Expert Panels: Can They Be Trusted?

To the Editor:

Van Meer et al.[1] recently compared the Knee Injury and Osteoarthritis Outcome Score (KOOS)[2] with the International Knee Documentation Committee (IKDC) Subjective Knee Form[3] to investigate which of the instruments is most useful after anterior cruciate ligament reconstruction (ACLR). Following the COSMIN (COnsensus-based Standards for the selection of health status Measurement INstruments) checklist, an expert panel reached consensus on a priori hypotheses regarding responsiveness and construct validity. This letter raises concerns about the use of expert panels in general and gives examples of bias introduced by the expert panel in this particular study.

van Meer et al. state: "Without specific hypotheses there is a risk of bias, because it is tempting to formulate explanations for the low and high correlation coefficients retrospectively instead of concluding that the questionnaire may not be valid. Conversely, the choice of magnitude of the hypotheses is arbitrary." If the expert panel had considered available literature when establishing a priori hypotheses, their results would have been different. Using an expert panel is thus not necessarily associated with less bias compared with relying on available literature. A minimum requirement for an expert panel must be awareness of existing literature when formulating a priori hypotheses.

Especially evident for KOOS, correlations *hypothesized* are not in line with existing literature. In contrast, the majority of the correlations actually *found* for both KOOS and IKDC are in line with existing literature. One example relates to effect size (ES). For the IKDC, van Meer et al. hypothesized an ES >0.8, found an ES of 1.36, and the hypothesis was confirmed, well in line with existing literature (ES, 0.76 to 2.11).[4] For KOOS, however, it seems that the predefined hypotheses were arbitrary. For the subscales Pain, activities of daily living (ADL), and quality of life, ESs less than 0.4 were hypothesized. The ESs found (0.58 to 1.51) are, however, in line with existing literature (ACLR, 0.84 to 1.65).[2,4]

Expert clinicians and patients evaluated content validity. This approach may introduce further bias because the perspectives of clinicians and patients differ and current standards consider the patient to be the expert. The importance of focusing on the patient's perspective is underscored by the findings in the van Meer study. Although the authors concluded that the KOOS subscales Pain and ADL were irrelevant, they reported ES > 0.50. A moderate improvement in Pain and ADL at 1 year after ACLR is not irrelevant.

Van Meer et al. state: "A disadvantage of the IKDC subjective is the use of one total score, which means it is impossible to see in which domain the patients have improved. According to our results, the IKDC subjective is more responsive to changes over time than is the KOOS." This statement is incorrect. The ES for the different KOOS subscales ranged from 0.55 to 1.51, reflecting the difference in improvement for the 5 domains at 1 year after ACLR, whereas the ES for the IKDC subjective was 1.36.

Finally, in Table 2 the correlation found between the IKDC and Lysholm score (0.47) is wrongly interpreted as being higher than the hypothesized correlation of 0.6. Overall, 75% of the hypothesized and the found correlations are not given in the tables and can thus neither be compared with the previous literature nor checked for correctness.

Ewa M. Roos, PT, Ph.D.
Carsten B. Juhl, PT, Ph.D.
L. Stefan Lohmander, M.D., Ph.D.
Odense, Denmark

Note: E.M. Roos is the developer of KOOS, HOOS, HAGOS, FAOS, RAOS. Instruments have been developed in an academic context and are freely available. No license is required for their academic or commercial use. No funding from commercial parties or non-for-profit organizations has been received.

http://dx.doi.org/10.1016/j.arthro.2013.05.016

References

1. van Meer BL, Meuffels DE, Vissers MM, et al. Knee Injury and Osteoarthritis Outcome Score or International Knee Documentation Committee Subjective Knee Form: Which questionnaire is most useful to monitor patients with an anterior cruciate ligament rupture in the short term? *Arthroscopy* 2013;29:701-715.
2. Roos EM, Roos HP, Lohmander LS, et al. Knee Injury and Osteoarthritis Outcome Score (KOOS)—Development of a self-administered outcome measure. *J Orthop Sports Phys Ther* 1998;28:88-96.

