SEMMELWEIS EGYETEM

DOKTORI ISKOLA

Ph.D. értekezések

3214.

OSZLÁNSZKY JUDIT

Fogorvostudományi kutatások

című program

Programvezető: Dr. Varga Gábor, egyetemi tanár

Témavezető: Dr. Hermann Péter, egyetemi tanár
Dr. Zrubka Zsombor

ASSESSMENT OF ORAL HEALTH-RELATED QUALITY OF LIFE: SYSTEMATIC REVIEW, PSYCHOMETRIC EVALUATION, AND VALIDATION OF THE HUNGARIAN GOHAI USING COSMIN METHODOLOGY

PhD thesis

Judit Oszlánszky, DMD

Semmelweis University Doctoral School
Dental Research Division

Supervisor: Prof. Péter Hermann, DMD, D.Sc/Zsombor Zrubka, MD,Ph.D

Official reviewers: Gyula Marada, DMD, Ph.D
László Simonffy, MD, DMD, Ph.D

Head of the Complex Examination Committee:
Prof. István Gera, DMD, Ph.D

Members of the Complex Examination Committee: Beáta Kerémi, Ph.D
Balázs Sándor, Ph.D

Budapest
2025

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ADD-GOHAI	Additive Score of General Oral Health Assessment Index
CFA	Confirmatory Factor Analysis
CFI	Comparative fit index
COSMIN	Consensus-based Standards for the Selection of Health Measurement Instruments
CTT	Classical Test Theory
CVR	Content Validity Ratio
DMFT	Decayed, Missing, and Filled Teeth Index
dPROM	Dental Patient-Reported Outcome Measure
DTN	Dental Treatment Need
EFA	Exploratory Factor Analysis
EMPRO	Evaluating Measures of Patient-Reported Outcomes
EQ-5D-5L	5-Level Version of the EQ-5D General Health-Related Quality of Life Instrument
EQ-VAS	EQ-5D Visual Analogue Scale
GFI	Goodness of Fit Index
GRADE	Grading of Recommendations Assessment, Development, and Evaluation
GOHAI	General Oral Health Assessment Index
GOHAI-HU	Hungarian Version of the General Oral Health Assessment Index
GHRQoL	General Health-related Quality of Life
HRQoL	Health-Related Quality of Life
ICC	Intra-Class Correlation Coefficient
IRT	Item Response Theory
LM	Lagrange multiplier
MID	Minimal Important Difference
OHIP	Oral Health Impact Profile
OHIDL	Oral Health Impact on Daily Living

OHRQoL	Oral Health-Related Quality of Life
OIDP	Oral Impacts on Daily Performances
OH-SQ	Oral Health Single Question
PCFA	Principal Component Factor Analysis
PCC	Patient-Centered Care
PROM	Patient-Reported Outcome Measure
PROSPERO	International Prospective Register of Systematic Reviews
QoL	Quality of Life
RMSEA	Root Mean Square Error of Approximation
SC-GOHAI	Simple Count Score of the General Oral Health Assessment Index
SD	Standard Error
SDC	Smallest Detectable Change
SEM	Standard Error of Measurement
SRMR	Standardized Root Mean Square Residual
SU	Semmelweis University
TLI	Tucker-Lewis index
TMD	Temporomandibular Disorders
VAS	Visual Analog Scale
WHO	World Health Organization

1. INTRODUCTION

1.1. Health, Quality of Life, and Oral Health-Related Quality of Life

The World Health Organization's (WHO) 1948 definition of health—as "a complete state of physical, mental, and social well-being, and not merely the absence of disease"—significantly broadened the concept of health (1). This marked the beginning of a paradigm shift, with researchers increasingly focusing on health as a multidimensional construct. Over time, quality of life (QoL) has gained prominence, and in 1995, the WHO defined it as individuals' "perceptions of their position in life in the context of the culture and value systems in which they live, and in relation to their goals, expectations, standards, and concerns" (2).

By the late 1990s, a noticeable shift in healthcare expectations had already begun to emerge. Henrietta L. Logan highlighted the increasing tensions between patients and healthcare providers, emphasizing that the dramatic expansion of patient choice—including treatment alternatives and providers—combined with unprecedented access to vast amounts of information about treatments, materials, and options, had created new challenges. Patients were no longer satisfied with competence and reliability alone; they began to demand involvement in their care, education about available options, and personalized attention (3). Over the past three decades, these expectations have only intensified, driven further by advances in digital technology, which have introduced innovative treatment alternatives, new materials, and varying costs (4). The concept of Patient-Centered Care (PCC), defined by the Institute of Medicine as "care that respects and responds to individual patient preferences, needs, and values," has become a cornerstone of modern healthcare (5). Medicine has increasingly shifted towards patient-centered, individualized treatments that rely on patient feedback and are facilitated by open, mutual communication. However, this approach cannot be fully realized without measuring Patient-Reported Outcome Measures (PROMs)—tools that capture patients' perspectives on their health, quality of life, and treatment outcomes. Today, the analysis of PROMs has become a critical focus across all areas of physical and mental healthcare, including dentistry and oral health. Although the principles of PCC are gradually gaining ground in dentistry, its application and research remain less

extensive than in general medicine (6). This gap indicates that dentistry, as a scientific discipline, is likely on the cusp of significant growth and advancement in adopting patient-centered approaches.

The most important dental patient-reported outcome measure (dPROM) is oral health-related quality of life (OHRQoL) (7). The multifaceted impact of oral health on overall well-being spans several dimensions. As Kleimann outlined in his 1989 biopsychosocial model, physical functioning, emotional health, and social well-being are inherently interconnected within this framework (8). Health-Related Quality of Life (HRQoL) is a highly subjective concept, and numerous researchers have sought to uncover its underlying constructs, often yielding partially divergent results. Nevertheless, these findings generally align with Kleimann's categorization of the primary domains of quality of life as physical/biological, psychological, and social components. Some authors further subdivide the physical domain into functional aspects and pain/discomfort, reflecting nuanced perspectives. Wilson and Cleary's model, for instance, emphasizes the influence of both individual and societal characteristics on these domains (9). Building on foundational models, researchers across diverse branches of medicine have developed specialized frameworks and measurement tools tailored to their respective fields, including oral health and dentistry (10). By the late 1990s, the accumulation of substantial knowledge in OHRQoL culminated in a landmark event: a major conference held in Chapel Hill in 1996, where prominent scholars convened to exchange insights (11). This pivotal gathering marked the beginning of a sustained expansion of research into OHRQoL, driving significant advancements in its measurement and application that continue to the present day.

Locker simplified the concept of quality of life by defining it as the answer to the question, "How good is your life?" He emphasized that health status is merely one aspect of the broader concept, cautioning against using the two terms interchangeably. The relationship between general health-related quality of life and OHRQoL remains a topic of frequent debate, although the significant connection between them is undeniable (11). However, their measurement presents a considerable challenge due to the complexity of the underlying construct and the high degree of subjectivity involved (12).

1.2. The Importance and Challenges in Measuring OHRQoL

Our perception of oral health undergoes significant changes throughout life. Expectations differ greatly during childhood, young adulthood, and old age, leading to varying evaluations of oral health conditions (13). Since oral health is strongly age-dependent, with substantial differences in OHRQoL requiring distinct measurement tools for children, it is important to highlight that both my dissertation and our research were exclusively focused on adults (7, 13, 14).

Similarly, cultural background and socio-cultural environment play a fundamental role in shaping these expectations, thereby influencing the overall impact of oral health on QoL. Numerous studies have examined the OHRQoL of specific patient groups (e.g., oral cancer patients, children with orofacial clefts, etc.), as these conditions also significantly influence an individual's subjective evaluation and experience of QoL (15, 16). Given these distinct characteristics, it is understandable that developing a measurement tool capable of providing universally applicable and comparable results across age groups and subpopulations poses a considerable challenge. OHRQoL measurement tools can be categorized into two main types: generic instruments and condition-specific instruments. Generic instruments assess overall OHRQoL, enabling comparisons across populations and health conditions with the help of normative data. However, they may lack the sensitivity of condition-specific tools, which focus on the unique impacts of particular conditions on QoL (17). Therefore, understanding the psychometric properties of different measurement tools and their performance across various populations is crucial. This knowledge helps identify which instrument's measurement properties are most appropriate for a given research or clinical context.

1.3. OHRQoL Measurement Tools

Deana et al., in a 2024 systematic review, examined OHRQoL instruments available for adults using the Evaluating Measures of Patient-Reported Outcomes (EMPRO) tool, which ensures standardized and comparable results. They identified 14 instruments (8 generic and 6 specific to treatments or conditions). One cancer-specific questionnaire (EORTC QLQ OH-15) achieved the highest score, followed by three generic OHRQoL questionnaires that were notably superior to the others: the Oral Health Impact Profile

(OHIP) (18), the General Oral Health Assessment Index (GOHAI) (19), and the Oral Health Impact on Daily Living (OHIDL).(20). (21) Riva et al.'s 2021 review, which examined OHRQoL instruments validated for adults, identified 42 original tools and similarly found that the 14-item OHIP (22), the GOHAI and Oral Impacts on Daily Performances (OIDP) (23) were the most commonly used (24).

1.3.1. Oral Health Impact Profile

The OHIP is a widely used OHRQoL measurement tool available in multiple language versions. Its development was based on methodologies previously applied in general health settings to assess the impact of medical care on functional and social well-being. The OHIP was designed to provide a comprehensive measure of self-reported dysfunction, discomfort, and disability attributed to oral conditions. Initially developed in 1994, it consisted of 49 items across seven domains, defined using Locker's model of oral health. These domains represent seven conceptual dimensions of impact: functional limitation, physical pain, psychological discomfort, physical disability, psychological disability, social disability, and handicap (25).

To improve usability, shorter versions like the OHIP-14 have been developed. Responses are scored on a five-point Likert-type scale, ranging from "never" to "very often," to quantify the frequency of impacts (22).

1.3.2. General Oral Health Assessment Index

GOHAI consists of 12 items assessing three hypothesized dimensions: ‘physical function’, ‘psycho-social function’, and ‘pain and discomfort’. The questions of GOHAI focus on the last 3 months, and its items are scored on either 3, 5, or 6-level Likert scales. The GOHAI score can be calculated via the additive (Add-GOHAI) and the simple counts methods (SC-GOHAI). Add-GOHAI is the sum of Likert scores after reversing the oppositely worded items. Its range is 12-36, 12-60, and 0-60 when using the 3, 5, and 6-level Likert items, respectively. Higher Add-GOHAI scores indicate better OHRQoL, but examples of reverse scoring (26, 27) warrant care when interpreting results. Add-GOHAI allows a nuanced assessment of OHRQoL, while SC-GOHAI is the count of items with responses ‘sometimes’, ‘often’ or ‘always’, ranging between 0 and 12. Higher SC-GOHAI scores indicate poorer OHRQoL.

