



Published in final edited form as:

Spine J. 2010 April ; 10(4): 330–332. doi:10.1016/j.spinee.2010.01.019.

Testing Minimal Clinically Important Difference: Additional Comments and Scientific Reality Testing

Robert J. Gatchel, Ph.D., ABPP* and Tom G. Mayer, M.D.

Over the years, a great number of clinical measures were developed and used to evaluate treatment outcomes. The developers of such self-report instruments focused primarily on the psychometric properties of reliability and validity. Unfortunately, the key issue of patient responsiveness (i.e., a measure's ability to detect change over time) was not seriously investigated¹. Recently, though, the topic of clinical relevance has begun to receive greater attention because both clinicians and researchers alike are seeking methods to demonstrate a treatment's efficacy by comparing post-treatment scores to pre-treatment baseline. Some, like Glassman and Carreon, seek to define clinical importance/significance, and not merely the statistical significance. A driving force for this has been the growing demand by the federal government for more evidence-based approaches to document the true effectiveness of high-cost procedures, such as spine fusions and/or implantable devices, etc. One such approach has been the concept of a *minimal clinically important difference* (MCID). However, in the effort to develop a useful MCID measure, there have been some major methodological and psychometric problems overlooked by various investigators. Indeed, as recently highlighted by Spratt²: “..Over the last 30 years, an array of approaches for assessing MCID has evolved with little concern on which approach applies in any given situation” (p. 1722). One major purpose of our article published in this issue of *TSJ*³ was to point out and discuss these problems. The most significant areas were the following:

- Two self-report measures obtained from the same individual are correlated (from a statistical perspective, this is defined as *correlated error terms*). When measuring the same construct (e.g., treatment effectiveness), it is psychometrically essential that the two instruments (measures) are independent and that the external criterion or anchor be an objective (non self-report) measure. This is a critical psychometric principle that needs to be strictly adhered to.
- A 30% improvement in scores, relative to baseline, was merely agreed upon by an “expert panel consensus” (i.e., the IMMPACT recommendation 4). It was not based on any comprehensive empirical study that documented its validity or utility.

In their Commentary, entitled “Thresholds for HRQOL Outcome Measures: Reality Testing,”⁵ Glassman and Carreon attempted to address the issue we originally raised about the necessity of using an objectively measured external criterion as the only appropriate anchor to use for determining an MCID. As they noted: “While this in theory is reasonable argument, the difficulty is that an ideal objective external criterion does not exist. In fact, if there was a reliable objective external criterion that defines success or value for patients with lumbar spinal disorders, there would be much less need for patient-based outcome measures in the first place”.

*Corresponding Author: College of Science, University of Texas at Arlington, Arlington, TX, gatchel@uta.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

They then take us to task for proposing return-to-work and work retention as the appropriate external criteria for defining an MCID threshold. However, such outcomes, as well as other socioeconomic outcomes that can be objectively evaluated (e.g., additional healthcare visits, additional surgeries, medication use change, etc.), are important for workers' compensation populations. Glassman and Carreon are correct in speculating that such outcomes may not be applicable to other sectors of the general population (although a disproportionately larger number of lumbar fusions and ADRs are performed in this population). We merely used this particular anchor criterion as an example, and not as the only objective external criterion that can be used. In fact, specific external anchors will have to be developed for specific groups. Just as we are moving away from the "homogeneity of pain patients myth," and towards attempts to match treatment to specific assessment outcomes of patients (e.g., 6,7), we will also have to match the type of external criterion used to the type of patient group being evaluated. Patients with the same medical diagnosis or set of symptoms (e.g., chronic low back pain) have traditionally been "lumped together" and then treated in the same manner, as though "one size fits all." However, it has been shown that pain patients with the same diagnosis have different responses to the same treatment. This will be similar for different subgroups of spinal lumbar patients (e.g., workers' compensation, private-pay insurance, older patients, etc.) when selecting the most appropriate objective anchor for MCID use. Certainly, regardless of whether patients have WC insurance or group health, if they were working before a treatment procedure and cannot work afterwards, or if the procedure fails to restore lost productivity the patients were used to, then this is a relevant outcome. Other outcomes that relate to persistent disability include: use of opioids or other habituating medications; additional surgeries to correct complications (or failure of expectations) of an initial treatment; persistent health-care utilization and doctor visits, as well as interventional procedures; etc. Moreover, other objective external anchors that can be used to measure *functional restoration* include gains in mobility, improvement in gait, increases in trunk strength and lifting capacity.

