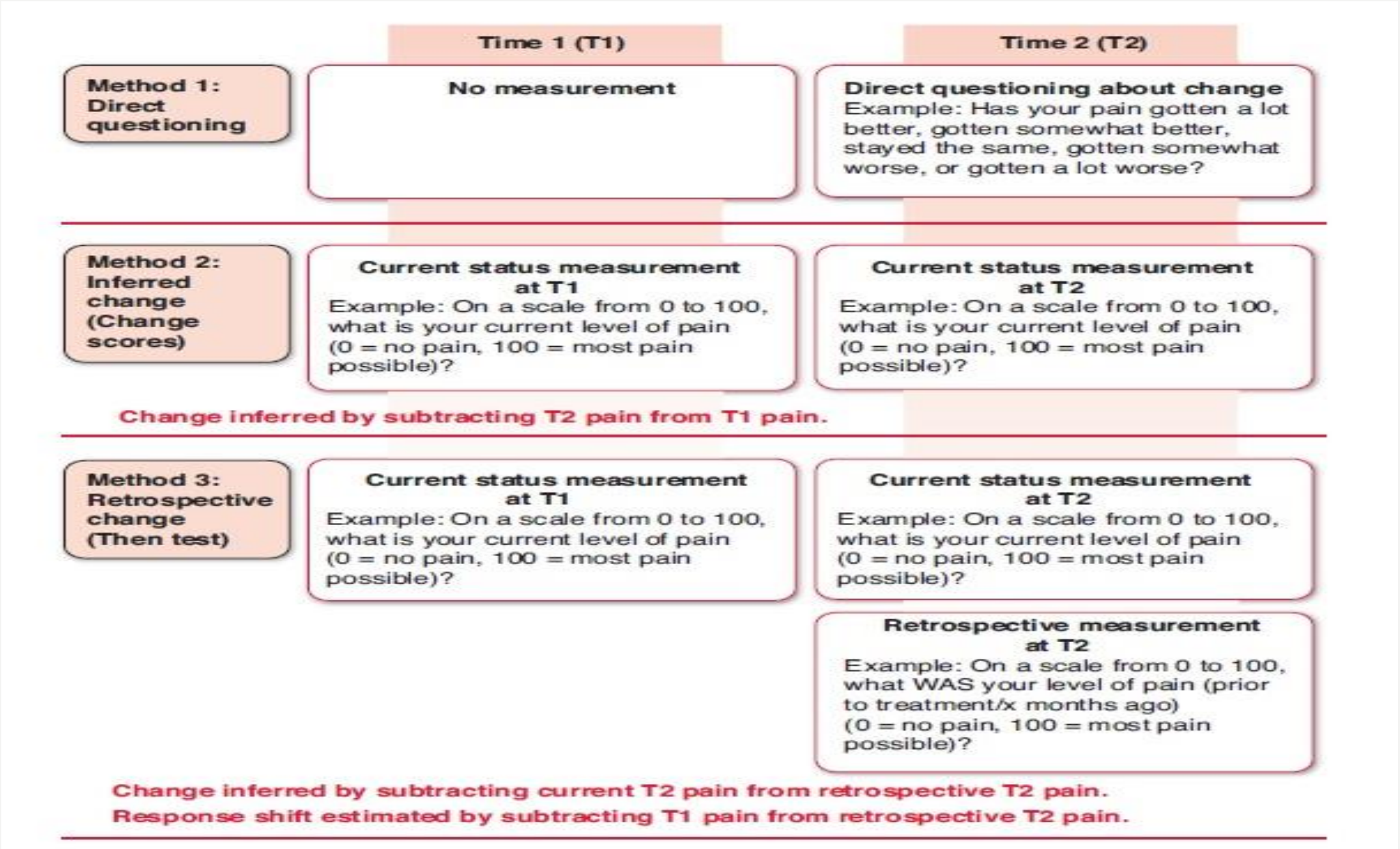
- برای برخی ویژگی های یک ابزار، جهت اندازه گیری تغییرات تنها یک انتخاب وجود دارد:
- Measuring it in two occasions and comparing the values;
- In other word:
- Subtracting one value from the other to calculate the amount of change.