1.3.3. Comparison of OHIP and GOHAI

The comparison between the OHIP and GOHAI questionnaires has been the subject of numerous studies, often addressing the question of which tool is superior (28-30). These studies consistently conclude that while both questionnaires exhibit excellent psychometric properties, they assess slightly different constructs, although they are strongly correlated (31). The OHIP places greater emphasis on the psychosocial domain, whereas the GOHAI is more strongly aligned with functionality and dental status. Locker once described the GOHAI as a "dental status" questionnaire (32), a characterization that has since been definitively refuted by several studies (31). As a result, the choice between these two instruments is primarily determined by the specific research objectives rather than any objective measure of superiority.

1.3.4. Hungarian-Language OHRQoL Measurement Tools for Adults

Currently, two general OHRQoL questionnaires designed for adults are available in Hungarian. These are the OHIP (33), including its 14-item and 5-item short forms (34), and the Hungarian GOHAI, which was developed as part of the work underlying this dissertation (35).

In 2006, Szentpétery and colleagues created the Hungarian version of the 49-item OHIP, followed by the publication of the two shorter versions in 2008. The questionnaire measures the frequency of symptoms, referencing the past month as the time frame. Until 2024, the OHIP remained the only general OHRQoL measurement tool available in Hungary. Thanks to our work, researchers and clinicians can now choose between these questionnaires based on the considerations discussed earlier.

1.4. The Role of Consensus-Based Standards in Enhancing Psychometric Research

The methodology of psychometrics has evolved alongside the development of the field, offering increasingly precise methods to demonstrate various characteristics of measurement tools. With these advancements, the expectations have also risen, particularly for newly developed language versions of instruments, which now need to adhere to much stricter protocols than in the past.

A common issue in the literature is the confusing and inconsistent terminology used, where different authors may use the same term to refer to completely different concepts (36). It is often observed that even so-called "validated" measurement tools have been subjected to only partial investigations of their properties, or that the methods employed were inadequate for testing their validity. To address this inconsistency and confusion, a Dutch research group set out to establish clarity, creating standardized evaluation protocols and providing comprehensive guidelines for researchers and users in the field.
In 1998, this effort culminated in the publication of the COSMIN (Consensus-based Standards for the Selection of Health Measurement Instruments), which raised the bar for methodological rigor in psychometrics (37-39). The COSMIN framework provides a structured approach to evaluating psychometric properties such as reliability, validity, and responsiveness, and it has become a gold standard in psychometric research. It includes recommendations for evaluating both the measurement properties, methodological compliance, and the quality of evidence, which will be discussed in detail later. Studies that follow the COSMIN protocol are considered more robust, as they are grounded in established best practices that help mitigate biases and inconsistencies.

1.5. Future Directions in OHRQoL Measurement

The need for QoL measurement in healthcare has never been as pressing as it is today in our rapidly evolving, digitalized world, which is also seeing rising costs (4). Numerous treatment options are available for various diseases, and financing these through social health insurance is a significant health economics issue (40, 41).
While the two most widely used OHRQoL measurement tools, GOHAI and OHIP-14, were developed in 1990 and 1997, several gaps remain in our understanding of their performance. Key areas such as responsiveness, measurement invariance (MI), and measurement error are underexplored, and our research aims to address some of these blind spots. An important consideration is the reference period of these tools, which refers to the time frame over which the frequency of problems is being assessed. GOHAI uses a three-month reference period, while OHIP allows flexibility, with one month being the most common internationally (31). Further research is needed to determine the optimal reference period, as comparing results across different time frames can be challenging. Given the significant changes in healthcare, technology, and societal expectations since

these tools were developed, a review of the factors influencing OHRQoL and its conceptual frameworks is necessary. Additionally, there is a growing need for modern measurement tools incorporating updated methodologies and current knowledge, ensuring they remain relevant and accurate in today's context.

2. OBJECTIVES

GOHAI is a widely accepted instrument for measuring OHRQoL, but it has not been available in the Hungarian language. While initially developed for older adults, GOHAI has been applied in all adult age groups. The COSMIN guidelines have been developed to standardize the psychometric assessment methods for outcome measurement instruments, guiding both clinicians and researchers in their development, selection, and validation processes. Also, COSMIN offers a robust framework for the systematic evidence synthesis of the psychometric properties of PROMs.

The primary objective of this PhD research is to achieve the cross-cultural validation of the Hungarian version of GOHAI, which was preceded by a systematic COSMIN review to critically evaluate pre-existing evidence on the psychometric properties of GOHAI and identify potential research gaps. Altogether, these efforts led to the formulation of the following research questions:

1. What insights does a systematic COSMIN review provide about the measurement properties of GOHAI as well as the quality and strength of the supporting evidence?
2. Are the psychometric properties of the GOHAI, including structural validity, construct validity, and reliability, uniformly established across different age groups to support its use as a general OHRQoL instrument?
3. Is the construct validity of the Hungarian GOHAI supported by at least 75% of the predefined hypotheses, as suggested by the COSMIN guidelines?
4. Does a single-factor structure of Hungarian GOHAI with secondary dimensions of physical function, psycho-social function, and pain and discomfort demonstrate adequate fit in confirmatory factor analysis (CFA)?
5. When tested for measurement invariance, does the Hungarian GOHAI exhibit at least metric invariance between the general and clinical populations, as well as different age groups?
6. What are the test-retest reliability, standard error of measurement (SEM), and the smallest detectable change (SDC) of the Hungarian GOHAI after repeated administrations?

3. METHODS

We conducted two studies, both following the COSMIN methodology. The first study summarized the psychometric properties of the GOHAI (31), while the second focused on validating the Hungarian version of the GOHAI (35).

3.1. The COSMIN guideline for systematic reviews

We followed the COSMIN guideline for systematic reviews (37, 38). It is designed to increase transparency and reduce duplication in research by publicly documenting the methodology and intended scope of systematic reviews before they are conducted. COSMIN reviews assess the psychometric measurement properties of PROM instruments in three domains: reliability (i.e., the degree to which the measurement is free from measurement error), validity (i.e., the ability to measure the intended construct), and responsiveness (i.e., the ability to measure change over time). PROMs are assessed in terms of the goodness of their measurement properties (i.e., quality) as sufficient, insufficient, or indeterminate. COSMIN reviews evaluate the methodological quality (i.e., risk of bias) of the included studies as very good, adequate, doubtful, or inadequate. The strength of supporting evidence is graded as high, moderate, low, or very low, following the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) for systematic reviews of clinical trials (37). For a detailed description of the assessment principles, please refer to the original publication (37, 38).

3.2. Methods of Study I.: Psychometric Properties of General Oral Health Assessment Index Across Ages: COSMIN Systematic Review

3.2.1 Literature Search and Eligibility Criteria

3.2.1.1. Inclusion and Exclusion Criteria

We included English peer-reviewed full-text journal articles reporting primary psychometric studies of GOHAI that described at least one item of the COSMIN checklist. We excluded non-primary research reports (e.g., reviews, guidelines, letters, editorials) or sources other than peer-reviewed journals (e.g., unpublished manuscripts, dissertations, reports, books or book chapters, conference proceedings or abstracts, etc.).

Furthermore, we excluded studies in which GOHAI was used for the validation of another instrument.

3.2.1.2. Literature Search

We performed a comprehensive search in PubMed, Web of Science, and EMBASE covering Jan 1990 (the first publication of GOHAI) through December 31, 2023. We applied Terwee's "Precise search filter for measurement properties" (42) in accordance with COSMIN guideline and a previous systematic review in the field (24). The detailed syntax is provided in the Supplementary Materials of the original publication (Supplementary Table 1 (31)).

3.2.2. Record Screening and Selection of Reports

Screening and the selection of eligible reports were performed by two reviewers. One reviewer (JO) examined twice all titles and abstracts. A second reviewer (ZZ) screened 20% of randomly selected records in duplicate. All differences were resolved by consensus. Full-text papers were checked for the pre-determined inclusion and exclusion criteria by one reviewer (JO). In case of ambiguity, full-text papers were evaluated by both reviewers, and a joint decision was made. The results are presented in a PRISMA flowchart; see Fig.3.

3.2.3. Data Extraction, Evaluation, and Quality Assessment

Prior to commencing data extraction, the protocol was registered in the PROSPERO database (CRD42022384132) (43). PROSPERO (International Prospective Register of Systematic Reviews) is an online database where researchers can register systematic review protocols.

3.2.3.1. General Information and Age Groups

For each study, we extracted the author, publication year, sample size, age, male/female ratio, country, as well as language version, and mode of administration. Based on the age of respondents, studies were categorized into four groups to evaluate the psychometric properties of GOHAI across different age ranges. The four age groups are defined as follows:: (1) elderly (≥60 years old), (2) all ages, (3) middle-aged or younger (<60 years old), and (4) young adults (≤45 years of age). While one study was assigned to only one

age group, it's important to note that there might be some overlap due to variations in age groups used across the studies. For the elderly group, we lowered the 65 years cut-off of the registered protocol to 60 years to better match the age distribution of the included studies. Studies that evaluated GOHAI using item-response theory (IRT) were assessed separately.

3.2.3.2. Content Validity

Content validity refers to the extent to which the content of a PROM adequately represents the construct it is intended to measure (36). The development of GOHAI was based on a comprehensive literature review of questionnaires addressing oral functional status, patient satisfaction, oral symptoms, self-esteem, and socialization (19, 44). In 2017, Campos et al. assessed the importance of items in measuring the construct by involving a group of fifteen dental experts. All items were unanimously regarded as essential, indicating a high level of consensus (Content validity ratio (CVR) > 0.506) (45). Given the decades-long international usage of GOHAI, we considered its content validity well-established and omitted its evaluation from this study

3.2.3.3. Internal Structure

Before examining the internal structure, it is important to clarify the relationship between the items of the measurement tool and the underlying construct being assessed. Depending on the underlying relationship between the items and the construct to be measured, two distinct models are identified. In a reflective model, the items are manifestations of the construct, meaning that they are expected to correlate with each other and may be interchangeable. On the other hand, in a formative model, each item contributes a distinct part to the overall construct, and together, the items form the complete construct. In this case, the items do not necessarily correlate with one another and are not interchangeable (46). However, we note that by addressing diverse and independent oral conditions that impact OHRQoL, GOHAI may have characteristics of a formative instrument, while its primary aim is to assess the single underlying construct of OHRQoL in a reflective model. Therefore, the same GOHAI score can be achieved by different combinations of oral health problems, which, depending on the underlying population, may cause inconsistent results when evaluating its internal structure. In the COSMIN taxonomy, internal structure comprises structural validity (i.e., the degree to

which an instrument's scores reflect the dimensionality of the underlying construct), internal consistency (i.e., the degree of interrelatedness between items within the dimensions of the construct) and cross-cultural validity / measurement invariance (i.e., the degree to which the original PROM's psychometric properties are retained in language adaptations or population subgroups).

For the evaluation of structural validity, we extracted the methods and results of factor analyses. Methodological quality was rated as very good if confirmatory factor analysis (CFA) was performed. Exploratory factor analysis (EFA) or principal component factor analysis (PCFA) were rated as adequate. The sample size with N≥100 rated as very good, 72-100 as adequate, 60-72 as doubtful, and <60 as inadequate.

