Published in final edited form as:

Sleep Med Rev. 2011 February ; 15(1): 7–17. doi:10.1016/j.smrv.2010.06.003.

Development of Pediatric Sleep Questionnaires as Diagnostic or Epidemiological Tools: A Brief Review of Do's and Don'ts

Karen Spruyt, Ph.D. and David Gozal, M.D.

Department of Pediatrics and Comer Children's Hospital, Pritzker School of Medicine, University of Chicago, Chicago, Illinois, USA

Abstract

Questionnaires are a useful and extensively used tool in clinical sleep medicine and in sleep research. The number of sleep questionnaires targeting the pediatric age range has tremendously increased in recent years, and with such explosion in the number of instruments, their heterogeneity has become all the more apparent. Here, we explore the theoretical and pragmatic processes required for instrument design and development, i.e., how any questionnaire, inventory, log, or diary should be created and evaluated, and also provide illustrative examples to further underline the potential pitfalls that are inherently embedded in every step of tool development.

Keywords

sleep; questionnaire; log; diary; child; reliability; validity; apnea; periodic limb movements; sleepiness

“Ask and ye shall be answered”

In 1884, the *quality of sleep*, or the inability to sleep, was perceived as a potential symptom of acute insanity and nervous exhaustion, whereas the *quantity of sleep*, or curtailing hours of sleep was suggested to imperil the integrity of the brain.⁽¹⁾ Ever since those days, attempts to document sleep by non-invasive and inexpensive methods have been intensively pursued. Consequently, questionnaires, scales, diaries, and logs have been and continue to be extensively applied. In the early ages such instruments were primarily developed for research data collection purposes, but lately some of the available tools have found their way to the clinical setting for diagnostic or outcome assessment purposes. Given their epidemiological or clinical value, customary use of these instruments as part of routine screening processes in clinical practice should enhance awareness of the importance of sleep to both the patient and the professional alike. Of note, it is still somewhat disconcerting how undervalued ‘nighttime’ behavior is when reflecting upon the abundance of standardized ‘daytime’ behavior assessment tools and the relative scarcity of such tools for the “dark

© 2010 Elsevier Ltd. All rights reserved

Correspondence to: David Gozal, M.D., Department of Pediatrics, Comer Children's Hospital, The University of Chicago, 5721 S. Maryland Avenue, MC 8000, Suite K-160 Chicago, IL 60637. Tel: (773) 702-6205; (773) 702-4523 – FAX; dgozal@uchicago.edu.

Karen Spruyt, Ph.D., Department of Pediatrics, Comer Children's Hospital The University of Chicago, 5721 S. Maryland Avenue, Chicago, IL 60637. Tel: (773) 702-6205; (773) 702-4523 – FAX; kspruyt@peds.bsd.uchicago.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of interest: The authors have no conflict of interest to declare pertaining to this manuscript.

hours". Similar to daytime diagnostic tools, the core principles on which development of tools aiming to obtain valid information on nighttime behaviors are similarly heavily dependent on the psychometric quality of the "subjective" `tool' (i.e., instrument, test, interview, measure, scale, diary, log, and questionnaire, hereafter referred to as "tool"). Gathering information on sleep quality or quantity might appear a priori a straightforward simple question-and-answer scenario, and indeed a multitude of custom-developed tools are currently used in the field of pediatric sleep. However, as will become apparent in this technical review, the design of a good tool is a complicated process that requires consideration of 11 major operational steps.

Development and design of tools

The development and design of tools (Figure 1 and 2) should be defined and described in sufficient detail and sufficient rigor to not only facilitate a fully-informed decision about whether to implement the tool in the particular task at hand, but further to enhance and enable accurate interpretation of the results collected using the tool. Therefore, it is imperative that researchers, practitioners and publishers should be as transparent as possible regarding the subjective tool(s) utilized and reported.

Step 1: Purpose

The use of a tool to measure knowledge, attitude, attribute, emotion, cognition, intention or behavior is cardinal to the social sciences. Additionally consideration needs to be given as to whether the tool will be used for data-collection, screening, or for diagnostic purposes. The very same considerations are needed in pediatric sleep research and practice because questionnaires, logs, diaries are considered as the source of subjective information, since they capture observations of the individual, and are commonly used to measure perceptions of sleep. To filter out as much as possible such `subjectiveness' one would need to adhere to the steps outlined in this review. In practice, researchers or practitioners usually generate questions, and responses to these questions are then converted into a numerical format to be statistically analyzed or interpreted. To achieve such a prior "easy" goals, researcher or practitioner either assumes or unequivocally wishes that underlying dimensions contained in the questions, such as language, similar meaning and interpretation of the wording, etc. will be identically perceived by the respondents, and therefore, minimize errors of comprehension and completion. As discussed in more detail below, the how – the which – the number – the type and the layout of questions included in the instrument being developed or evaluated, need to be reliably operationalized, and be relevant and acceptable to the targeted group. Thus, in the process of defining the purpose of the tool, it is critical to define the targeted sample as well, and to delineate the use of the tool as the first step of instrument development (or evaluation).

Step 2: Research Question

The use of a subjective tool is only suitable when the collected data can be applied to test or analyze the research question. In other words, is a questionnaire the best method of acquiring this information? The prerequisite of an instrument is being clear, unambiguous, and uniformly workable (2), which results from a time consuming, stringent and scrupulous preparation process, that needs to be highly adherent to ethical considerations as well.

Steps 3 and 4: Response format and Generation of items

Step 3 & 4A. The `how' of a question is a critical consideration in tool development, i.e., closed-ended questions or open-ended questions, or a mixture (see Figure 3 for several examples).