- برای تعیین صحت اندازه گیری تغییرات توسط ابزار سه روش وجود دارد:

الف) روش سئوال مستقیم (Direct Questioning)

ب) روش استنباط تغییرات یا نمره تغییر(Inferred change Or Change scores)

ج) روش ارزیابی تغییرات گذشته نگر یا روش «سپس آزمون» ((Retrospective change (Then test)


Time 1 (T1)
Time 2 (T2)
Method 1: Direct questioning
No measurement
Direct questioning about change
Example: Has your pain gotten a lot better, gotten somewhat better, stayed the same, gotten somewhat worse, or gotten a lot worse?
Method 2: Inferred change (Change scores)
Current status measurement at T1
Example: On a scale from 0 to 100, what is your current level of pain (0 = no pain, 100 = most pain possible)?
Current status measurement at T2
Example: On a scale from 0 to 100, what is your current level of pain (0 = no pain, 100 = most pain possible)?
Change inferred by subtracting T2 pain from T1 pain.
Method 3: Retrospective change (Then test)
Current status measurement at T1
Example: On a scale from 0 to 100, what is your current level of pain (0 = no pain, 100 = most pain possible)?
Current status measurement at T2
Example: On a scale from 0 to 100, what is your current level of pain (0 = no pain, 100 = most pain possible)?
Retrospective measurement at T2
Example: On a scale from 0 to 100, what WAS your level of pain (prior to treatment/x months ago) (0 = no pain, 100 = most pain possible)?
Change inferred by subtracting current T2 pain from retrospective T2 pain.
Response shift estimated by subtracting T1 pain from retrospective T2 pain.

What is Measurement Error

- Differences between observed scores and true scores reflect measurement error
- **Measurement error** is the systematic and random error that occurs when a score obtained through a measuring process differs from a hypothetical "true" score of a latent trait.
- Measurement error is a component within the reliability domain
- Th e concepts of measurement error and reliability are inextricably connected, and yet measurement error parameters yield information that reliability coefficients alone cannot provide
- Unless a reliability coefficient is 1.0 (which is virtually never the case), measurement error is present.

Very Important Tips!!

- Reliability can be reasonably high even when measurement error is not negligible
- Conversely, low measurement error cannot guarantee an acceptable reliability parameter

The explanation for this seeming paradox concerns the

heterogeneity of the sample

Types of Measurement Error Petameters

1. Standard Error of Measurement
2. Bland–Altman Limits of Agreement
3. Coefficient of Variation (CV)

- Note-1: CV is not often encountered in connection with reliability assessments in most health fields, except in laboratory assays.
- Note-2: The CV is the standard deviation divided by the mean, multiplied by 100 to result
- in a percentage.
- Note-3: Standard error of measurement and Bland–Altman limits of agreement, which are relevant for measures of any type— reflective and formative patient-reported outcomes (PROs), observations, performance measures, and biophysiological measures.

Relative Reliability (or Stability)

- The Intraclass Correlation Coefficient (ICC) provides an estimate of relative reliability for consistency of measurement when the population under study is heterogeneous.
- The ***ICC*** reflects a test's ability to differentiate between participants and hence, the position of the individual relative to others in the group.
- **However, the *ICC* does not *provide* information about the accuracy of the scores for an individual.**

Absolute Reliability

- The ICC **does not provide** information about the ***accuracy*** of the scores for an individual.

So

The standard error of measurement (***SEM***), a measure of **absolute reliability**, quantifies reliability of scores within individual participants on different occasions.

Absolute Reliability

- The standard error of measurement (SEM) is an index tell us about the **Precision** of a given measurement;
- The lower the SEM value : the grater the degree of precision.

- مثال:
- فرض کنید داده های مربوط به ابزار افسردگی مرکز مطالعات اپیدمیولوژیک (CES-D) در اختیار ماست.
- نتایج – که در ادامه ملاحظه خواهید کرد– نشان خواهد داد که ممکن است علی رغم وجود پایایی نسبی مطلوب، پایایی مطلق به دلیل واریانس نمونه ها و نمرات ایشان مطلوب نباشد.
- پایایی مطلق را می توان از طریق **SEM** و نیز نمودار **Bland-Altman Limits of Agreement** تعیین نمود.

