Development of a Quality-Assessment Tool for Experimental Bruxism Studies: Reliability and Validity

Andreas Dawson, DDS, Odont Dr
PhD Student
Department of Orofacial Pain and Jaw Function
Malmö University, Malmö, Sweden

Karen G. Raphael, PhD
Professor, Oral and Maxillofacial Pathology, Radiology and Medicine
College of Dentistry
New York University, New York, USA

Alan Glaros, PhD
Professor, Kansas City University of Medicine and Biosciences
Kansas City, Missouri, USA

Susanna Axelsson, DDS, PhD
Associate Professor, Swedish Council on Health Technology Assessment
Stockholm, Sweden

Taro Arima, DDS, PhD
Assistant Professor
Department of Oral Rehabilitation
Graduate School of Dental Medicine
University of Hokkaido, Sapporo, Japan

Malin Ernberg, DDS, PhD
Professor
Section of Orofacial Pain and Jaw Function
Department of Dental Medicine
Karolinska Institutet, Huddinge, Sweden

Mauro Farella, DDS, PhD
Professor, Department of Oral Sciences
Faculty of Dentistry
University of Otago, Dunedin, New Zealand

Frank Lobbezoo, DDS, PhD
Professor, Department of Oral Kinesiology
Academic Centre for Dentistry Amsterdam (ACTA)
University of Amsterdam and VU University
Amsterdam, The Netherlands

Daniele Manfredini, DDS, PhD
Assistant Professor, TMD Clinic
Department of Maxillofacial Surgery
University of Padova, Italy

Ambra Michelotti, DDS, PhD
Professor, Section of Orthodontics and Clinical Gnathology
School of Dentistry, Oral, Dental and Maxillo-Facial Sciences
University of Naples "Federico II", Naples, Italy

Peter Svensson, DDS, PhD, Dr Odont
Professor, Section of Clinical Oral Physiology
Department of Dentistry
Aarhus University, Aarhus, Denmark

Thomas List, DDS, PhD
Professor, Department of Orofacial Pain and Jaw Function
Malmö University, Malmö, Sweden

Correspondence to:
Dr Andreas Dawson
Department of Orofacial Pain and Jaw Function
Faculty of Odontology, Malmö University
SE-205 06 Malmö, Sweden
Fax: +46 40 6658420
Email: andreas.dawson@mah.se

Aims: *To combine empirical evidence and expert opinion in a formal consensus method in order to develop a quality-assessment tool for experimental bruxism studies in systematic reviews.* ***Methods:*** *Tool development comprised five steps:* (1) *preliminary decisions,* (2) *item generation,* (3) *face-validity assessment,* (4) *reliability and discriminitive validity assessment, and* (5) *instrument refinement. The kappa value and phi-coefficient were calculated to assess inter-observer reliability and discriminative ability, respectively.* ***Results:*** *Following preliminary decisions and a literature review, a list of 52 items to be considered for inclusion in the tool was compiled. Eleven experts were invited to join a Delphi panel and 10 accepted. Four Delphi rounds reduced the preliminary tool—Quality-Assessment Tool for Experimental Bruxism Studies (Qu-ATEBS)— to 8 items: study aim, study sample, control condition or group, study design, experimental bruxism task, statistics, interpretation of results, and conflict of interest statement. Consensus among the Delphi panelists yielded good face validity. Inter-observer reliability was acceptable (k = 0.77). Discriminative validity was excellent (*phi coefficient *1.0;* P < *.01). During refinement, 1 item (no. 8) was removed.* ***Conclusion:*** *Qu-ATEBS, the seven-item evidence-based quality assessment tool developed here for use in systematic reviews of experimental bruxism studies, exhibits face validity, excellent discriminative validity, and acceptable inter-observer reliability. Development of quality assessment tools for many other topics in the orofacial pain literature is needed and may follow the described procedure.* J OROFAC PAIN 2013;27:111–122. doi: 10.11607/jop.1065

Key words: bruxism, Delphi technique, masticatory muscles, pain measurement

Systematic reviews are a cornerstone in evidence-based medicine. A systematic review is a compilation of all published research over a defined time period that addresses a carefully formulated question. Results of included studies are collected based upon predetermined inclusion and exclusion criteria; data are then critically analyzed and synthesized so that evidence-based conclusions on the benefits or disadvantages of a certain treatment or test of the issue at hand can be drawn.[1] Systematic reviews have been used in numerous areas to assess, for example, treatment efficacy, diagnostic accuracy, education, and experimental human research.[1–8]

In pain research, human experimental pain models are essential for improving our understanding of pain mechanisms and pathogenesis, with the final goal of translating these findings into improved patient care.[9] Experimental pain models that mimic aspects of the clinical pain condition make it possible to study pain as an isolated phenomenon under controlled settings.[10] However,

human experimental pain models have limitations. Because the complexity of the clinical pain condition is much higher than in experimental pain, human experimental pain models are generally unable to capture the total complexity of clinical pain.[11] Manfredini et al observed an association between pain and psychosocial disorders in patients with clinical pain, such as temporomandibular disorders (TMD).[12] Other researchers have identified associations between TMD and anxiety[13] and depression.[14] Psychological factors are an important aspect of clinical pain that are not possible to capture with human experimental pain models.

Several human experimental pain models with special emphasis on jaw muscle pain after exercise have been developed to gain better understanding of how muscle exercise affects jaw muscle pain.[15–26] A literature review of experimental bruxism studies revealed large methodological variations in the type of jaw muscle exercise, intensity, duration of bruxism task, outcome measures, exercise days, and number of follow-up days.[6] No criterion standard currently exists for assessing bruxism experiments, but one is needed because findings between studies are difficult, if not impossible, to compare. An experimental bruxism model should *(1)* represent a standardized technique for inducing jaw muscle pain that mimics the clinical pain conditions, ie, pain-related variables should be similar to those in patients with persistent muscle pain in the orofacial region; *(2)* exhibit good reproducibility; and *(3)* have good within- and between-session variability. If variability in the pain-related outcome measures is high, then the effect of a particular experimental bruxism task could be camouflaged and it would be difficult to interpret the effects of the task; this would limit generalizability.