Closed-ended questions occur where the possible range of responses is pre-determined, and usually require a check of the range of answers, a numerical value, thereby allowing for creation of many options that include the close-ended question format. For example, *dichotomous* (or binary) *category items* are used for questions allowing only two answer choices, e.g., yes – no, and *multiple category items* offer three or more choices to a question (e.g., checklists). Alternatively, a respondent might be asked to *rank choices*, in which case, it is advisable to limit ranking to 10 choices. However, category items approaches may obscure the possible range within the answer category, and are therefore seen as fixed-category items, forcing the respondent to choose only between the offered categories. Less fixed approaches which allow the response-range to move along a continuum or linear scale, and still allow for control over the responses, include *scaled questions or items* (commonly known as rating 'scales'). The continuum can be detailed (e.g., several choice categories) or have few responses scattered along the linear scale (e.g., extremes are labeled) (see Figure 3a and b). Furthermore, this ordinal scaling creates a shift away from the more descriptive (percentage of subjects selecting each option within the item or question) towards a more quantitative analysis (responses to the item are treated as a continuous variable). The latter approach would then allow for expressing differences in responses in terms of the order and magnitude of a response. This trend towards potentially parametric data, and thus parametric analyses, has raised significant debate among statisticians with respect to the optimal amount of response categories along the continuum. As a rule of thumb, 5-, 7- and 9-point scales have been designated as preferable.⁽³⁾ Nonetheless, the numerical values within the ordinal scale refer to categories rather than to points along a continuum, as with interval or numeric data. This mistake is unfortunately a common, yet serious pitfall in many questionnaires (especially with ordinal scales that include fewer points). The Item-Response Theory (IRT) provides a framework on how the latent pattern (or scale) among the categories or items should be dealt with appropriately See Figure 4: Box 1 explaining the importance of response categories with respect to the validity and reliability of tools. A very frequently used and a familiar variant of such rating scales is the Likert scale item;¹ sometimes, numerical values are absent (i.e., only choice category is printed) but appended later on during data-entry. Alternatively, the quantitative equivalents are shown in the questionnaire, such that a latent continuum is 'inferred'. Usually with a Likert-type item, an odd number of response categories (typically 5) with a neutral central category are typically chosen [for example: positive-to-negative scale (Figure 3c) and 1-to-5 scale (Figure 3a)].

Some key comments:

- 1) If the item is perceived as interval data, the positive-to-negative approach is preferred since the zero point stands for 'absence'. In fact, and as already alluded to, keeping the 1-to-5 numerical value would implicitly indicate that value 3 reflects 'a stronger' response than values 2 and 1, although the differences in strength among response categories may not be equal. [see Montgomery-Downs et al. (4) for a clinical illustration of the presumed continuum, and the effect of 'arbitrary' cut-offs on a number of response categories with respect to sensitivity-specificity].
- 2) The item is perceived as ordinal whereby the numbers are treated as categories (unless IRT analyses are performed).
- 3) Labels, numerical values and how missing responses need to be treated are all items that warrant discussion in tool development and evaluation. For the same reason, changing response formats can not be done without critical reflection. Expert panels can also help in identification of the best response format. Other examples of ordinal

¹Likert type is commonly used to measure attitudes where the response is an indication of a degree of (dis)agreement. Consequently, this type often has a neutral category allowing negative – to – positive scaling. However frequency (ordered) scales (i.e., how often a target behavior occurred) are sometimes approached as a Likert-type scales.

scales are: *the Thurstone scales* where the behavior being measured is spaced along a continuum with equal spacing (Figure 3d), whereas *the Guttman scaling*, a hierarchical technique, ranks the measured behavior (when agreeing with one item you also agree with items of lower rank)(Figure 3e), and so forth (see Research and Methods handbooks within social sciences). Allow “don't know” and “not applicable” responses where appropriate because sometimes “don't know” or “not applicable” will really represent some respondents' most honest answers to some of the questions. Providing this answer choice might prevent non-response or abandonment of the survey.

The type of scale needs to be carefully considered since the forced choice may not only lead to dramatically different results, but additionally may prompt respondent irritation, or non-response bias. Furthermore regarding scaling, providing numerical (e.g., 1–5) values may indeed prevent the respondent to choose extreme numbers, but may also draw to the ‘referent’ (i.e., the middle value, known as central tendency or end aversion), or may lead to uncertainty as to how the response can or will be converted into a result (e.g., respondent circling 2 numbers). Presenting choice categories or labels (e.g., agree – disagree) might alleviate some of these issues, yet a comparable solution is *semantic differentials* where a number of bipolar categories are offered, and the choice of pole expresses the strength of response (see Figure 3f). Another familiar item is the *visual analogue scale*, in which an horizontal or vertical line of 100 millimeters in length is present along with(or without) an extreme being anchored at each end of the line (see Figure 3g). The response is converted into a numerical value via an ordinary ruler.

When the respondent is free to give any response, the item is considered an *open-ended question*. Also here variants exist, but we should keep in mind that such responses reflect the respondents' reference-frame. Unstructured items allow for answers to be provided in an unlimited fashion, which in turn, create a situation where anything can be expected. Using *word associations* or *sentence completion items*, in which the respondent is probed, and responses remain unstructured and uncontrolled constitutes another viable approach. Of special interest here are *interviews*, since the open-ended format may elicit a deeper response necessitating adequate training of the interviewers, as well as adequate coding or quantification of the answers. An advantage to the interview approach is that a completely unambiguous question to the researcher, and which might be interpreted differently by the respondent, can be clarified or corrected in a standardized fashion when an interviewer is present (for example, when double negatives are being used or the negating word was not captured by the respondent as may occur when filling out questionnaires).

