IJAYUSH
International Journal of AYUSH
AYURVEDA • YOGA • UNANI • SIDDHA • HOMEOPATHY
https://internationaljournal.org.in/journal/index.php/ijayush/
ISSN: 2349-7025
OPEN ACCESS
★ PEER REVIEWED ★
MONTHLY

Review Article	Volume 15 Issue 11	November 2026

WHAT DOES AN *AGNI* QUESTIONNAIRE MEASURE? A CRITICAL REVIEW OF CONSTRUCT DEFINITION, RELIABILITY AND CLINICAL INTERPRETATION

Bhavendra Goyal[1], Dr. Yogesh Yadav[2*], Dr. Pushpendra Kumar Pandey[3], Dr. Ashish Choyal[4]

[1]Undergraduate Student, R.N. Kapoor Memorial Ayurvedic Medical College and Hospital, Dr. A.P.J. Abdul Kalam University, Indore, Madhya Pradesh, India.

[2]Assistant Professor, Department of *Shalya Tantra*, R.N. Kapoor Memorial Ayurvedic Medical College and Hospital, Dr. A.P.J. Abdul Kalam University, Indore, Madhya Pradesh, India.

[3]Professor and HOD, Department of *Sahita Siddhanta*, R.N. Kapoor Memorial Ayurvedic Medical College and Hospital, Dr. A.P.J. Abdul Kalam University, Indore, Madhya Pradesh, India.

[4]Associate Professor, Department of *Panchkarma*, R.N. Kapoor Memorial Ayurvedic Medical College and Hospital, Dr. A.P.J. Abdul Kalam University, Indore, Madhya Pradesh, India.

***Corresponding Author:** Dr. Yogesh Yadav, **Email:** *dryogeshyadav00@gmail.com*

DOI: https://doi.org/10.5281/zenodo.24505884

Abstract

Agni questionnaires translate a complex Ayurvedic construct into responses that can be scored, but a numerical score is useful only when its meaning is clear. This critical narrative review examines selected development studies to distinguish content coverage, internal consistency, reproducibility and responsiveness. Existing tools differ in length, item selection and classification approach. A recent *Agni* Assessment Scale adds exploratory factor analysis to this literature, while also reporting variation in internal consistency between subscales. These developments support continued validation rather than interchangeable use of instruments. We discuss why symptoms, dietary behaviour and a theoretical classification should not be collapsed into a single unexamined score. Researchers should specify the construct, recall period, language, scoring algorithm and intended use before selecting a tool. Changes during treatment require evidence of measurement error and responsiveness. The review proposes an interpretation framework for studies using *Agni* as an outcome, without claiming a physiological biomarker or universal reference standard.

Keywords: *Agni*; *Agnibala*; questionnaire; psychometrics; measurement validity; Ayurveda

Introduction

Agni is frequently invoked in Ayurveda research as an explanation of health and disease. Measuring it requires more than attaching numbers to familiar symptoms. Investigators need to decide whether they are classifying an Ayurvedic state, describing digestive complaints or tracking a response to treatment. These purposes overlap, but they are not identical.

Eswaran and colleagues developed a questionnaire to assess *Jatharagni*, while Singh and colleagues developed a self assessment tool for *Agnibala*.[1,2] Their existence makes structured assessment possible.

It does not make their scores interchangeable. A tool should be selected because its content and measurement properties fit the research question, rather than because its title contains *Agni*.

Review approach

Targeted searches of PubMed indexed records and publisher pages were conducted on 19 and 20 November 2026 using *Agni*, *Agnibala*, *Jatharagni*, questionnaire, validity and reliability. Selected instrument development studies were read alongside COSMIN measurement principles. This focused critical review examines interpretation and does not rank all available instruments or reproduce the results of a formal systematic review.

