Validation of questionnaires, with emphasis on the Strengths and Difficulties Questionnaire (SDQ)

Introduction

We will focus on validation of questionnaires in this teaching. This is because several of us have translated questionnaires into another language or are planning to do so, and we know that there is a need to investigate whether the translated version of the instrument works well. The question is: How do we do this validation process? We will use the Strengths and Difficulties Questionnaire, the SDQ, as an example. Several of us have used this questionnaire to assess mental problems in children. It has been translated into almost 90 languages, and is extensively used, including in many child and adolescent psychiatric outpatient departments here in Norway.

Then, what is validation? Let me give you Last´s definition from "A dictionary of epidemiology" (Last 2001): Validation is "The process of establishing that a method is sound." This shows that validation is a broad concept in epidemiology. That the method is "sound" raises the next question: "Sound for what?" That is, we need to consider the use of an instrument to be able to tell if it is well validated. As Kane (2006) states it, one of the authorities on validation: "Validation involves the evaluation of the proposed interpretations and uses of measurements…It is not the test that is validated and it is not the test scores that are validated. It is the claims and decisions based on the test results that are validated…" We will return to this statement later on. But now, I just want to underline that validation of the SDQ is closely linked to how we use it, and how we interpret the results. Most times the SDQ is used as a screening instrument – we use it to screen for psychiatric problems in children and adolescents. Then it is very important to validate it as a screening instrument and know exactly which conclusions we can draw from the results we get.

Plan for the teaching

First, I will give you an overview of the validation process. Then I will give you some details about each step, before we will summarize what would be a good strategy for validation of a questionnaire like the SDQ.

Then, we will take a look at how the SDQ has been validated in different countries, and particularly in England by Goodman, where it was made. We will not go through all the SDQ validation studies available, as there are a lot of them. But those I have picked will at least give you an impression of the quality of the validation of this questionnaire around the world.

Overview of the validation process

Let me just mention that when I searched for relevant papers for our theme, I came across a paper from CIH. Peter Chipimo and Knut Fylkesnes published a paper in 2010 on Comparative validity of screening instruments for mental distress in Zambia (Chipimo & Fylkesnes 2010). Actually, this is one of the best validation studies I have come across so far. I recommend you to read it.

What are the main ingredients of a validation study? Let us think about the SDQ. Its original version is in English, and let us imagine that we consider translating it into Nepalese, because we need a screening instrument for mental problems among children and adolescents in Nepal, and as yet there is no authorized version of the SDQ in Nepalese. Let us look at the process:

First stage: Election and translation/adaptation of the instrument

1. Evaluate the usefulness of the SDQ for our need
2. Consider the aim of the translation/adaptation process
3. The translation/adaptation process itself

Second stage: Examining the validity of the instrument

1. Samples for validation
2. Different types of validity and reliability and how they are tested
    a) Criterion validity
    b) Content validity
    c) Construct validity
    d) Internal consistency
    e) Reproducibility
    f) Responsiveness
    g) Floor or ceiling effects
    h) Interpretability

A table summarizing the validation process

First stage: Election and translation/adaptation of the instrument

1. Evaluate the usefulness of the SDQ for our need

In this evaluation process we should answer the following questions:

a) Exactly what do we need an instrument for?

Often in a country like Nepal, one of the first instruments we would need is an effective screening tool, as professional health workers are scarce and resources in general are few. If we could screen for mental problems, then we could select for treatment those children who are most heavily affected.

b) Is there any instrument already translated into Nepalese that can serve the purpose?

If this is the case, we would save a lot of work and problems, as the translation and validation process is pretty demanding. We would be willing to compromise some regarding the qualities of the instrument, if it was already translated and tested.

c) If not, which are the "candidate instruments" available in English or in other languages that may be translated?

There are other instruments than the SDQ available, although nowadays it does not seem like it ☺ For example, more extensive questionnaires like the very well validated ASEBA system (Achenbach, CBCL, TRF, YSR). We should consult the literature as well as experts in the field before we decided which questionnaire to use. But it should be noted that the SDQ seems to be as good as the ASEBA system to identify children and adolescents with common psychiatric diagnoses.

d) Which of these candidate instruments will be the most useful for our purpose?

Again, what do we need the instrument for? Is the SDQ really the most useful questionnaire available for our purpose? The SDQ is useful, but it has its weaknesses. Are there other instruments that could be more useful now in Nepal?

e) Has the original version of this instrument been well enough validated?
When we have found the most promising instrument, then we need to take a careful look at the validation process. That is, in our case, has the SDQ been well enough validated in England? If not, we should be careful to choose it, because we don´t know enough about it – we don´t know how it really works.