Due to varying standards among published articles, quality assessment could be said to be the keystone of a systematic review. Study results cannot be judged to have a high level of evidence or contribute substantially to a review's conclusions and recommendations if quality standards defined as lack of bias, applicability, and good reporting and design are not met. There is currently a lack of systematically developed and evaluated tools for assessing experimental bruxism. This has created a need for a platform of measures that can be used in clinical experimental bruxism trials and which would allow results to be compared between studies and general conclusions drawn. This study aimed to combine empirical evidence and expert opinion in a formal consensus method in order to develop a quality-assessment tool for experimental bruxism studies in systematic reviews.

Materials and Methods

Streiner and Norman's five-step method for developing quality-assessment tools—*(1)* preliminary decisions, *(2)* item generation, *(3)* face-validity assessment, *(4)* assessment of reliability and discriminative validity, and *(5)* refinement of the final instrument—was followed in this study.[27]

Preliminary Decisions

The steering group comprised two of the authors (AD and SA), who defined the desired characteristics and purpose of a quality-assessment tool: the tool should *(1)* be suited for use in systematic reviews of experimental bruxism studies, *(2)* be able to assess the methodological quality of the study in generic terms (relevant to all experimental bruxism studies), *(3)* be easy to understand and respond to and quick to use, and *(4)* contain a maximum of 10 items.

Item Generation

A search of PubMed was conducted with the following MeSH terms: Masseter Muscle AND Pain Measurement AND Bite Force OR Isometric Contraction AND Masticatory Muscles. The search was limited to articles published in English from 1970 to January 11, 2011. A hand search was also done. Following a review of 16 articles, one author (AD) compiled the first preliminary tool.[15–26,28–31] These articles were selected because they represented various experimental bruxism models for jaw muscle pain, ie, tasks that consisted of clenching, grinding, or protrusive movements, with electromyography or a bite force transducer.

Face-Validity Assessment: Delphi Procedure

A Delphi procedure was chosen to assess face validity. Eleven experts were asked to participate on a Delphi panel. Participants with varying backgrounds and perspectives, and with extensive clinical and research experience for at least 10 years in this field, were considered. Two authors, AD and TL, compiled a list of researchers to be invited to participate on the Delphi panel. Invitations to participate were sent by email and included detailed information about the study.

The Delphi procedure continued for 4 rounds when a consensus was reached on the 10 or fewer items to be included in the quality-assessment tool, their meaning, and how to rate them.[32] Figure 1 is a schematic flow diagram of the study design.

Round 1. The steering group sent to the panelists the preliminary tool, created in Item Generation, and instructions which included a description of the ideal properties of a quality-assessment tool.

• *Panelists*. Delphi members were asked to read the instructions and then, while keeping in mind the ideal properties of a quality-assessment tool, rate the items proposed for inclusion in a preliminary tool on a 5-point Likert scale (1, Strongly Disagree; 2, Moderately Disagree; 3, Neutral; 4, Moderately Agree; 5, Strongly Agree). The panel members were also asked to make free-text comments, rephrase items, or add new items as needed.

• *Steering Group*. Feedback and questionnaire results were tabulated and used to revise the preliminary tool. Items with ≥ 80% agreement (≥ 8/10 panelists rated the item Strongly Agree) remained on the tool. These items were rephrased according to the feedback and included. Items for which there was ≥ 80% disagreement (≥ 8/10 panelists rated the item Strongly Disagree) were excluded. All other items were considered undecided and put on the undecided list for re-rating in subsequent rounds, together with any free-text suggestions for new items.

Round 2. The steering group sent the revised preliminary tool, the undecided list, a tabulated summary of panel member responses and free-text feedback, and instructions to the panelists.

• *Panelists*. All members were asked to read through the tabulated summary and then *(1)* approve or suggest new phrasing for the items on the preliminary tool, *(2)* reassess the items on the undecided list by using the response options Yes, No, or Undecided, and *(3)* make free-text comments as needed.

• *Steering Group*. Feedback and questionnaire results were tabulated and used to revise the preliminary tool. Items with ≥ 80% agreement (≥ 8/10 panelists rated the item Yes) were included on the tool. Items with 70% agreement (7/10 panelists rated the item Yes) were placed on the undecided list. Items included on the tool and on the undecided list were merged according to feedback. Remaining items were struck from the undecided list. The steering group suggested a name for the tool: Quality-Assessment Tool for Experimental Clenching studies (Q-TEC). A Yes, No, or Unclear rating system was proposed for the tool. Yes answers would be assigned 1 point; No and Unclear answers would receive no points.

Round 3. The steering group sent the revised preliminary tool, suggestions for a tool name and a rating system, a tabulated summary of panel member responses and free-text feedback, and instructions to the panelists.

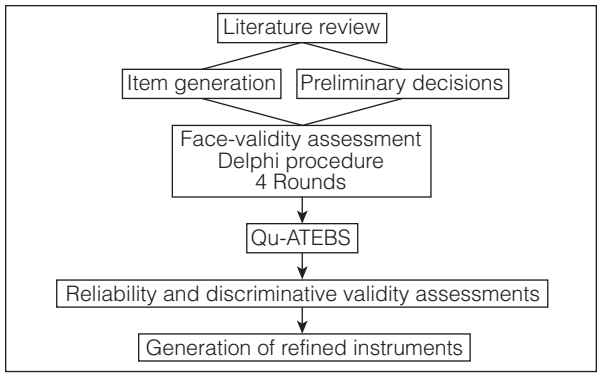

Fig 1 Schematic flow diagram of the study design.

• *Panelists*. As in previous rounds, panel members were encouraged to read through the tabulated summary before: *(1)* approving or suggesting new phrasing for the items on the revised preliminary tool, *(2)* approving the proposed name, Q-TEC, or suggesting a new one, *(3)* accepting the suggested rating system or proposing a new one, and *(4)* making free-text comments.

• *Steering Group*. Feedback and questionnaire results were tabulated and used to revise the preliminary tool. Items were rephrased and merged, item rating was revised, and a new tool name was suggested based on the feedback. Items with high agreement (≥ 80% panelists rated the item Yes) were not rephrased. Remaining items were changed according to feedback.