The internal consistency statistic is meaningful only when the items are interrelated and together form a reflective model. For evaluating internal consistency, we extracted the Cronbach's alpha (α) values of the entire instrument, excluding instances where the items of GOHAI were modified or deleted.

We did not find studies assessing cross-cultural validity or measurement invariance.

3.2.3.4. Other Measurement Properties

Other measurement properties in COSMIN include reliability (i.e. degree to which the measurement is free from measurement error), measurement error (the part of a patient's score that is not related to true changes in the measured construct), criterion validity (i.e., the degree to which measurements correspond to a gold standard), hypothesis testing for construct validity (i.e., alignment with the assumption the instrument is a valid measure of the underlying construct) and responsiveness (i.e., the ability to detect change over time in the measured construct).

The extended definition of test-retest reliability is the degree to which scores for patients who have not changed remain consistent across repeated measurements over time. For the reliability assessment, we extracted the methods and results of the test-retest reliability. For good methodological quality, stable patient status and consistent test conditions are required between repeated measurements. If these preconditions were not met, we assigned 'doubtful' rating. A time interval of 1 to 4 weeks was considered appropriate. Intra-class correlation coefficient (ICC) was considered a very good statistical method by COSMIN. For measurement error, we sought reports of the SEM, SDC, or the limits of agreement from Bland-Altman analyses (47).

Due to the lack of a reasonable 'gold standard' for PROMS, we did not evaluate the criterion validity.

Construct validity assesses how accurately a PROM captures the concept it was designed to measure. This validation process generally relies on hypothesis testing to determine whether the PROM behaves as theoretically expected (36). Hypothesis testing for construct validity comprises convergent validity (i.e., the correlation with related instruments) and known-groups validity (i.e., differences in scores of subgroups with apparently different levels of the underlying construct). All eligible hypothesis tests were assessed, regardless of the authors' explicit intention to test construct validity. We extracted the parameters that allowed effect size calculation: Cohen's d for group comparisons or correlation coefficient for continuous variables. If effect sizes could not be calculated, or in the case of U-shaped associations in multiple group comparisons, we assessed the hypothesis as 'indeterminate (?)'.

Hypotheses were grouped into broader categories, as shown in Table 1. We expected a statistically significant positive correlation with each instrument, albeit with varying magnitude. When evaluating known-groups validity, we included hypotheses based on primary oral cavity symptoms but excluded general conditions (e.g., diabetes, obesity, or mental disorders) and demographic variables, which were indirectly related to oral health (38). We hypothesized that patients with poor dental, periodontal, and prosthetic status, as well as poor oral hygiene and risky behaviors such as smoking or excessive alcohol use, would have worse GOHAI scores. The details of the hypothesis testing are available in the Supplementary Materials of the original publication (Supplementary Tables 6–9 (31))

When assessing responsiveness, we extracted the methods and results of hypothesis tests comparing GOHAI scores before and after intervention. We expected that the authors determine both the direction and the expected magnitude of change resulting from the intervention. Without this information, the review team attempted to formulate a hypothesis based on COSMIN recommendations, utilizing available literature data, expert experience, and the hypotheses used in the articles included in this review (37).

Table 1. Hypothesis groups and the expected effect size (31)

Property	Group	Hypothesis	Expected effect-size
Convergent validity	**A) Oral health-related**	H_{A1}: OHIP-14/OHIP-EDENT	Strong correlation
		H_{A2}: self-rated oral health	Strong correlation
		H_{A3}: satisfaction with oral health/teeth/mouth	Moderate correlation
		H_{A4}: denture satisfaction questionnaire	Weak correlation
		H_{A5}: self-rated need for dental treatment	Moderate correlation
	B) General/Mental Health-related	H_{B1}: self-rated general health	Weak correlation
		H_{B2}: Medical Outcomes Study Short Form General Health Survey (MOS-20)	Weak correlation
		H_{B3}: Medical Outcomes Study Short Form Health Survey (SF-36)	Weak correlation
		H_{B4}: life stress/perceived stress questionnaire	Weak correlation
		H_{B5}: morale index/Philadelphia Geriatric Center Morale Scale	Weak correlation
		H_{B6}: Hospital Anxiety and Depression scale	Weak correlation
		H_{B7}: satisfaction with life situation	Weak correlation
Known-groups validity	**C) Physician-reported**	H_{C1}: dental status	Small difference
		H_{C2}: prosthetic status	Small difference
		H_{C3}: periodontal and mucosal health, objective oral hygiene	Small difference
	D) Patient-reported	H_{D1}: behavioral variables	Small difference
		H_{D2}: oral symptoms	Small difference

3.2.4. Evidence Synthesis and Analysis

Results were summarized by each psychometric property for all included studies (overall) and by the four age groups. Structural validity factor analyses were tabulated by the number of extracted factors (Table 2) and summarized qualitatively. Internal consistency and reliability effect sizes were summarized via random-effects meta-analysis.

Correlation coefficients were entered in the meta-analysis after Fisher's Z transformation, and the overall results were back-transformed to a correlation coefficient (48). Construct validity and responsiveness were assessed by calculating effect sizes and summarizing the results descriptively by their magnitude in Table 3, indicating whether a hypothesis was confirmed (+), not confirmed (-), or indeterminate (?).

3.2.5. Additional Considerations

3.2.5.1. Item Response Theory/ Rasch analyses

Measurement theory explains how item scores represent the construct being measured. The two most widely recognized measurement theories are Classical Test Theory (CTT) and Item Response Theory (IRT), both of which are applicable to reflective measurement models (46). IRT is a family of models that describe the relationship between a latent trait (e.g., OHRQoL) and the probability of a particular response to an item. In 1960, Georg Rasch published a specific type of IRT model, now widely known as Rasch Analysis, which focuses on the relationsfhip between a respondent's ability (or level of a latent trait) and the difficulty of an item. (49). Due to their methodological characteristics, studies using IRT models were evaluated separately, and during this evaluation, we followed the recommendations of the COSMIN Guidelines.

3.2.5.2. Scoring Variants of GOHAI

For all included studies, we recorded and summarized the scoring methods used, the number of response options provided, and any reported modifications to the scoring system.

3.2.5.3. Handling of Missing Data and Floor and Ceiling Effects

Floor and ceiling effects help determine if a measurement tool can capture the full range of a construct. A floor effect happens when a large number of respondents pick the lowest score and a ceiling effect when many choose the highest score. These effects can limit the tool's ability to track changes over time. For instance, a strong ceiling effect restricts the detection of improvement, and a floor effect makes it hard to show further deterioration for those scoring low (36, 37). Whenever studies provided information on the handling of missing data or reported on floor and ceiling effects, we documented and summarized the respective approaches and findings.

3.3. Methods of Study II.: Validation of the Hungarian version of the General Oral Health Assessment Index (GOHAI) in clinical and general populations

3.3.1. Linguistic Adaptation of GOHAI

We adhered to the guidelines provided by Beaton et al. (50) throughout the translation process. Forward translation from English to Hungarian was performed by two bilingual translators and consolidated by a third translator. Backward translation was performed by one native English speaker without access to the original questionnaire. After consolidation by the translator team and 15 clinicians, the draft instrument was piloted by 20 randomly selected patients via think-aloud interviews. As a result, we adjusted the polarity of items 3, 5, and 7, ensuring that a higher score for each answer now indicates fewer problems. The Hungarian GOHAI (GOHAI-HU) inquiries about problems in the last 3 months, and participants respond on a 5-point Likert-type frequency scale (1=always; 2=often; 3=sometimes; 4=seldom; 5=never), resulting in ADD-GOHAI-HU scores ranging from 12 to 60. The stages of cross-cultural adaptation for the Hungarian version of GOHAI are detailed in the Supplementary Materials of the original publication (see Appendix, Fig. S1). Both the Hungarian and original English versions of GOHAI are provided in the appendix (see Table S1) of this publication (35).

3.3.2. Participants

The research was approved by Semmelweis University Regional and Institutional Committee of Science and Research Ethics (permit number 61/2023), and all participants provided their informed consent before their enrolment in the study. The study sample comprised two groups: participants from the general population without primary dental concerns and clinical patients with primary dental issues. The general patient cohort was recruited in Budapest from a nursing home, and the general population attended mobile health screening kiosks on 06/05/2023 and 20/05/2023. Clinical patients were sourced from Semmelweis University (SU), Department of Oral Diagnostics, Temporomandibular Disorders (TMD) Care Unit, and Department of Prosthodontics. All patients underwent examination by a dental medical doctor (KM or JO). For those patients whose stable oral health status was affirmed, the questionnaire was repeatedly administered a week after the initial interview by the same doctor personally or via phone.

See Figure 1. for study sample recruitment details. The sample size was determined using the COSMIN modified GRADE approach, which recommends a total sample size exceeding 100 for high quality of evidence on hypothesis tests, the evaluation of test-retest reliability, and conducting CFA. We recruited a sample to meet at least 100 respondents in each subgroup, including clinical and general populations, age groups, and the retest sample.

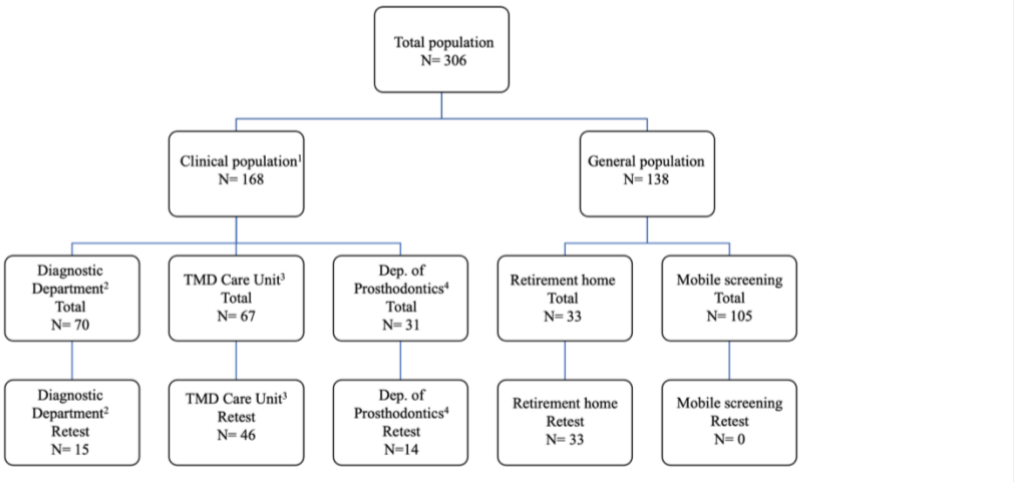

Fig1. Study sample recruitment (35)

[1]Clinical population from Semmelweis University, Faculty of Dentistry's outpatient services: [2] Department of Oral Diagnostics, [3] Temporomandibular Disorders Care Unit, [4] Department of Prosthodontics

3.3.3. Data

The questionnaire had five main parts. First, basic demographic information was gathered. In the second section, participants answered questions regarding their oral health, which included self-assessed oral health using the Oral-Health Single Question (OH-SQ) based on the previous week (responses were rated as 1. Excellent, 2. Very Good, 3. Good, 4. Average, 5. Poor, 6. Very Poor, 7. Don't Know). Additionally, participants reported their dental treatment needs (DTN) (current status: Yes/No/Don't Know), gingival bleeding (during the last week: Yes/No/Don't Know), chewing ability (during the last week: Yes/No/Don't Know), oral pain (during the last week: Yes/No/Don't Know), dry mouth (during the last week: Yes/No/Don't Know), halitosis (during the last week: Yes/No/Don't Know), and satisfaction with the appearance of their teeth (current

status: Yes/No/Don't Know). Third, participants completed the GOHAI-HU section. Fourth, they responded to the OHIP-14 section and the 5-level version of the EQ-5D GHRQoL instrument (EQ-5D-5L). The EQ-5D questionnaire consists of two parts: the descriptive system inquiries about the level of problems in five domains (i.e., mobility, self-care, usual activities, pain/discomfort, and anxiety/depression), while the EQ-VAS is a visual analogue scale inquiring the respondents' subjective health status between the best (100) and worst (0) imaginable health (51).