3. Irrgang JJ, Anderson AF, Boland AL, et al. Development and validation of the International Knee Documentation Committee Subjective Knee Form. *Am J Sports Med* 2001;29:600-613.
4. Collins NJ, Misra D, Felson DT, et al. Measures of knee function: International Knee Documentation Committee (IKDC) Subjective Knee Evaluation Form, Knee Injury and Osteoarthritis Outcome Score (KOOS), Knee Injury and Osteoarthritis Outcome Score Physical Function Short Form (KOOS-PS), Knee Outcome Survey Activities of Daily Living Scale (KOS-ADL), Lysholm Knee Scoring Scale, Oxford Knee Score (OKS), Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), Activity Rating Scale (ARS), and Tegner Activity Score (TAS). *Arthritis Care Res (Hoboken)* 2011;63(suppl 11): S208-S228.

Author's Reply

We thank Roos, Juhl, and Lohmander for their interest in our article and their remarks. They raise concern about the use of an expert panel for determining a priori hypotheses for the assessment of construct validity and responsiveness. They argue that the expert panel should have considered existing literature when formulating a priori hypotheses. In contrast, we do not think that hypotheses should be based on the results of previous validation studies. Classic psychometric works state that hypotheses should be based on theoretical considerations about expected relations among constructs. Cronbach and Meehl[1] proposed to start from theories about the construct and then formulate hypotheses to test whether the scores of the instrument are consistent with a theoretical model of the construct. Our hypotheses were based on theoretical models such as the International Classification of Functioning and the model of Wilson and Cleary.[2] Based on these theories, we hypothesized that scales that intend to measure similar constructs (e.g., visual analogue scale Pain and KOOS Pain) should correlate at least 0.60, scales that intend to measure closely related constructs (e.g., Short Form 36 physical functioning and KOOS sport/recreation) should correlate between 0.40 and 0.60, and scales that intend to measure different constructs (e.g., Short Form 36 bodily pain and KOOS activities of daily living) should correlate $\leq$0.40.[1]

The expected correlations may not always be similar to those found previously in the literature. That is because when setting up a validation study we are not sure a priori that the instrument is valid (otherwise, the study is unnecessary).

Regarding content validity, we agree with Roos et al. that the current standard considers the patient to be the expert. However, according to the COSMIN guidelines for good content validity,[3] the items of a questionnaire should be relevant for the construct to be measured, for the population being assessed, and for the purpose for which the instrument is being used. Although we consider the patient to be the expert with regard to relevance for the population, we consider the opinion of clinicians and methodologists of additional value for assessing relevance for the construct and purpose.[4]

Roos et al. raised their concerns about the conclusion of responsiveness of the KOOS and IKDC. We agree with Roos et al. that the effect sizes for the KOOS reflect the difference in improvement for the 5 domains at follow-up, whereas for the IKDC it is impossible to see in which domain the patients have improved. This is an important advantage of the KOOS. However, our conclusions on responsiveness were based on all hypotheses tested, not only on the effect size.

Finally, we would like to thank Roos et al. for their attention regarding Table 2. Unfortunately, the correlation between the IKDC subjective and the Lysholm score (0.47) was interpreted as in agreement with the predefined hypothesis of $\geq$0.6. This incorrect interpretation means that the confirmed hypotheses A should be 67% instead of 83%. However, the confirmed hypotheses A + B are still higher than 75% (17 of 21 = 81% instead of 86%).

Belle L. van Meer, M.D.
Duncan E. Meuffels, M.D., Ph.D.
Sita M. Bierma-Zeinstra, Ph.D.
Max Reijman, Ph.D.
Rotterdam, The Netherlands

Caroline B. Terwee, Ph.D.
Amsterdam, The Netherlands

http://dx.doi.org/10.1016/j.arthro.2013.05.018

References

1. Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull* 1955;52:281-302.
2. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. *JAMA* 1995;273:59-65.
3. Mokkink LB, Terwee CB, Patrick DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: An international Delphi study. *Qual Life Res* 2010;19:539-549.
4. De Vet HCW, Terwee CB, Mokkink LB, Knol DL. *Measurement in medicine. A practical guide*. Cambridge: Cambridge University Press, 2011.