Sources were organised by the question they could answer: direct clinical evidence, formulation or measurement evidence, and general methodological guidance. Clinical reports were interpreted within their population, comparator and follow up; laboratory findings were used only for the measured property. Proposed explanations and research priorities are identified as interpretation rather than established treatment effects. The synthesis does not claim comprehensive retrieval, duplicate independent screening or a formal risk of bias assessment. Consequently, the examples support a critical argument but cannot determine the frequency of a reporting problem or a pooled estimate of benefit.

Table 1 Selected evidence and its interpretation

Source	Design or material	Direct contribution	Boundary of inference
Eswaran and colleagues[1]	Questionnaire development study	Operationalises *Jatharagni* domains and reports internal consistency.	Internal consistency alone does not establish responsiveness.
Singh and colleagues[2]	Self assessment tool development	Examines *Agnibala* assessment in healthy adults.	Performance cannot automatically be transferred to every clinical population.
Ashwathykutty and colleagues[3]	Scale development with exploratory factor analysis	Reports a 26 item scale and differing subscale performance.	Independent and clinical validation remain separate requirements.

Define the construct before choosing the score

A construct is the concept that an instrument is intended to measure. For an *Agni* questionnaire, the authors should state whether the construct concerns present functioning, a usual pattern or a classification within Ayurvedic theory. Questions about current appetite and questions about habitual food preferences may refer to different time frames. Combining them requires a clear rationale.

Item wording also determines what a respondent can report reliably. A question about a recent experience differs from one requiring knowledge of Ayurvedic terminology. Researchers should explain whether the questionnaire is intended for general patients, students trained in Ayurveda or clinician administration. Familiarity with the expected categories can influence responses.

The *Jatharagni* questionnaire study reported internal consistency for its domains.[1] That is one component of evaluation. Internal consistency alone cannot establish that an instrument captures the full theoretical construct or that its classification corresponds to a reproducible clinical judgement.

Internal consistency is not repeatability

Items can correlate because they ask similar questions. That does not prove that a person will receive a stable score when the underlying state has not changed. Test retest reliability requires repeated assessment under conditions where stability is plausible. Inter assessor reliability addresses a different issue when clinician judgement contributes to scoring.

The *Agnibala* development study illustrates the effort to make self assessment explicit.[2] Its findings should be interpreted in the population and setting used for development. Validation is not a permanent property attached to a questionnaire regardless of language, age or clinical condition. Evidence supporting one use may be insufficient for another.

Table 2 Measurement properties and the questions they answer

Property	Question addressed	Inference not justified alone
Content validity	Are the intended domains represented appropriately	Numerical precision
Structural validity	Does the observed structure support the proposed scoring	Clinical importance of a change
Internal consistency	Do relevant items behave coherently within a scale	Stability over time
Test retest reliability	Are scores reproducible in stable participants	Correctness of a diagnostic label
Measurement error	How much variation can occur without real change	Meaningfulness to the patient
Responsiveness	Can the tool detect the intended change	Causal treatment efficacy

COSMIN distinguishes these properties and provides a framework for evaluating their study methods.[4] A paper should therefore report which property was tested and avoid the blanket statement that a tool is fully validated.

What newer development adds

The 2026 *Agni* Assessment Scale study reported a 26 item structure with four subscales, using exploratory factor analysis and reliability testing.[3] The reported internal consistency differed between subscales. This is useful information because it draws attention to the performance of separate domains rather than only a pooled total.

Exploratory factor analysis is a development step. Independent confirmation, cross language assessment and examination in clinical populations answer additional questions. A factor structure that fits one sample should not be treated as proof that the construct has been fully established across settings. Nor should a low performing subscale be ignored because the theoretical categories are familiar.

Interpreting change during treatment

A reduction in digestive complaints can be meaningful without proving restoration of *Agni* as a whole. Investigators should report the affected items or domains and explain the intended interpretation. If the outcome is a shift between categories, the classification rule must be prespecified and the stability of that rule examined.