Following the interviews, an oral examination was conducted. The number of teeth (up to 28), missing teeth, decayed teeth, filled/crowned teeth, mobile teeth, the type of dentures worn (fixed/removable), and any observed mucosal changes were recorded.

During the second session, conducted one week after the first, participants were first asked two questions of opposite polarity to verify the stability of their oral health status. Ensuring consistent test conditions, a critical requirement for reliability testing was also prioritized. Once oral health stability was confirmed, the same interviewer re-administered the GOHAI-HU questionnaire along with questions about the persistence of symptoms originally recorded during the first session.

3.3.4. Statistical Analysis

Incompletely filled questionnaires were excluded from the evaluation. For every psychometric property, we considered the terminology and reference values defined and proposed by COSMIN (37).

3.3.4.1. Descriptive Analyses

The proportion of female respondents, mean and standard deviation (SD) of age, ADD-GOHAI, SC-GOHAI, OHIP, EQ-5D-5L index, and EQ-VAS values were reported by each subsample. The proportions of respondents by the chosen level of each GOHAI item were tabulated. The floor and ceiling effect of ADD-GOHAI and SC-GOHAI was assessed against a 15% threshold (52). Subgroups were compared using the independent t-test and two-sided significance test or one-way ANOVA in case of multiple categories.

3.3.4.2. Content Validity

Given the widespread use of GOHAI over decades in several countries, we considered its content validity as established (31). Therefore, content validity was not assessed in this study.

3.3.4.3. Structural Validity

When developing GOHAI, Atchison proposed a three-dimensional structure for it (19). Following Kressin et al. (53), we classified the items into these three domains as follows: Physical function (items 1-4), Psychosocial function (items 6, 7, 9-11), and Pain and discomfort (items 5, 8, 12). Our COSMIN systematic review showed that the factor structure of the GOHAI exhibits significant variability. While studies identify structures of one to five factors, an overall unidimensional structure is more plausible (31).

We compared two CFA models for ADD-GOHAI scores (Fig. 2.).

In the three-factor model, we allowed correlations between the "Physical function," "Psychosocial function," and "Pain and discomfort" factors, assuming that these three primary dimensions reflect a single underlying OHRQoL construct (i.e., a secondary factor).

In contrast, the one-factor model assumed a single OHRQoL construct with correlated errors between items within each of the three factors. While both models reflect a similar structure, they differ in parameter constraints and degrees of freedom (i.e., the number of freely estimable parameters), resulting in different factor loadings (54, 55). The adequacy of the CFA model was evaluated using the following fit indices and threshold values: comparative fit index (CFI) > 0.95, standardized root-mean-square residual (SRMR) < 0.08, and root-mean-square error of approximation (RMSEA) < 0.06. Given the non-normal distribution of GOHAI scores, we applied the Satorra-Bentler estimator using a scaled chi^2 statistic when calculating the fit-indices (56).
We also calculated modification indices to identify potential improvements in model fit via adding additional error correlations. By evaluating the modification indices, we allowed minimal and justifiable amendments of the theoretical models (57).

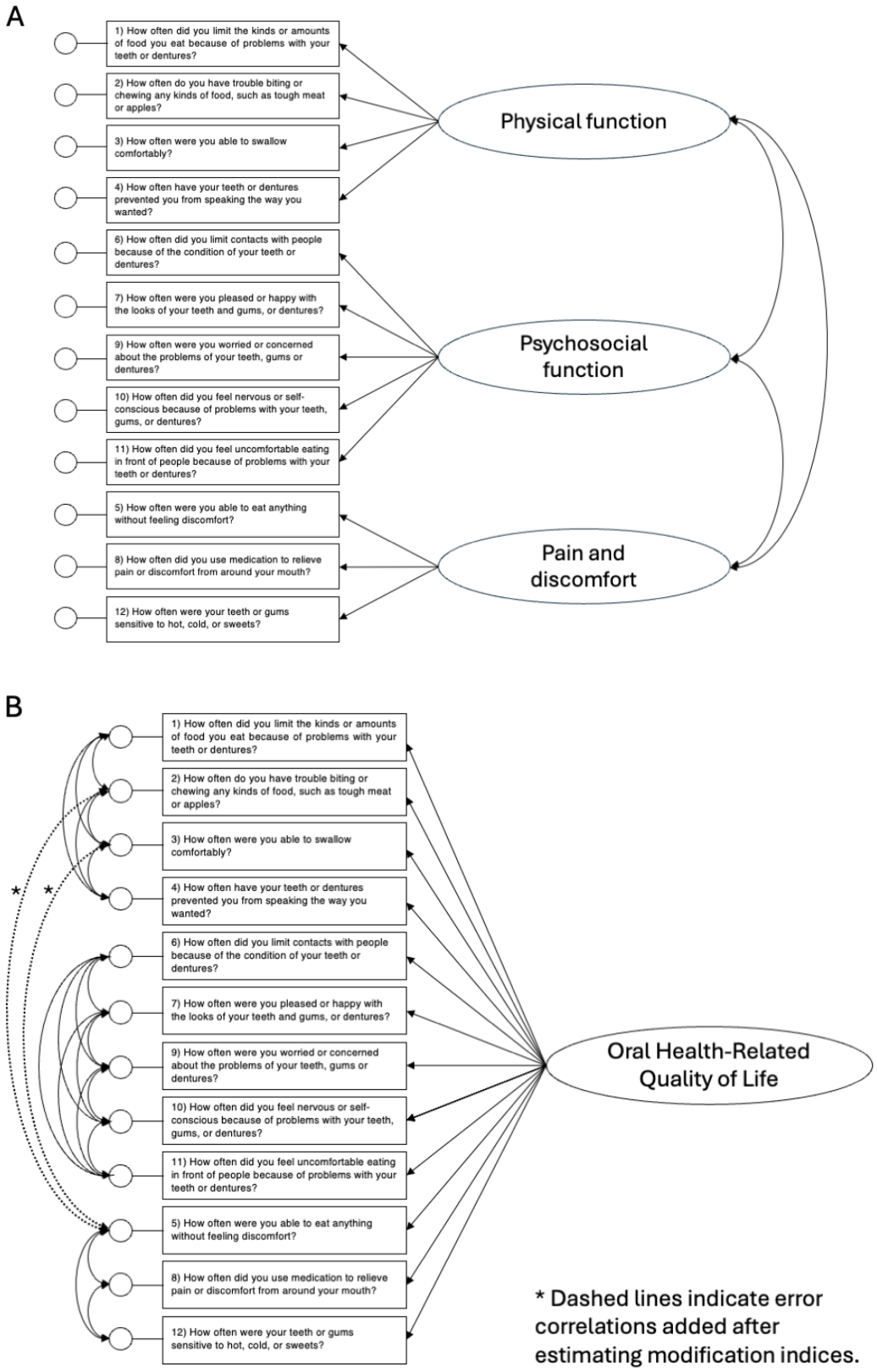

Fig.2. ***Path diagrams of the A) three-factor and B) one-factor Confirmatory Factor Analysis (CFA) models*** **(35)**

We also conducted EFA for ADD-GOHAI. We computed varimax orthogonal rotated factor loadings. The number of extracted factors was determined by the Kaiser-Guttman criterion (i.e., factors with eigenvalues >1 were retained), observing the elbow of the Scree plot, and the lowest Bayesian Information Criterion (58). Factor loadings of 0.32 and 0.5 were considered as minimum and adequate / strong, respectively (59). CFA and EFA were conducted on the total sample, general and clinical population subgroups, age subgroups, and retest data. On the total sample we also explored inter-item Spearman correlations.

3.3.4.4. Internal Consistency

To assess internal consistency, Cronbach's alpha coefficient was calculated for the three dimensions of GOHAI, as well as for the entire questionnaire. Alpha values ≥ 0.7 were regarded as sufficient.

3.3.4.5. Measurement Invariance

Measurement invariance examines whether individuals from different groups with the same latent trait level respond similarly to individual items (60). It is essential in psychometric analysis, as it ensures that a tool measures the construct consistently across various groups. This consistency is crucial for comparing groups, ensuring that any observed score differences truly reflect variations in the underlying construct rather than biases in item interpretation or response. Measurement invariance encompasses three hierarchical levels: (1) Configural invariance is the baseline level that confirms that the measurement structure (e.g., the factor structure) is similar across groups. For configural invariance to hold, the same items must load onto the same factors across all groups, although loading magnitudes may differ. Establishing configural invariance confirms that the construct is perceived similarly across groups, validating that participants interpret the items in comparable ways. (2) Building on configural invariance, metric invariance ensures that factor loadings are consistent across groups. Establishing metric invariance indicates that each item's contribution to the construct is equivalent, enabling meaningful comparisons of the relationships between items and the underlying latent construct. This level is essential for comparing factor scores reliably across groups. (3) Strong invariance is a stricter level, which requires equivalency of item intercepts (or means) in addition to

factor loadings across groups. Achieving strong invariance suggests that any group differences in observed scores reflect actual differences in the underlying latent trait rather than biases in item interpretation. This level is crucial for comparing mean scores across groups reliably. If measurement invariance fails at any of these levels, it indicates possible measurement biases or differences in how groups interpret certain items, impacting the validity of comparisons.

We tested the measurement invariance of both ADD-GOHAI and SC-GOHAI using multi-group CFA by the clinical and general population subgroups, as well as the 18-64 years old and 64+ years old subgroups. Following the hierarchy of measurement invariance (60), first, we ensured configural invariance by checking the fit indices of the one- and three-dimensional CFA models for each subgroup. We assumed that the same factor structure shows best fit in each subgroup. Then, metric invariance was tested for each GOHAI item via the chi^2 test to compare the factor loadings by subgroups. Overall metric invariance was examined using a joint Lagrange multiplier (LM) test of all factor loadings. Strong invariance assuming equal factor loadings and intercepts was also tested by each item, and overall. With p values < 0.05, the hypothesis of equal factor loadings and equal intercepts can be rejected, suggesting that the contribution of items to the overall GOHAI score differs between subgroups. Sample heterogeneity beyond the tested subgroups (i.e., clinical heterogeneity within age groups, or the clinical or general population samples) may interfere with the testing of measurement invariance. Hence, our measurement invariance results have to be interpreted with caution.

