Journal of Clinical Epidemiology 140 (2021) 159–162

Journal of Clinical Epidemiology

KEY CONCEPTS IN CLINICAL EPIDEMIOLOGY

# Key concepts in clinical epidemiology: Responsiveness, the longitudinal aspect of validity

Lidwine Mokkink*, Caroline Terwee, Henrica de Vet

*Amsterdam UMC, Vrije Universiteit Amsterdam, Departement of Epidemiology and Data Science, Amsterdam Public Health research institute, De Boelelaan 1117, Amsterdam, The Netherlands*

Received 12 May 2021; Received in revised form 26 May 2021; Accepted 3 June 2021; Available online 8 June 2021

**Abstract**

**Objectives:** Responsiveness is one of nine measurement properties that reflect the quality of outcome measurement instruments.

**Methods:** In this article, we explain that responsiveness is considered longitudinal validity, which refers to the degree to which an instrument is able to measure change in the construct to be measured.

**Results:** Responsiveness should be assessed in a longitudinal design, where hypotheses are tested about (1) the expected direction and magnitude of correlations between change scores on the instrument of interest and change scores of other instruments; (2) expected differences in change scores between different subgroups (i.e. known groups); or (3) the magnitude of the change in score that is expected on the construct of interest after a treatment with known efficacy.

**Conclusion:** Responsiveness cannot be proven, though, it is an ongoing process of testing hypotheses. 

*Keywords:* Responsiveness; Measurement instruments; Longitudinal validity; COSMIN; hypotheses testing; outcomes

## 1. Background

Scores and changes in scores of measurement instruments are used in research and clinical care, for example, to draw conclusions about the effectiveness of treatment and a patient's progress. As a clinician or researcher we must be able to rely on those scores. Therefore the measurement instruments that we use should be of good quality. Responsiveness is one of the nine measurement properties that reflect the quality of outcome measurement instruments. Other measurement properties are content validity, structural validity, internal consistency, cross-cultural validity, reliability, measurement error, hypotheses testing for construct validity, and criterion validity [1]. Responsiveness is especially important when patients are followed over time. It refers to the ability of a measurement instrument to detect change over time in the construct to be measured [1]. Responsiveness is often considered an aspect of validity [2],[3]. In this paper we will explain how responsiveness relates to validity and how it should be assessed.

## 2. Validity and responsiveness

When we use a score obtained at one moment in time, we need to know whether this score is valid, meaning that the scores obtained with an instrument reflect the construct the instrument purports to measure [1]. Validity refers to the degree to which the scores of an instrument are consistent with hypotheses about the construct to be measured. When we measure change over time, we need to know whether the change in score within a person—for example, after an intervention—is valid. This means that the instrument should measure change in the purported construct, but also that it should measure the right amount of change, that is, it should not under- or overestimate the real change in the construct that has occurred [4]. This is called responsiveness, and can be considered as "longitudinal validity."

Conflict of Interest: The royalties for the book to which we refer (De Vet HC, Terwee CB, Mokkink L, Knol DL: Measurement in medicine. Cambridge: Cambridge University Press; 2011) are transferred to our university department.

* Corresponding author.
*E-mail address:* w.mokkink@amsterdamumc.nl (L. Mokkink).

https://doi.org/10.1016/j.jclinepi.2021.06.002

## 3. Study design and approaches for evaluating responsiveness

Responsiveness refers to the longitudinal context in which scores are used. Therefore, to assess responsiveness the design should be longitudinal, with at least two measurements. The time points should be chosen in such a way that it can be expected that at least part of the study population will change in the construct of interest between the two time points. A priori hypotheses should be formulated about expected change scores. Essential is that we have a clear definition of the construct an instrument purports to measure. This definition helps us to consider validity and responsiveness and formulate hypotheses about the direction and magnitude of expected change scores. Two approaches for assessing responsiveness, the criterion and construct approach, are generally distinguished.

In the rare situation were a gold standard is available to measure to construct of interest, we can use the hypothesis that we expect a very strong correlation (>0.7) between the change score on the instrument of interest and the change score on the gold standard. As the change score is based on two measurements, both with some degree of measurement error included, a correlation of 0.7 is already quite challenging. This is called the criterion approach to assess responsiveness, in analogy to criterion validity.