2. Consider the aim of the translation/adaptation process.
You have already become aware that I use "translation/(slash) adaptation". This is because it is pretty common to talk about adaptation of an instrument instead of translation. We cannot just translate an instrument word by word into the new language and think that it will work. We need to adapt it to the new language and culture. This takes much more than simple translation – we have to work on it until the validation process shows that it works the same way as it does in the original version.

Therefore, before we move on to the practical things, we will need to take a look at some prerequisites for using a questionnaire in a different culture. We need to consider the aim of the translation. For example, the SDQ was designed in Britain. We want a Nepalese version that works exactly the same way in Nepal as the English version does in England.

What are the prerequisites for being able to use the SDQ in other cultures, such as the Nepalese? In this context we talk about equivalence of the questionnaire in the English and Nepalese cultures. I will list the four levels of equivalence that Hui and Triandis (Hui & Triandis, 1985) mention. When I have used their terminology, I need to inform you that there are a number of different concepts used about the different equivalence levels, and this is, of course, confusing. But I hope that you will grasp the main points.

1. Conceptual/functional equivalence
This is the first and most necessary requirement for cross-cultural comparison. Then, what does it mean? "A construct that can be meaningfully discussed in the cultures concerned is said to have cross-cultural conceptual equivalence." This means that a construct exists in both cultures, and that this construct is relevant, and acceptable and has similar meaning in the two cultures.

The example the authors give is the construct "weight". Weight lacks conceptual equivalence if you intend to compare a bushel (skjeppe, 36,4 liter) of oranges and love, because weight is irrelevant as an attribute of love. In our example, does peer problems as a construct exist in the Nepalese culture, and is it relevant and acceptable? Does it have similar meaning as in the British culture?

Conceptual equivalence is closely tied with functional equivalence, which in psychological research has to do with the similarity between the goals of different behaviours. Two acts of aggression are functionally equivalent across two cultures if people of both cultures emit such behaviours in certain situations, to achieve certain purposes. For example, do SDQ behavioural problems such as disobedience, stealing, lying and fighting have the same functions in the British and the Nepalese culture? Or peer problem items such as "rather solitary, tends to play alone", "has at least one good friend", "generally liked by other children", "picked on or bullied by other children" and "gets on better with adults than with other children"?

2. Equivalence in construct operationalization
Operationalization is the transition from theory to measurement. If a construct, such as

childhood peer problems, is operationalized in the same way in Britain and in Nepal, then the instrument that is made is equivalent in construct operationalization across these two cultures. In addition, the operationalization should be equally meaningful in the two cultures. Let me mention an example of lacking equivalence: Operationalizing aggression in terms of verbal insults would lack equivalence when the objective is to study aggressive behaviour between mute people and the general population.
In our case, the operationalization of peer problems in the SDQ should be equally meaningful in Britain and in Nepal. This also means that peer problems as operationalized in the SDQ is paid the same amount of attention in the two cultures.

3. Item equivalence

This more concrete and micro-level of equivalence presupposes the two mentioned types of equivalence. Let us assume that a construct has similar meaning in two cultures, and that it is operationalized in similar ways. Then the next consideration is that the construct has to be measured by the same instrument, in our case the SDQ. Only by doing this can cultures be compared numerically. That is, only then is it possible to compare levels of anxiety, and hyperactivity between Nepalese and British children. On the item level, the instruments used in the different cultures have to be identical. For example, each of the 25 items of the SDQ should mean the same thing to subjects in Britain and Nepal. If this is not the case, then the SDQ in effect represents two separate tests, one for each culture. If this happens, direct comparison of test scores is misleading and illegitimate.

My friend Kamran Salayev in Azerbaijan examined the factor structure of the parent version of the SDQ and found that the peer problems item "picked on or bullied by other children" loaded highest on the conduct problems subscale. Why? Probably because most parents thought that a child is picked on or bullied because it behaves badly towards other children. That is, it was perceived as a behavioural problem more than a peer problem. My question is: Does the Azeri version of the SDQ show conceptual, operationalization and/or item equivalence with the English version?

4. Scalar equivalence

Scalar equivalence is only present if the 3 other types of equivalence are present and if it can be demonstrated that the construct is measured on the same metric. This means that a numerical value on the SDQ scale refers to the same degree, intensity, or magnitude of the construct in both Britain and in Nepal. This type of equivalence is ideal for concrete cross-cultural comparison, but it is the most difficult to achieve.

Torbjørn Torsheim and I (Sanne et al, 2009) examined the parent and teacher versions of the Norwegian SDQ for measurement invariance. First of all, we had to find out if it was possible to do so, that there was scalar equivalence. With some reservations, we found that it was safe to compare the two versions, that there was scalar equivalence. However, congruent with an earlier study, we found that parents and teachers differed substantially in their ability to discriminate on one Peer problem item, namely "has at least one good friend". The same was the case for the Prosocial item "shares readily with other children". Parents answered more favourably than the teachers, probably because of what we call social desirability. The parents answered more favourably because they desired to give the more socially acceptable answer. They wanted to be able to communicate that their children had a friend, and that they shared readily with other children. But taking these two items into account, we found it sufficiently safe to compare the ratings of teachers and parents. And when we did that, it showed that the mean SDQ score was lower for the teacher version compared with the parent version for all factors. Kyrre Breivik and I are also working on a paper where we compare the scores of the British and the Norwegian SDQ. But my point here is just to stress that if we want to compare SDQ-scores between two countries or cultures, then we have to have all 4 levels of equivalence present. If not, a comparison of scores does not make sense.

