6.1 Introduction

Psychological constructs such as depression or anxiety, and health-related measures such as pain or physical functioning, can be reliably assessed today by means of standardized tests. In fact, such tests are now well established as being an important part of clinical practice. Over the last few years, the number of bio-medical publications citing the word questionnaire has risen exponentially (Figure 6.1).

As a result, a considerable item burden is often placed on patients. In this context, the application of computer adaptive tests (CATs) seems promising. However, most of the theoretical and practical contributions to the application of CATs are still in the area of ability and achievement testing. Although efforts have been made to develop CATs for health-related measures, there have been very few reports on using CATs as a means of psychometric assessment in a medical setting. This situation is about to change. A prominent example is a joint initiative working on building a “Patient-Reported Outcomes Measurement Information System” (PROMIS) sponsored by the U.S. National Institutes of Health (NIH). The aim of this network is to develop a large bank of items that measures patient-reported outcomes and to create a computerized adaptive testing system that allows for efficient assessment in clinical research of a wide range of chronic diseases. These tools are expected to be available to the general medical community in 2008 (Fries, Bruce & Cella, 2005; Cella et al., 2007).

Given the advantages that the application of CATs promises, this large-scale effort on the part of the NIH to advance the development of CATs in patient-reported outcomes measurement is not surprising. Many of the advantages of CATs seem to be well suited to assessments in clinical psychology (Embretson, 1996) or medicine.

A particularly attractive property of CATs is the possibility of determining the measurement precision conditional upon the level of the underlying latent trait \( \theta \). A low measurement precision often occurs for extreme (high or low) \( \theta \)-values. In CATs constructed within the framework of item response theory (IRT), situations...
with undesirably low measurement precision can be identified and, if necessary, corrected by the administration of additional items. These additional items are not required when the measurement precision is already high. Substantial item savings may result, which would help to reduce the item burden placed on patients while still ensuring an efficient and precise measurement.

Response formats used in clinical psychology and medicine differ from those employed in ability and achievement tests. Typically, such clinical tests involve polytomous items, whereas a dichotomous coding of responses (correct vs. incorrect) is often a more natural choice for ability and achievement tests. These differences in response formats might have contributed to the delay in the use of CATs in medicine (compared to its use in educational contexts) even though many results holding for the dichotomous case have been generalized to polytomous items (e.g., Samejima, 1993).

As far as CATs are concerned, polytomous response formats open up the opportunity for substantial item savings. With the exception of highly discriminating dichotomous items, information in responses to polytomous items is usually considerably higher than in the dichotomous case. As the asymptotic standard error (SE) of a response pattern is the reciprocal of the square root of the sum of item information of this response pattern, a CAT algorithm that terminates as soon as the current error falls below a given error bond will terminate sooner if item information is high.

For instance, the Anxiety-CAT presented by Walter et al. (2005, 2007) requires, on average, about seven items to reach a predefined measurement precision of $SE \leq 0.32$ (stopping criterion, corresponding to a reliability $\rho \geq 0.9$) for latent trait values within two standard deviations around the population mean. Similar reports on substantial item savings have been published by Ware, Bjorner, and Kosinski (2000), Ware et al. (2003), Fliege et al. (2005, 2009), or Haley et al. (2006).

If IRT, as a general framework, and computer adaptive testing, as a specific application of this framework, are recognized as being beneficial for overcoming the shortcomings of classical test theory (e.g., Hambleton, 2000; Hornke, 1999), the question remains as to why there are relatively few working applications of CATs