Repeated administration introduces practical concerns. A participant can remember earlier answers or change the way symptoms are interpreted after counselling. Investigators should specify the recall period, assessment schedule and assistance provided. If the language is translated, translation alone is insufficient evidence that scores have the same meaning.

Table 3 Information required when *Agni* is a study outcome

Domain	Reporting requirement
Purpose	Classification, symptom description or treatment response
Instrument	Exact name, version, language and source

Domain	Reporting requirement
Administration	Respondent, assistance, recall period and setting
Scoring	Item coding, missing responses and classification algorithm
Applicability	Evidence relevant to the recruited population
Change interpretation	Measurement error and basis for a meaningful difference

A theoretical construct needs an explicit measurement model

The first task is to decide whether the items are intended to reflect one underlying state or to describe several features that jointly define a clinical category. These are different measurement models. In a reflective model, related items are expected to behave coherently because they express a common construct. In a formative or descriptive index, diverse features may contribute without being highly correlated. Applying an internal consistency statistic without clarifying this choice can lead to misleading interpretation. A low correlation between clinically distinct features is not automatically a reason to delete them, and a high correlation is not sufficient reason to retain redundant questions.

Agni also appears in research at different levels of abstraction. A theory of digestive and metabolic functioning is broader than a questionnaire about recent hunger, bowel pattern and postmeal symptoms. The instrument necessarily samples an observable part of that theory. Its title should not be treated as proof that it captures the entire construct. A useful methods section names the domains represented and identifies important domains not measured. This makes the operational definition transparent and permits later researchers to test whether the selected items behave as intended rather than debating an undefined global score.

The intended decision should follow this definition. A tool used to describe a sample does not necessarily need the same evidence as a tool used to classify an individual or determine treatment. Classification requires a reproducible rule and attention to errors near category boundaries. Monitoring requires evidence that change can be detected and interpreted. A research instrument can be promising for one purpose while remaining unsuitable for another. COSMIN distinguishes measurement properties precisely because different uses require different questions to be answered.[4] The phrase validated questionnaire should therefore always be followed, implicitly or explicitly, by validated for what use and in whom.

Content coverage and participant understanding

Content validity begins before statistical analysis. The item pool should reflect the intended construct and the experiences of the people who will answer it. Expert consultation can establish relevance to Ayurvedic theory, but experts cannot fully determine whether patients understand the wording as intended. Patient interviews or cognitive testing can reveal ambiguous terms, double questions and assumptions about eating patterns. These steps are especially important when the questionnaire moves between professional and general populations. A student familiar with the expected *Agni* categories may interpret an item differently from a patient encountering the language for the first time.

Questions can inadvertently combine several observations. An item asking about hunger and comfortable digestion may be difficult to answer when only one is present. A frequency response may also be unclear if the underlying experience is episodic or related to particular foods. The recall period should be short enough to be answerable and long enough to represent the intended state. Neither a general impression nor a recent symptom count is inherently preferable. The choice must fit the construct. The report should explain that choice rather than using an available response format without considering its meaning.

Language adaptation requires more than literal translation. Food terms, expressions for appetite and descriptions of bowel experience can carry different meanings across settings. A translation process

should examine conceptual equivalence and participant comprehension. Even a careful translation does not automatically establish that the same score threshold has the same interpretation. Cross language measurement work is a separate task. Investigators should identify the exact version used and avoid modifying wording or response options while continuing to cite the original validation as if nothing changed. Small changes in phrasing can alter the question the participant actually answers.

What the development studies establish

The *Jatharagni* questionnaire report describes a 64 item instrument assessed in 500 volunteers, with items grouped into four *Agni* domains.[1] This is a substantial attempt to turn clinical concepts into structured responses. The report's emphasis on internal consistency is informative, but it should not be read as a complete demonstration of all measurement properties. A broad total coefficient can be difficult to interpret when the instrument contains theoretically distinct domains. Readers should examine the proposed structure and domain level results rather than assume that a high overall coefficient proves one coherent scale.

