A major problem facing veterinary surgeons is the lack of high-quality evidence available for optimal decision making. Ideally, fair and direct head-to-head comparisons of diagnostic and therapeutic interventions would allow us to make well-informed decisions for our patients. This type of information, however, is rarely available for many reasons, not the least of which is a lack of consensus about what it is exactly we wish to compare. Ideally, comparisons would be made based on clinically relevant, widely available, easy to use, well validated, and objective outcome measures. Unfortunately, no such measures currently exist and investigators are forced to choose which attributes of an outcome measure are most important for their study.

Historically, when testing the efficacy of an intervention, veterinary orthopedists have relied heavily on the numbers generated from force plate gait analysis. Properly collected, gait analysis data offers an objective measure that can be reliably followed temporally; however, it can be very time consuming, requires specialized equipment, varies in execution among users, and relies on relatively strict inclusion criteria. It only evaluates the animal at one specific point in time and weight bearing on an affected limb is only one part of the larger clinical picture of response to an intervention. Whether the veterinarian and/or the owner believe that the animal has benefited from an intervention are clinically relevant questions and could be addressed with carefully designed outcome measures that would be readily available, easy to use, and standard across studies. The fact that an owner or veterinarian assessment is inherently subjective does not preclude its use as an outcome measure. If appropriate methodology is used, hard measures of subjective outcomes can be developed and subjective states can be reliably quantified. Physician and patient outcome assessments are a mainstay in determining the efficacy of interventions in human orthopedics. The short form 36 (SF-36), Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), Knee injury and Osteoarthritis Outcome Score (KOOS), and International Knee Documentation Committee (IKDC) Subjective Knee Form are all examples of validated outcome measures that provide reliable answers to clinically relevant questions in human orthopedics.\textsuperscript{1–7}

**The Stepwise Process of Developing a Health Measurement Instrument\textsuperscript{8–17}**

**Step 1—Devising the Items**

The first step in developing an instrument (questionnaire) involves devising the items (questions) themselves. Whereas this seems obvious, it is far from trivial, because no amount of statistical manipulation after the fact can compensate for questions that are poorly worded, ambiguous, or irrelevant. Items are generated through multiple focus groups or interviews with owners and/or veterinarians of animals that have the disease or condition for which we are hoping to determine the efficacy of interventions. Once the set of questions is formulated, a method by which responses will be obtained must be chosen. Will the responses can be categorical or continuous; visual analog, adjectival, or specific? Should there be an even or odd number of categories? Whereas there may not be a “right answer” to these questions, choice of response scale can impact the data collected.
ple, when using a bipolar scale (like "strongly agree" to "strongly disagree"), if an odd number of categories is used, the rater is allowed the choice of expressing no opinion (i.e. strongly agree, agree, no opinion, disagree, strongly disagree). However, if an even number of categories is chosen, the rater is forced to commit themselves, one way or the other. Whether or not it is desirable to allow a neutral position depends on the goals of the instrument being developed. So while devising the items of a questionnaire appears on the surface to be straightforward, it is a very involved and time consuming process that can greatly impact the validity, reliability, and ultimate utility of the questionnaire.

Step 2—Selecting the Items

Typically, not all of the items that are developed are ultimately included in the new instrument. Some may be confusing, interpreted differently by different respondents, or not deliver the desired information. Various criteria can be used to determine which of the developed items should be retained for the preliminary instrument.

Items that are difficult to interpret are removed or rephrased. It is generally recommended that instruments not require reading skills beyond a 6th grade level. There are a number of methods by which the readability of an instrument can be measured, but readability alone does not ensure the absence of ambiguity. Even seemingly straightforward questions like "Does your dog have difficulty going up and down the stairs?" can pose problems. This is a double-barreled question, one that asks 2 questions at the same time. How would one answer if their dog has difficulty going up the stairs, but not down? How does one answer the question if the dog lives in an environment with no stairs? In addition there is ambiguity of time frame, is the question referring to difficulty today, in the past week, past month, past year? Questions must also be screened for the use of jargon (e.g. "NSAID" or "range-of-motion") that may not be understood by all respondents or value-laden words (e.g. "trivial" or "too much") that could prejudice the respondents and distract from clear interpretation of a question.

Once the items have been screened for readability and ambiguity, they can be presented to a panel of "experts" to determine whether the items look reasonable. This is typically a handful of people with experience managing the condition we are hoping to capture with the questionnaire. They can make an assessment of whether the items appear, on the surface, to be measuring what we are intending (face validity) and whether the questions cover all of the relevant aspects of the condition (content validity). Once a preliminary set of items is agreed upon, the instrument can be pretested.

Pretesting involves the administration of the preliminary instrument to a small group of respondents to determine whether some items may not perform well. In most situations, when measuring a trait, the instrument should be homogeneous. That is, all of the items should be tapping into different aspects of the same trait, not different aspects of different traits. Therefore, the items should be moderately correlated with each other. An inter-item correlation matrix can be analyzed from the data collected from this group of respondents to identify items that have consistently low correlations with other items in the instrument. These items can either be revised and the preliminary instrument pretested again in a different group of respondents, or the items can be removed from the instrument before moving onto large scale testing for reliability and validity.