3.3.4.6. Reliability

Reliability refers to the consistency of a measurement tool in distinguishing “true” differences in a trait or construct among individuals, rather than random variations or errors. According to Classical Test Theory , each observed score comprises a "true" score and an error component. Therefore, reliability represents the proportion of the total score variance that can be attributed to actual differences in the construct being measured, rather than random error. This "true" score is not an absolute value but rather the average score that would be expected if the measurement were administered repeatedly an infinite number of times under the same conditions. Reliability is essential for determining if a

PROM can consistently differentiate between individuals (36, 38).To ensure accurate reliability assessments, certain conditions must be met: (1) Stability of the construct: Participants should remain stable in the construct being measured across repeated administrations. Stability is often verified using a global rating of change or confirmed if no significant interventions occurred between measurements. In assessing the reliability of the Hungarian GOHAI, we included two questions with opposing polarity during the second assessment to confirm stability. One question asked if participants noticed any changes in their oral health since the previous visit, and another asked if they would consider their oral health stable since the last meeting. This approach encouraged reflection and reducing automatic responses. (2) Appropriate time interval: The interval between repeated administrations should be long enough to prevent recall bias but short enough to avoid real changes in the construct. For many PROMs, a two-week interval is often appropriate. Following COSMIN guidelines, we used a one- to two-week interval between the two administrations in our study, balancing the need to avoid recall bias while ensuring that participants' oral health remained relatively unchanged. (3) Consistency of test conditions: Test conditions should remain as similar as possible to avoid influencing responses due to external factors. During our data collection, we ensured that follow-up interviews were conducted by the same interviewer for consistency. For elderly participants in nursing homes, the second data collection was conducted in person, mirroring the initial setting. For some TMD patients, the follow-up was conducted over the phone, with careful attention to replicate the tone, instructions, and atmosphere as closely as possible to the initial interview.

For assessing the reliability of continuous scores, the intraclass correlation coefficient (ICC) is generally preferred. The recommended ICC model is the two-way random effects model, which accounts not only for variance within individuals but also for systematic differences between time points. This approach captures both random and systematic variability, providing a comprehensive measure of reliability across repeated assessments.

Using the dataset with repeatedly administered questionnaires, we measured test-retest reliability for ADD-GOHAI and SC-GOHAI through the ICC agreement formula (ICC

model 2.1) (61). For each participant, repeated administrations of GOHAI were performed by the same dental doctor, so the systematic difference between test-retest scores due to raters was negligible. ICC≥ 0.7 was regarded as a threshold for sufficient test-retest reliability.

3.3.4.7. Standard Error of Measurement (SEM) and smallest detectable change (SDC)

SEM calculation was based on the SEM agreement formula for repeatedly administered GOHAI scores (61, 62):

$$SEM = \sqrt{\sigma_{rater}^2 + \sigma_{residual}^2} \quad (1)$$

where σ_{rater}^2 is the variance due to raters (i.e., differences between test and retest administrations by the same doctor in this study) and $\sigma_{residual}^2$ is the residual error variance. When measuring GOHAI for a respondent, due to the measurement error, the true GOHAI score lies within the 95% confidence interval (95%CI) of the measured value calculated as $\pm 1.96 * SEM$. The SDC (i.e., the change in the GOHAI score of a respondent, which is not attributable to measurement error) was calculated by the following formula:

$$SDC = 1.96 * \sqrt{2} * SEM. \quad (2)$$

3.3.4.8. Construct Validity

During the assessment of construct validity, we followed the expected effect size magnitudes and directions defined in our COSMIN review of GOHAI (63), and we followed Cohen's criteria when interpreting the magnitude of effects (64). Convergent validity was tested via Spearman correlation due to the non-normal distribution of GOHAI scores. We expected a strong correlation (r >0.50) between GOHAI and instruments measuring similar constructs (OHIP-14 for OHRQoL, and OH-SQ for self-assessed oral health). For instruments measuring related but dissimilar self-reported constructs (i.e., DTN, EQ-5D-5L index, EQ-VAS), the expected correlation was moderate (r ~ 0.30-0.50). For physician-reported objective measures of dental status (e.g., Decayed, Missing, and Filled Teeth (DMFT) Index or number of teeth), we expected a weak (r ~ 0.10-0.30) correlation, as dental problems are usually alleviated by prosthodontic or dental treatments, and dental status does not reflect the entire spectrum of oral health.

We assessed discriminative (i.e., known groups) validity using independent t-tests and one-sided p values. The effect size was reported as standardized mean difference. We hypothesized that patients with poor dental status, bleeding gum, oral pain, chewing problems, xerostomia, halitosis, and aesthetical problems would have worse GOHAI scores, and the magnitude of the effect size would be at least small. Altogether, for convergent validity and known-groups validity, we evaluated six and ten hypotheses, respectively.

3.3.4.9. Criterion Validity and Responsiveness

As for all PROMs, there is no available gold standard measure for OHRQoL instruments, so we did not investigate criterion validity (i.e., the agreement with a gold standard) in this study (65). The responsiveness of GOHAI was not investigated either, as it can be measured in the context of an intervention, and our study was not intervention-based.

4. RESULTS

4.1. Results of Study I.: Psychometric Properties of General Oral Health Assessment Index Across Ages: COSMIN Systematic Review

4.1.1. Search

A total of 497 records were initially identified in the electronic databases. After removing duplicates, 283 records were screened based on title and abstract. On the 57 randomly selected records screened in duplicate, the agreement between the two reviewers was 86.0% (kappa = 0.714). A joint decision was made on 24 records; 72 articles underwent full-text assessment, and 60 were included in the review [Fig. 3.]. The full-text paper of a study abstract identified in the examined databases was retrieved from other sources. The lists of excluded and included reports can be found in the Supplementary Materials of the original publication (Supplementary Tables 2. and 4., respectively (31)).

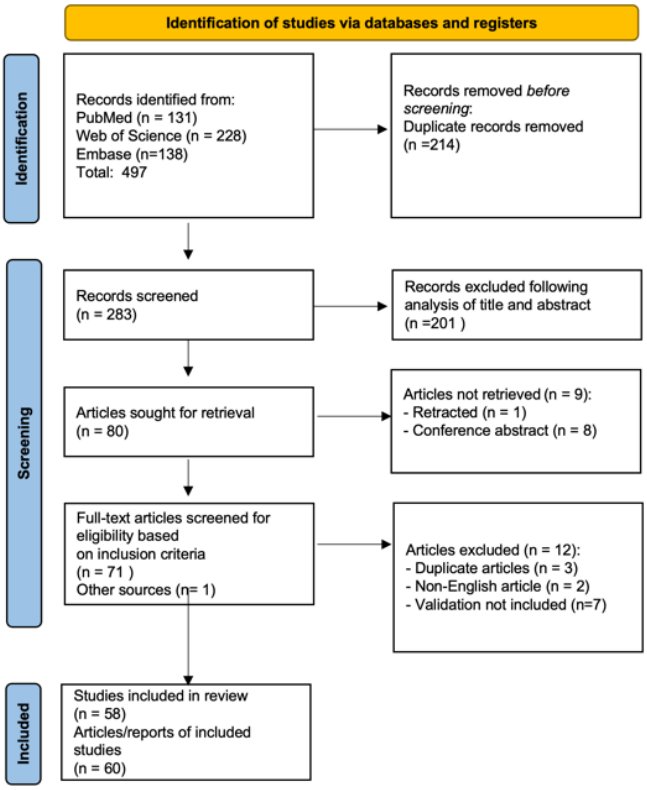

Fig. 3.: PRISMA Flow chart (31)

4.1.2. Internal Structure

4.1.2.1. Structural Validity

Structural validity was evaluated in 27 studies. CFA was employed in four studies (with very good methodological quality), while EFA/PCFA was conducted in 23 studies (with adequate methodological quality). Most studies (n=19, 70%) were conducted in the elderly (≥ 60 years old). Structural validity was not assessed in the ≤45-year-old age group, while 5 studies (20%) were performed in patients under 60 years of age.

Results were inconsistent across all age groups and overall. Neither a one-factor nor a three-factor model fitted optimally in the four CFA studies (45, 66-68). However, using a three-factor model, the Brazilian version (33) fit acceptably (CFI=0.942) according to COSMIN standards (CFI of 0.95). The results of EFA/PCFA showed a wide range of factor structures, varying from one to five factors [Table 2.].

Although the results differed in terms of the number of factors, overall, the interpretation of the findings can be considered consistent. Atchison's study (n=1755) (19), along with five other studies (69-72), identified a single-factor structure. In line with this, most of the studies that identified multiple factors (Gutiérrez et al. n=7200 (73), Sánchez-García et al. n=696 (74)) concluded as well, that OHRQoL should be interpreted as a single construct rather than being divided into distinct dimensions. The lack of clear separation between different dimensions suggests that OHRQoL measured by GOHAI should be viewed as a unidimensional construct. However, we downgraded the quality of evidence to low in all age groups and overall due to significant inconsistency in the results [Table 3.].

For more information, refer to the Supplementary Materials of the original publication (Supplementary Table 5 (31)).