Step 3 & 4 **B**. Another imperative is ‘**which**’ questions to ask. Expert or opinion panels, brainstorming sessions, interviews often preceded by appropriate literature review, as well as piloting the preliminary subjective tools are all cardinal elements of the due process required for tool development. Manuals or journal articles should discuss the process of item development with an expert panel (i.e., also disclose the experts and their expertise) and the process of validity and reliability measures. Equally informative can be groundwork data from documents, records, standardized scale items, and other miscellaneous sources. All of these should increase content validity, but mostly will uncover aspects previously not considered by the researcher or practitioner. Once a pool of relevant indicators that will address the research question being investigated has been assembled, the question wording, the question type, and the question order are all issues that need to be addressed. Although it would be ideal to ask about every aspect of sleep, this is clearly impractical. Therefore, to allow for valid prediction or description of nighttime behavior in children, the sample of items or questions must be representative of pediatric sleep, i.e. Step 3 & 4 **C**.; in terms of ‘**type**’ and ‘**number**’ of questions. As a rule of thumb, less than 20 minutes should be

necessary to complete the survey instrument, and thus prevent response fatigue or carelessness. (5) Questions should generally be less than 20 words in length, and should be simply and unambiguously worded (*see do's and don'ts synopsis* below) (5–9). Reliance on a single-item response is not recommended, and multi-item questionnaires are advocated and preferable such as to prevent bias, misinterpretation and measurement error. 'Tapping' the concept being investigated using several items gives a more comprehensive, reliable and representative view of the behavior being measured. On the other hand, attention should be given to whether the item is deemed as absolutely necessary, and where it should be inserted in the tool. In effect, step 3 & 4 A–C should be devised in the realm of the research question being asked, and paying attention to all of these details should keep the instrument close to target. Following these steps, it would be anticipated that a representative number of expected subscales or factors should be available, and can be evaluated in Step 7.

Do's:

1. Consider literacy (e.g., Reading Difficulty Formula (10), or Flesch-Kincaid reading)
2. Consider comprehension (e.g., FOG Index (11))
3. Consider ethnic and cultural background
4. Consider the age
5. Consider the target population
6. Consider alternative meanings of words selected, and beware of implicit assumptions
7. Word questions positively (e.g., are you satisfied with your child's sleep), but avoid 'leading' questions (e.g., in order to lengthen your adolescent child's sleep, should school times start later?) and avoid 'socially desirable responses' (e.g., Do you attach great importance to your child's sleep?)
8. If fixed format is chosen, provide responses that reflect potential opinion/responses exactly. Thus, consider a comprehensive list of alternative responses, although 'please specify' open category might sometimes be a useful contingency measure, or for factual questions the 'do not know' option can be provided. Foremost, is the selected response format the most suitable?
9. Make the tool aesthetically pleasing (e.g., large font, sufficient spacing, do not split questions over pages etc.)
10. Aim for a smooth, logical flow of items but place important questions at the beginning (as later responses might be influenced by previous questions or reduced interest results in non-response or carelessness)
11. Adding filter and/or funnel questions might maintain flow and interest (note that sometimes falsely the fast-track might be chosen) (e.g., Do you have a roommate or bed partner? Yes/ No. If you have a roommate or bed partner, ask him/her how often in the past month you have) or block by topic (12))

Don'ts:

1. Use jargon (e.g., apnea, hypnagogic, cataplexy)
2. Use unusual words (e.g., sleep hygiene, latency or assist (better to use 'help'))
3. Use acronyms, abbreviations or colloquialisms (e.g., PSG, biweekly)
4. Patronize by oversimplification (e.g., How well did you sleep last night?)

5. Make questions elitist (e.g., Taking into account the last week, how would you rate your child's sleep quality compared to the previous week?)
6. When possible avoid double barreled questions, or questions containing double negatives, or hypothetical questions. Questions being 'double barreled' or with an 'and' or an 'or' may be confusing and eventually the sentence parts may not be perceived equally, nor will the response help you in understanding to what part of the sentence it pertains (e.g., the child has frequent twitching or jerking of legs while asleep or often changes position during the night or kicks the covers off the bed)
7. Avoid controversial or emotive items at the beginning, conversely demographic and clinical data may be presented at the end²
8. Avoid asking unnecessary and irrelevant questions
9. Avoid being vague; i.e., 'regularly', 'often', 'rarely' etc. do not have to same meaning to us all
10. Avoid calculations

In short, the aim is to obtain unbiased information! It is advisable to take into consideration the fallibility of the human memory, and therefore the length and the proximity of the time period elapsed, as well as the saliency of the items.

Lay-out of tools

Needless to say presentation or the visual impact affects response, but also data entry. Ideally the natural reading style should be kept, and a mixture of many different response formats should also be avoided. The same applies to fonts. Some rule-of-thumbs: leave white space for open questions – do not print lines-, use colored sheets if several questionnaires are used, provide contact information, leave space for open comments, etc.. (for example, (8,13)). The same applies for computer-based administered tools.(14)

Order of questions

Since many questionnaires have closed-ended questions, their order is of greater importance than commonly assumed. Especially with binary category responses their number should be limited, and their order should be rigorously reviewed to reduce the likelihood of successful guessing. Implementing reversed items, or if needed a negatively worded items or items screening for social desirability or consistency could be inserted.

Modification of existing tools

When copying already standardized items from elsewhere, even the slightest change of scale (for example, 3- to 4-point scale) or change of wording (for example, 'compared to others' vs. 'overall', similarly with respect to response, e.g., 'often' vs. 'regularly') may cause not only difficulties in comparison of scales especially with respect to differences in time frame questioned, but more importantly may dramatically change the response, and affect the validity and reliability (as an illustration on how for example sensitivity and specificity can change depending upon cut-offs see, and Figure 4 (see Box 1 or reference (4)). Thus, even the smallest adaptations of existing questionnaires need to undergo pre-test of its structure, validity and reliability.

