Review Article

Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine

ABSTRACT

The task of developing a new questionnaire or translating an existing questionnaire into a different language might be overwhelming. The greatest challenge perhaps is to come up with a questionnaire that is psychometrically sound, and is efficient and effective for use in research and clinical settings. This article provides guidelines for the development and translation of questionnaires for application in medical fields, with a special emphasis on perioperative and pain medicine. We provide a framework to guide researchers through the various stages of questionnaire development and translation. To ensure that the questionnaires are psychometrically sound, we present a number of statistical methods to assess the reliability and validity of the questionnaires.

Key words: Anesthesia; development; questionnaires; translation; validation

Introduction

Questionnaires or surveys are widely used in perioperative and pain medicine research to collect quantitative information from both patients and health-care professionals. Data of interest could range from observable information (e.g., presence of lesion, mobility) to patients' subjective feelings of their current status (e.g., the amount of pain they feel, psychological status). Although using an existing questionnaire will save time and resources,[1] a questionnaire that measures the construct of interest may not be readily available, or the published questionnaire is not available in the language required for the targeted respondents. As a result, investigators may need to develop a new questionnaire or translate an existing one into the language of the intended respondents. Prior work has highlighted the wealth of literature available on psychometric principles, methodological concepts, and techniques regarding questionnaire development/translation and validation. To that end, this article is not meant to provide an exhaustive review of all the related statistical concepts and methods. Rather, this article aims to provide straightforward guidelines for the development or translation of questionnaires (or scales) for use in perioperative and pain medicine research for readers who may be unfamiliar with the process of questionnaire development and/or translation. Readers are recommended to consult the cited references to further examine these techniques for application.

Access this article online	
Website: www.saudija.org	Quick Response Code
DOI: 10.4103/sja.SJA_203_17	

How to cite this article: Tsang S, Royse CF, Terkawi AS. Guidelines for developing, translating, and validating a questionnaire in perioperative and pain medicine. Saudi J Anaesth 2017;11:S80-9.

Siny Tsang, Colin F. Royse[1,2], Abdullah Sulieman Terkawi[3,4,5]

Department of Epidemiology, Columbia University, New York, NY, [3]Department of Anesthesiology, University of Virginia, Charlottesville, VA, [5]Outcomes Research Consortium, Cleveland, OH, USA, [1]Department of Surgery, University of Melbourne, Melbourne, [2]Department of Anesthesia and Pain Management, The Royal Melbourne Hospital, Parkville, Victoria, Australia, [4]Department of Anesthesiology, King Fahad Medical City, Riyadh, Saudi Arabia

Address for correspondence: Dr. Siny Tsang, Department of Epidemiology, Columbia University, New York, NY, USA.
E-mail: st2989@cumc.columbia.edu

This article is divided into two main sections. The first discusses issues that investigators should be aware of in developing or translating a questionnaire. The second section of this paper illustrates procedures to validate the questionnaire after the questionnaire is developed or translated. A model for the questionnaire development and translation process is presented in Figure 1. In this special issue of the Saudi journal of Anesthesia we presented multiple studies of development and validation of questionnaires in perioperative and pain medicine, we encourage readers to refer to them for practical experience.

Preliminary Considerations

It is crucial to identify the construct that is to be assessed with the questionnaire, as the domain of interest will determine what the questionnaire will measure. The next question is: How will the construct be operationalized? In other words, what types of behavior will be indicative of the domain of interest? Several approaches have been suggested to help with this process,[2] such as content analysis, review of research, critical incidents, direct observations, expert judgment, and instruction.

Once the construct of interest has been determined, it is important to conduct a literature review to identify if a previously validated questionnaire exists. A validated questionnaire refers to a questionnaire/scale that has been developed to be administered among the intended respondents. The validation processes should have been completed using a representative sample, demonstrating adequate reliability and validity. Examples of necessary validation processes can be found in the validation section of this paper. If no existing questionnaires are available, or none that are determined to be appropriate, it is appropriate to construct a new questionnaire. If a questionnaire exists, but only in a different language, the task is to translate and validate the questionnaire in the new language.