As we finish this part about equivalence, let me just add that a clear demarcation of these four types of equivalence is not easy – there is considerable overlap between them.

3 and 4: How to demonstrate item and scalar equivalence?

Hui & Triandis (1985) as well as others (eg. Acquadro et al 2008) argue for the use of the item response theory (IRT) approach in the validation of an instrument. In this way it is possible to bypass the problem of selecting a relevant and unbiased criterion for judging an instrument. We will shortly come back to this as we talk about the different kinds of validity. But let me just say here that IRT uses item parameters derived "internally", and in this way we avoid the use of "external" criteria such as a "gold standard". Item characteristic curves (ICC), which represent the probabilities of responding to an item in a certain specified manner at different levels of the latent trait to be measured, are obtained from different cultures. Such differences can point to the lack of equivalence between the two cultures on a particular item. On the other hand, an instrument that has similar ICCs across cultures has, at least in part, demonstrated its item equivalence and scalar equivalence.

3. The translation/adaptation process itself

Then we have come so far that we want to translate the SDQ from English into Nepalese. This is a pretty resource-demanding and difficult part, that is, if we want to make a good translation. But a thorough translation will pay off, both in clinical practice and in research. However, we have to be aware that an instrument which has a very accurate translation is not foolproof. Therefore we have to test it afterwards, to find out whether the translation works as it should or not.

The possibility of non-equivalence on an abstract level, such as conceptual equivalence, cannot be neglected. Difference in social desirability levels and motivation to respond on the respondents´ part, and inconsistent test administration procedures are some other potential problems that can plague a study. Kimberlin & Winterstein (2008) reminds us that self-reports of behaviour are particularly subject to problems with social desirability biases. This is important to keep in mind when dealing with the S-SDQ.

Several authors have written about the translation process. I have used a review paper by Acquadro et al (2008) to summarize the most important elements. This means that I will not go into detail here, as it is an extensive theme. But every team deciding to translate an instrument should spend considerable time planning the process.

Most authors recommend a multistep approach involving a centralized review process. This means that there are some key people planning, leading and reviewing the whole process. However, Acquadro et al found that each research group in reality proposes its own sequence of translation events and weights each step differently. Some have argued for a shorter, faster and more inexpensive translation process in some situations (Mathias et al 1994), as most experts suggest pretty resource demanding procedures. However, most agree with Acquadro et al, arguing that a rigorous and multistep procedure leads to better translations. But they also underline that there is no empirical evidence in favour of one specific of these thorough methods. They give different examples of procedures that they think are acceptable and recommendable. Let me summarize one of these, in order to show you the most important ingredients and give you an example of how it can be done:

1. Permission from the author

In our case, we need to contact the creator of the SDQ, Robert Goodman. Goodman is a good man, but he is also a very busy man. Authorizing translated instruments is a great deal of work. Let me tell you what we experienced in Herat, Afghanistan. The SDQ has been translated into two of the main languages in Afghanistan, Dari and Pashto. In the city of Herat, most people speak Dari, but a different dialect than what they speak in the capital, Kabul. This means that there are certain words in the Dari version of the SDQ that people in Herat don´t understand. However, Herat is close to Iran, where they speak Farsi, and the Herat dialect of Dari is closer to Farsi than it is to Dari spoken in Kabul. But, even in the Farsi version there are words that people don´t understand in Herat. We would like to use the SDQ in research in Herat, and therefore I asked Goodman for permission to translate it into Dari, Herat dialect. But he did not give this permission, because he sees it as a problem if the SDQ is translated into various dialects. Also, the work for him during the translation process would be too much for him to accept. So, what do we do now?

2. Forward translation

At least two different translations done by bilingual translators are needed. The mother tongue of the translators should be the target language. That is, if we want to translate the SDQ from English to Nepalese, then the translators should be fluent in both languages and have Nepalese as their mother tongue.

In general, completing a questionnaire should not require reading skills beyond that of a 12-year-old. I know that Goodman has made sure that this is the case for the English version of the SDQ. However, when translating the SDQ into other languages, we have to make sure that the same is the case in the new culture. This is often even more important than in England and other Western countries, as the educational level in a country like Nepal is low, and we can expect large parts of the population to have a limited vocabulary in comparison to highly educated people.

