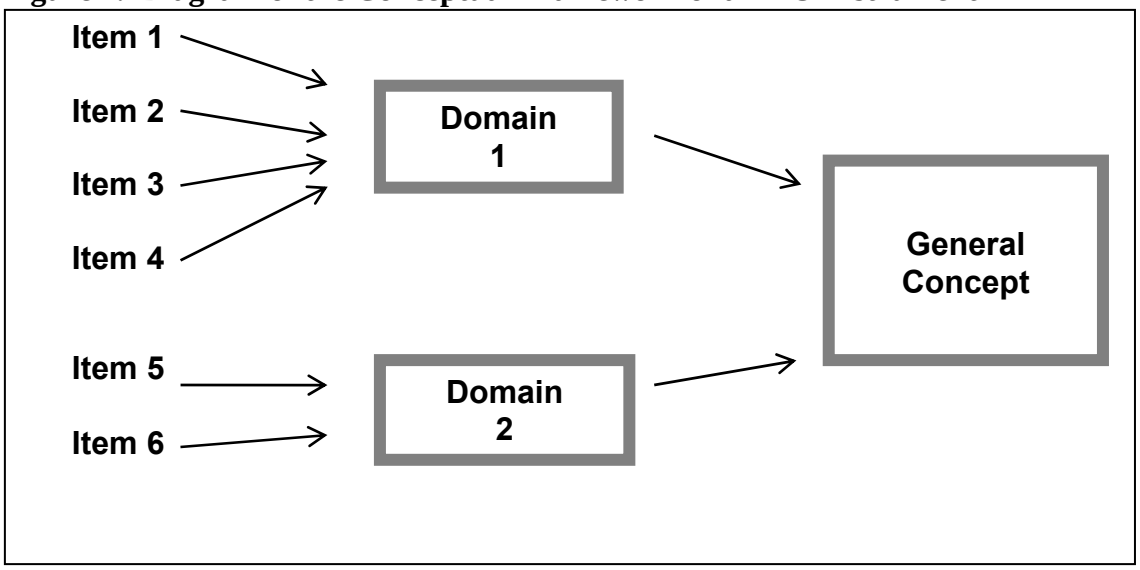
Documentation of the instrument development process should reveal the means by which the items and domains were identified. The exact words used to represent the concepts measured by domain or total scores should be derived using patient input to ensure the conclusions drawn using instrument scores are valid.

For measures of general concepts, we intend to review how individual items are thought to be associated with each other, how items are associated with each domain, and how domains are associated with each other and the general concept of interest based on the conceptual framework of the PRO instrument. The diagram in Figure 4 depicts a generic example of a conceptual framework of a PRO instrument where Domain 1, Domain 2, and General Concept each represent related but separate concepts. Items in this diagram are aggregated into domains. The final framework is derived and confirmed by measurement property testing.

Figure 4. Diagram of the Conceptual Framework of a PRO Instrument



The conceptual framework of a PRO instrument may be straightforward if a single item is a reliable and valid measure of the concept of interest (e.g., pain intensity). If the concept of interest is general (e.g., physical function), a single-item PRO instrument does not provide a useful understanding of the treatment's effect because a stand-alone single item does not capture the domains of the general concept. For this reason, single-item questions about general concepts that include multiple items or domains rarely provide sufficient evidence to support claims about that general concept. For example, in clinical trials of functional disorders defined by clusters of specific symptoms and signs, a PRO instrument consisting of a single-item global question usually would be inadequate as an endpoint to support labeling claims and would be uninformative about the effects on each specific symptom and sign. Instead, the effect of treatment on each of the appropriate symptoms and signs should be adequately measured.

The conceptual framework for PRO instruments intended to measure a general concept will be complex because identifying all of the appropriate domains and items of the general concept can be difficult. Multidomain PRO instruments can be used to support claims about a general concept if the PRO instrument has been developed to measure the important and relevant

domains of the general concept contained in the claim. However, the complex nature of multidomain PRO instruments often raises significant questions about how to interpret and report results in a way that is not misleading. For example, if improvement in a score for a general concept (e.g., symptoms associated with a certain condition) is driven by a single responsive item (e.g., pain intensity improvement) whereas other important items (e.g., other symptoms) did not show a response, a general claim about the general concept (e.g., improvements in symptoms associated with the condition) cannot be supported. However, that single responsive item or domain may support a claim specific to that item or domain.