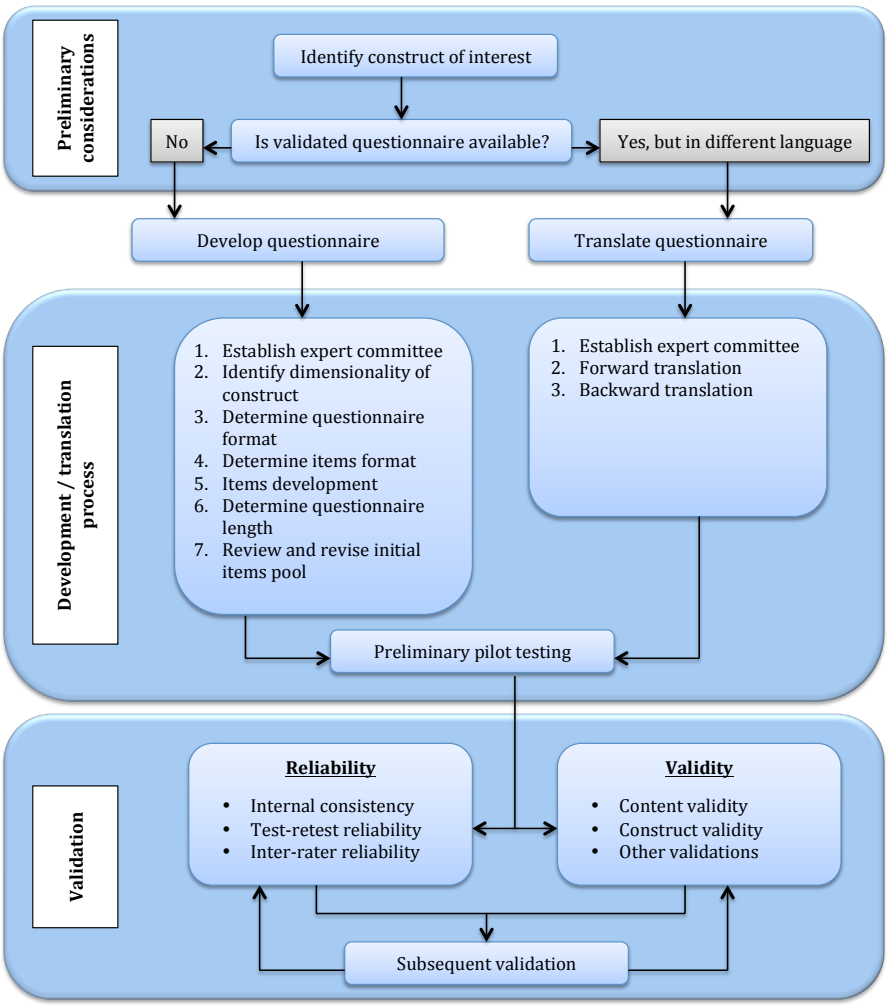

Figure 1: Questionnaire development and translation processes

Developing a Questionnaire

To construct a new questionnaire, a number of issues should be considered even before writing the questionnaire items.

Identify the dimensionality of the construct

Many constructs are multidimensional, meaning that they are composed of several related components. To fully assess the construct, one may consider developing subscales to assess the different components of the construct. Next, are all the dimensions equally important? or are some more important than others? If the dimensions are equally important, one can assign the same weight to the questions (e.g., by summing or taking the average of all the items). If some dimensions are more important than others, it may not be reasonable to assign the same weight to the questions. Rather, one may consider examining the results from each dimension separately.

Determine the format in which the questionnaire will be administered

Will the questionnaire be self-administered or administered by a research/clinical staff? This decision depends, in part, on what the questionnaire intends to measure. If the questionnaire is designed to measure catastrophic thinking related to pain, respondents may be less likely to respond truthfully if a research/clinical staff asked the questions, whereas they may be more likely to respond truthfully if they are allowed to complete the questionnaire on their own. If the questionnaire is designed to measure patients' mobility after surgery, respondents may be more likely to overreport the amount of mobility in an effort to demonstrate recovery. To obtain a more accurate measure of mobility after surgery, it may be preferable to obtain objective ratings by clinical staff.

If respondents are to complete the questionnaire by themselves, the items need to be written in a way that can be easily understood by the majority of the respondents, generally about Grade 6 reading level.[3] If the questionnaire is to be administered to young respondents or respondents with cognitive impairment, the readability level of the items should be lowered. Questionnaires intended for children should take into consideration the cognitive stages of young people[4] (e.g., pictorial response choices may be more appropriate, such as pain faces to assess pain[5]).

Determine the item format

Will the items be open ended or close ended? Questions that are open ended allow respondents to elaborate upon their responses. As more detailed information may be obtained using open-ended questions, these items are best suited for situations in which investigators wish to gather more information about a specific domain. However, these responses are often more difficult to code and score, which increases the difficulty of summarizing individuals' responses. If multiple coders are included, researchers have to address the additional issue of inter-rater reliability.

Questions that are close ended provide respondents a limited number of response options. Compared to open-ended questions, these items are easier to administer and analyze. On the other hand, respondents may not be able to clarify their responses, and their responses may be influenced by the response options provided.

If close-ended items are to be used, should multiple-choice, Likert-type scales, true/false, or other close-ended formats be used? How many response options should be available? If a Likert-type scale is to be adopted, what scale anchors are to be used to indicate the degree of agreement (e.g., strongly agree, agree, neither, disagree, strongly degree), frequency of an event (e.g., almost never, once in a while, sometimes, often, almost always), or other varying options? To make use of participants' responses for subsequent statistical analyses, researchers should keep in mind that items should be scaled to generate sufficient variance among the intended respondents.[6,7]

Item development

A number of guidelines have been suggested for writing items.[7] Items should be simple, short, and written in language familiar to the target respondents. The perspective should be consistent across items; items that assess affective responses (e.g., anxiety, depression) should not be mixed with those that assess behavior (e.g., mobility, cognitive functioning).[8] Items should assess only a single issue. Items that address more than one issue, or "double-barreled" items (e.g., "My daily activities and mood are affected by my pain."), should not be used. Avoid leading questions as they may result in biased responses. Items that all participants would respond similarly (e.g., "I would like to reduce my pain.") should not be used, as the small variance generated will provide limited information about the construct being assessed. Table 1 summarizes important tips on writing questions.