3. Synthesis of the translations

The two translators should produce one translation out of the two versions.

4. Back-translation

Two back-translations are seen as a minimum, and those doing it should be blinded to the original version. They should have the source language as their mother tongue. That is, in our example, two persons with English as their mother tongue should back-translate the Nepalese version of the SDQ into English.

Let me add that it has been much discussed whether back-translation is necessary. As some state: Back-translation is “not the infallible quality control tool it is purported to be”. However, although this step has its weak sides, most authorities recommend including it in the translation process.

The Norwegian version of the SDQ was made by a team translating it from English into Norwegian, and by some others doing the back-translation. But then the adaptation process of this first version stopped, and the first back-translation became the final version. Some years ago, when I started doing factor analysis of the Norwegian SDQ, my first analysis showed that something was wrong with the first of the conduct problems items. In English this item says “Often has temper tantrums or hot tempers”. This had been translated “Har ofte raserianfall eller dårlig humør”. This item loaded higher on Emotional problems than on Conduct problems. Why? If we look at the last part of the item, "hot tempers" in English and "dårlig humør" in Norwegian, the Norwegian part can be back-translated "bad mood." That is, quite a few Norwegian respondents perceived this as an emotional problem more than a behavioural problem. If someone had done a simple examination of the factor structure in a small sample, this translation weakness would have been discovered. But when I found out, it was too late to change the translation, because it had already been used in the Bergen Child Study.

By the way, this item even in the original version is not optimal. Generally, items should be short, clear and simple, and should not consist of two questions at the same time. However, this item reads: "Often has temper tantrums or hot tempers". That is, this item is not as clear as it should have been.

5. Expert committee

This committee should be composed of methodologists, health professionals, language professionals, and all the translators involved in the process. The original developers of the questionnaire should be in close contact with the committee. The committee should study the process so far, and agree on a version that we could call the pre-final version. They should aim at achieving the two first levels of equivalence at this point, that is, conceptual equivalence and equivalence in construct operationalization.

6. Test of the pre-final version

Ideally, 30-40 persons should be tested. Each subject should complete the questionnaire and be interviewed about the meaning of each item. This stage provides a rough evaluation of content validity. We will return to the concept of content validity soon.

7. Submission of documentation to the developers or coordinating committee for appraisal of the adaptation process

This is a process to ensure that all steps have been performed and fully documented.

As I said, this is one example of a thorough translation process. There are many ways to do it, but one of the main points in all procedures is that each step in the translation is thoroughly reported in writing. In this way it will be possible to go back and identify weak spots in the translation process at a later stage.

The authors emphasize that the translation is only the start of the adaptation process. As mentioned, the pre-final version needs to be tested. That is, the examination of the psychometric properties starts.

References

Acquadro C, Conway K, Hareendran A, Aaronson N, for the European Regulatory Issues and Quality of Life Assessment (ERIQA) Group. Literature review of methods to translate health-related quality of life questionnaires for use in multinational clinical trials. Value in Health 2008, 3: 509-521.

Chipimo PJ, Fylkesnes K. Comparative validity of screening instruments for mental distress in Zambia. Clinical Practice & Epidemiology in Mental Health, 2010, 6: 4-15.

EMGO+ (Institute for Health and Care Research). Questionnaires: selecting, translating and validating. 01.01.2010. http://www.emgo.nl/kc/preparation/research%20design/8%20Questionnaires%20selecting,%20translating%20and%20validating.html

Fayers PM, Machin D. Quality of life. The assessment, analysis and interpretation of patient-reported outcomes. 2nd ed. Wiley, 2007.

Hui CH, Triandis HC. Measurement in cross-cultural psychology. A review and comparison of strategies. Journal of cross-cultural psychology 1985, 16: 131-152.

Kane MT (2006). Validation. In Brennan RL (Ed), Educational Measurement, 4th ed. Westport: Praeger Publishers, pp 17-64.

Kimberlin CL, Winterstein AG. Validity and reliability of measurement instruments used in research. Am J Health-Syst Pharm 2008, 65: 2276-2284.

Last JM. A dictionary of epidemiology, 4th ed. Oxford University Press, 2001.

Mathias SD, Fifer SK, Patrick DL. Rapid translation of quality of life measures for international clinical trials: avoiding errors in the minimalist approach. Quality of Life Research 1994, 3, 403-412.

Sanne B, Torsheim T, Heiervang E, Stormark KM. The Strengths and Difficulties Questionnaire in the Bergen Child Study: A conceptually and methodically motivated structural analysis. Psychological Assessment 2009; 21: 352-364.

Schmidt ME, Steindorf K. Statistical methods for the validation of questionnaires. Discrepancy between theory and practice. Methods Inf Med 2006, 4: 409-413.

Terwee CB, Bot SDM, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. Journal of Clinical Epidemiology 2007; 60: 34-42.