When there is no gold standard available, a construct approach can be used for testing responsiveness, in analogy to construct validity. With a construct approach many different hypotheses can be tested. We distinguish between three types of hypotheses: (1) hypotheses about the expected direction and magnitude of correlations between change scores on the instrument of interest and change scores of instruments that measure similar constructs (i.e., strong relationships, e.g., above 0.5) or instrument that measure unrelated constructs (i.e., weaker relationships, e.g., below 0.3) [5]; (2) hypotheses about expected differences in change scores between different subgroups (i.e., known groups). For example, we expect that people who receive an intervention of known efficacy change more after 2 weeks than people on the waiting list for that intervention, and we specify the expected difference in the change scores between the two groups; or (3) hypotheses about the magnitude of the change in score that we can expect after undergoing a treatment with known efficacy on the construct of interest. For example, if we *know* from previous research that a specific treatment will have a moderate effect on a specific outcome, we expect a medium effect size (between 0.3 and 0.5) using an instrument that purports to measure the specific outcome in patients before and after having received this treatment.

In some situations, for example, when we are evaluating a newly developed instrument, we don't have clear expectations about how much change is expected. Instead of hypotheses about the absolute magnitude of the relations we can formulate hypotheses about the relative magnitude of change that we expect. For example, we can expect the instrument of interest to correlate at least 0.1 higher with an instrument measuring the same construct compared to an instrument measuring a slightly different constructs.

## 4. An example

We want to use the Animated Activity Questionnaire (AAQ; see Table 1) which measures how a person performs basic activities in daily life [6] to evaluate an intervention for patients with hip or knee osteoarthritis. The AAQ aims to measure activity limitations. Activity limitations are defined as "any difficulties an individual may have in executing daily activities" [7]. Before we can use the AAQ, we want to assess the responsiveness of the AAQ.

To assess the responsiveness of the AAQ, a longitudinal study [8] was performed in 94 patients with hip or knee osteoarthritis who completed the AAQ at admission and 6-month after treatment (see Table 1). Because we know already that the treatment will increase the ability to perform daily activities, it was expected that at least part of the included patients would change in their level of activity limitations as measured by the AAQ. In addition to the AAQ, a patient-reported outcome measurement instrument (PROM) and three performance-based tests were administered (see Table 1). A Global Rating Scale (GRS) of change was administered at follow-up. Several hypotheses were formulated about expected changes in the AAQ in relation to the Global Rating Scale of change, the PROM and the performance-based tests (see Table 1, and Peter et al. [8] for a rationale). All these hypotheses concerned expected correlations of the change in AAQ scores with changes in other instruments. An example of a known group hypothesis would be that the change on AAQ after surgery would be larger (at least five points) than after physiotherapy, and a hypothesis that the change in AAQ after surgery would be a large effect size (ES > 0.8) would be an example of an expectation of the magnitude of the change score.

Based on the study results [8] we should decide whether we accept that the responsiveness is sufficient. As a rule of thumb, we use the criterion that 75% of the results should be in accordance with the hypotheses [5]. It was concluded that the AAQ was sufficiently responsive at 6-month follow-up [8].

## 5. Pointers

Responsiveness is considered longitudinal validity, and refers to the degree to which an instrument is able to measure change in the construct to be measured. As the content and the number of hypotheses are inexhaustible and arbitrary, and the quality of comparator instruments (including the gold standard) is never perfect, responsiveness, just like