²Some authors prefer to commence with them to facilitate the process or increase confidence.

Translation of tools

Translating tools to other languages requires a process that will fulfill conceptual equivalence (e.g., how sleep is conceptualized, see (15) as an illustration) and item/content equivalence (i.e., validity or relevance of items, see (16) as an illustration), but also semantic equivalence, operational equivalence (i.e., the way data was collected), measurement (i.e., psychometric properties) and functional (i.e., equivalence throughout the development steps) equivalence need to be evaluated (see (17,18)). Forward-backward translation is a necessity, with ideally a bilingual expert panel evaluating the equivalences. Establishing an equivalent instrument is a tedious job, since language is subjective and highly culturally colored.

Step 5: Pilot

Pilot trials are mandatory for validation and reliability of any tool. Repeating the pilot runs to define and to refine the instrument is usually necessary. Additionally, pilot work allows assessment of other aspects related to the subjective tool; for example, bias in the administration or scoring, problems with distribution, difficulties in motivating respondents, data-entry issues, and most often piloting simply prevents 'disasters' with the actual data-gathering. Some authors suggest piloting the questionnaire in a sample of the targeted population while observing them, such that body language, facial expression might be informative.(6) The size of pilot sample is determined by several factors (19), that include the size of target population, the complexity of topic, the estimated refusal rate and other issues that may arise in the process. As a rule of thumb $n \leq 100$ is generally used for the final pilot study (or 5–10% of targeted population)(20–22).

Whilst step 3 & 4 A-C-results will help to judge whether or not a tool is adequate for its designated purposes or will suggest which tool may be the most appropriate, the pilot studies unequivocally will help to attune the tool as much as possible to the audience by tackling the 'how', the 'which', the 'type' and the 'number of items in the tool. Survey handbooks (7,8,23–29) guide and discuss these aspects; e.g., as simple as the questionnaire format, such that horizontal response format might be preferred over vertical response format, or the alignment of the response (e.g., check on the left or right, top or bottom of the response), while piloting may also indicate that for example a juxtaposed scale (Figure 3h), where multiple responses to one item are expected, should be discarded in favor of separate scales addressing the item. Data from pilot studies will further help screening for comprehension, language and clarity, as well as, the 'hello-goodbye effect' or faking bad to qualify, or desirability or 'faking good', which is responding to the items as perceived by the responder as to what constitutes desirable by the researcher. The latter relates to the fact that respondents learn throughout filling out questionnaires or interviews, or [(un)consciously] guess the hypothesis, which will undoubtedly affect their responses. Other aspects of 'subjective' tools, which should be considered, include cultural influences, and the primacy or recency effects in recall. Piloting and the next two steps which allow for amendments in the development of the tool should result in an unbiased tool subserving its designated purpose.

Step 6: Item-analysis and non-response analyses

This step is used to establish if items are indeed representative of the expected factors. Throughout piloting, items that lack clarity or discriminative power should be identified, on a sufficient sample size; e.g., comparisons of wording of items, comparisons between groups for certain items. Item-analyses and non-response analyses should be analyzed as well.

Item-analyses are performed to assess the performance of individual test items, separately and as a whole, based on the assumption that the overall quality of a tool depends on the quality of its items (being the question and response format). Descriptive analyses are key-

elements here (25–28,30,31), such as for example: mean and SD being the 'average' respondents response to an item calculated by adding the number of points earned by all respondents for that item, and dividing the total by the number of respondents, whereas the SD will express the dispersion of respondents' scores in that item; item-difficulty or discriminatory power might be the endorsement rate with $\geq 80\%$ and $\leq 20\%$ being weak discriminators, but also comments provided during the pilot studies should aid to this step. Some handbooks suggest that inter-item correlations or item-scale correlations (see further) should be used. If for example, certain subsamples, e.g., low income groups, are the non-responders, one can adjust by weighting up the results (i.e., assign a weight to items), or if item non-response remains low and at random, then imputation (i.e., the substitution of a value for a missing data point) could be a solution. However, every effort should be undertaken to have a representative pilot sample. The interested reader may find more detailed information in Item-Response Theory analyses handbooks with respect to details on item-analyses. (for example see reference (32)) Evidently the boundaries of item-difficulty (i.e., its differentiating power) depend upon the purpose of the tool. A rule of thumb is 85% 'easy' item, between 51–85% 'moderate' and $<50\%$ 'difficult' (31), although items could be incorporated, if despite being relevant to the research question they lacked in psychometric criteria. If such items are included, then this should be very clearly outlined in the manual.

Because the respondent does not want to appear as being ignorant or disappoint the researcher or practitioner, an undesirable response may then be provided, or alternatively, a non-response may occur due to the fact that the respondent has little or no knowledge of the item. Similar considerations with respondents versus non-respondents should be made for early to late responders, volunteer and non-volunteer respondents, and so forth. Of note, descriptive analyses of the characteristics of responders and non-responders are highly informative, and should be documented. Non-response or missing data-analyses should reveal answer-patterns, inadequate formulated items, etc..

Since maximizing the response rates is crucial, especially with mailed questionnaires, a suggestion is to code the questionnaires allowing for non-response analyses, as well as for efficient and low-cost follow-up or reminder mailings. Additionally, incentives, additional assurance of confidentiality, clear instructions and aims of the study, follow-up calls, and many other strategies should be considered to maximize response rates. As a reminder, non-respondents are as important as respondents at any stage of tool development. In fact, generalization of results is fraught with difficulties or impossibility to defend such assumptions, depending on the level of response (and non-response).