Round 4. The steering group sent the revised preliminary tool, a detailed explanation of all revisions made to the tool, and instructions to the panelists.

• *Panelists*. As in previous rounds, panel members were encouraged to read through the explanation of the revisions before: *(1)* accepting the items on the preliminary tool or suggesting new phrasing, *(2)* approving or suggesting a new name for the tool, *(3)* approving or suggesting a new rating system, and *(4)* making free-text comments.

• *Steering Group*. Feedback was used to revise the preliminary tool. No changes were made to items for which there was high agreement (≥ 8/10 panelists rated the items Yes). All other items were rephrased according to the feedback.

Assessment of Reliability and Discriminative Validity

The steering group decided to include a small sample of published articles in the reliability and discriminative validity testing. To assess discriminative validity, two investigators (TL and ME) reviewed 11 studies that were randomly chosen from the PubMed search results and the results of the manual

Table 1 Characteristics of the Delphi Panel Members

	No.
Sex	
Female	3
Male	7
Profession*	
Psychologist	2
Orofacial pain specialist	8
Orthodontist	3
Orofacial pain researcher	10

*A panel member may have more than one profession.

handsearch.[15,16,18,24,26,33–38] Five of these 11 studies were also a part of the item generation. TL and ME used the checklist for manuscript review of the *Journal of Orofacial Pain* and considered the following questions: *Is the study design clearly expressed? Is the study design acceptable? Is the material presented logically and technically accurate? Is the experimental material adequate and/or the study population appropriate? Is there a control group? Are the conclusions reflective and appropriate of the results?* and *Are the authors' thoughts clearly expressed?* Study quality was rated as high or low.

Two other authors (AD and SA) assessed the same articles using the present study's quality-assessment tool. The results from the two teams were compared (TL and ME vs AD and SA), and discriminative validity was assessed. Inter-observer reliability was assessed by comparing the results of AD and SA. One of the authors (AD) read the articles on two different occasions with a time interval of 3 months in order to assess the intra-observer reliability.

Refinement of the Final Instrument

Based on the results of the reliability and discriminative validity assessments, the instrument was refined. If the instrument was not able to discriminate between high- and low-quality studies, or if the inter-observer reliability was low, then the instrument was refined.

Statistical Analyses

Inter-observer reliability was assessed by calculating the kappa value between AD and SA. To assess the intra-observer reliability of Qu-ATEBS, the kappa value was calculated. The maximum possible score was 80 points. In a consensus discussion between AD and SA, it was decided that a score between 0 and 50 could be considered low quality and a score between 51 and 80 considered high quality. To determine discriminative validity, the phi coefficient was calculated to measure the degree of association between the results from TL and ME and those of AD and SA. Statistical analyses were two-tailed and set at the 5% significance level. All statistical calculations were performed using IBM SPSS, Windows, version 20.

Results

Item Generation

Based on the literature review, 52 items were generated, each phrased as a question. The items were subdivided into 6 categories: study sample, study design, statistics, instruments, baseline characteristics, and outcome measures.

Face-Validity Assessment

Ten of the 11 experts invited to join the Delphi panel agreed. The number of years of experience in orofacial pain research ranged from 10 to 35 years, based upon publications in scientific peer-reviewed journals. Table 1 presents the characteristics of the Delphi panel members.

Round 1. All panelists completed and returned the questionnaire. Two items had ≥ 80% disagreement and were omitted. Five items had ≥ 80% agreement and became the preliminary tool: 13, 16, 23, 29, and 36 (Table 2).

Forty-five items had < 80% agreement and < 80% disagreement. These formed the undecided list, along with 14 new items suggested by the panelists: items 53 through 66 (Table 2). Due to the length of the undecided list (59 items), the steering group decided not to rephrase or merge any items until round 3.

Round 2. Nine of 10 panelists completed and returned the questionnaire. Four items on the undecided list had ≥ 80% (≥ 7/9 panelists) agreement and were added to the preliminary tool: 53, 54, 65, and 66 (Table 2).

Forty-eight items had < 70% (< 6/9) agreement and were excluded. No new items were suggested, which left seven items on the undecided list with 70% agreement: items 1, 3, 5, 6, 19, 27, and 60 (Table 2).

In their free-text comments, the panelists suggested merging items within a category and making the items more general and less specific. The steering group revised the preliminary tool according to these suggestions and thus transferred items from the undecided list to the preliminary tool, so that it contained 10 items:

Table 2 Results from Delphi Rounds 1 and 2

No.	Item	Round 1			Round 2		
		In†	Un	Ex‡	In§	Un	Ex
1	Was recruitment of participants sufficiently described?		X			X	
2	Did participants receive financial compensation for participation?			X			
3	Was the study sample sufficiently described?		X			X	
4	Was mean age presented?		X				X
5	Did participants undergo a standardized clinical examination?		X			X	
6	When necessary, were participants given a diagnosis?		X			X	
7	Was a control group used?		X				X
8	Was the control group matched according to age and sex?		X				X
9	If there were drop-outs, were reasons for drop-outs described?		X				X
10	Did participants sign an informed-consent form?		X				X
11	Was it stated that Helsinki Declaration Guidelines were followed?		X				X
12	Was the study approved by the local ethics committee?		X				X
13	Were inclusion and exclusion criteria sufficiently described?	X					
14	Was a power analysis done?		X				X
15	Were repeated measures statistics done?		X				X
16	Was the study design clearly described?	X					
17	Was a single session protocol used in the study?		X				X
18	Was a multisession protocol used in the study?		X				X
19	Was maximal voluntary clenching (MVC) measured, and was the procedure described?		X			X	
20	Was a bite force transducer used during experimental tooth clenching exercises?		X				X
21	Was electromyography (EMG) used during tooth clenching exercises?		X				X
22	If EMG was used, what voltage (μV) was used?		X				X
23	Was the experimental tooth clenching exercise clearly described?	X					
24	How many bouts of clenching were used?		X				X
25	Was duration of each bout of clenching described?		X				X
26	Did participants clench until exhaustion?		X				X
27	How many experimental days were involved in the study?		X			X	
28	What percent of maximal voluntary clenching was used in the tooth clenching exercise?		X				X
29	Was the experimental tooth clenching exercise described in such detail that the procedure can be reproduced?	X					
30	Was the methodology for each instrument used in the study clearly described?		X				X
31	Was a reference index test performed, such as index finger?			X			
32	Were questionnaires used to evaluate pain-related variables?		X				X
33	Were questionnaires used to evaluate jaw function?		X				X
34	Were questionnaires used to evaluate psychosocial-related variables?		X				X
35	Were baseline characteristics for participants clearly described?		X				X
36	Was pain intensity reported at baseline?	X					
37	Was intensity of fatigue reported at baseline?		X				X
38	Were detection thresholds reported at baseline?		X				X
39	Were pain thresholds reported at baseline?		X				X
40	Were pain tolerance levels reported at baseline?		X				X
41	Was pain distribution reported at baseline?		X				X
42	Were baseline characteristics for questionnaires reported?		X				X
43	Were baseline characteristics for pain questionnaires reported?		X				X
44	Were baseline characteristics for jaw function questionnaires reported?		X				X
45	Were baseline characteristics for psychosocial questionnaires reported?		X				X