The *Agnibala* self assessment study represents another operational approach.[2] Its relevance should be judged through its actual items, scoring and intended population. It cannot be assumed equivalent to the *Jatharagni* questionnaire merely because both relate to digestive strength. A study using one tool and a later study using another may be measuring overlapping but different aspects of experience. Direct comparisons or linking studies would be needed before their scores could be combined. Converting each score to a percentage does not create equivalence between instruments with different content and category rules.

The 2026 *Agni* Assessment Scale development included an initial phase and exploratory analysis in a larger second sample, leading to a 26 item structure with four subscales.[3] The reported subscale coefficients ranged from 0.57 to 0.83, with the *Samagni* subscale below the authors' conventional reliability criterion. This variation deserves explicit discussion rather than being concealed by a favourable global description. It suggests that different parts of the instrument may require different levels of caution. Further confirmation in independent and clinically relevant samples would address questions that exploratory development cannot resolve on its own.

Reliability depends on the conditions of repetition

Test retest assessment assumes that the underlying state is sufficiently stable during the interval. That assumption is difficult if recent diet, illness, counselling or treatment changes the reported experience. The study should justify the interval and describe how stability was assessed. An interval that is too short may encourage recall of previous answers; an interval that is too long may allow genuine change. No single interval is suitable for every *Agni* construct. The design should follow whether the instrument aims to measure a habitual pattern, current functioning or a short term symptom state.

Reliability coefficients also depend on the variation within the sample. A heterogeneous group can yield a high relative reliability even when repeated individual scores differ enough to matter clinically. Absolute measurement error provides complementary information. For individual change, researchers need to know the likely range of variation when no true change has occurred. Reporting a correlation alone does not provide that answer. GRRAS guidance is useful for describing participants, raters, repeated measurements and the analytical approach in studies of agreement.[5] It complements the broader COSMIN framework rather than replacing instrument specific judgement.

Clinician assisted administration introduces another source of variation. Explanations can help participants complete a questionnaire, but different explanations can change its meaning. A standard assistance protocol should distinguish permitted clarification from interpretation of the answer. If clinicians assign categories from interview responses, interrater agreement becomes relevant. If

participants complete the form independently, readability and missing item patterns become more important. The administration mode should remain consistent across assessment occasions or any change should be documented. Otherwise, an apparent treatment response may partly reflect a change in how questions were asked.

Structural validity and the limits of a total score

Exploratory factor analysis can identify patterns of association among items, but interpretation depends on the sample, extraction approach and theoretical model. A factor should not be accepted solely because its label resembles a familiar Ayurvedic category. The items loading on it need a coherent clinical interpretation. Independent confirmation can then examine whether the proposed structure holds in another sample. If the structure changes markedly, researchers should investigate why rather than assuming the original arrangement is universally correct. Different populations may interpret the same experiences differently or display different symptom combinations.

A single total score may be inappropriate when subscales represent distinct or opposing patterns. Summing them can conceal a mixed profile or produce the same total from clinically different responses. If the instrument's published algorithm specifies separate domain scores, researchers should not invent a new combined score for convenience. Any alternative scoring rule requires justification and evaluation. The same caution applies to creating local mild, moderate and severe categories from arbitrary percentage ranges. Such thresholds may look orderly in a results table while lacking evidence that the boundaries correspond to meaningful differences.

Missing responses also affect scoring. A participant may skip a question because it is unclear, inapplicable or sensitive. Treating every missing answer as absence of the symptom can bias the result. The rule for incomplete forms should be prespecified and consistent with the instrument's guidance. Researchers should report how many forms were incomplete and whether missingness clustered in particular items. That pattern can reveal a content or administration problem. A complete dataset created through undocumented imputation is less informative than a transparent account of what participants could and could not answer.