Table 2. Results of factor analyses (31)

N: Number of studies; n: pooled sample size

		Exploratory										Confirmatory	
		1 factor		2 factors		3 factors		4 factors		5 factors		1 / 3 factors	
Age group	Quality	N	n	N	n	N	n	N	n	N	n	N	n
Overall	Very good											4	2525
	Adequate	3	633	5	9148	8	2971	3	1106	1	197		
	Doubtful	3	2148										
≥60	Very good											1	613
	Adequate	2	345	4	8642	6	2729	2	841	1	197		
	Doubtful	3	2148										
All ages	Very good											1	211
	Adequate					2	242						
	Doubtful												
<60	Very good											2	1701
	Adequate	1	288	1	506			1	265				
	Doubtful												
≤45	information not available												

4.1.2.2. Internal Consistency

Internal consistency was assessed in 50 studies and in all age groups. With the exception of one study (75), all authors assessed the internal consistency (α) for the entire scale, which, by accepting the one-dimensional OHRQoL model, we considered as methodologically appropriate. The random-effects meta-analysis estimate of overall α=0.81 and α-values between 0.80 and 0.86 in the four age groups suggested that the internal consistency of GOHAI is sufficient. α was highest among ≤45-year-old adults, and the difference between the four age groups was significant (χ_3^2=9.59, p=0.02). The heterogeneity of α estimates was high with overall I^2=96.05% and I^2 values between 86.55%-97.15% across the age groups. We assigned a rating of "sufficient" in every age group, and the quality of evidence was determined as "high" due to the presence of at least one very good study in each group without the need for downgrading [Table 3.]. For more information, refer to the Supplementary Materials of the original publication (Supplementary Fig 1., Supplementary Table 5. (31))

Table 3. Summary of Findings (31)

n: number of studies; N: pooled sample size; (+): hypothesis confirmed, (-): hypothesis not confirmed, (?): indeterminate result; [a]multiple very good studies, consistent results; [b] there is one very good study available; [c] multiple doubtful studies, consistent results; [d]one doubtful study, imprecision (n<50); [e] multiple very good studies, inconsistent results; [f] multiple very good studies, very serious inconsistency; [g] multiple inadequate studies; inconsistent results; [i] multiple adequate studies, very serious inconsistency

	Age group	**n**	**N**	**Exploratory (n)**	**Confirmatory (n)**	**Overall rating**	**Evidence quality**
Structural validity	Overall	27	18728	23	4	sufficient	Low[i]
	≥60	19	15515	18	1	sufficient	Low[i]
	All ages	3	453	2	1	sufficient	Low[i]
	<60	5	2760	3	2	sufficient	Low[i]
	≤45	-	-	-	-	-	-
	Age group	**n**	**N**	**Cronbach α [95% CI]**	**I²**	**Overall rating**	**Evidence quality**
Internal consistency	Overall	50	21961	0.81 [0.78-0.83]	96.05%	sufficient	High[a]
	≥60	33	18215	0.80 [0.77-0.83]	97.15%	sufficient	High[a]
	All ages	12	1543	0.80 [0.76-0.84]	86.55%	sufficient	High[a]
	<60	4	1973	0.83 [0.78-0.87]	92.58%	sufficient	High[a]
	≤45	1	230	0.86 [0.83-0.89]	-	sufficient	High[b]
	Age group	**n**	**N**	**r [95%CI]**	**I²**	**Overall rating**	**Evidence quality**
Reliability	Overall	24	1262	0.84 [0.79-0.87]	78.09%	sufficient	Moderate[c]
	≥60	16	1014	0.84 [0.79-0.89]	83.31%	sufficient	Moderate[c]
	All ages	6	186	0.81 [0.71-0.88]	58.11%	sufficient	Moderate[c]
	<60	1	30	0.72 [0.49-0.86]	-	sufficient	Very low[d]
	≤45	1	32	0.87 [0.75-0.94]	-	sufficient	Very low[d]
	Age group	**n**	**(+)**	**(-)**	**(?)**	**Overall rating**	**Evidence quality**
Construct validity	Overall	49	234	79	48	sufficient	Moderate[e]
	≥60	31	151	49	30	sufficient	Moderate[e]
	All ages	12	45	21	13	sufficient	Low[f]
	<60	5	30	8	5	sufficient	High[a]
	≤45	1	8	1	0	sufficient	High[b]
	Age group	**n**	**(+)**	**(-)**	**(?)**	**Overall rating**	**Evidence quality**
Responsiveness	Overall	7	5	1	2	indeterminate	Very low[g]
	≥60	3	2	0	1	indeterminate	Very low[g]
	All ages	4	3	1	1	indeterminate	Very low[g]
	<60	-	-	-	-	-	-
	≤45	-	-	-	-	-	-

4.1.3. Other Measurement Properties

4.1.3.1. Reliability

Test-retest reliability was evaluated in 24 studies, with 16 (66%) studies in the ≥60-year-old age group and 6 (25%) studies including participants of all ages. Only one study was available in each of the <60-year-old and ≤45-year-old groups. The overall random-effect meta-analysis estimate of test-retest reliability was r=0.84, with values ranging between 0.79-0.87 across the four age groups with no significant between-group difference (χ_3^2=3.16, p=0.37). The detailed results, including heterogeneity statistics, are provided in the Supplementary Materials of the original publication (Supplementary Fig. 2. (31)) Reliability was consistently sufficient, both overall and across all age groups. The overall, ≥60-year-old, and all-age subgroups had "moderate" quality of evidence for reliability, downgraded due to methodological limitations. The <60-year-old and ≤45-year-old age groups had "very low" quality of evidence, further downgraded due to serious limitations in precision, specifically sample sizes < 50 [Table 3.].

4.1.3.2. Construct Validity

Out of the 60 included studies, construct validity was assessed in 49, resulting in the evaluation of a total of 361 hypotheses. Of these, 135 (37.4%) were evaluated for convergent validity and 226 (62.6%) for known-groups validity. Altogether 313 (86.7%) hypotheses were quantifiable, while reporting was insufficient to calculate the effect size for 48 hypotheses (13.3%), rendering the results indeterminate. Among the quantifiable hypotheses, the predefined effect size criteria were confirmed by overall 74.7% (234/313, 95%CI 69.6%-79.5%), supporting the construct validity of GOHAI. The percentage of confirmed hypotheses in the ≥ 60-year-old age group, all ages, <60-year-old age group, and ≤ 45-year-olds were respectively 75.5% (151/200), 66.7% (44/66), 78.9% (30/38) and 88.9% (8/9) [Table 3.].

4.1.3.2.1. Convergent Validity

The convergent validity of GOHAI was supported by 83.2% (99/119) of the quantifiable hypotheses. We found a consistently strong correlation between GOHAI and OHIP-14 scores across all studied age groups (H_{A1}), with 94.4% (17/18) of predefined hypotheses

confirmed. Only 67.6% (23/34) of hypotheses expecting a strong correlation between GOHAI and self-rated oral health (H_{A2}) were confirmed due to moderate correlation in 9 (26.4%) and low correlation in 2 (5.9%) studies. However, in the three studies conducted on <60-year-olds, the correlation between GOHAI and self-reported oral health was strong. The predefined hypotheses concerning satisfaction (H_{A3-A4}), self-rated need for dental treatment (H_{A5}), self-rated general health (H_{B1-B3}), and self-rated mental health (H_{B4-B7}) were confirmed in 92.3% (12/13), 71.4% (15/21), 96% (24/25) and 100% (8/8) of the studies [Fig.4.]. Please refer to the Supplementary Materials of the original publication for detailed results (Supplementary Table 6. (31)) and for a summary of the findings related to convergent validity (Supplementary Table 7 (31)).

4.1.3.2.2. Known-groups Validity

Known-groups validity was supported by 69.5% (135/194) of the quantifiable hypotheses, which is below the 75% threshold proposed by COSMIN. Hypotheses concerning dental status (H_{C1}), prosthetic status (H_{C2}), periodontal and mucosal health / objective oral hygiene (H_{C3}), behavioral variables (H_{D1}), and oral symptoms (H_{D2}) were confirmed by 71.9% (64/89), 76.9% (20/26), 63% (17/27), 16.7 (2/12), 80% (32/40) of studies, respectively [Fig.4.]. Please refer to the Supplementary Materials of the original publication for detailed results (Supplementary Table 8. (31)) and for a summary of the findings related to known-groups validity (Supplementary Table 9 (31)).

Since convergent validity and known-groups validity hypothesis tests are both related to the same psychometric property, construct validity, we summarized the results by age group and overall, presenting the number of confirmed (+), rejected (-), and indeterminate (?) hypotheses. Our construct validity summary shows sufficient ratings in all age groups and overall [Table 3.]. However, due to inconsistency, evidence quality was downgraded, varying from ”moderate” ratings for overall and ≥60-year-old, a ”low” rating for all-age, and ”high” ratings for <60-year-old and ≤45-year-old subgroups.

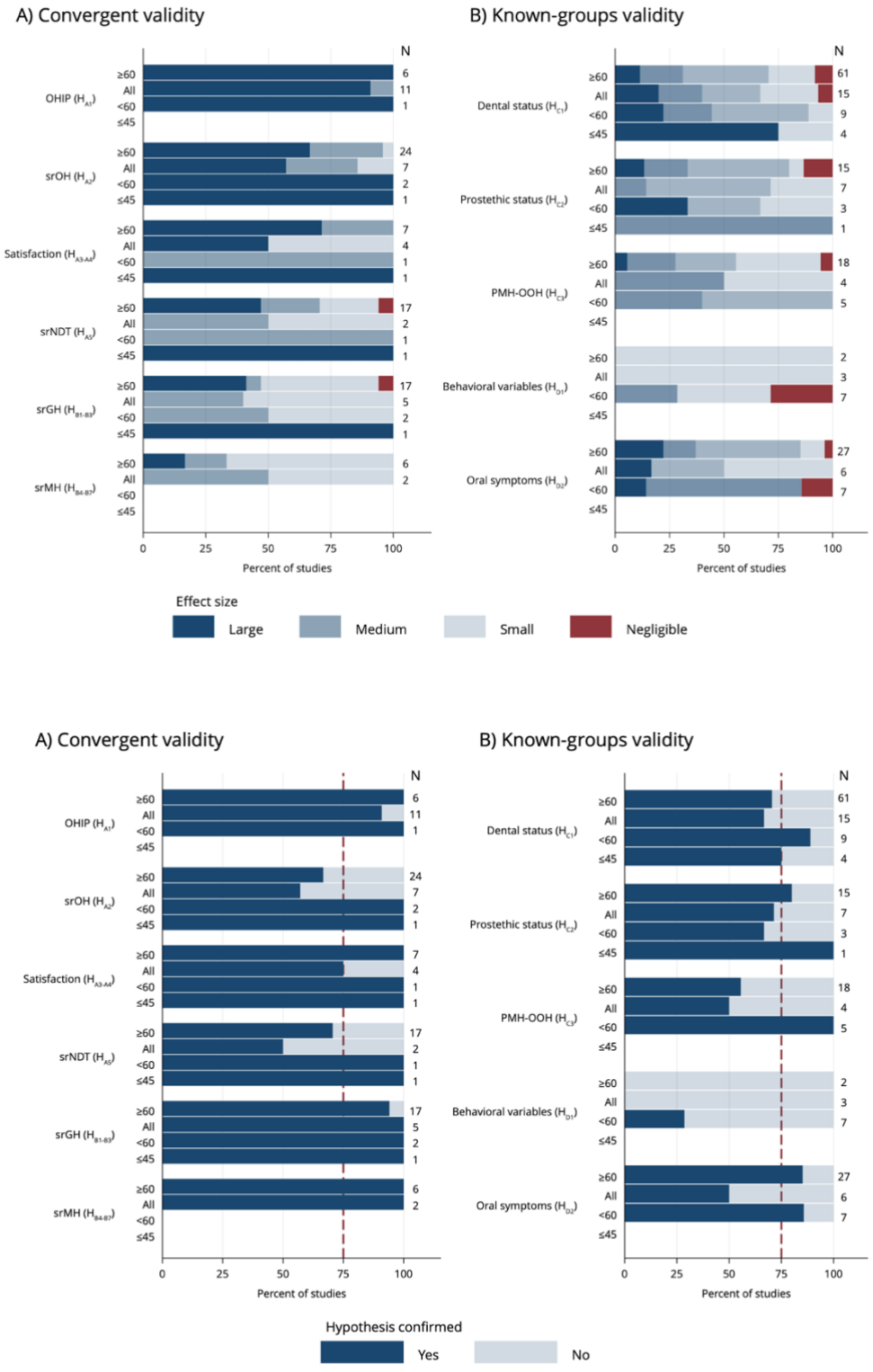

Fig. 4.: Summary of hypothesis testing for convergent-, and known-groups validity (31)

N: number of studies; H_{A1}: OHIP-14/OHIP-EDENT, H_{A2}: self-rated oral health, H_{A3}: satisfaction with oral health/teeth/mouth, H_{A4}: denture satisfaction questionnaire, H_{A5}: self-rated need for dental treatment, H_{B1}: self-rated general health, H_{B2}: Medical Outcomes Study Short Form Health Survey (SF-20), H_{B3}: Medical Outcomes Study Short Form Health Survey (SF-36), H_{B4}: life stress/perceived stress questionnaire, H_{B5}: morale index/Philadelphia Geriatric Center Morale Scale, H_{B6}: Hospital Anxiety and Depression scale, H_{B7}: satisfaction with life situation, H_{C1}: dental status, H_{C2}: prosthetic status, H_{C3}: periodontal and mucosal health, objective oral hygiene, H_{D1}: behavioral variables, H_{D2}: oral symptoms (The detailed categorization and numbering of the hypotheses can be found in Table 1.)