**Table 1.** Design and hypotheses of a study on assessing responsiveness of the Animated Activity Questionnaire (AAQ)

| | |
|---|---|
| *AAQ*: The developers operationalized the construct by a series of videos of different levels of difficulty one could have when performing the activity. The patient watches a series of videos per activity in which an animated person performs the activity in different ways, each video showing a different level of difficulty of performing the activity. Patients are asked to select the video that best represents their way of performing the activity [6]. In the AAQ, 17 basic activities were included, such as walking, climbing stairs, rising, and sitting down. The total score ranges between 0 and 100, with higher scores corresponding to higher levels of functioning. | *Results* |
| *Patients*: 94 patients with hip or knee osteoarthritis and who were scheduled for total joint arthroplasty or participated in a strengthening exercise program | |
| *Measurements at baseline and 6 months follow-up*:<br>The AAQ [6]; the Activities of Daily Living subscale of the Hip disability or Knee injury and Osteoarthritis Outcome Score (H/KOOS) [9]; and three performance-based tests, i.e., the 30 s chair-stand test [10],[11], the timed up-and-go test [10],[12] and nine-step stair climbing test [10],[13]; Global Rating Scale (GRS) of change only administered at follow-up, i.e., a single question asking how much limitation in ADL tasks were experienced in the last week owing to their hip or knee problems, as compared with baseline (worse, the same, or improved) [14]. | |
| *Hypothesis 1*: The change in the total score of the AAQ will have a positive correlation of at least 0.5 with the GRS of change at 6 months. | 0.59, in accordance [8] |
| *Hypothesis 2*: The change in the total score of the AAQ will have a positive correlation of at least 0.5 with the change in the H/KOOS ADL subscale [9] at 6 months. | 0.73, in accordance [8] |
| *Hypothesis 3*: The change in the total score of the AAQ will have a positive correlation of at least 0.5 with the change in the average score of three performance-based tests, i.e., the 30 s chair-stand test [10],[11], timed up-and-go test [10],[12]; and nine-step stair climbing test [10],[13] at 6 months. | 0.46, not in accordance [8] |
| *Hypothesis 4*: The change in the total score of the AAQ (0–6 months) will have a correlation at least 0.10 higher with the changes in scores of the H/KOOS ADL subscale [9] than with the change on the average score of the performance-based tests. | [8]<br>H/KOOS ADL 0.27 higher, in accordance |

validity, cannot be proven, though, it is an ongoing process of testing hypotheses.

## Further reading

De Vet HC, Terwee CB, Mokkink L, Knol DL. Measurement in medicine. Cambridge: Cambridge University Press; 2011.

Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010, 63(7):737–745.

Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Qual Life Res 2003, 12(4):349–362.

## References

[1] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 2010;63(7):737–45.

[2] Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Qual Life Res 2003;12(4):349–62.

[3] de Vet HC, Terwee CB, Mokkink L, Knol DL. Measurement in medicine. Cambridge: Cambridge University Press; 2011.

[4] COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs) – user manual. Accessed 19 April 2021. Available at https://cosmin.nl.

[5] Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res 2018;27(5):1147–57.

[6] Peter WF, Loos M, de Vet HC, Boers M, Harlaar J, Roorda LD, et al. Development and preliminary testing of a computerized Animated Activity Questionnaire in patients with hip and knee osteoarthritis. Arthritis Care Res (Hoboken) 2015;67(1):32–9.

[7] Organization WH ICF: international classification of functioning, disability and health. Geneva: World Health Organization; 2001.

[8] Peter WF, Poolman RW, Scholtes VAB, de Vet HCW, Terwee CB. Responsiveness and interpretability of the Animated Activity Questionnaire for assessing activity limitations of patients with hip or knee osteoarthritis. Musculoskeletal Care 2019;17(4):327–34.

[9] de Groot IB, Favejee MM, Reijman M, Verhaar JA, Terwee CB. The Dutch version of the Knee Injury and Osteoarthritis Outcome Score: a validation study. Health Qual Life Outcomes 2008;6:16.

[10] Dobson F, Hinman RS, Hall M, Terwee CB, Roos EM, Bennell KL. Measurement properties of performance-based measures to assess physical function in hip and knee osteoarthritis: a systematic review. Osteoarthritis Cartilage 2012;20(12):1548–62.

[11] Gill S, McBurney H. Reliability of performance-based measures in people awaiting joint replacement surgery of the hip or knee. Physiother Res Int 2008;13(3):141–52.

[12] Alghadir A, Anwer S, Brismee JM. The reliability and minimal detectable change of Timed Up and Go test in individuals with grade 1-3 knee osteoarthritis. BMC Musculoskelet Disord 2015;16:174.

[13] Kennedy DM, Stratford PW, Wessel J, Gollish JD, Penney D. Assessing stability and change of four performance measures: a longitudinal study evaluating outcome following total hip and knee arthroplasty. BMC Musculoskelet Disord 2005;6:3.

[14] Kamper SJ, Maher CG, Mackay G. Global rating of change scales: a review of strengths and weaknesses and considerations for design. J Man Manip Ther 2009;17(3):163–70.