Step 7: Structure

The item-analysis and non-response analyses go together with pre-analysis checks such as skewness (the asymmetry in distribution of answers) and kurtosis (the "peakedness" in distribution of answers) of variables, correlation matrix (i.e., correlation >0.5), scatterplots, histograms, and principal component analysis (a mathematical procedure to reduce the number of variables and to detect structure in the relationships between variables). A rule of thumb here is that about 5 respondents are needed per item or >100 respondents for this step. (31) It is only rarely that a questionnaire is uni-dimensional, and one can identify this issue with principal component analysis (PCA), such that when unrotated, the majority of items should 'load' on the first component. This analytical technique is commonly used as a variable-reduction method. Weak items with loading scores <0.40 should be excluded, and guided by scree plots (i.e., a line segment plot that shows the fraction of total variance in the data as explained) and Kaiser criterion of eigenvalue (i.e., a scalar associated with a linear system of equations) >1 components can be defined. In fact, PCA is seen as an iterative process prior to factor analysis. Subsequent identification of underlying concepts, or subscales and verifying their stability should be done via exploratory factor analyses (EFA),

which is the next step in instrument development (33). Some authors use PCA and EFA interchangeably; however, the latter approach technically assumes an underlying construct. In the case of very large samples, random sub-sampling or bootstrapping (i.e., re-sampling methods) analysis are recommended towards psychometric evaluation. In technical terms, making sense of a factor structure can be achieved via varimax rotation, which is an orthogonal orientation in which the factors are not correlated, or conversely via oblimin rotation when they are thought to be related, all of which do not need to be mutually exclusionary, and depend on the purpose and research question.

When an adequate item set is found, the tool has now reached the final preparatory stages, i.e., steps 8–11. Indeed, any newly created psychometric tool undergoes a lengthy process with respect to its design as construction (i.e., content, format and looks), and subsequently, it should unequivocally meet the standards of *Validity and Reliability*, *Standardization*, and *Norms*, as discussed below.

Steps 8 and 9: Reliability and Validity

Although Validity and Reliability are recurring terms and type of analyses included in most manuscripts, they entail substantial effort in tool development, and are more often than not assigned relative secondary importance in a large number of publications. Of note, Validity is potentially more important than Reliability.

Reliability

Reliability does not imply validity, although a tool cannot be considered valid if it is not reliable. Reliability entails the consistency of measurement (e.g., Kuder-Richardson, Cronbach Alpha, split-half coefficients or Spearman-Brown coefficient). This can be *internal consistency* or the consistency within itself (e.g., Cronbach Alpha). The sensitivity of a tool in measuring time after time the same attribute under investigation should be quantified. In other words, the tool is consistent in its measurement when no change has taken place in the attribute, and similarly, the tool is consistent in measurement of the change that has taken place. Change then will reflect (true) change in the attribute and certainly not in the tool when applied in its standardized fashion. The simplest way of assessing reliability of a tool is test-retest (i.e., temporal stability), although such approach is not straightforward, because one needs to think about practice effects, recollection of answers by respondent (hence this would test memory and not the stability of the test), consideration that prior exposure sensitizes respondents to the attribute, and that the time interval between test-retest needs to be appropriate (e.g., to prevent the attribute from being changed when retested). Solutions to this problem can include *equivalent-forms reliability* (i.e., alternate form with immediate or delayed assessment), which are not always easy to find, and, *split-half reliability*, even if for the latter the administration (e.g. simultaneously), the scoring factors and limited number of items might jeopardize reliability assessment. Likewise, *the reliability of raters or scorers* (i.e., interrater reliability via kappa, product-moment correlation, intraclass correlation coefficient), who will be using the tools needs to be established. A rule of thumb regarding reliability coefficients should reach at least .80, whereas in preliminary stages of tool development .70 can be accepted.(34) Sattler (35) suggests that tests with reliabilities ≤ 0.6 are unreliable, >0.6 are marginally and >0.7 are relatively reliable, however this again is interpreted by authors depending upon the purpose. As a reminder, Cronbach Alpha is affected by the number of items; i.e., the alpha will increase when the numbers of items increase, such that one should be cautious when comparing alpha values between scales with different number of items and one should also avoid overreliance on alpha.

Validity

When a tool measures the attribute for which it is designed to measure, it is valid. In plain words, the resulting score has a meaning. Different types of validity exist each subserving different aspects of the 'it is doing what it claims to do'-criterion; namely, content validity, [face validity], criterion validity (further subdivided into concurrent validity and predictive validity), and construct validity. *Content validity* demonstrates the extent to which the tool includes or represents the attribute under investigation (i.e., the representativeness of the questions) and hence nothing else, which can be established via expert panels. In contrast, face validity which is not a true indicator of validity, and is certainly different from content validity, in that it only expresses what the tool appears to measure and not what is actually being measured. *Criterion-related validity* is demonstrated by comparison of the tool scores to another criterion (i.e., indicator of what the tool is trying to measure), and therefore includes a temporal relationship between the tool and the criterion (i.e., refers to the use of another measuring tool *that has been validated*). Concurrent validity indicates a simultaneous measuring in individuals who are known to show the attribute being measured, whereas for predictive validity, an independent follow-up measurement was performed to validate the tool. For example, if we want to measure success of sleep therapy, we would expect a high correlation between a total sleep problem questionnaire score and NPSG normal sleep diagnosis. Likewise, discriminant function analysis might be used to determine which variables discriminate between two or more occurring groups. For example, this approach will help predict who was versus who was not successfully treated. *Construct validity*, refers to the construct, that is the extent to which the tool measures the theoretical construct or concept (e.g., sleep duration) and is the most difficult to assess when developing a tool (see (36)).

There are no fast rules towards validity evaluation, and more recently classification accuracy statistics (e.g., Receiver Operator Curves, ROC) are additionally applied to guide this psychometric criterion, and is advocated when diagnosis is the purpose. Nonetheless, a minimum validity of >0.5 is expected, although some authors stress the importance of defining this in the realm of the purpose of the tool (see handbooks (34)). Therefore we recommend implementation of an already validated instrument (i.e., measuring a similar or contrasting construct) against which we can then test the instrument under development.