به جدول داده های ساختگی اسلاید بعد توجه کنید

Table 10.1 Fictitious Data for Test–Retest Reliability Example for CES-D Scores

ID	Week 1	Week 2	Change
1	16	19	3
2	5	8	3
3	8	14	6
4	20	14	−6
5	9	13	4
6	13	19	6
7	17	18	1
8	26	29	3
9	2	5	3
10	6	3	−3
Mean	12.20	14.20	2.00
SD	7.54	7.67	3.80

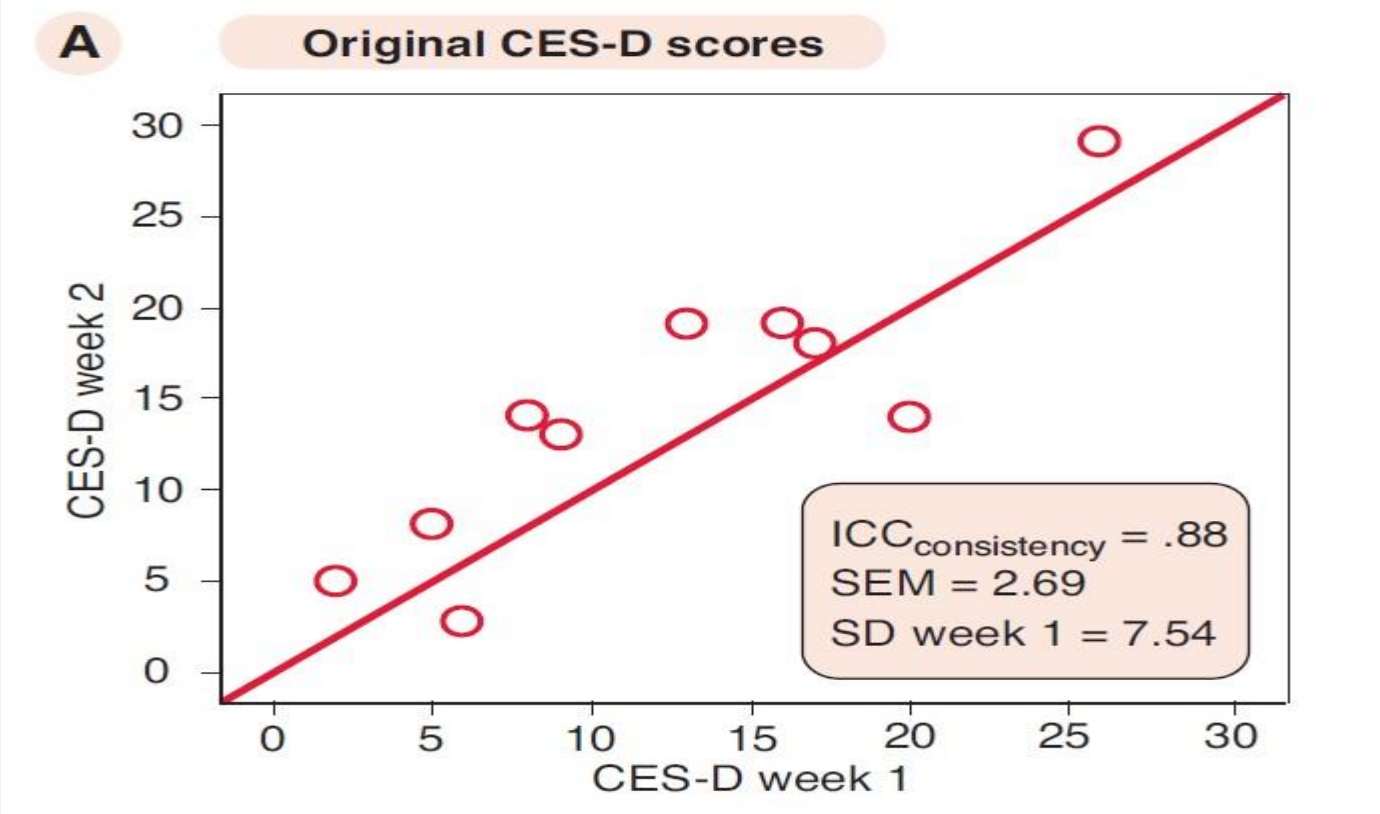
SD, standard deviation.

چرا محاسبه پایایی مطلق مهم است؟

When scores for each person across two Stages in a retest study are close in value, the dots cluster close to the diagonal line and the ***SEM*** tends to be ***low***.