The issue of whether reverse-scored items should be used remains debatable. Since reverse-scored items are negatively worded, it has been argued that the inclusion of these items may reduce response set bias.[9] On the other hand, others have found a negative impact on the psychometric properties of scales that included negatively worded items.[10] In recent years, an increasing amount of literature reports problems with reverse-scored items.[11-14] Researchers who decide to include negatively worded items should take extra steps

Table 1: Tips on writing questions[15,16]

Use short and simple sentences
Ask for only one piece of information at a time. Example: Have you had nausea and vomiting in the last 24 h? Someone may have nausea, but may not have vomiting , thus this question should be divided into two questions
Avoid negatives if possible. Example: In the last 24 h, how many times did you not have pruritus? The better format would be; in the last 24 h, how many times did you have pruritus?
Ask precise questions. Example: Have you had pain before? Better question would be; what was your worst pain in the last 24 h?
Ensure that those you ask have the necessary knowledge. Example: Have you had neuropathic pain before? Many patients may not know what "neuropathic" means, a better question(s) would be to ask about the symptoms of neuropathic pain, for example, "have you had episodes of piercing pain like hot needles into your skin, before"
Avoid unnecessary details, as people are usually less inclined to complete long questionnaires, however make sure to ask for all the essential details
Avoid asking direct questions on sensitive issues. Example: "Are you obese?" can be better written as " do you think you have a weight issue"
Minimize bias. Example: "I was satisfied with the pain management that I had (yes or no)?" Better question is to ask about the level of satisfaction in a scale from 0 to 10. As many patients may choose yes to please you
Avoid weasel words such as commonly, usually, some, and hardly ever. Example: "Do you commonly have pain?" is better written as "How often do you have pain?"
Avoid using statements instead of questions
Avoid using agreement response anchors. Example: Your postoperative pain was the main concern to you before surgery (with Likert scale options). A better question would be what was your main concern before surgery? (with listing some options)
Avoid using too few or too many response anchors. Use five or more response anchors to achieve stable participant responses
Verbally label each response option, use only verbal labels, maintain equal spacing between response options, and use additional space to visually separate nonsubstantive response options from the substantive options
Arrange the questions. Always go from general to particular, easy to difficult, and factual to abstract
Consider adding some contradictory questions, to detect the responders' consistency, as some tend to tick whether "agree" or "disagree"

to ensure that the items are interpreted as intended by the respondents, and that the reverse-coded items have similar psychometric properties as the other regularly coded items.[7]

Determine the intended length of questionnaire

There is no rule of thumb for the number of items that make up a questionnaire. The questionnaire should contain sufficient items to measure the construct of interest, but not be so long that respondents experience fatigue or loss of motivation in completing the questionnaire.[17,18] Not only should a questionnaire possess the most parsimonious (i.e., simplest) structure,[19] but it also should consist of items that adequately represent the construct of interest to minimize measurement error.[20] Although a simple structure of questionnaire is recommended, a large pool of items is needed in the early stages of the questionnaire's development as many of these items might be discarded throughout the development process.[7]

Review and revise initial pool of items

After the initial pool of questionnaire items are written, qualified experts should review the items. Specifically, the items should be reviewed to make sure they are accurate, free of item construction problems, and grammatically correct. The reviewers should, to the best of their ability, ensure that the items do not contain content that may be perceived as offensive or biased by a particular subgroup of respondents.

Preliminary pilot testing

Before conducting a pilot test of the questionnaire on the intended respondents, it is advisable to test the questionnaire items on a small sample (about 30–50)[21] of respondents.[17] This is an opportunity for the questionnaire developer to know if there is confusion about any items, and whether respondents have suggestions for possible improvements of the items. One can also get a rough idea of the response distribution to each item, which can be informative in determining whether there is enough variation in the response to justify moving forward with a large-scale pilot test. Feasibility and the presence of floor (almost all respondents scored near the bottom) or ceiling effects (almost all respondents scored near the top) are important determinants of items that are included or rejected at this stage. Although it is possible that participants' responses to questionnaires may be affected by question order,[22-24] this issue should be addressed only after the initial questionnaire has been validated. The questionnaire items should be revised upon reviewing the results of the preliminary pilot testing. This process may be repeated a few times before finalizing the final draft of the questionnaire.

Summary

So far, we highlighted the major steps that need to be undertaken when constructing a new questionnaire. Researchers should be able to clearly link the questionnaire items to the theoretical construct they intend to assess. Although such associations may be obvious to researchers who are familiar with the specific topic, they may not be apparent to other readers and reviewers. To develop a questionnaire with good psychometric properties that can subsequently be applied in research or clinical practice, it is crucial to invest the time and effort to ensure that the items adequately assess the construct of interest.