Table 2 Results from Delphi Rounds 1 and 2 (continued)

No.	Item	Round 1			Round 2		
		In†	Un	Ex‡	In§	Un	Ex
46	Was pain intensity measured after each bout of clenching?		X				X
47	Was intensity of fatigue measured after each bout of clenching?		X				X
48	Was detection threshold measured after each bout of clenching?		X				X
49	Was pain threshold measured after each bout of clenching?		X				X
50	Were pain tolerance levels measured after each bout of clenching?		X				X
51	Were pain drawings used to measure pain distribution after each bout of clenching?		X				X
52	Were outcome variables measured at follow-ups?		X				X
Additional items from Delphi round 1*							
53	Was voluntary clenching from a relaxed baseline used?				X		
54	Was the study hypothesis driven?				X		
55	Was statistical analysis appropriate?						X
56	Was a primary outcome variable identified?						X
57	Was facial morphology considered in the analysis of data?						X
58	Was elicited pain (on palpation) assessed at baseline and follow-up?						X
59	Was a randomization procedure used to allocate participants to different clenching or control conditions?						X
60	Were measurements taken under blind conditions?					X	
61	Was stability of force over time measured?						X
62	Were operators blinded?						X
63	Were reliability and validity of outcome measures and diagnosis described?						X
64	Was there a conflict of interest?						X
65	Were the aims and hypothesis clearly described, was "a priori" design used?				X		
66	Were the conclusions appropriately formulated by the authors?				X		

In, included; Un, undecided; Ex, excluded.
*Items that were added in round 1 and assessed in round 2.
†Items included after round 1.
‡Items that were excluded in round 1, and not re-assessed.
§Items that were included after round 2.

1. Were the study's aims and hypothesis clearly described, and was an "a priori" design used?
2. Were the study sample and the informed-consent procedure sufficiently described?
3. Were inclusion and exclusion criteria clearly described?
4. Was the study design described in sufficient detail to permit replication?
5. Was a randomization procedure used to choose a control group or define a control condition?
6. Were location, setting, and instructions of the experimental clenching task clearly described?
7. Was the experimental clenching task described in such detail that replication is possible?
8. Were statistical methods sufficiently described?
9. Were study conclusions appropriately formulated?
10. Was a conflict of interest statement made?

Round 3. Nine panelists responded to the material from round 2. No items on the preliminary tool had ≥ 80% agreement (7/9 panelists). Based on the free-text comments, the steering group excluded one item (item 6) and merged two items (items 2 and 3). Other comments in the feedback stated that it was not clear whether the items assessed quality of design or quality of reporting. It was suggested that each item should be a two-barreled question, designed to specifically test quality of reporting and quality of design. Accordingly, the steering group rephrased each item on the preliminary tool, which was sent to the Delphi members in round 4:

1. *Quality of Reporting:* Were the study's aims and hypotheses clearly described?
   Quality of Design: Were the aims and hypothesis based on relevant theory?

2. *Quality of Reporting:* Were the eligibility criteria used to select participants sufficiently described?
   Quality of Design: Were the eligibility criteria appropriate for the objectives of this study?
3. *Quality of Reporting:* Was it clearly described whether a control group, control condition, or an experimental condition was used?
   Quality of Design: Were the control group, control condition, or experimental condition appropriate for this study, and was a randomization procedure used to randomly allocate subjects to different groups or conditions?
4. *Quality of Reporting:* Was the study design described in sufficient detail to permit replication?
   Quality of Design: Was the study design appropriately selected for the objectives of this study?
5. *Quality of Reporting:* Was the experimental bruxism task described in such detail that replication is possible?
   Quality of Design: Was the experimental bruxism task appropriately selected for the objectives of this study?
6. *Quality of Reporting:* Were statistical methods and data sufficiently described?
   Quality of Design: Were statistical methods and data appropriate for the objectives of this study?
7. *Quality of Reporting:* Were study conclusions appropriately formulated?
   Quality of Design: Were aims and hypothesis clearly addressed in the conclusion and relevant to the objectives?
8. *Quality of Reporting:* Was a conflict of interest statement made?
   Quality of Design: Were the level of involvement and influence in this study for each funder described in detail?

The panelists disagreed with the proposed rating system; a five-point Likert scale (anchor definitions Strongly Disagree and Strongly Agree) was preferred. Although 80% agreement was found for the proposed name of the tool (Q-TEC), the panelists indicated that this instrument would not only be applicable to experimental clenching studies but also to experimental grinding and bracing studies. Quality-Assessment Tool for Experimental Bruxism Studies (Qu-ATEBS) was suggested.