4.1.3.3. Responsiveness

Limited evidence exists for responsiveness, with few studies (n=7) and no information for the <60-year-old and ≤45-year-old subgroups. The methodological quality and results of hypotheses testing before and after intervention for responsiveness can be found in the Supplementary Materials of the original publication, see Supplementary Table 10 (31). In every instance, a suitable hypothesis was lacking, and the reviewer team was unable to formulate one. As a result, the rating of responsiveness overall, as well as in the ≥60-year-old and all-age groups, is 'indeterminate' with 'very low' quality of evidence due to significant risk of bias and inconsistency. [Table 3.].

4.1.4. Other Findings

4.1.4.1. Item Response Theory/ Rasch analyses

Two reports utilized item response theory or Rasch analysis, but both had limitations. The first study (76) had a small sample size of only 85 individuals, which is insufficient for Rasch models (37). We omitted the second report (77), which used the same sample as two previous studies (78, 79).

4.1.4.2. Scoring Variants of GOHAI

The studies evaluated revealed a prevalent use of a 5-point Likert scale (n=43, 71.7%), with a smaller proportion utilizing a 6-point Likert scale (n=8, 13.3%), or a 3-point Likert scale (n=8, 13.3%) (80, 81). Several studies opted to combine response options into fewer

categories, while in two instances, the 6-point and 5-point scale was simplified to a 3-point scale. (81). For more detailed information, please refer to Supplementary Table 4 in the original publication (31). We note that while using fewer Likert items may seem more convenient, the reliability and validity of scales with 5 or 6 response categories are usually considered superior in the psychometric literature. (82)

4.1.4.3. Handling of Missing Data, "Floor and Ceiling" Effect

Four articles addressed the "floor and ceiling" effect, and none reported experiencing this phenomenon (30, 83-85). Regarding missing data, 11 studies mentioned their handling: one study excluded participants with only one missing data (19), five studies allowed one missing data (71, 83, 84, 86, 87), and five studies allowed two (53, 69, 72, 88, 89). Missing data were replaced using mean, median, or multiple linear regression analysis methods.

4.2. Results of Study II.: Validation of the Hungarian version of the General Oral Health Assessment Index (GOHAI) in clinical and general populations

4.2.1. Descriptive Analyses

4.2.1.1. Sample Characteristics

A total of 315 participants completed the questionnaire, of which 9 (2.8%) provided incomplete data. Therefore, data from 306 participants without missing information were analyzed. The questionnaire was administered twice to 108 individuals. For the TMD population, the second interviews were conducted via telephone (n=46), while for all other cases, they were conducted in person, in the same location (n=62). Mean (SD) age of the total sample was 57.3 (20.4) years; 71.2% were female; 16.0%, 43.1%, and 40.9% had primary, secondary, and tertiary education respectively; 75.5% lived in cities, 16.7% in towns, and 7.8% in rural areas. Over half of the participants (54.9%) were recruited from clinical populations, and 45.1% were 65+ years old. Table 4 displays the descriptive statistics for each subgroup. For the median and interquartile range values of the scales, see the Supplementary Material of the original publication (Table S2 (35)). As expected, the clinical population showed a lower average ADD-GOHAI score than the general population (t-test p=0.002). The clinical population's GHRQoL was somewhat better than

the general population, which is explained by the age difference. The ADD-GOHAI scores did not differ between the younger and older age groups (t-test p=0.127). Despite similar mean age, GOHAI scores of respondents with primary education showed worse OHRQoL than those with tertiary education (ANOVA p=0.001). The ADD-GOHAI scores were similar for the urban and rural residents (ANOVA p=0.150).

4.2.1.2. Distribution of GOHAI Scores

Appendix Fig. S2 of the original publication illustrates the distribution of ADD-GOHAI scores by subgroup, including the proportion of respondents with the highest and lowest possible scores (35). While no respondents scored 12 (worse OHRQoL), indicating a lack of floor effect, 18.8% in the total sample and 20.3% in the 65+ years old subgroup scored 60 (best OHRQoL), indicating that ADD-GOHAI has a ceiling effect. Appendix Fig. S3 of the original publication shows the distribution of SC-GOHAI scores by subgroup (35). In the total sample and all subgroups, over 15% of respondents had a 0 score (best OHRQoL), indicating a floor effect for SC-GOHAI. Due to a greater proportion of respondents without problems in OHRQoL, both ADD-GOHAI and SC-GOHAI have skewed distribution.

4.2.1.3. Responses by Item

Most respondents (59.5%) indicated the presence of problems on item 9 (worry about the problems with teeth, gums, or dentures), while problems occurred least often (9.5%) on item 6 (limiting contacts with people due to the condition of teeth or dentures). On item 7 (problems with the looks of teeth, gums, or dentures), 20.3% of respondents always had problems, while on item 6, only 0.7% indicated the always option. The proportion of responses by item is shown in Figure 5; further details are provided in the Appendix of the original publication (Table S3 (35)). Furthermore, the proportion of sometimes / often / always responses contributing to SC-GOHAI is provided in the Appendix of the original publication (Fig. S4 (35)).

Table 4. Sample Characteristics and Mean Values with Standard Deviations for Add-GOHAI, SC-GOHAI, OHIP, EQ-5D-5L, and EQ-VAS by Subgroups (35)

[a] Semmelweis University, Department of Oral Diagnostics; [b] Semmelweis University, Department of Prosthodontics; [c] Semmelweis University, Temporomandibular Disorders Care Unit

	Population	Site	N	Female %	Age mean (SD) years	ADD-GOHAI mean (SD)	SC-GOHAI mean (SD)	OHIP mean (SD)	EQ-5D-5L index mean (SD)	EQ VAS mean (SD)
Retest No	General	Screening kiosk	105	67.6%	61.3 (15.6)	51 (8.7)	2.7 (2.6)	65.6 (6.9)	0.86 (0.23)	69.9 (17.7)
	Clinical	SU-OD[a]	55	67.3%	48.9 (19.3)	48.5 (9.1)	3.6 (2.8)	62.1 (9.6)	0.91 (0.13)	70.5 (17.9)
		SU-DP[b]	17	76.5%	61.7 (17.4)	46.4 (7.9)	4.5 (2.3)	61.9 (8.1)	0.87 (0.13)	73.2 (16)
		SU-TMD[c]	21	90.5%	43.9 (19.4)	49.2 (8.6)	3.4 (2.6)	61.5 (9.2)	0.9 (0.14)	79 (12)
	Subtotal	-	198	70.7%	56 (18.5)	49.7 (8.8)	3.2 (2.7)	63.9 (8.2)	0.88 (0.19)	71.3 (17.2)
Retest Yes	General	Retirement home	33	69.7%	87.8 (4.4)	55.4 (5.9)	1.4 (1.9)	67.4 (3.8)	0.76 (0.29)	67.6 (20.6)
	Clinical	SU-OD[a]	15	60%	49.7 (18.8)	51.9 (7.5)	2.7 (2.7)	62.8 (10.5)	0.93 (0.12)	77.9 (19.4)
		SU-DP[b]	14	92.9%	62.4 (11.9)	49.4 (9.1)	3.3 (2.8)	62.4 (10)	0.93 (0.07)	73.6 (16.2)
		SU-TMD[c]	46	71.7%	42.1 (14.6)	49.2 (8.7)	3.5 (2.8)	60.8 (9.9)	0.91 (0.15)	74.3 (16.6)
	Subtotal	-	108	72.2%	59.8 (23.4)	51.5 (8.2)	2.7 (2.7)	63.3 (9)	0.87 (0.2)	72.6 (18.4)
Population	General	-	138	68.1%	67.6 (17.9)	52.1 (8.3)	2.4 (2.5)	66 (6.3)	0.84 (0.25)	69.4 (18.4)
	Clinical	-	168	73.8%	48.9 (18.4)	48.9 (8.6)	3.5 (2.7)	61.8 (9.5)	0.91 (0.13)	73.8 (16.7)
Age-group	Age:18-64	-	168	70.8%	41.8 (12.9)	49.7 (8.1)	3.3 (2.6)	63.1 (8.7)	0.92 (0.14)	75.7 (15.9)
	Age: 64 +	-	138	71.7%	76.2 (8.2)	51.2 (9.2)	2.7 (2.7)	64.5 (8.1)	0.82 (0.23)	67.1 (18.5)
Education	Primary	-	49	63.3%	56.6 (19.6)	46.8 (10.9)	4.1 (3.2)	60.5 (11.3)	0.83 (0.22)	66.4 (19.4)
	Secondary	-	132	80.3%	57.7 (20.3)	50.2 (8.5)	3.0 (2.6)	63.3 (8.6)	0.87 (0.2)	70.7 (17.5)
	Tertiary	-	125	64.8%	57.3 (21)	51.9 (7.3)	2.5 (2.4)	65.3 (6.5)	0.90 (0.17)	75.1 (16.4)
Population	City	-	231	70.6%	59.6 (20.5)	50.9 (8.6)	2.8 (2.7)	64.3 (8.1)	0.87 (0.19)	70.8 (17.2)
	Town	-	51	78.4%	50.0 (18)	48.9 (8.1)	3.5 (2.6)	62.1 (8.9)	0.9 (0.16)	75.9 (17.7)
	Rural area	-	24	62.5%	51.4 (19.9)	48.2 (10.1)	3.8 (3)	61.2 (10.3)	0.81 (0.27)	72.3 (20.3)
Total	**-**	**-**	**306**	**71.2%**	**57.3 (20.4)**	**50.3 (8.6)**	**3 (2.7)**	**63.7 (8.5)**	**0.87 (0.19)**	**71.8 (17.6)**