Translating a Questionnaire

The following section summarizes the guidelines for translating a questionnaire into a different language.

Forward translation

The initial translation from the original language to the target language should be made by at least two independent translators.[25,26] Preferably, the bilingual translators should be translating the questionnaire into their mother tongue, to better reflect the nuances of the target language.[27] It is recommended that one translator be aware of the concepts the questionnaire intend to measure, to provide a translation that more closely resembles the original instrument. It is suggested that a naïve translator, who is unaware of the objective of the questionnaire, produce the second translation so that subtle differences in the original questionnaire may be detected.[25,26] Discrepancies between the two (or more) translators can be discussed and resolved between the original translators, or with the addition of an unbiased, bilingual translator who was not involved in the previous translations.

Backward translation

The initial translation should be independently back-translated (i.e., translate back from the target language into the original language) to ensure the accuracy of the translation. Misunderstandings or unclear wordings in the initial translations may be revealed in the back-translation.[25] As with the forward translation, the backward translation should be performed by at least two independent translators, preferably translating into their mother language (the original language).[26] To avoid bias, back-translators should preferably not be aware of the intended concepts the questionnaire measures.[25]

Expert committee

Constituting an expert committee is suggested to produce the prefinal version of the translation.[25] Members of the committee should include experts who are familiar with the construct of interest, a methodologist, both the forward and backward translators, and if possible, developers of the original questionnaires. The expert committee will need to review all versions of the translations and determine whether the translated and original versions achieve semantic, idiomatic, experiential, and conceptual equivalence.[25,28] Any discrepancies will need to be resolved, and members of the expert committee will need to reach a consensus on all items to produce a prefinal version of the translated questionnaire. If necessary, the process of translation and back-translation can be repeated.

Preliminary pilot testing

As with developing a new questionnaire, the prefinal version of the translated questionnaire should be pilot tested on a small sample (about 30–50)[21] of the intended respondents.[25,26] After completing the translated questionnaire, the respondent is asked (verbally by an interviewer or via an open-ended question) to elaborate what they thought each questionnaire item and their corresponding response meant. This approach allows the investigator to make sure that the translated items retained the same meaning as the original items, and to ensure there is no confusion regarding the translated questionnaire. This process may be repeated a few times to finalize the final translated version of the questionnaire.

Summary

In this section, we provided a template for translating an existing questionnaire into a different language. Considering that most questionnaires were initially developed in one language (e.g., English when developed in English-speaking countries[25]), translated versions of the questionnaires are needed for researchers who intend to collect data among respondents who speak other languages. To compare responses across populations of different language and/or culture, researchers need to make sure that the questionnaires in different languages are assessing the equivalent construct with an equivalent metric. Although the translation process is time consuming and costly, it is the best method to ensure that a translated measure is equivalent to the original questionnaire.[28]

Validating a Questionnaire

Initial validation

After the new or translated questionnaire items pass through preliminary pilot testing and subsequent revisions, it is time to conduct a pilot test among the intended respondents for initial validation. In this pilot test, the final version of the questionnaire is administered to a large representative sample of respondents for whom the questionnaire is intended. If the pilot test is conducted for small samples, the relatively large sampling errors may reduce the statistical power needed to validate the questionnaire.[2]

Reliability

The reliability of a questionnaire can be considered as the consistency of the survey results. As measurement error is present in content sampling, changes in respondents, and differences across raters, the consistency of a questionnaire can be evaluated using its internal consistency, test-retest reliability, and inter-rater reliability, respectively.

Internal consistency

Internal consistency reflects the extent to which the questionnaire items are inter-correlated, or whether they are consistent in measurement of the same construct. Internal consistency is commonly estimated using the coefficient alpha,[29] also known as Cronbach's alpha. Given a questionnaire *x*, with *k* number of items, alpha (α) can be computed as:

$$\alpha = \frac{\kappa}{\kappa - 1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_x^2}\right)$$

Where, σ_i^2 is the variance of item *i*, and σ_x^2 is the total variance of the questionnaire.

Cronbach's alpha ranges from 0 to 1 (when some items are negatively correlated with other items in the questionnaire, it is possible to have negative values of Cronbach's alpha). When reverse-scored items are [incorrectly] not reverse scored, it can be easily remedied by correctly scoring the items. However, if a negative Cronbach's alpha is still obtained when all items are correctly scored, there are serious problems in the original design of the questionnaire), with higher values indicating that items are more strongly interrelated with one another. Cronbach's $\alpha = 0$ indicates no internal consistency (i.e., none of the items are correlated with one another), whereas $\alpha = 1$ reflects perfect internal consistency (i.e., all the items are perfectly correlated with one another). In practice, Cronbach's alpha of at least 0.70 has been suggested to indicate adequate internal consistency.[30] A low Cronbach's alpha value may be due to poor inter-relatedness between items; as such, items with low correlations with the questionnaire total score should be discarded or revised. As alpha is a function of the length of the questionnaire, alpha will increase with the number of items. In addition, alpha will increase if the variability of each item is increased. It is, therefore, possible to increase alpha by including more related items, or adding items that have more variability to the questionnaire. On the other hand, an alpha value that is too high ($\alpha \geq 0.90$) suggests that some questionnaire items may be redundant;[31] investigators may consider removing items that are essentially asking the same thing in multiple ways.