Round 4. Nine of 10 panelists completed the questionnaire, and 100% agreement (9/9 panelists) was found for the preliminary tool. Minor suggestions were made for refinement. Item 3, quality of design, asked whether *(1)* the control group, control condition, or experimental condition was appropriate for the study and *(2)* a randomization procedure was applied to allocate participants to different groups or conditions. A suggested refinement was to omit the second part concerning randomization, since that aspect is covered by the word "appropriate" in the first part of the item. The steering group revised the preliminary tool according to this feedback. The panelists accepted the proposed rating system (five-point Likert scales) and the name Qu-ATEBS.

Assessment of Reliability and Discriminative Validity

Following quality assessment of the 11 selected articles, inter-observer reliability between AD and SA was found to be acceptable ($k = 0.77$). Intra-observer reliability was found to be excellent ($k = 1.0$). Discriminative validity of the instrument between TL and ME, and AD (*phi coefficient* 0.79; $P < .01$) and between TL and ME, and SA (*phi coefficient* 1.0; $P < .01$) was high.

Refinements of the Final Instrument

After reliability and discriminative validity testing, item no. 8 was removed because it was not found to be applicable to any of the reviewed studies. No further changes to the instrument were made. Table 3 shows the final Qu-ATEBS.

Discussion

This study developed an evidence-based quality-assessment tool (Qu-ATEBS) for use in systematic reviews of experimental bruxism studies. Each of the instrument's seven items has two dimensions: quality of reporting and quality of design. Items are phrased as questions and rated on a five-point Likert scale. The maximum attainable score is 70 points; a score between 0 and 50 is considered low quality and a score between 51 and 70 is considered high quality.

The steering group made preliminary decisions about the feature of the instrument: The instrument should be easy to understand and respond to, be quick to use, and contain 10 or fewer items. A greater number of items is considered to lead to higher reliability,[27] but there are exceptions. Studies have shown that global rating scales have higher reliability and validity than detailed checklists.[39–42] A limit of 10 items was chosen.

According to the Swedish Council on Health Technology Assessment, high-quality research is defined as the scientific quality of a study and its ability to resolve the research question reliably.[43]

Table 3 Final Version of the Quality Assessments Tool for Experimental Bruxism Studies (Qu-ATEBS)

		Not applicable	Strongly disagree				Strongly agree
1	Quality of reporting Were the study's aims or hypotheses clearly described?	N/A	1	2	3	4	5
	Quality of design Were the aims or hypothesis based on relevant theory?	N/A	1	2	3	4	5
	What is meant by this item This refers to whether the objectives of the study were clearly defined, and based on relevant theory or careful clinical observation, and one or more working hypothesis that were resolved in the course of the study. There should be a scientific justification of the study, ie, what gap in existing knowledge does the study attempt to fill?						
2	Quality of reporting Were the eligibility criteria, used to select participants, sufficiently described?	N/A	1	2	3	4	5
	Quality of design Were the eligibility criteria appropriate for the objectives of this study?	N/A	1	2	3	4	5
	What is meant by this item This item refers to whether the eligibility criteria were well-defined and relevant to the objectives of the study. A clear description of the study sample's characteristics should be provided, eg. at least age, sex, and when relevant, the diagnosis of participants and their disease duration. A justification for the selected study sample should also be given, eg why only females were included. Reasons for drop- outs, if any, should be described.						
3	Quality of reporting Was it clearly described whether a control group, control condition, or an experimental condition was used?	N/A	1	2	3	4	5
	Quality of design Were the control group, control condition, or experimental condition appropriate for this study?	N/A	1	2	3	4	5
	What is meant by this item This item refers to whether a control group, control condition or an experimental condition was used in this study, and its appropriateness for the objectives. When appropriate, a randomization procedure should be applied to allocate participants to different groups or conditions.						
4	Quality of reporting Was the study design described in sufficient detail to permit replication?	N/A	1	2	3	4	5
	Quality of design Was the study design appropriately selected for the objectives of this study?	N/A	1	2	3	4	5
	What is meant by this item This item refers to the procedure used to achieve the objectives of the study. Clinical and self-reported easures should be relevant to the research question. Variables should be operationally defined. Types of variables and procedure used to measure the variables should be described. Instruments used to collect information should be appropriate for this study, and briefly described (including model and manufacturer's name and location), and if available, references for the instrument's reliability and validity provided, as appropriate. Study type should be defined, and study design should be explained in detail so that whether the study design was appropriate for the proposed objective can be determined. Strategies to reduce or eliminate confounding factors should be described.						
5	Quality of reporting Was the experimental bruxism task described in such detail that replication is possible?	N/A	1	2	3	4	5
	Quality of design Was the experimental bruxism task appropriately selected for the objectives of this study?	N/A	1	2	3	4	5

Table 3 cont.

		Not applicable	Strongly disagree				Strongly agree
	What is meant by this item This item refers to whether reported information on the experimental bruxism task was sufficient to allow reproduction of the task, and if it was appropriate for the objectives of the study. The following should be described: (i) the degree to which the subjects engaged in the experimental bruxism task, ie, was the bruxism exercise related to a force level (percent of maximal voluntary clenching, MVC, or voltage, μV), (ii) the control condition or control group, (iii) duration and number of bruxism bouts, and (iv) the instruments for measuring bite force: a bite force transducer, an electromyographic recording (EMG), or both.						
6	Quality of reporting Were statistical methods and data sufficiently described?	N/A	1	2	3	4	5
	Quality of design Were statistical methods and data appropriate for the objectives of this study?	N/A	1	2	3	4	5
	What is meant by this item This item refers to the methods and models used for statistical analysis of the data. It should be stated how the variables were presented and calculated (eg, mean or median). The statistical method should be appropriate and based on the objectives and type of variables (qualitative or quantitative variables). The computer software used for statistical analysis should be described briefly. For non-significant results addressing important experimental hypothesis post-hoc, power analyses should be performed to determine whether the study was sufficiently well-powered to answer the experimental questions.						
7	Quality of reporting Were the study's conclusions appropriately formulated?	N/A	1	2	3	4	5
	Quality of design Were aims and hypothesis clearly addressed in the conclusions and relevant to the objectives?	N/A	1	2	3	4	5
	What is meant by this item This item refers to whether the conclusions properly summarize the results and were relevant to the objectives of the study. The hypotheses should be clearly addressed in the conclusions.						