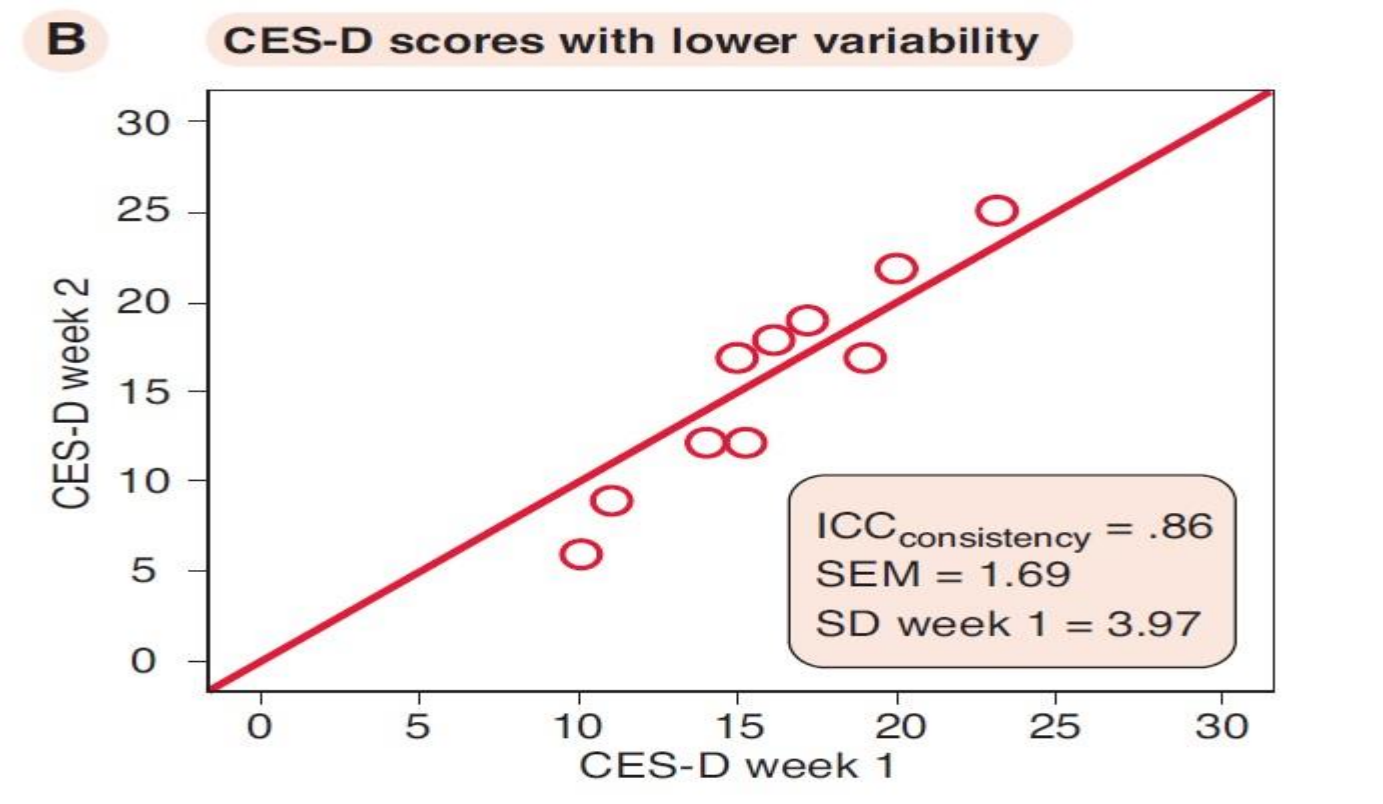
چرا محاسبه پایایی مطلق مهم است؟...

- In Graph A, the score values across the test and retest **are not strongly similar**, and yet the reliability is very respectable**: 0.86**.
- The reason for the high ICC in Graph A is that variability in the sample is high. In Week 1, patients' scores range from a low of 2 (virtually no risk of depression) to 26 (high risk of depression), and the *SD* is 7.54.



چرا محاسبه پایایی مطلق مهم است؟...