It is important to note that Cronbach's alpha is a property of the responses from a specific sample of respondents.[31] Investigators need to keep in mind that Cronbach's alpha is not "the" estimate of reliability for a questionnaire under all circumstances. Rather, the alpha value only indicates the extent to which the questionnaire is reliable for "a particular population of examinees."[32] A questionnaire with excellent reliability with one sample may not necessarily have the same reliability in another. Therefore, the reliability of a questionnaire should be estimated each time the questionnaire is administered, including pilot testing and subsequent validation stages.

Test-retest reliability

Test-retest reliability refers to the extent to which individuals' responses to the questionnaire items remain relatively consistent across repeated administration of the same questionnaire or alternate questionnaire forms.[2] Provided the same individuals were administered the same questionnaires twice (or more), test-retest reliability can be evaluated using Pearson's product moment correlation coefficient (Pearson's *r*) or the intraclass correlation coefficient.

Pearson's *r* between the two questionnaires' responses can be referred to as the coefficient of stability. A larger stability coefficient indicates stronger test-retest reliability, reflecting that measurement error of the questionnaire is less likely to be attributable to changes in the individuals' responses over time.

Test-retest reliability can be considered the stability of respondents' attributes; it is applicable to questionnaires that are designed to measure personality traits, interest, or attitudes that are relatively stable across time, such as anxiety and pain catastrophizing. If the questionnaires are constructed to measure transitory attributes, such as pain intensity and quality of recovery, test-retest reliability is not applicable as the changes in respondents' responses between assessments are reflected in the instability of their responses. Although test-retest reliability is sometimes reported for scales that are intended to assess constructs that change between administrations, researchers should be aware that test-retest reliability is not applicable and does not provide useful information about the questionnaires of interest. Researchers should also be critical when evaluating the reliability estimates reported in such studies.

An important question to consider in estimating test-retest reliability is how much time should lapse between questionnaire administrations? If the duration between time 1 and time 2 is too short, individuals may remember their responses in time 1, which may overestimate the test-retest reliability. Respondents, especially those recovering from major surgery, may experience fatigue if the retest is administered shortly after the first administration, which may underestimate the test-retest reliability. On the other hand, if there is a long period of time between questionnaire administrations, individuals' responses may change due to other factors (e.g., a respondent may be taking pain

management medications to treat chronic pain condition). Unfortunately, there is no single answer. The duration should be long enough to allow the effects of memory to fade and to prevent fatigue, but not so long as to allow changes to take place that may affect the test-retest reliability estimate.[17]

Inter-rater reliability

For questionnaires in which multiple raters complete the same instrument for each examinee (e.g., a checklist of behavior/symptoms), the extent to which raters are consistent in their observations across the same group of examinees can be evaluated. This consistency is referred to as the inter-rater reliability, or inter-rater agreement, and can be estimated using the kappa statistic.[33] Suppose two clinicians independently rated the same group of patients on their mobility after surgery (e.g., 0 = needs help of 2+ people; 1 = needs help of 1 person; 2 = independent), kappa (κ) can be computed as follows:

$$\kappa = \frac{P_0 - P_e}{1 - P_e}$$

Where, P_o is the observed proportion of observations in which the two raters agree, and P_e is the expected proportion of observations in which the two raters agree by chance. Accordingly, κ is the proportion of agreement between the two raters, after factoring out the proportion of agreement by chance. κ ranges from 0 to 1, where $\kappa = 0$ indicates all chance agreements and $\kappa = 1$ represents perfect agreement between the two raters. Others have suggested $\kappa = 0$ as no agreement, $\kappa = 0.01 - 0.20$ as poor agreement, $\kappa = 0.21 - 0.40$ as slight agreement, $\kappa = 0.41 - 0.60$ as fair agreement, $\kappa = 0.61 - 0.80$ as good agreement, $\kappa = 0.81 - 0.92$ as very good agreement, and $\kappa = 0.93 - 1$ as excellent agreement.[34,35] If more than two raters are used, an extension of Cohen's κ statistic is available to compute the inter-rater reliability across multiple raters.[36]

Validity

The validity of a questionnaire is determined by analyzing whether the questionnaire measures what it is intended to measure. In other words, are the inferences and conclusions made based on the results of the questionnaire (i.e., test scores) valid?[37] Two major types of validity should be considered when validating a questionnaire: content validity and construct validity.

Content validity

Content validity refers to the extent to which the items in a questionnaire are representative of the entire theoretical construct the questionnaire is designed to assess.[17] Although the construct of interest determines which items are written and/or selected in the questionnaire development/translation phase, content validity of the questionnaire should be evaluated after the initial form of the questionnaire is available.[2] The process of content validation is particularly crucial in the development of a new questionnaire.