In systematic reviews, it is important to be able to quantify the relation between quality and outcome of the included studies, so the authors decided to incorporate a scoring system in the instrument, a five-point Likert scale. A systematic review is based on a carefully formulated question and uses systematic methods to identify, select, and critically evaluate, analyze, and synthesize data from relevant studies.[1] Qu-ATEBS can be used in systematic reviews of experimental bruxism studies. The benefit of using this instrument is that it is reliable and valid, and it can quantify the quality of experimental bruxism studies. Clarke and Oxman emphasized the importance of using a quality score that quantifies the relation between quality and outcome.[44] In quality assessment, it cannot be excluded that raters' subjectivity affects the quality score. High inter-observer reliability might indicate a low level of subjectivity. However, if inter-observer reliability is low, the instrument's quality score should be used with caution. When reporting the results of a quality-assessment process, it is essential to present the results in detail so that readers can estimate study quality for themselves.[45]

In narrative reviews, quality and outcome are based on expert opinion. It has been pointed out that authors' opinions often bias narrative reviews, and so do not reflect an evidence-based approach.[46] In contrast, systematic reviews are based on a comprehensive literature search, and the studies that meet the inclusion criteria are objectively evaluated and analyzed so that conclusions can be drawn, methodological issues identified, and areas that require more original research highlighted. To our knowledge, no systematic review on experimental bruxism studies exists. The authors consider that a systematic review with Qu-ATEBS would contribute to existing knowledge on experimental jaw muscle pain by adding an objective, quality-assessed summary of studies' levels of evidence and by identifying methodological issues and areas that need more research.

Initially, the thought was to develop a tool for quality assessment of experimental clenching studies. As the Delphi process progressed, panelist feedback caused a change in the instrument's perceived end-use. It is known that daytime tooth clenching is a risk factor for persistent muscle pain in the orofacial region,[17,47,48] but it is also known that other jaw muscle activities, eg, eccentric contraction, can contribute to persistent muscle pain. For this reason, Qu-ATEBS focuses on experimental bruxism studies. During the development of this tool, a range of articles was needed from low to high quality. The initial literature search provided a sufficient number of articles to assess the inter-observer reliability and the discriminative validity of the instrument. Because the initial thought was to develop a tool for experimental clenching studies, the MeSH term *bruxism* was not included. But in a systematic review and investigations to assess all published studies, the MeSH term *bruxism* should be included in the PubMed search.

The five-step method suggested by Streiner and Norman was used for developing Qu-ATEBS, and this method was chosen since it has been widely used.[27,49,50] Likewise, the Delphi method is being increasingly applied in medicine and dentistry to assess face validity, step 3 in Streiner and Norman's method.[50–53] The Delphi method derives its name from the Oracle in Delphi. The Oracle pronounced its truths by using a network of informants, similar to the current study's Delphi panel.[54] This method was chosen because it aims to achieve the most reliable consensus amongst a group of experts. The Delphi procedure allows the opinion of each member to be heard, since every round is completed independently. Consensus on items to be included in the quality-assessment tool, on items that need rephrasing, or on how to rate an item, can be reached in a series of Delphi rounds with controlled anonymous feedback.[55]

The most important step in the Delphi procedure is panelist selection. The degree of expertise on the panel correlates directly with the generated results.[56] It must be pointed out that knowledge in a particular field does not necessarily make an individual an expert.[57] After careful consideration, two of the authors, AD and TL, suggested a list of researchers to be invited to join a Delphi panel. Because the area of experimental bruxism studies is new, small, and specialized, only 11 experts were invited.

A Delphi panel should also exhibit heterogeneity.[58] One strength in the present study is that the varying backgrounds and perspectives of the Delphi panel members contributed to the heterogeneity of the panel. Thus, the consensus reached by the expert panel seems credible and with a high discriminative validity.[59]

One benefit of using the Delphi method is that a complex problem is anonymously communicated among a group of experts, thus reducing social influence effects and preventing the more prominent researchers from overwhelming the consensus process.[60] To the authors' knowledge, no formal definition of *high level of agreement* in consensus discussions has been agreed upon, but 51% to 80% is common.[61] In the present study, 10 panel members participated, and 80% agreement (≥ 8/10 panelists) was considered to be a high level of consensus. Too many members on a Delphi panel can make consensus difficult.[62] Baker et al have stated that Delphi panels containing a maximum of 20 members are most reliable.[63] In contrast, too small a sample might affect heterogeneity.[64] The orofacial pain research field is small, but the authors' believe that their Delphi panel was representative and comprised sufficient members, without compromising heterogeneity. Support for this is that a consensus on the quality assessment tool was reached in round 4.

It is important to limit the number of Delphi rounds, since panel members could become fatigued and thus compromise the results.[32] For this reason, the study limited the number of rounds to 4, which later was shown to be enough to reach consensus on a preliminary tool. As the Delphi procedure progressed, there was a gradual increase in agreement and decrease in comments, which is in line with the observations of others.[32] During the Delphi process, 52 items were reduced down to 7. It must be emphasized that the excluded items were not necessarily unimportant; it might be that these items were excluded because they were not applicable to all experimental bruxism studies. In the first round, all 10 panelists completed and returned the questionnaires. In each of the remaining Delphi rounds, 9 of 10 panelists completed and returned the questionnaires; however, it was not always the same person who dropped out in these rounds. A strength of this study is the stability of the response rate throughout the procedure, which has been shown to be a reliable indicator of consensus.[65] As the Delphi process progressed, a basic problem with the preliminary tool was identified in round 3. A majority of panelists indicated that the tool assessed quality of reporting rather than quality of design. However, one cannot be assessed without the other. The feedback proposed making each item a double-barreled question, which assessed both quality of reporting and quality of design. If the quality of reporting is poor, a well-performed study might receive a low score. The tool was revised according to this feedback, and a consensus for Qu-ATEBS was reached in round 4 (see Table 3).