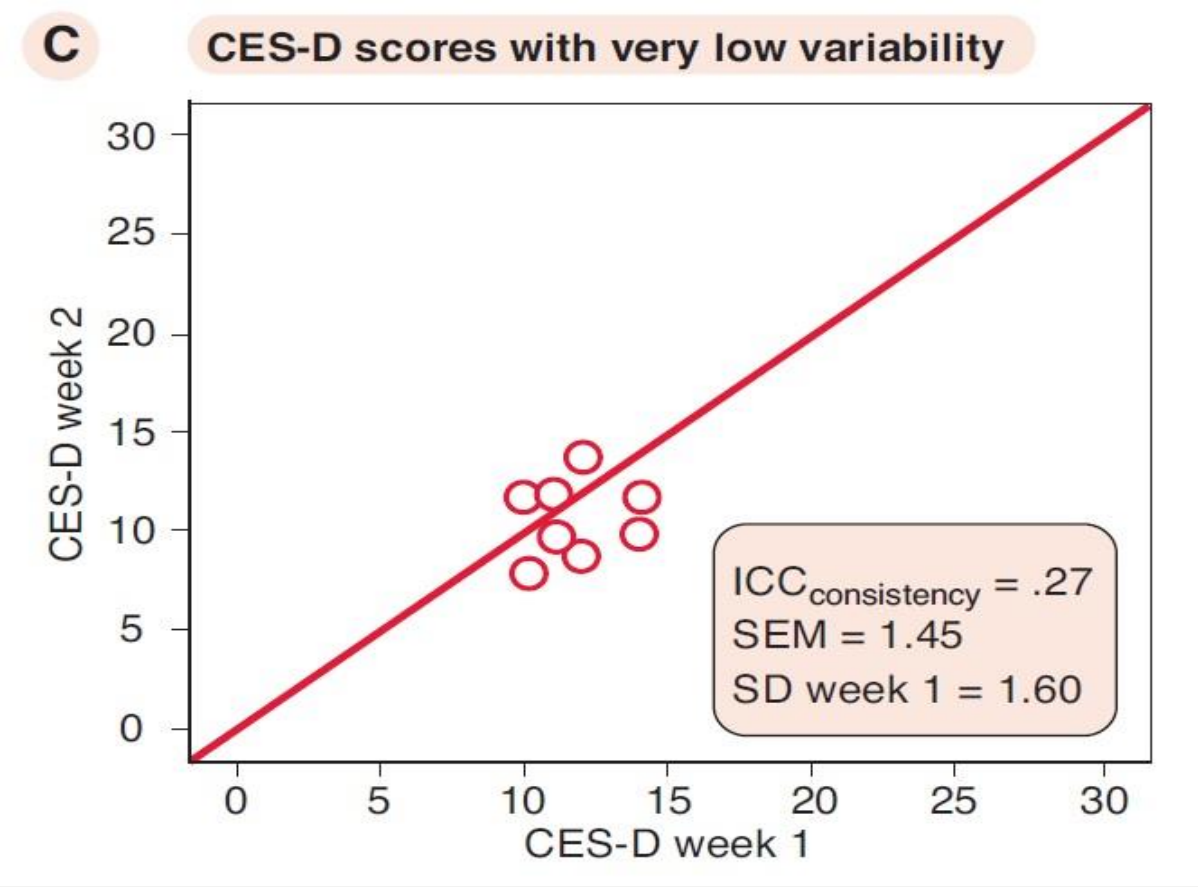
- In Graph B, precision has increased, as illustrated by tight clustering around the diagonal.
- Th e *SEM* is now only 1.69, substantially lower than the *SEM* in Graph A.
- Yet, the ICC value remains the essentially same (.86).
- The reason that the ICC has not improved despite lower measurement error is that sample heterogeneity declined. The range of scores in Graph B for Week 1 is only 10 to 23, and the SD for all scores has dropped to 3.97.



چرا محاسبه پایایی مطلق مهم است؟...

- In Graph C, precision has again increased, with an ***SEM*** of **1.45**.
- However, variability is now severely restricted, with a score range in Week 1 of only 5 points, from 10 to 14, and the *SD* is 1.60. As a result, the ICC is a discouraging .27.

This example illustrates why it is difficult to interpret ICC values without information about sample variability.



لذا:

- ضروری است در کنار پایایی نسبی، پایایی مطلق نیز محاسبه شود.