Step 3—Reliability and Validity Testing

Once a data-gathering instrument is developed, it must be established that it will target what it is supposed to measure. This is defined as the validity of the instrument. In addition, the instrument must measure what it is supposed to measure in a consistent manner. The tendency toward consistency is referred to as reliability.

Before we can obtain evidence that an instrument is measuring what it is intended, it is first necessary to gather evidence that it is measuring something reliably. There are a number of ways in which an instrument can be tested for reliability. An assessment of internal consistency can be based on the data collected from a single administration of the instrument to a large group of respondents. Whereas there are a number of methods of consistency calculation, all represent the average of the correlations among all of the items. An assessment of internal consistency alone; however, is not sufficient to declare an instrument reliable, because it relies on only a single administration.

To account for the day-to-day variability in responses, an assessment of the stability (i.e. reproducibility) of the instrument must also be made. By administering the instrument to the same population of respondents on 2 different occasions, the test–retest reliability of the instrument can be assessed. For example, if an instrument designed to measure chronic pain in dogs with osteoarthritis is reliable, it should deliver very similar results when administered on 2 different occasions, 1 week apart, to owners of dogs that have stable disease and no change in treatment. If instead, the results between the 2 administrations are very different, the instrument would not be useful for testing the efficacy of an intervention. How would one interpret a change in instrument score after an intervention, when a change can be documented when there is no intervention at all? It is only when we can
demonstrate that the instrument is measuring something reliably, that we can begin the process of determining what that something is.

To determine that the instrument is measuring what is intended requires more than peer judgments (face validity). Validating an instrument is a process by which we determine the degree of confidence we can place on conclusions we draw about an animal based on their score from that instrument. If other validated instruments designed to measure the same attribute exist, then an obvious approach is to administer the experimental instrument along with the existing one and see whether there is a strong correlation between the 2. More likely, however, no other measure exists and developers must test “construct” validity.

Construct validity is evaluated when the attribute being measured cannot be directly observed. For example, chronic pain cannot be “seen,” but behaviors can be observed which, according to our theories about chronic pain in companion dogs, result from it. There is no one single experiment or statistic, which can unequivocally “prove” a construct. It is through multiple analyses and assessments that a construct appears to be valid. For example, based on a construct of “Chronic Pain,” dogs with osteoarthritis that score high on a newly developed instrument differ from dogs that score low on it in terms of attributes such as the results of force plate gait analysis, an assessment of health related quality of life, and daily activity monitoring. There is no one single experiment to “prove” the validity of the instrument, but multiple, well-designed, hypothesis driven studies can build the body of evidence that the instrument is measuring what we have intended.

It is necessary to conduct validation studies for each new instrument that is developed and the task is an on-going one. If we want to use the instrument in a group of animals that it was not initially validated on (i.e. dogs with bone cancer), we must first demonstrate that the inferences made for them are as valid as for the original population (i.e. dogs with osteoarthritis). In addition, modifications of the instrument such as changes in wording of items or responses, order of items, removal or addition of items, often requires new validity studies.

APPLICABILITY TO VETERINARY SURGERY

Very few attempts have been made in veterinary medicine to develop valid and reliable instruments. It is not uncommon to find published efficacy studies where one of the outcomes is the summative score of a set of items that were made up for the study, but underwent no tests for reliability or validity. It is impossible to know for sure, therefore, whether the results (positive or negative) of such studies are because of the intervention rather than inconsistencies or inaccuracies associated with the outcome score. One reason for the prevalent use of unvalidated outcomes may be a lack of awareness of the necessity for the process and the fact that sound methodology for development and validation exists. Another reason may be that, as outlined above, the process of instrument development and validation is an onerous one. It takes a great amount of time (years), expertise, and financial support. Whereas the up front investment of resources to the development of validated outcomes for use in veterinary surgery would be large, the benefits of having well-validated, quantifiable, veterinarian and owner efficacy assessments would be invaluable to investigators designing studies, as well as to practitioners that use the results of such studies to guide their clinical decision making. Using validated outcomes in well-designed studies to objectively determine the efficacy of the diagnostic and therapeutic interventions we prescribe every day is not just something that “would be nice to do if we could,” but is an obligation to ourselves, to our patients and to their owners.

ACKNOWLEDGMENTS

The author wishes to acknowledge the important contributions of Drs. Jimi Cook, Amy Kapatkin, Kurt Schulz, Bob Taylor, John Innes, and John Pascoe to this Outcomes Measures Program and the preparation of this manuscript.

REFERENCES

7. McHorney CA, Ware JE Jr: Construction and validation of an alternate form general mental health scale for the Med-
ical Outcomes Study Short-Form 36-Item Health Survey. Medical Care 33:15–28, 1995