Responsiveness and clinically meaningful change

Responsiveness concerns the ability to detect change in the intended construct, not merely the presence of a statistically significant before and after difference. A questionnaire may change because participants learn its vocabulary, reinterpret symptoms or wish to report improvement. These possibilities do not negate all observed change, but they require an appropriate design and interpretation. A responsiveness study should specify hypotheses about expected relationships with other relevant changes. It should also consider whether the instrument can distinguish stability from improvement. A treatment trial alone does not automatically provide a complete responsiveness evaluation.[4]

Measurement error and clinical importance are separate. A change can exceed expected random variation while remaining too small to matter to the participant. Conversely, a meaningful individual experience may be difficult to distinguish from error with an imprecise instrument. Establishing a meaningful change threshold requires an appropriate anchor and population. Researchers should not derive it solely from the instrument's maximum score or from an arbitrary percentage. Any threshold adopted from another setting should be justified. The result should remain an estimate with uncertainty rather than a permanent dividing line for every clinical use.

Category shifts are especially vulnerable near boundaries. A participant can move from one category to another after a very small score change, while another has a larger change without crossing a threshold. Reporting only category conversion can therefore distort the magnitude of response. Where appropriate, investigators should present the underlying domain scores and explain the classification algorithm. The

interpretation should identify whether the aim is symptom improvement, a change in theoretical pattern or a treatment decision. These outcomes should not be treated as interchangeable merely because the same questionnaire supplies the numerical data.

Relationship with physiological and clinical measures

An *Agni* score can be compared with dietary intake, gastrointestinal symptoms or a physiological measurement, but the expected relationship should be stated before analysis. A correlation with one measure does not establish identity with it. If a broad construct contains several dimensions, weak association with one laboratory value may be unsurprising. Selecting only favourable correlations after examining many candidates can create an exaggerated impression of validation. Studies should distinguish planned hypotheses from exploratory findings and report the latter as questions for confirmation rather than proof of a universal biological equivalent.

There may also be no accepted gold standard for the full *Agni* construct. In that situation, construct validation relies on a network of justified relationships rather than a single decisive comparison. A clinician's judgement can be informative, but it has its own reproducibility limits and should not automatically be treated as error free. Agreement with an untested reference does not solve the measurement problem. A transparent study describes the reference assessment, its independence from the questionnaire and the reasons it is relevant. This prevents circular validation in which the same symptoms define both the index tool and its supposed external standard.

A practical agenda for future studies

An immediate priority is publication of exact instrument versions, administration instructions and scoring rules in accessible supplements. This would reduce unacknowledged modifications and permit replication. Independent validation should then examine content, structure, reproducibility and change interpretation in the populations where the tool will be used. Studies should report unsuccessful items and uncertain findings alongside favourable coefficients. A shorter instrument may reduce burden, but shortening requires reassessment because removing items can alter content and scoring. Convenience alone cannot establish that the abbreviated version measures the same construct.

For treatment research, the safest interpretive language remains specific. A study can state that a named domain score changed under a defined intervention and describe the uncertainty. It should reserve broader statements about restoration of *Agni* for evidence that genuinely supports that interpretation. This is not a restriction on theoretical enquiry. It is a way to make theoretical claims testable and cumulative. When construct definition, administration and scoring are explicit, later studies can confirm, challenge or refine the result instead of producing incomparable numbers under a shared traditional label.

Data availability can further support this work when consent and privacy arrangements permit it. Deidentified item level data allow independent examination of scoring and missingness, while a published analysis plan helps distinguish confirmation from exploration. Where sharing is restricted, investigators can still provide the item dictionary and analytical decisions. These materials make measurement research cumulative and reduce reliance on a single favourable coefficient repeated in later clinical papers.

Instrument burden should be reported as part of feasibility. Completion time, requests for help and the proportion of unanswered items can reveal whether a questionnaire works in the intended clinic. A shorter form may be attractive for repeated assessment, but convenience does not establish equivalence to a longer instrument. Its content and scoring require evaluation after shortening. This is a practical measurement question that can be studied independently of whether a treatment changes the resulting score.