برای محاسبه پایایی مطلق، بایستی شاخص زیر محاسبه شود:

Standard Error of Measurement (SEM)

خطای استاندارد اندازه گیری

- مهمترین شاخص خطای اندازه گیری SEM است.
- ضریب پایایی به تنهایی یک شاخص نسبی (Relative Index) است که از نمونه ای به نمونه دیگر و در جمعیت های مختلف متفاوت خواهد بود.
- در کنار ضریب پایایی، بایستی SEM هم گزارش شود.
- SEM یک شاخص مطلق (Absolute Index) در اندازه گیری است.
- SEM، نشاندهنده این مهم است که یک نمره ابزار تا چه اندازه دقیق است و محاسبه CI حول آن می تواند به تفسیر دقیق تر کمک کند
- اگر بخواهیم بطور کاملاً مستقیم SEM را محاسبه نماییم، بایستی از جمعیت هدف، نمونه های متعددی، نمونه گیری شود و انحراف معیار نمرات حاصله از تمام نمونه ها محاسبه شود تا SEM حاصل شود (امری غیر ممکن)
- برای حل این مشکل بایستی SEM تخمین زده شود

The Standard Error of Measurement (SEM)

- Reliability coefficient typically range from 0 to 1, with higher values indicating grater reliability.
- Note: ICCs and coefficient Alpha values is RELATIVE index that varies from sample to sample across population.
- BUT: SEMs are in the measurement units of the measure (ABSOLUTE).
- The SEM values are not as affected as reliability coefficient by the sample within which the estimate is computed.
- It is an index of how precise a score is, and can be used to compute confidence intervals (CI) around obtained scores.

Formulas for Standard Error of Measurement

- SEMs are estimated using the same data as those used to compute reliability coefficients.
- For example: data from a test-retest or inter-rater situation can be used in equations for estimating the SEM.
- When K=2 (two rater, or a test then a retest) :

$$SEM = SD_{Difference} \div \sqrt{2}$$

- Where:
- $SD_{Difference} = \sqrt{\frac{\sum(Change\ Differnce - \bar{X}\ of\ Difference)^2}{n-1}}$

مثالی از داده ها جهت انجام پایایی باز – آزمایی ابزار CES-D

ID	Week-1	Week-2	Change(D)	D-(Mean of D)	D-(Mean of D)*2
1	16	19	3	1	1
2	5	8	3	1	1
3	8	14	6	4	16
4	20	14	-6	-8	64
5	9	13	4	2	4
6	13	19	6	4	16
7	17	18	1	-1	1
8	26	29	3	1	1
9	2	5	3	1	1
10	6	3	-3	-5	25
Mean	12.20	14.20	2.00	∑D=0	∑=130
SD	7.54	7.67	**3.80**		

SPSS-Output

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	week2 - week1	2.000	3.801	1.202	-.719	4.719	1.664	9	.130

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	week2 & week1	10	.875	.001

$$SEM = SD_{Difference} \div \sqrt{2}$$

$$SEM = 3.801 \div 1.414 = 2.69$$

Second SEM Formula (Widely used)

$$SEM = SD \times \sqrt{(1 - R)}$$

Where R is the reliability estimate (ICC or Alpha Coefficient)

SD Pooled= (SD1 +SD2) /2

SD Pooled=(7.45+7.67)/2=7.605

SEM= 7.605 √1-0.875 =2.69

Confidence Intervals Around Observed Scores

- تفسیر ساده SEM با استفاده از CI این خواهد بود که اگر فرض کنیم نمره ۱۶ در مقیاس CES-D، نقطه برش جهت تعیین در معرض خطر بودن افسردگی است، لذا فردی با نمره ۱۴، دارای فاصله اطمینان بالاتر از ۱۶ خواهد شد که به معنای "در معرض خطر افسردگی بودن" است.

- Formula, for a 95% CI:

$$95\%\ CI\ around\ X = X \pm 1.96\ (SEM)$$

- X is an individual score
- 1.96 is the z value (95% CI)
- In CES-D example, Patient 6 had a score of 13 at Week 1, so we would estimate that the 95% CI around that score would be:

$$95\%\ CI = 13 \pm 1.96\ (2.69) = 13 \pm 5.3 = 7.7\ to\ 18.27$$

نکته: با افزایش ضریب پایایی نسبی و کاهش SEM، تصمیم گیری بالینی بر اساس نمرات حاصله از یک ابزار، دقیق تر خواهد بود و طبعاً احتمال خطا در تصمیم گیری بالینی کمتر

Standard Error of Estimate (SEE)

- به جز در زمانیکه پایایی کامل باشد، هر نمره حاصله، تقریبی از نمره واقعی است.
- ما هیچگاه نمی توانیم نمره واقعی را بدانیم اما؛ می توانیم نمره واقعی را تخمین بزنیم.
- The formula is:

$$X_{ET} = R\,(X - M) + M$$

- X_{ET}: is an estimated true score
- R: is the reliability coefficient
- X: is an observed score
- M: is the mean of scores