Of importance, the field is hampered by the lack of uniformity not only in terms of subjective tools, but also in terms of objective assessments. For example, no clear standard is determined for sleep disordered breathing or periodic leg movement disorder, and different centers commonly apply different diagnostic criteria, even if polysomnography has been considered as the golden standard. When the psychometric quality of a subjective tool is being discussed, one should certainly reflect on the criterion selected or applied, and hence question the fallibility of the objective tool with respect to the characteristic. Indeed, even though the golden standard might be objective, the cut-offs of such golden standard are chiefly based on clinical judgments and a 1-night assessment; e.g., maybe a tool-item concurs better with a different cut-off of the objective tool.

Step 10: Confirmatory analyses

Confirmatory analysis mostly refers to the process of verifying the stability of the aforementioned results. This is commonly done via an additional pilot study or in a random subsample(s). Sometimes a confirmatory factor analysis is regarded as a pre-validity check. Most often it is the step where the tool final psychometric quality is outlined. Once this last step is completed successfully, the tool development process is now ready for widespread usage.

Step 11: Standardize and develop norms

Through Standardization we aim to expel the notorious 'other factors' or extraneous variables affecting response, such as coding systems, question or response order, wording of the question etc.. In fact, this requirement pertains to the administration of the instrument, but also to the sampling plan, and the scoring of the tool. Since within pediatric sleep there is no pre-determined 'cut-off' of what constitutes normal sleep behavior, considering developmental and sociocultural factors, and therefore establishing norms would be beneficial for the field (note that this can be extrapolated to suggested subscale cut-offs, as they rely on the equivalence of items in diverse populations). However, despite the worldwide large number of questionnaire studies published, and the existence of objective 'golden standard' measures such as the PSG, 'a norm' or what is typical sleep duration for a child remains a vague notion. Using a normative sample that is reflective of the census demographics for the decade, and is also compatible with the place in which the tool was developed will enhance the likelihood of obtaining reliable incidence rates. Norms within clinical populations might be equally informative towards evidence-based practice.

Standardization, Norms, Validity and Reliability can be seen as the 'performance grades' of a tool, as such tools can pass or fail this grading and should *always be used in the realm of this grading*. In a nutshell, "subjective" tools can be time-effective, reaching many subjects at once, relatively inexpensive, mostly anonymous and thus confidential, etc. but respondents can misinterpret questions or answers. Other difficulties encountered with subjective tools potentially affecting response rates are the length, the time required to respond, lack of feedback (for the respondent unless self-addressed postcard to allow for additional information is provided), the possibility of incomplete data, the level of literacy of the respondent, just to name a few. If our goal is to validly question nighttime behavior, our "subjective" tools need to adhere to the same criteria as the daytime tools. In fact, the manuals accompanying most tools or manuscripts should provide sufficient details regarding their psychometric quality (or the outlined 11 steps).

In summary, by closely subscribing to the psychometric rules, we attempt to make the "subjective" tool as "objective" as possible, and guarantee that anyone using it will obtain reliable and valid results. Finally, when addressing the 'how', the 'which', the 'type', and the 'number' of items, and steps in development of a tool, consideration and planning on how the data will be analyzed is of paramount importance. Any statistical analyses should begin with normality of distribution of the data, and whether the data meet the underlying assumptions of the statistical technique or test chosen.

Armed with the above set of principles, the aim of this review was therefore:

- 1) To raise awareness to the issues pertaining to tool development and their psychometric evaluation, and to provide strategies that will enable researchers to design, develop or critically reflect on subjective tools within pediatric sleep, and
- 2) To describe the psychometric diversity of current subjective tools within pediatric sleep based on the 11 steps (see next manuscript), and to promote implementation of a collaborative study that would enable improved tool development in pediatric sleep and thus benefit children all over the world.

To summarize some practice points

In evidence based medicine it is vital that assessment tools adhere to a stringent psychometric evaluation, such as for example item-analyses, validity, reliability and norms. This approach will ultimately prevent difficulties and uncertainties in interpreting results, and will concomitantly favorably impact clinical or educational practices.

During the process of considering the use of an available instrument in either clinical or research projects or during the process of developing a custom instrument, careful adherence to the following principles will enable correct data acquisition, scoring, interpretation and comparison of results:

1. Assess the psychometric criteria based on which 'a tool' was developed and used
2. Apply adept statistics and adhere to guidelines of tool development
 - a. Fulfill the preliminary work - literature searches, piloting, equivalence, etc.
 - b. Decide on: how, which, number and type
 - c. Select the method of sampling (or distribution of your tool)
 - d. Pilot and revise, until the tool fulfill desired psychometric properties
 - e. Launch your study, and discuss initial findings
3. Remember that an unreliable tool cannot be valid; *however*, a reliable tool may still be invalid
4. Some items are by nature 'statistically' weak but might be clinical relevant, and need to be interpreted accordingly
5. Interpret results of the tool by its psychometric properties (strengths and weaknesses)
6. Remain critical (e.g., response format)
7. When reporting any findings using an instrument, the author should also provide tool and endorsement rates in a supplement

Research Agenda

To develop universally adopted and validated tools for sleep in children to:

- compile internationally collected data
- data-mine; perform item-analyses and meta-analyses on collected data
- launch international multi-center studies
 - form a international-expert panels
 - conduct pilot studies on pilot versions
- develop and provide a web-based questionnaire pool
- develop self-reporting tools
- develop sleep interventions