The seven items in Qu-ATEBS cover these areas: study aim, study sample, control condition or group, study design, experimental bruxism task, statistics, and interpretation of results. With appropriate modification, this instrument could be applied to other types of experimental pain studies. In instrument development, the Delphi method is a beneficial face-validity assessment method because it allows a group of experts to communicate a complex problem anonymously. If the Delphi method comprises a sufficient number of panelists, a sufficient number of Delphi rounds, and exhibits heterogeneity, a reliable consensus will be achieved. In this study, the Delphi method focused on experimental bruxism, but this method could also be applied for instrument development in other research areas.

Conclusions

This study developed Qu-ATEBS—a seven-item, evidence-based quality-assessment tool—for use in systematic reviews of experimental bruxism studies. The instrument has good inter-observer reliability and excellent discriminative validity. The procedure leading the development of this specific tool may serve as a template for other quality-assessment tools in orofacial pain research.

Acknowledgments

This study was supported by grants from the Faculty of Odontology at Malmö University, the Swedish Dental Society, and the Swedish National Graduate School in Odontological Science.
The authors have no conflicts of interest.

References

1. Glasziou P, Irwig L, Bain C, Colditz G. Systematic reviews in health care: A practical guide. Cambridge: Cambridge University Press, 2001.
2. List T, Axelsson S, Leijon G. Pharmacologic interventions in the treatment of temporomandibular disorders, atypical facial pain, and burning mouth syndrome. A qualitative systematic review. J Orofac Pain 2003;17:301–310.
3. Racine M, Tousignant-Laflamme Y, Kloda LA, Dion D, Dupuis G, Choiniere M. A systematic literature review of 10 years of research on sex/gender and experimental pain perception—Part 1: Are there really differences between women and men? Pain 2012;153:602–618.
4. Manfredini D, Guarda-Nardini L, Winocur E, Piccotti F, Ahlberg J, Lobbezoo F. Research diagnostic criteria for temporomandibular disorders: A systematic review of axis I epidemiologic findings. Oral Surg Oral Med Oral Pathol Oral Radiol Endod 2011;112:453–462.
5. Jung A, Shin BC, Lee MS, Sim H, Ernst E. Acupuncture for treating temporomandibular joint disorders: A systematic review and meta-analysis of randomized, sham-controlled trials. J Dent 2011;39:341–350.
6. Manfredini D, Lobbezoo F. Relationship between bruxism and temporomandibular disorders: A systematic review of literature from 1998 to 2008. Oral Surg Oral Med Oral Pathol Oral Radiol Endod 2010;109:e26–50.
7. List T, Axelsson S. Management of TMD: evidence from systematic reviews and meta-analyses. J Oral Rehabil 2010;37:430–451.
8. Manfredini D, Lobbezoo F. Role of psychosocial factors in the etiology of bruxism. J Orofac Pain 2009;23:153–166.
9. Charlton E. Ethical guidelines for pain research in humans. Committee on Ethical Issues of the International Association for the Study of Pain. Pain 1995;63:277–278.
10. Staahl C, Drewes AM. Experimental human pain models: A review of standardised methods for preclinical testing of analgesics. Basic Clin Pharmacol Toxicol 2004;95:97–111.
11. Svensson P. What can human experimental pain models teach us about clinical TMD? Arch Oral Biol 2007; 52:391–394.
12. Manfredini D, Marini M, Pavan C, Pavan L, Guarda-Nardini L. Psychosocial profiles of painful TMD patients. J Oral Rehabil 2009;36:193–198.
13. Madland G, Feinmann C, Newman S. Factors associated with anxiety and depression in facial arthromyalgia. Pain 2000;84:225–232.
14. Manfredini D, Winocur E, Ahlberg J, Guarda-Nardini L, Lobbezoo F. Psychosocial impairment in temporomandibular disorders patients. RDC/TMD axis II findings from a multicentre study. J Dent 2010;38:765–772.
15. Torisu T, Wang K, Svensson P, De Laat A, Fujii H, Arendt-Nielsen L. Effects of muscle fatigue induced by low-level clenching on experimental muscle pain and resting jaw muscle activity: Gender differences. Exp Brain Res 2006;174:566–574.
16. Torisu T, Wang K, Svensson P, De Laat A, Fujii H, Arendt-Nielsen L. Effect of low-level clenching and subsequent muscle pain on exteroceptive suppression and resting muscle activity in human jaw muscles. Clin Neurophysiol 2007;118:999–1009.
17. Glaros AG, Burton E. Parafunctional clenching, pain, and effort in temporomandibular disorders. J Behav Med 2004;27:91–100.
18. Svensson P, Burgaard A, Schlosser S. Fatigue and pain in human jaw muscles during a sustained, low-intensity clenching task. Arch Oral Biol 2001;46:773–777.
19. Arima T, Svensson P, Arendt-Nielsen L. Experimental grinding in healthy subjects: A model for postexercise jaw muscle soreness? J Orofac Pain 1999;13:104–114.
20. Glaros AG, Tabacchi KN, Glass EG. Effect of parafunctional clenching on TMD pain. J Orofac Pain 1998;12:145–152.
21. Glaros AG, Baharloo L, Glass EG. Effect of parafunctional clenching and estrogen on temporomandibular disorder pain. Cranio 1998;16:78–83.
22. Plesh O, Curtis DA, Hall LJ, Miller A. Gender difference in jaw pain induced by clenching. J Oral Rehabil 1998;25:258–263.
23. Scott DS, Lundeen TF. Myofascial pain involving the masticatory muscles: An experimental model. Pain 1980; 8:207–215.
24. Bowley JF, Gale EN. Experimental masticatory muscle pain. J Dent Res 1987;66:1765–1769.
25. Christensen LV. Jaw muscle fatigue and pains induced by experimental tooth clenching: A review. J Oral Rehabil 1981; 8:27–36.