A panel of experts who are familiar with the construct that the questionnaire is designed to measure should be tasked with evaluating the content validity of the questionnaire. The experts judge, as a panel, whether the questionnaire items are adequately measuring the construct intended to assess, and whether the items are sufficient to measure the domain of interest. Several approaches to quantify the judgment of content validity across experts are also available, such as the content validity ratio[38] and content validation form.[39,40] Nonetheless, as the process of content validation depends heavily on how well the panel of experts can assess the extent to which the construct of interest is operationalized, the selection of appropriate experts is crucial to ensure that content validity is evaluated adequately. Example items to assess content validity include:[41]

- The questions were clear and easy
- The questions covered all the problem areas with your pain
- You would like the use of this questionnaire for future assessments
- The questionnaire lacks important questions regarding your pain
- Some of the questions violate your privacy.

A concept that is related to content validity is face validity. Face validity refers to the degree to which the respondents or laypersons judge the questionnaire items to be valid. Such judgment is based less on the technical components of the questionnaire items, but rather on whether the items appear to be measuring a construct that is meaningful to the respondents. Although this is the weakest way to establish the validity of a questionnaire, face validity may motivate respondents to answer more truthfully. For example, if patients perceive a quality of recovery questionnaire to be evaluating how well they are recovering from surgery, they may be more likely to respond in ways that reflect their recovery status.

Construct validity

Construct validity is the most important concept in evaluating a questionnaire that is designed to measure a construct that is not directly observable (e.g., pain, quality of recovery). If a questionnaire lacks construct validity, it will be difficult to interpret results from the questionnaire, and inferences cannot be drawn from questionnaire responses to a

behavior domain. The construct validity of a questionnaire can be evaluated by estimating its association with other variables (or measures of a construct) with which it should be correlated positively, negatively, or not at all.[42] In practice, the questionnaire of interest, as well as the preexisting instruments that measure similar and dissimilar constructs, is administered to the same groups of individuals. Correlation matrices are then used to examine the expected patterns of associations between different measures of the same construct, and those between a questionnaire of a construct and other constructs. It has been suggested that correlation coefficients of 0.1 should be considered as small, 0.3 as moderate, and 0.5 as large.[43]

For instance, suppose a new scale is developed to assess pain among hospitalized patients. To provide evidence of construct validity for this new pain scale, we can examine how well patients' responses on the new scale correlate with the preexisting instruments that also measure pain. This is referred to as convergent validity. One would expect strong correlations between the new questionnaire and the existing measures of the same construct, since they are measuring the same theoretical construct.

Alternatively, the extent to which patients' responses on the new pain scale correlate with instruments that measure unrelated constructs, such as mobility or cognitive function, can be assessed. This is referred to as divergent validity. As pain is theoretically dissimilar to the constructs of mobility or cognitive function, we would expect zero, or very weak, correlation between the new pain questionnaire and instruments that assess mobility or cognitive function. Table 2 describes different validation types and important definitions.

Subsequent validation

The process described so far defines the steps for initial validation. However, the usefulness of the scale is the ability to discriminate between different cohorts in the domain of interest. It is advised that several studies investigating different cohorts or interventions should be conducted to identify whether the scale can discriminate between groups. Ideally, these studies should have clearly defined outcomes where the changes in the domain of interest are well known. For example, in subsequent validation of the Postoperative Quality of Recovery Scale, four studies were constructed to show the ability to discriminate recovery and cognition in different cohorts of participants (mixed cohort, orthopedics, and otolaryngology), as well as a human volunteer study to calibrate the cognitive domain.[46-49]

Sample size

Guidelines for the respondent-to-item ratio ranged from 5:1[50] (i.e., fifty respondents for a 10-item questionnaire), 10:1,[30] to 15:1 or 30:1.[51] Others suggested that sample sizes of 50 should be considered as very poor, 100 as poor, 200 as fair, 300 as good, 500 as very good, and 1000 or more as excellent.[52] Given the variation in the types of questionnaire being used, there are no absolute rules for the sample size needed to validate a questionnaire.[53] As larger samples are always better than smaller samples, it is recommended that investigators utilize as large a sample size as possible. The respondent-to-item ratios can be utilized

Table 2: Questionnaire-related terminology[16,44,45]

Terminology	Definitions
Construct	A model, idea, or theory that the researcher is attempting to assess (e.g., quality of postoperative recovery)
Validity	The ability of a questionnaire to truly measure what it purports to measure
Reliability	Reliability or reproducibility is the ability of a questionnaire to produce the same results when administered at two different points of time
Content validity	The extent to which a questionnaire measure includes the most relevant and important aspects of a concept in the context of a given measurement application
Face validity	The ability of an instrument to be understandable and relevant to the targeted population
Construct validity	The degree to which scores on the questionnaire measure relate to other measures (e.g., patient reported or clinical indicators) in a manner that is consistent with theoretically derived *a priori* hypotheses concerning the concepts that are being measured
Diagnostic validity	The accuracy of a questionnaire in diagnosing certain conditions (e.g., neuropathic pain)
Known-group validity	The ability of a questionnaire to be sensitive to differences between groups of patients that may be anticipated to score differently in the predicted direction
Criterion validity	The ability of a questionnaire to measure how well one measure predicts an outcome for another measure
Concurrent validity	The association of an instrument with accepted standards
Predictive validity	The ability of a questionnaire to predict future health status or test results. Future health status is considered a better indicator than the true value or a standard
Internal consistency	The degree of the inter-relatedness among the items in a multi-item questionnaire measure. It is usually measured by Cronbach's alpha
Repeatability (test-retest reliability)	The ability of the scores of an instrument to be reproducible if it is used on the same patient while the patient's condition has not changed (measurements repeated over time)
Responsiveness	The extent to which a questionnaire measure can detect changes in the construct being measured over time. It is applicable only for questionnaires that are designed to assess changes in the construct within a short period of time