We intend to examine the final version of an instrument in light of its development history, including documentation of the complete list of items generated and the reasons for deleting or modifying items, as illustrated in Table 1. We will determine from empiric evidence provided whether the PRO instrument's final conceptual framework (e.g., the hypothesized relationships among items, domains, and concepts measured) is confirmed in the appropriate study population and is consistent with the endpoint model of the planned clinical trials.

Table 1. Common Reasons for Changing Items during PRO Instrument Development

Item Property	Reason for Change or Deletion
Clarity or relevance	• Reported as not relevant by a large segment of the target population • Generates an unacceptably large amount of missing data points • Generates many questions or requests for clarification from patients as they complete the PRO instrument • Patients interpret items and responses in a way that is inconsistent with the PRO instrument's conceptual framework
Response range	• A high percent of patients respond at the floor (response scale's worst end) or ceiling (response scale's optimal end) • Patients note that none of the response choices applies to them • Distribution of item responses is highly skewed
Variability	• All patients give the same answer (i.e., no variance) • Most patients choose only one response choice • Differences among patients are not detected when important differences are known
Reproducibility	• Unstable scores over time when there is no logical reason for variation from one assessment to the next
Inter-item correlation	• Item highly correlated (redundant) with other items in the same concept of interest
Ability to detect change	• Item is not sensitive (i.e., does not change when there is a known change in the concepts of interest)
Item discrimination	• Item is highly correlated with measures of concepts other than the one it is intended to measure • Item does not show variability in relation to some known population characteristics (i.e., severity level, classification of condition, or other known characteristic)
Redundancy	• Item duplicates information collected with other items that have equal or better measurement properties
Recall period	• The population, disease state, or application of the instrument can affect the appropriateness of the recall period

2. Intended Population

Using documentation of the process described in Figure 3 and of the measurement properties as described in Table 2, we plan to compare the patient population studied in the PRO instrument development process to the population enrolled in the clinical trial to determine whether the instrument is applicable for that population. See the Appendix for a description of the types of information sponsors should provide for FDA discussion and review of PRO instruments.

Specific measurement considerations posed by pediatric, cognitively impaired, or seriously ill patients are discussed in section III.G., PRO Instruments Intended for Specific Populations.

Table 2. Measurement Properties Considered in the Review of PRO Instruments Used in Clinical Trials

Measurement Property	Type	What Is Assessed?	FDA Review Considerations
Reliability	Test-retest or intra-interviewer reliability (for interviewer-administered PROs only)	Stability of scores over time when no change is expected in the concept of interest	• Intraclass correlation coefficient • Time period of assessment
	Internal consistency	• Extent to which items comprising a scale measure the same concept • Intercorrelation of items that contribute to a score • Internal consistency	• Cronbach's alpha for summary scores • Item-total correlations
	Inter-interviewer reliability (for interviewer-administered PROs only)	Agreement among responses when the PRO is administered by two or more different interviewers	• Interclass correlation coefficient
Validity	Content validity	Evidence that the instrument measures the concept of interest including evidence from qualitative studies that the items and domains of an instrument are appropriate and comprehensive relative to its intended measurement concept, population, and use. Testing other measurement properties will not replace or rectify problems with content validity.	• Derivation of all items • Qualitative interview schedule • Interview or focus group transcripts • Items derived from the transcripts • Composition of patients used to develop content • Cognitive interview transcripts to evaluate patient understanding
	Construct validity	Evidence that relationships among items, domains, and concepts conform to *a priori* hypotheses concerning logical relationships that should exist with measures of related concepts or scores produced in similar or diverse patient groups	• Strength of correlation testing *a priori* hypotheses (discriminant and convergent validity) • Degree to which the PRO instrument can distinguish among groups hypothesized *a priori* to be different (known groups validity)
Ability to detect change		Evidence that a PRO instrument can identify differences in scores over time in individuals or groups (similar to those in the clinical trials) who have changed with respect to the measurement concept	• Within person change over time • Effect size statistic