Abbreviations

ROC Receiver Operator Curves

References

1. Baker LW. SLEEP. Journal of Education 1884;20(18):2393–293.

2. Cohen, L.; Mannion, L. *Research Methods in Education*. 2nd ed. Croom Helm; New Hampshire: 1985.
- 3*. Lozano LM, Garca-a-Cueto E, Muiz J. Effect of the number of response categories on the reliability and validity of rating scales. *Methodology* 2008;4(2):73–9.
4. Montgomery-Downs HE, O'Brien LM, Holbrook CR, Gozal D. Snoring and sleep-disordered breathing in young children: Subjective and objective correlates. *Sleep* 2004;27(1):87–94. [PubMed: 14998242]
5. Rothman, KJ.; Greenland, S. *Modern Epidemiology*. Lippincott-Raven; Philadelphia: 1998.
6. Murray P. Fundamental issues in questionnaire design. *Accid Emerg Nurs* Jul;1999 7(3):148–53. [PubMed: 10693384]
- 7*. Choi BC, Pak AW. A catalog of biases in questionnaires. *Preventing chronic disease [electronic resource]* 2005;2(1)
8. DeVilles, R. *Scale development: Theory and Applications*. 2ed. 1991.
- 9*. Boynton PM, Greenhalgh T. Hands-on guide to questionnaire research: Selecting, designing, and developing your questionnaire. *Br Med J* 2004;328(7451):1312–5. [PubMed: 15166072]
10. Gunning, R. *The Technique of Clear Writing*. McGraw Hill; New York: 1952.
11. Flesh, R. *The Art of Readable Writing*. Harper; New York: 1949.
12. McColl E, Meadows K, Barofsky I. Cognitive aspects of survey methodology and quality of life assessment. *Quality of Life Research* 2003;12(3):217–8. [PubMed: 12769133]
- 13*. Bradburn, NM.; Sudman, S.; Wansink, B. *Asking Questions: The Definitive Guide to Questionnaire Design -- For Market Research, Political Polls, and Social and Health Questionnaires (Research Methods for the Social Sciences)*. Revised ed. Jossey-Bass; 2004.
14. Van Schaik P, Ling J. Design parameters of rating scales for web sites. *ACM Transactions on Computer-Human Interaction* 2007;14(1)
15. Howlett BC, Ahmad WIU, Murray R. An Exploration of White, Asian and Afro-Caribbean Peoples' Concepts of Health and Illness Causation. *New Community* 1992;18(2):281–92.
16. Herdman M, Fox-Rushby J, Badia X. A model of equivalence in the cultural adaptation of HRQoL instruments: The universalist approach. *Quality of Life Research* 1998;7(4):323–35. [PubMed: 9610216]
- 17*. Meadows K, Bentzen N, Touw-Otten F. Cross-cultural issues: An outline of the important principles in establishing cross-cultural validity in health outcome assessment. *Cross-cultural Health Outcome Assessment: A User's Guide*. 1996
18. Meadows K, Wisher S. Establishing cross-cultural validity in health surveys. *Survey Methods Newsletter* 2000;20:16–22.
19. Newell R. Sampling and distribution issues. *Nurse Researcher* 1993;1(2):33–43.
20. Lehtonen, R.; Pahkinen, E. *Practical Methods for Design and Analysis of Complex Surveys*. Wiley; Chichester: 1995.
21. Lyberg, LE.; Biemer, P.; Collins, M.; De Leeuw, ED.; Dippo, C.; Schwarz, N., et al. *Survey Measurement and Process Quality*. Wiley; New York: 1997.
22. Kish, L. *Survey Sampling*. Wiley; New York: 1995.
23. Fink, A. *How to conduct surveys: a step-by-step guide*. 4th ed. SAGE; Los Angeles: 2009.
24. Miller PV. Alternative Question Forms for Attitude Scale Questions in Telephone Interviews. *The Public Opinion Quarterly* 1984;48(4):766–78.
25. Ferketich S. Focus on psychometrics. Internal consistency estimates of reliability. *Research in nursing & health* 1990;13(6):437–40. [PubMed: 2270309]
26. Ferketich S. Focus on psychometrics. Aspects of item analysis. *Research in nursing & health* 1991;14(2):165–8. [PubMed: 2047538]
27. Ferketich S, Phillips L, Verran J. Focus on psychometrics. Development and administration of a survey instrument for cross-cultural research. *Research in nursing & health* 1993;16(3):227–30. [PubMed: 8497675]
28. Ferketich SL, Mercer RT. Focus on psychometrics. Aggregating family data. *Research in nursing & health* 1992;15(4):313–7. [PubMed: 1496155]

- 29*. Jenkins CR, Dillman DA. Towards a theory of self-administered questionnaire design. *Survey Measurement and Process Quality* 1997;165–96.
30. Feldt LS. The relationship between the distribution of item difficulties and test reliability. *Applied Measurement in Education* 1993;6(1):37–48.
31. Priest J, McColl E, Thomas L, Bond S. Developing and refining a new measurement tool. *Nurse Researcher* 1995;2(4):69–81.
32. Embretson SE, Reise SP. *Item response theory for psychologists*. Mahwah Lawrence Erlbaum Associates. 2000
33. Ferguson E, Cox T. Exploratory factor analysis: A user's guide. *International Journal of Selection and Assessment* 1993;1(2):84–94.
34. Strauss, E.; Sherman, EMS.; Spren, O. *Psychometrics in Neuropsychological Assessment*. 3 ed.. University Press; Oxford: 2006.
- 35*. Sattler JM. Issues related to the measurement and change of intelligence. *Assessment of Children: Cognitive Applications* 2001:160–82.
36. Messick S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist* 1995;50(9):741–9.
37. LeBourgeois, MK. Validation of the Children's Sleep-Wake Scale [Ph.D.]. The University of Southern Mississippi; United States -- Mississippi: 2003.
38. Sohn H, Rosenfeld RM. Evaluation of sleep-disordered breathing in children. *Otolaryngol Head Neck Surg* 2003;128(3):344–52. [PubMed: 12646836]
39. Maldonado CC, Bentley AJ, Mitchell D. A pictorial sleepiness scale based on cartoon faces. *Sleep* May 1;2004 27(3):541–8. [PubMed: 15164912]
40. Schriesheim CA, Castro SL. Referent effects in the magnitude estimation scaling of frequency expressions for response anchor sets: An empirical investigation. *Educational and Psychological Measurement* 1996;56(4):557–69.
41. Osinski IC, Bruno AS. Response categories in Likert-scale. *Categorias de respuesta en escalas tipo likert* 1998;10(3):623–31.