26. Clark GT, Adler RC, Lee JJ. Jaw pain and tenderness levels during and after repeated sustained maximum voluntary protrusion. Pain 1991;45:17–22.
27. Streiner DL, Norman GR. Health Measurement Scales—A practical guide to their development and use, ed 4. Oxford, England: Oxford University Press, 2008.
28. Svensson P, Arendt-Nielsen L. Effects of 5 days of repeated submaximal clenching on masticatory muscle pain and tenderness: An experimental study. J Orofac Pain 1996;10:330–338.
29. Arima T, Svensson P, Arendt-Nielsen L. Capsaicin-induced muscle hyperalgesia in the exercised and non-exercised human masseter muscle. J Orofac Pain 2000;14:213–223.
30. Glaros AG, Forbes M, Shanker J, Glass EG. Effect of parafunctional clenching on temporomandibular disorder pain and proprioceptive awareness. Cranio 2000;18:198–204.
31. Hedenberg-Magnusson B, Brodda Jansen G, Ernberg M, Kopp S. Effects of isometric contraction on intramuscular level of neuropeptide Y and local pain perception. Acta Odontol Scand 2006;64:360–367.
32. Schmidt RC. Managing Delphi surveys using non-parametric statistical techniques. Decision Sciences 1997;28:763–774; 1997.
33. Christensen LV. Some subjective-experiential parameters in experimental tooth clenching in man. J Oral Rehabil 1979;6:119–136.
34. Christensen LV. Influence of muscle pain tolerance on muscle pain threshold in experimental tooth clenching in man. J Oral Rehabil 1979;6:211–217.
35. Christensen LV. Progressive jaw muscle fatigue of experimental tooth clenching in man. J Oral Rehabil 1981; 8:413–420.
36. Christensen LV, Mohamed SE, Harrison JD. Delayed onset of masseter muscle pain in experimental tooth clenching. J Prosthet Dent 1982;48:579–584.
37. Delcanho RE, Kim YJ, Clark GT. Haemodynamic changes induced by submaximal isometric contraction in painful and non-painful human masseter using near-infra-red spectroscopy. Arch Oral Biol 1996;41:585–596.
38. Farella M, Soneda K, Vilmann A, Thomsen CE, Bakke M. Jaw muscle soreness after tooth-clenching depends on force level. J Dent Res 2010;89:717–721.
39. Cohen DS, Colliver JA, Marcy MS, Fried ED, Swartz MH. Psychometric properties of a standardized-patient checklist and rating-scale form used to assess interpersonal and communication skills. Acad Med 1996;71:S87–S89.
40. Norcini JJ, Diserens D, Day SC, et al. The scoring and reproducibility of an essay test of clinical judgment. Acad Med 1990;65:S41–S42.
41. Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Acad Med 1998;73:993–997.
42. Rubin HR, Redelmeier DA, Wu AW, Steinberg EP. How reliable is peer review of scientific abstracts? Looking back at the 1991 Annual Meeting of the Society of General Internal Medicine. J Gen Intern Med 1993;8:255–258.
43. SBU. Methods of treating chronic pain. In: Axelsson S, Boivie J, Eckerlund I, et al (eds). SBU-report no 177:1. Stockholm: The Swedish Council on Technology Assessment in Health and Care (SBU), 2006.
44. Clarke M, Oxman AD (eds). Cochrane Reviewers Handbook 4.0. In Review Manager (computer programme), version 4.0. Oxford, England: The Cochrane Collaboration, 1999.
45. Moncrieff J, Churchill R, Drummond DC, McGuire H. Development of a quality assessment instrument for trials of treatments for depression and neurosis. Int J Methods Psychiatr Res 2001;10:126–133.
46. Mulrow CD. The medical review article: State of the science. Ann Intern Med 1987;106:485–488.
47. Chen CY, Palla S, Erni S, Sieber M, Gallo LM. Nonfunctional tooth contact in healthy controls and patients with myogenous facial pain. J Orofac Pain 2007;21:185–193.
48. Huang GJ, LeResche L, Critchlow CW, Martin MD, Drangsholt MT. Risk factors for diagnostic subgroups of painful temporomandibular disorders (TMD). J Dent Res 2002;81:284–288.
49. Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Control Clin Trials 1996;17:1–12.
50. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: A tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003;3:25.
51. John MT. Improving TMD classification using the Delphi technique. J Oral Rehabil 2010;37:766–770.
52. Lambe P, Bristow D. What are the most important nonacademic attributes of good doctors? A Delphi survey of clinicians. Med Teach 2010;32:e347–e354.
53. Zafar SY, Currow DC, Cherny N, Strasser F, Fowler R, Abernethy AP. Consensus-based standards for best supportive care in clinical trials in advanced cancer. Lancet Oncol 2012;13:e77–e82.
54. Kennedy HP. Enhancing Delphi research: Methods and results. J Adv Nurs 2004;45:504–511.
55. Kerr M. The Delphi Process. The Delphi Process 2002 City: Remote and Rural Areas Research Initiative, NHS in Scotland, 2001.
56. Robert CJ. Use of Delphi methods in higher education. Technological Forecasting and Social Change 1972;4: 173–186.
57. Keeney S, Hasson F, McKenna HP. A critical review of the Delphi technique as a research methodology for nursing. Int J Nurs Stud 2001;38:195–200.
58. Moore CM. Delphi technique and the mail questionnaire. In Moore CM (ed). Group Techniques for Idea Building: Applied Social Research Methods Series, vol 9. Newbury Park, CA: Sage, 1987: 50–77.
59. Goodman CM. The Delphi technique: a critique. J Adv Nurs 1987;12:729–734.
60. Jairath N, Weinstein J. The Delphi methodology (Part one): A useful administrative approach. Can J Nurs Adm 1994;7:29–42.
61. Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. J Adv Nurs 2000;32: 1008–1015.
62. Reid NG. The Delphi technique: Its contribution to the evaluation of professional practice. In Ellis R (ed). Professional Competence and Quality Assurance in the Caring Professions. Beckenham, Kent, UK: Croom-Helm, 1988.
63. Baker J, Lovell K, Harris N. How expert are the experts? An exploration of the concept of 'expert' within Delphi panel techniques. Nurse Res 2006;14:59–70.
64. Berquez AE, Cook FM, Millard SK, Jarvis E. The Stammering Information Programme: A Delphi study. J Fluency Disord 2011;36:206–221.
65. Crisp J, Pelletier D, Duffield C, Adams A, Nagy S. The Delphi method? Nurs Res 1997;46:116–118.