Doing good science is not an easy task. Our scientific tradition does not allow us to settle for imperfection in the area of MCID just because, as stated by Glassman and Carreon, there is a "...lack of an appropriate external criterion..." Just because they have difficulty defining an objective external anchor does not mean that psychometrically invalid correlated error terms (by comparing two subjective self-report measures) becomes scientifically valid. We need to conduct systematic research for such appropriate criteria to use with specific disorders and for specific populations. It should also be kept in mind that, in science, we routinely use inferential constructs (such as pain, stress, satisfaction, general health, etc.) in order to help us predict certain behaviors/phenomena. Rather than being an actual entity or thing, "patient satisfaction of general health," for example, is a construct which is inferred in order to predict some form of behavior (e.g., returning for additional treatment; subsequently returning to pre-injury activity levels; etc.). Using such a construct necessitates the need to go through the process of construct validation. In the case of using a single question from a much larger instrument for rating change (the Health Transition Item) of the SF-36, in which patients are asked to describe their response to a procedure, drug or treatment regimen on a scale from "much better" to "much worse," relative to their prior recollection of health (as suggested by Glassman and Carreon), is not an appropriate MCID anchor. Moreover, another puzzling aspect of their approach was the use of only two points of the five points on this SF-36 Item Question (about the same, and much better) in determining their MCID threshold for change due to treatment. In addition, Schwartz and Finkelstein⁸ have highlighted the fact that patient "response shift phenomena" can likely significantly affect the measurement properties of a standard self-report outcome measure from pre-treatment to post-treatment. This accounts for the often found inconsistencies in patient-reported outcomes after spine treatments. It has not been proven to be a valid predictor of anything. If they are insistent on using it, they will have to go through the process of construct validation to demonstrate its reliability, predictive validity and construct validity. Because this will be extremely difficult to do for a single question dealing

with a construct that is inherently difficult to quantify, their goal of evaluating effectiveness of spinal procedures would be far better served if they helped to develop consensus on truly objective external anchors of functional, physical and socioeconomically relevant criteria (e.g., subsequent surgeries, medication use changes, etc.)

Finally, we strongly disagree with the statement that "...the reality is that thresholds, while imperfect, are necessary to facilitate widespread application of patient-based outcome measures." Such thresholds are certainly desirable, but they are currently not *appropriate* if they are statistically and psychometrically flawed. We certainly do not need anymore "junk science" to determine treatment effectiveness. At this point in time, we must all "roll up our sleeves" and develop psychometrically-sound measures of the constructs we use. Using the wrong measure, just for the sake of expediency and ease, is not going to advance the science of spinal care and its effectiveness.

References

1. Worzer W, Theodore BR, Rodgerson M, Gatchel RJ. Interpreting clinical significance: A comparison of effect sizes of commonly used patient self-report pain instruments. *Practical Pain Management* 2008;8:16–29.
2. Spratt, KF. Outcomes assessment: Overview and specific tools. In: Herkowitz, HN.; Dvorak, J.; Bell, G.; Nordin, M.; Grob, D., editors. *The Lumbar Spine*. 3rd. Philadelphia: Lippincott Williams & Wilkins; 2004.
3. Glassman SD, Carreon L. Thresholds for HRQOL outcome measures: reality testing. *Spine J*. 2009 this issue.
4. Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. *Pain* 2005;113:9–19. [PubMed: 15621359]
5. Gatchel RJ, Mayer TG. Testing minimal clinically important difference: consensus or conundrum? *Spine J*. 2009 this issue.
6. Turk, DC.; Okifuji, A. Matching treatment to assessment of patients with chronic pain. In: Turk, DC.; Melzack, R., editors. *Handbook of Pain Assessment*. 2nd. New York: Guilford; 2001.
7. Turk, DC.; Monarch, ES. Biopsychosocial perspective on chronic pain. In: Turk, DC.; Gatchel, RJ., editors. *Psychological Approaches to Pain Management: A Practitioner's Handbook*. 2nd. New York: Guilford; 2002.
8. Schwartz CE, Finkelstein JA. Understanding inconsistencies in patient-reported outcomes after spine treatment: Response shift phenomenon. *Spine J* 2009;9:1039–45. [PubMed: 19574107]