to further strengthen the rationale for the large sample size when necessary.

Other considerations

Even though data collection using questionnaires is relatively easy, researchers should be cognizant about the necessary approvals that should be obtained prior to beginning the research project. Considering the differences in regulations and requirements in different countries, agencies, and institutions, researchers are advised to consult the research ethics committee at their agencies and/or institutions regarding the necessary approval needed and additional considerations that should be addressed.

Conclusion

In this review, we provided guidelines on how to develop, validate, and translate a questionnaire for use in perioperative and pain medicine. The development and translation of a questionnaire requires investigators' thorough consideration of issues relating to the format of the questionnaire and the meaning and appropriateness of the items. Once the development or translation stage is completed, it is important to conduct a pilot test to ensure that the items can be understood and correctly interpreted by the intended respondents. The validation stage is crucial to ensure that the questionnaire is psychometrically sound. Although developing and translating a questionnaire is no easy task, the processes outlined in this article should enable researchers to end up with questionnaires that are efficient and effective in the target populations.

Financial support and sponsorship

Siny Tsang, PhD, was supported by the research training grant 5-T32-MH 13043 from the National Institute of Mental Health.

Conflicts of interest

There are no conflicts of interest.

References

1. Boynton PM, Greenhalgh T. Selecting, designing, and developing your questionnaire. BMJ 2004;328:1312-5.
2. Crocker L, Algina J. Introduction to Classical and Modern Test Theory. Mason, Ohio: Cengage Learning; 2008.
3. Davis TC, Mayeaux EJ, Fredrickson D, Bocchini JA Jr., Jackson RH, Murphy PW. Reading ability of parents compared with reading level of pediatric patient education materials. Pediatrics 1994;93:460-8.
4. Bell A. Designing and testing questionnaires for children. J Res Nurs 2007;12:461-9.
5. Wong DL, Baker CM. Pain in children: Comparison of assessment scales. Okla Nurse 1988;33:8.
6. Stone E. Research Methods in Organizational Behavior. Glenview, IL: Scott Foresman; 1978.
7. Hinkin TR. A brief tutorial on the development of measures for use in survey questionnaires. Organ Res Methods 1998;2:104-21.
8. Harrison DA, McLaughlin ME. Cognitive processes in self-report responses: Tests of item context effects in work attitude measures. J Appl Psychol 1993;78:129-40.
9. Price JL, Mueller CW. Handbook of Organizational Measurement. Marshfield, MA: Pitman; 1986.
10. Harrison DA, McLaughlin ME. Exploring the Cognitive Processes Underlying Responses to Self-Report Instruments: Effects of Item Content on Work Attitude Measures. Academy of Management Annual Meetings; 1991. p. 310-4.
11. Embretson SE, Reise SP. Item Response Theory for Psychologists. Mahwah, N.J.: Lawrence Erlbaum Associates, Publishers; 2000.
12. Lindwall M, Barkoukis V, Grano C, Lucidi F, Raudsepp L, Liukkonen J, *et al.* Method effects: The problem with negatively versus positively keyed items. J Pers Assess 2012;94:196-204.
13. Stansbury JP, Ried LD, Velozo CA. Unidimensionality and bandwidth in the Center for Epidemiologic Studies Depression (CES-D) Scale. J Pers Assess 2006;86:10-22.
14. Tsang S, Salekin RT, Coffey CA, Cox J. A comparison of self-report measures of psychopathy among non-forensic samples using item response theory analyses. Psychol Assess. [In press].
15. Leung WC. How to design a questionnaire. Stud BMJ 2001;9.
16. Artino AR Jr., La Rochelle JS, Dezee KJ, Gehlbach H. Developing questionnaires for educational research: AMEE Guide No 87. Med Teach 2014;36:463-74.
17. Schultz KS, Whitney DJ. Measurement Theory in Action: Case Studies and Exercises. Thousand Oaks, CA: Sage; 2005.
18. Schmitt NW, Stults DM. Factors defined by negatively keyed items: The results of careless respondents? Appl Psychol Meas 1985;9:367-73.
19. Thurstone LL. Multiple-Factor Analysis. Chicago, IL: University of Chicago Press; 1947.
20. Churchill GA. A paradigm for developing better measures of marketing constructs. J Mark Res 1979;16:64-73.
21. Perneger TV, Courvoisier DS, Hudelson PM, Gayet-Ageron A. Sample size for pre-tests of questionnaires. Qual Life Res 2015;24:147-51.
22. Bowling A, Windsor J. The effects of question order and response-choice on self-rated health status in the English Longitudinal Study of Ageing (ELSA). J Epidemiol Community Health 2008;62:81-5.
23. Lee S, Schwarz N. Question context and priming meaning of health: Effect on differences in self-rated health between Hispanics and non-Hispanic Whites. Am J Public Health 2014;104:179-85.
24. Schwarz N. Self-reports: How the questions shape the answers. Am Psychol 1999;54:93-105.
25. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: Literature review and proposed guidelines. J Clin Epidemiol 1993;46:1417-32.
26. Beaton D, Bombardier C, Guillemin F, Ferraz M. Recommendations for the Cross-Cultural Adaptation of the DASH and Quick DASH Outcome Measures. Toronto: Institute for Work and Health; 2007.
27. Hendricson WD, Russell IJ, Prihoda TJ, Jacobson JM, Rogan A, Bishop GD, *et al.* Development and initial validation of a dual-language English-Spanish format for the Arthritis Impact Measurement Scales. Arthritis Rheum 1989;32:1153-9.
28. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine (Phila Pa 1976) 2000;25:3186-91.
29. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297-334.
30. Nunnally J. Psychometric Theory. New York: McGraw-Hill; 1978.
31. Streiner DL. Starting at the beginning: An introduction to coefficient alpha and internal consistency. J Pers Assess 2003;80:99-103.
32. Wilkinson L, the Task Force on Statistical Inference. Statistical methods in psychology journals: Guidelines and explanations. Am Psychol