- برای مثال بیمار شماره ۱ با نمره ۱۶

$$X_{ET} = 0.875\,(16 - 12.20) + 12.20 = 15.53$$

Standard Error of Estimate (SEE)

- In every case, unless *R* = 1.0, the estimated true score will be closer than the observed score to the mean of the overall distribution.
- Because this reflects regression to the mean, centering confidence intervals on the estimated true score is often referred to as a ***regression-based approach***.
- In this true-score method, the confidence intervals are not based on the *SEM* but rather on another index called the **Standard Error of Estimate (SEE)**.
- The formula:

$$SEE = SD\sqrt{R(1 - R)}$$

- **با استفاده از نمرات جدول اسلاید های قبلی:**

$$SEE = 7.605\sqrt{.875\ (1 - .875)} = 7.605\sqrt{.109} = 2.52$$

Standard Error of Estimate (SEE)

- فاصله اطمینان برای بیمار با نمره ۱۶ از ابزار CES-D:

95% CI of 15.53 ± (1.96 * 2.52) or 95% CI = 15.53 ± 4.93

- The 95% CI will be narrower using the *regression-based approach* using estimated true scores than those using the traditional approach using obtained scores.
- Both the traditional and regression-based methods are acceptable, the interpretation is different.
- Thus, the two approaches to computing CIs around scores can be used for different purposes.
- The traditional approach can be used to interpret an individual person's obtained score, and the regression-based approach can be used to interpret the scores of all people with a given score.
- Manuals for standardized tests created within a CTT model typically use the regression-based approach.

Smallest Detectable Change

Reliable change for continuous data is often estimated using an index called the **Smallest Detectable Change (SDC)** or **Minimal Detectable Change(MDC)**.

(Smallest detectable difference, minimal detectable difference)

- SDC definition:

  Change in scores that is beyond measurement error

- SDC or MDC Formula:

$$MDC_{95} = SEM \times \sqrt{2} \times 1.96$$

SDC & Limit of Agreement(LOA)

عملیاتی تر و دقیق تر :

- SDC has been defined as a change score that falls outside the **Limits of Agreement (LOA)**
- If a change score falls outside the LOA, there can be greater confidence that change is "real"

$$LOA = d \pm 1.96 \times SD_{difference}$$

- Where d is the mean difference in T1-T2
- $SD_{difference}$: Standard deviation of difference score
- If the standard error of measurement (SEM) has been calculated, an alternative Formula is:

$$LOA = d \pm 1.96 \times (\sqrt{2} \times SEM)$$

مثالی از داده ها جهت انجام پایایی باز– آزمایی ابزار CES-D

ID	Week-1	Week-2	Change(D)	D	D*2
1	16	19	3	3	9
2	5	8	3	3	9
3	8	14	6	6	36
4	20	14	-6	-6	36
5	9	13	4	4	16
6	13	19	6	6	36
7	17	18	1	1	1
8	26	29	3	3	9
9	2	5	3	3	9
10	6	3	-3	-3	9
Mean	12.20	14.20	**2.00**	∑D=20	∑D*2=170
SD	7.54	7.67	3.80		

$$SEM = SD\sqrt{(1-R)}$$

Where R is the reliability estimate (ICC or Alpha Coefficient)

SD Pooled= ($SD1$ +$SD2$) /2

SD Pooled=(7.45+7.67)/2=7.605

$$SEM = 7.605\ \sqrt{1-0.875} = 2.69$$

$$LOA = d \pm 1.96 \times (\sqrt{2} \times SEM)$$

- Upper LOA = 2 + 1.96 (1.41 × 2.69) = 9.434
- Lower LOA = 2 + (-1.96) ×(1.41 × 2.69) = -5.434

- **اگر هر نمره تغییری خارج از این محدوده باشد، با احتمال ۹۵٪ ، تغییر ایجاد شده کاملاً دقیق و واقعی است.**
- **این روش سختگیرانه است**

Reliable Change Index(RCI)

- Jacobson et al. (1984) argue that a change score on outcomes must pass the test of being " real" –that is, a CHANGE BEYOND measurement error.
- The RCI is calculated by dividing the difference between an observed individual posttest score (X POST) and pretest score (X PRE) by a factor representing the degree of measurement error:

$$RCI = \frac{X_{POST} - X_{PRE}}{\sqrt{2 \times SEM^2}}$$

Where X is the score of an individual and SEM is the standard error of measurement

$$SEM = SD \times \sqrt{(1 - R)}$$

Note: R is the reliability of the instrument & SD (in context of the RCI formula) is standard deviation of a "control group, normal population OR pretreatment experimental group (Jacobson& Traux,1991,p.14)