Choosing the evidence needed for a measurement claim

Terwee and colleagues proposed separate criteria for properties such as validity, reproducibility, responsiveness and interpretability.[6] The later COSMIN content validity work further specifies the importance of relevant, comprehensive and understandable items.[7] For an *Agni* questionnaire, these approaches favour a clearly defined intended use and direct investigation of how participants understand its language. A high internal consistency coefficient cannot establish that every clinically important aspect of the construct has been represented.

Agreement and reliability answer related but distinct questions. De Vet and colleagues explain that reliability depends partly on the variation between the people being measured, whereas agreement concerns the closeness of repeated measurements.[8] Bland and Altman's method comparison work likewise shows why correlation alone is inadequate for assessing interchangeability.[9] Repeated *Agni* scores should therefore be analysed according to the scale and the actual question, with attention to whether the participant's state was expected to remain stable.

When an intraclass correlation coefficient is appropriate, its model, definition and whether it represents a single rating or an average must be stated.[10] The Prototype *Prakriti* Analysis Tool provides a neighbouring Ayurvedic example of instrument development and pilot validation.[11] Its findings cannot validate an *Agni* questionnaire, but they reinforce the need to identify the specific tool, population and assessment process. Validation belongs to an interpretation of a score in a defined setting, not to an Ayurvedic term in the abstract.

Discussion

The strongest contribution of *Agni* measurement research is making assumptions visible. An investigator should be able to explain what a score represents and why it is appropriate for the intended use. This standard allows theoretical development and empirical testing to proceed together without forcing *Agni* into an unsupported single biochemical equivalent.

This review is selective and does not provide a head to head assessment of every instrument. Its conclusions concern interpretation rather than the superiority of one questionnaire. Future studies should investigate measurement properties in the populations where instruments are actually used and report uncertainty at the domain level.

Conclusion

An *Agni* questionnaire measures an operationalised construct through specific items and scoring rules. Its validity depends on the intended use and population. Clear construct definition and evidence of reproducibility and responsiveness are necessary before treatment related score changes can be given a broad clinical meaning.

References

1. Eswaran HT, Kavita MB, Tripaty TB, Shivakumar. Formation and validation of questionnaire to assess Jāṭharāgni. Anc Sci Life. 2015;34(4):203-209. doi:10.4103/0257-7941.159829.
2. Singh A, Singh G, Patwardhan K, Gehlot S. Development, validation, and verification of a self-assessment tool to estimate Agnibala digestive strength. J Evid Based Complementary Altern Med. 2017;22(1):134-140. doi:10.1177/2156587216656117.
3. Ashwathykutty V, Singh K, Shabdapriti G, Rao BC, Srikanth N, Acharya R. Development and preliminary psychometric evaluation of Agni Assessment Scale a measure of functional digestive capacity based on Ayurveda. Complement Med Res. 2026;33(3):199-210. doi:10.1159/000551010.

4. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res. 2010;19(4):539-549. doi:10.1007/s11136-010-9606-8.

5. Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol. 2011;64(1):96-106. doi:10.1016/j.jclinepi.2010.03.002.

6. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34-42. doi:10.1016/j.jclinepi.2006.03.012.

7. Terwee CB, Prinsen CAC, Chiarotto A, Westerman MJ, Patrick DL, Alonso J, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual Life Res. 2018;27(5):1159-1170. doi:10.1007/s11136-018-1829-0.

8. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol. 2006;59(10):1033-1039. doi:10.1016/j.jclinepi.2005.10.015.

9. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1(8476):307-310. doi:10.1016/s0140-6736(86)90837-8.

10. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med. 2016;15(2):155-163. doi:10.1016/j.jcm.2016.02.012.

11. Rastogi S. Development and validation of a Prototype Prakriti Analysis Tool (PPAT): Inferences from a pilot study. Ayu. 2012;33(2):209-218. doi:10.4103/0974-8520.105240.