Figure 1.
Steps in tool development

Figure 2.
Design of tool

Figure 3a.
Example 1: I like to sleep

22. Mark the OVERALL physical activity level of your child during the 30 minutes before he/she falls asleep:

OSA-18 Quality of life survey (38)

I am sleepy:

Figure 3b.
Example 2: from Children's Sleep Wake Scale (37) [dissertation – General Sleep Information]

strongly agree	agree	neutral	disagree	strongly disagree
5	4	3	2	1
(+2)	(+1)	(0)	(-1)	(-2)

Figure 3c.
Example 3: I like to sleep

Figure 3d.

Example 4: Thurstone scale

A Thurstone scale mostly applies an equal-appearing interval which is accomplished by judges rating a large set of statements in terms of how much they agree with them. Each rated statement is plotted and, median score and interquartile range is determined. Subsequently the list of statements is sorted and one selects the set of statements that are at equal positions across the range of medians, and has the lowest interquartile range for each position. Maldonado et al (39) made a pictorial version.

I like to sleep

I like to sleep at home

I like to sleep at home in bed

I like to sleep at home in my bed

I like to sleep only at home in my bed

Figure 3e.

Example 5: Guttman scale

Items in a Guttman scale gradually increase in specificity. The respondent will agree with items up to a point. Responses can be dichotomous or Likert-type. Again judges will rate and subsequently one sorts the statements and judges agreeing, and select from that pool.

2. For each of the pairs below, mark the response that best describes how you felt overall today.

happy	① ② ③ ④ ⑤	sad
tense	① ② ③ ④ ⑤	relaxed
sleepy	① ② ③ ④ ⑤	alert
even tempered	① ② ③ ④ ⑤	mood swings
irritable	① ② ③ ④ ⑤	easy going
poor concentration	① ② ③ ④ ⑤	good concentration
tired	① ② ③ ④ ⑤	energetic
worried	① ② ③ ④ ⑤	care free
calm	① ② ③ ④ ⑤	jittery

Sleeping is:

Good	Bad
Easy	Difficult
Useful	Worthless
Challenging	Problematic
Interesting	Boring

Figure 3f.

Example 6: Sleep-wake Diary (Carskadon, Acebo, Manber 1992)

Neutral Extreme

Figure 3g.

Example 7: Please indicate with a cross on the line how sleepy your child currently is

Figure 3h.

Example 8: Indicate below how important and how satisfied you are with

Figure 4.

Box 1: The importance of response categories with respect to the validity and reliability of tools.(3)

Left: Printed from Lozano et al (3) the interrelation between number of response categories (here: number of alternatives: 2 – 9) and sample size in relation to Cronbach alpha (α). From the graph it is clear that reliability of the unidimensional scale decreases with fewer responses; i.e., fewer responses decreases the variability and as such the reliability (here expressed by Cronbach alpha). The percentage of variance explained by the first factor decreases as the number of responses is reduced, regardless of the correlations between the items and sample size.(3) This might be extrapolated to multiple dimensions or factors.

Right: Hypothetical visualization of the 'perceived' range of response (never to always/often/usually) with number of response categories.(40,41) The fewer response categories, the more forced a choice might 'sense'. In addition, assigning a label and a numerical value influences the psychometrics of a tool (e.g., the perceived range, the summated score – e.g., a zero). For instance, the label might blur the hypothetical equal spacing; when 3 responses are given 'poor' – 'good' – 'very good' the differentiation between good and very good might not always be clear-cut, for example respondents circling both (possible solution: replace 'very good' by 'excellent'). Another example, 'often', 'occasional' and 'sometimes' might have a different value (interpretation) from person to person, or even culturally (e.g., add the frequency when possible to avoid misperception and thus inaccurate responses). With hypothetically equal spacing (or perceived value) among responses 'sometimes'(*) would fall in the middle, but depending on the perception of the labels or the assigned numerical value substantial shifts might occur; i.e., horizontally being less or more distinctive, and vertically being gradual or abrupt. It is of utmost importance that each person in each study perceives the responses in identical ways.

An analogy would be a grade at school; i.e., mathematical they are at equal distance but some might perceive a 7 closer to 8 and 9, whereas other see it more towards 6, or the difference between a 6 and 7 on a 10 point grade might be perceived large whereas on a 20 point grade it might be perceived small.

A: 7-point scale from 1–7 with 1: never, 2: rarely, 3: occasionally, 4: sometimes*, 5: often, 6: almost always, and 7: always.

B: 5-point scale from 1 – 5 with 1: never, 2: just a few times, 3: sometimes*, 4: quite often, and 5: always.

C: 5-point scale from 0 – 4 with 0: never, 1: just a few times, 2: sometimes*, 3: quite often, and 4: always.

D: 5-point scale from 0 – 3 with 0: never, 0: rarely, 1: occasionally*, 2: often, and 3: very often.

E: 3-point scale from 1 – 3 with 1: never/rarely, 2: sometimes*, and 3: usually.

F: 3-point scale from 0 – 2 with 0: never/rarely, 1: sometimes*, and 2: usually.