Table 17.1 Fictitious Scores for 10 Patients on the CES-D, Illustrating the Reliable Change Index

Patient	Pretest		12-Week Intervention	Posttest
	T1 Week 1	T2 Week 2		T3 Week 14[a]
1	18	17		14
2	16	19		**13**
3	27	24		20
4	20	21		**15**
5	19	22		**14**
6	19	19		16
7	17	18		**11**
8	26	29		**13**
9	22	20		16
10	26	24		**17**
Mean	21.0	21.3		14.9
SD	4.03	3.59		2.51

Test–Retest

Assessment of Reliable Change

Mean (*SD*) change for retest (T1–T2)	=	$-.30$ (2.263), $t = 0.42$, $p = .69$
Mean (*SD*) change for trial (T2–T3)	=	6.4 (3.806), $t = 5.32$, $p < .001$
Test–retest reliability (T1–T2)	=	.824 ($ICC_{Consistency}$)
$SEM = 3.592 \times \sqrt{(1 - .824)}$	=	1.507
Denominator for RCI	=	$\sqrt{2 \times 1.507^2} = 2.131$
RCI score to exceed at 95% CI	=	$1.96 \times 2.131 = 4.18$

[a]Bolded values indicate that the pretest–posttest change was reliable (60% of the sample).

محاسبه RCI بر اساس جدول اسلاید قبل

- SEM = 3.59 × √ 1- 0.824) = 1.507
- مخرج کسر:
- $\sqrt{2 \times 1.507^2} = 2.131$
- We can use this value to assess whether any particular patient had a reliable improvement in CES-D score following the intervention.
- For example for the first two patients:

Patient 1: (14-17) ÷ 2.131 = -1.0408

Patient 2: (13-19) ÷ 2.131 = -2.816

تفسیر:

- These RCI value compared to the *z* value for desired criterion, which is usually ±1.96.
- In this case, the reduced score on the CES-D for patient 1 is not reliable, but the change for patient 2 is reliable according to this RCI

راه ساده تری جهت تفسیر

- A conclusion about a reliable change for an entire sample is:

  To calculate the change score value that must be exceeded:

- In this example, the cutoff value for 95% confidence is

$$\pm 1.96 \times 2.131 = \pm 4.18$$

لذا تفاضل بین پره تست و پست تست را محاسبه می کنیم اگر فراتر از عدد ±4.18 بود، پایا محسوب خواهد شد

سپس درصد افرادی که پایا هستند تعیین می شود که بایستی بالاتر از ۵۰٪ باشد.

Table 17.1 Fictitious Scores for 10 Patients on the CES-D, Illustrating the Reliable Change Index

Patient	Pretest T1 Week 1	Pretest T2 Week 2	12-Week Intervention	Posttest T3 Week 14[a]
1	18	17		14
2	16	19		**13**
3	27	24		20
4	20	21		**15**
5	19	22		**14**
6	19	19		16
7	17	18		**11**
8	26	29		**13**
9	22	20		16
10	26	24		**17**
Mean	21.0	21.3		14.9
SD	4.03	3.59		2.51

Test–Retest | Assessment of Reliable Change

Mean (*SD*) change for retest (T1–T2)	=	$-.30\ (2.263), t = 0.42, p = .69$
Mean (*SD*) change for trial (T2–T3)	=	$6.4\ (3.806), t = 5.32, p < .001$
Test–retest reliability (T1–T2)	=	$.824\ (ICC_{Consistency})$
$SEM = 3.592 \times \sqrt{(1 - .824)}$	=	1.507
Denominator for RCI	=	$\sqrt{2 \times 1.507^2} = 2.131$
RCI score to exceed at 95% CI	=	$1.96 \times 2.131 = 4.18$

[a]Bolded values indicate that the pretest–posttest change was reliable (60% of the sample).

نتیجه گیری

- When a measurement is made with a multi-item scale, it may be advantageous to estimate both internal consistency and test–retest reliability and to compare the resulting *SEM*s.

- When the reliability of change scores is of interest, then it is likely to be important to distinguish short-term fluctuations from true change, and so test–retest reliability should be estimated.


موفق باشید