1999;54:594-604.
33. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37-46.
34. Dawson B, Trapp RG. Basic and Clinical Biostatistics. 3rd ed. Norwalk, Conn.: Lange Medical Books; 2001.
35. Grootscholten C, Bajema IM, Florquin S, Steenbergen EJ, Peutz-Kootstra CJ, Goldschmeding R, *et al*. Inter-observer agreement of scoring of histopathological characteristics and classification of lupus nephritis. Nephrol Dial Transplant 2008;23:223-30.
36. Berry KJ, Mielke PW. A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. Educ Psychol Meas 1988;48:921-33.
37. Murphy KR, Davidshofer CO. Psychological Testing: Principles and Applications. Upper Saddle River, NJ: Prentice Hall; 2001.
38. Lawshe CH. A quantitative approach to content validity. Pers Psychol 1975;28:563-75.
39. Barrett RS. Content validation form. Public Pers Manage 1992;21:41-52.
40. Barrett RS, editor. Content validation form. In: Fair Employment Strategies in Human Resource Management. Westport, CT: Quorum Books/Greenwood; 1996. p. 47-56.
41. Alnahhal A, May S. Validation of the arabic version of the quebec back pain disability Scale. Spine (Phila Pa 1976) 2012;37:E1645-50.
42. Cronbach L, Meehl P. Construct validity in psychological tests. Psychol Bull 1955;52:281-302.
43. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Erlbaum; 1988.
44. Anthoine E, Moret L, Regnault A, Sébille V, Hardouin JB. Sample size used to validate a scale: A review of publications on newly-developed patient reported outcomes measures. Health Qual Life Outcomes 2014;12:176.
45. Reeve BB, Wyrwich KW, Wu AW, Velikova G, Terwee CB, Snyder CF, *et al.* ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. Qual Life Res 2013;22:1889-905.
46. Newman S, Wilkinson DJ, Royse CF. Assessment of early cognitive recovery after surgery using the Post-operative Quality of Recovery Scale. Acta Anaesthesiol Scand 2014;58:185-91.
47. Royse CF, Newman S, Williams Z, Wilkinson DJ. A human volunteer study to identify variability in performance in the cognitive domain of the postoperative quality of recovery scale. Anesthesiology 2013;119:576-81.
48. Royse CF, Williams Z, Purser S, Newman S. Recovery after nasal surgery vs. tonsillectomy: Discriminant validation of the Postoperative Quality of Recovery Scale. Acta Anaesthesiol Scand 2014;58:345-51.
49. Royse CF, Williams Z, Ye G, Wilkinson D, De Steiger R, Richardson M, *et al.* Knee surgery recovery: Post-operative Quality of Recovery Scale comparison of age and complexity of surgery. Acta Anaesthesiol Scand 2014;58:660-7.
50. Gorusch RL. Factor Analysis. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1983.
51. Pedhazur RJ. Multiple Regression in Behavioral Research: Explanation and Prediction. Fort Worth, TX: Harcourt Brace College Publishers; 1997.
52. Comfrey AL, Lee HB. A First Course in Factor Analysis. Hillsdale, NJ: Lawrence Erlbaum Associates; 1992.
53. Osborne JW, Costello AB. Sample size and subject to item ratio in principal components analysis. Pract Assess Res Eval 2004;9:8.