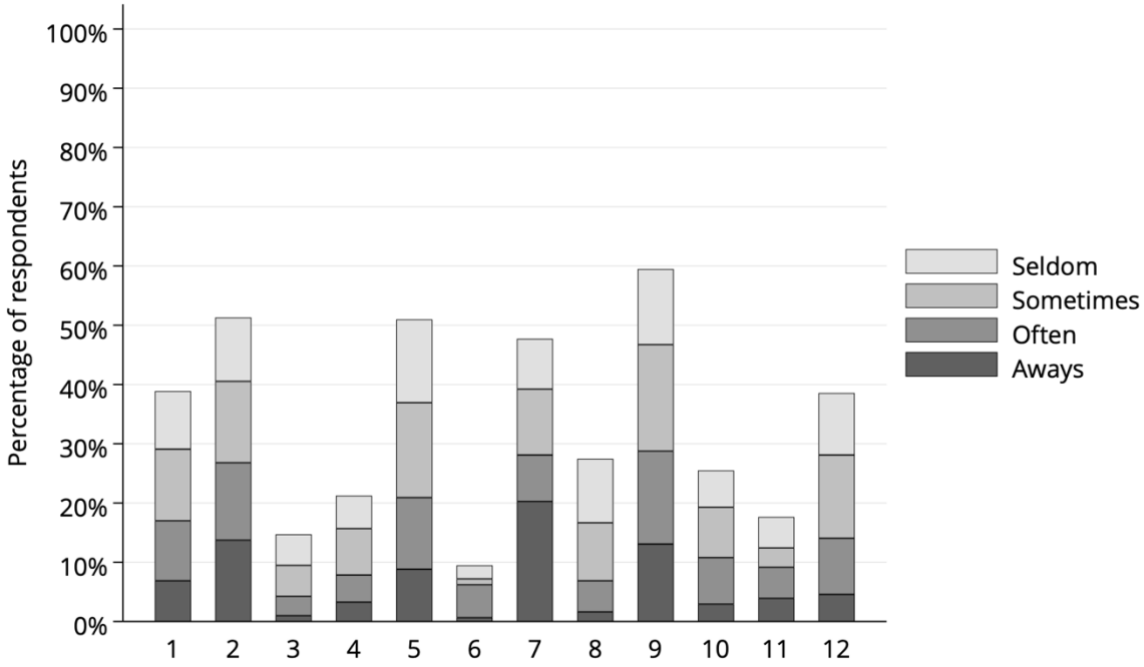

Fig. 5. The frequency distribution of problems by the items of GOHAI (35)

4.2.2. Internal Structure

4.2.2.1. Structural Validity

Most inter-item correlations were weak or negligible, and only 23 out of 66 (34.8%) inter-item correlations were moderate or strong (for details, see Fig. S5. in the Appendix of the original publication (35).). The probable cause for this, namely the presence of a mixed model structure containing both reflective and formative elements, has been discussed earlier.

Contrary to the three-factor model, the single-factor model showed good fit across all subgroups for both ADD-GOHAI and SC-GOHAI, with adequate internal consistency (Table 5.). The loadings on the main OHRQoL factor in both CFA models and the EFA model followed a similar pattern, with only minor differences (see Fig. S7. in the Appendix of the original publication (35)). All items but 4, 6, and 12 had at least minimal

loading on the single factor in both the total sample and retest data (see Fig. S8. in the Appendix of the original publication (35)).

The one-factor model met all good fit criteria of COSMIN in the total sample, the clinical subsample, and both age groups. Also, the one-factor model showed an acceptable fit in all criteria on the retest data. In addition to the correlated errors between the items of the three proposed factors, we allowed error correlation between items 2 (trouble with biting or chewing) and 3 (problems with swallowing) as well as items 3 and 5 (eating with discomfort). Although these items are from the domains of "physical function" and "pain and discomfort", they are all conceptually related to problems with eating, which justifies the amendment of the model.

In EFA, the eigenvalues and Scree plot suggested a single factor structure, while BIC suggested a three and two factor solution in the total sample and retest data, respectively (see Fig. S9. and Fig. S10. in the Appendix of the original publication (35)). In the nonrotated solution, the first factor explained 85.4% and 81.1% of the variance in the total sample and retest data, respectively. In the rotated solution in both datasets, items 1, 2, and 5 (problems with eating) had high loadings on the main factor, and items 6, 10, and 11 (psychosocial problems) had high loadings on the minor second factor while the loadings on a minor third factor were mostly weak and inconsistent. Several items had high loadings on multiple factors, and the theoretical pain/discomfort domain did not emerge as an independent dimension (see Fig. S11.- Fig. S14. in the Appendix of the original publication (35)).

Altogether, the single factor structure shown by both CFA and EFA analyses supports the use of a single OHRQoL score instead of forming subscales for various OHRQoL domains.

4.2.2.2. Internal Consistency

The Cronbach α values for the total model and each subgroup also indicated better internal consistency for a single-factor model for both ADD-GOHAI and SC-GOHAI (Table 5). In all subgroups, the single-factor model had adequate internal consistency with α values greater than 0.70.

Table 5. Fit indices of the CFA of the three-factor and one-factor model and Cronbach α values (35)

[a]Comparative fit index; [b]Tucker-Lewis index; [c]Goodness of Fit Index; [d]Root Mean Square Error of Approximation; [e]Standardized Root Mean Residuals; [f]Cronbach alpha, [g]Physical function: items 1-4; [h]Psychosocial function: items 6, 7, 9-11, [i]Pain and discomfort: items 5, 8, 12. [j]LM test p value for equal factor loadings; [k] LM test p value for equal intercepts

													Measurement invariance	
									Total	Phys[g]	Psy[h]	PD[i]		
	Model	Subgroup	n	CFI[a]	TLI[b]	GFI[c]	RMSEA[d]	SRMR[e]	α[f]	α	α	α	Loadings[j]	Intercepts[k]
ADD-GOHAI	Three-factor	Total	306	0.93	0.91	0.87	0.059	0.062	-	0.70	0.74	0.50	-	-
		Retest	108	0.89	0.86	0.77	0.077	0.086	-	0.65	0.77	0.42	-	-
		General	138	0.85	0.81	0.75	0.084	0.103		0.73	0.70	0.47	-	-
		Clinical	168	0.93	0.91	0.84	0.064	0.063	-	0.68	0.76	0.33	-	-
		18-64 years old	168	0.93	0.91	0.83	0.058	0.073	-	0.70	0.70	0.41	-	-
		64+ years old	138	0.95	0.93	0.84	0.053	0.071	-	0.70	0.78	0.56	-	-
	One-factor with correlated errors	Total	306	0.98	0.95	0.94	0.042	0.041	0.83	-	-	-	-	-
		Retest	108	0.96	0.92	0.88	0.059	0.068	0.84	-	-	-	-	-
		General	138	0.95	0.89	0.87	0.062	0.070	0.83	-	-	-	0.002	0.003
		Clinical	168	0.99	0.97	0.93	0.036	0.039	0.82	-	-	-		
		18-64 years old	168	0.97	0.95	0.91	0.068	0.051	0.81	-	-	-	0.303	<0.001
		64+ years old	138	0.97	0.95	0.90	0.047	0.050	0.85	-	-	-		
SC-GOHAI	Three-factor	Total	306	0.93	0.91	0.86	0.055	0.061	-	0.63	0.68	0.46	-	-
		Retest	108	0.90	0.88	0.76	0.066	0.081	-	0.60	0.70	0.45	-	-
		General	138	0.82	0.76	0.71	0.90	0.110	-	0.65	0.66	0.57	-	-
		Clinical	168	0.91	0.88	0.80	0.062	0.065	-	0.62	0.68	0.34	-	-
		18-64 years old	168	0.86	0.82	0.76	0.076	0.081	-	0.64	0.65	0.39	-	-
		64+ years old	138	0.91	0.88	0.78	0.063	0.079	-	0.62	0.71	0.49	-	-
	One-factor with correlated errors	Total	306	0.99	0.99	0.95	0.022	0.035	0.78	-	-	-	-	-
		Retest	108	0.95	0.91	0.86	0.058	0.068	0.80	-	-	-	-	-
		General	138	0.93	0.87	0.86	0.067	0.065	0.79	-	-	-	0.115	0.026
		Clinical	168	0.98	0.96	0.91	0.037	0.040	0.76	-	-	-		
		18-64 years old	168	0.97	0.94	0.90	0.045	0.051	0.76	-	-	-	0.142	<0.001
		64+ years old	138	0.96	0.99	0.90	0.021	0.049	0.81	-	-	-		

4.2.2.3. Measurement Invariance

Configural invariance for ADD-GOHAI and SC-GOHA could be concluded because the single-factor structure best fits all studied subgroups.

For ADD-GOHAI, the overall LM test showed significant difference between the factor loadings of the clinical and general populations, with the largest difference between items 11 (feeling uncomfortable with eating in front of others) and 12 (sensitivity of teeth or gums to hot, cold or sweets). However, there was no difference in the factor loadings by age group. The intercepts differed between the clinical and general populations and the subgroups by age. For SC-GOHAI, the difference between factor loadings was not significant, but item 12 differed between the clinical and general populations, and item 11 between age groups. The difference between intercepts was significant for both subgroups.

Altogether, strong measurement invariance could not be demonstrated for ADD-GOHAI or SC-GOHAI, suggesting that mean score differences across groups may be due to measurement bias and do not necessarily reflect true differences in OHRQoL. However, SC-GOHAI showed metric invariance between clinical and general populations as well as age groups, indicating that changes or differences in SC-GOHAI scores reflect similar differences in the latent construct (OHRQoL) across these subgroups. For ADD-GOHAI, metric invariance was shown only across age groups but not between general and clinical populations.

The potentially different measurement properties of ADD-GOHAI should be considered when comparing or synthesizing results across subpopulations. For instance, similar effect sizes may reflect different changes in OHRQoL in different populations.

Additionally, associations between ADD-GOHAI scores and other variables may vary between populations due to these measurement differences rather than true differences in the underlying constructs. Since SC-GOHAI demonstrated metric invariance across clinical and general populations and age groups, it is a suitable tool for tracking changes within these groups and for assessing the relationship of OHRQoL with other variables.

4.2.3. Remaining Measurement Properties

4.2.3.1. Reliability and Measurement Error

The test-retest reliability of both ADD-GOHAI and SC-GOHAI was adequate in the total sample as well as all subgroups, with ICC values ranging between 0.87-0.96 (Table 6). The SDC is approximately 5 points using ADD-GOHAI and 2-3 points using SC-GOHAI. Compared to the measurement range, ADD-GOHAI scores are able to detect more nuanced changes in individual patients' OHRQoL than SC-GOHAI. However, when using ADD-GOHAI for individual follow-up, users should be aware that, despite a score range of 48 points, a minimum change of 5 points is required to indicate a true change in an individual's OHRQoL, while smaller differences may be indistinguishable from measurement error.

Table 6. Reliability and measurement error in subsamples with repeat measurements (35)

Score	Sample	N	ICC	SEM	±95%CI of true score	SDC
ADD-GOHAI	Total	108	0.95	1.8	±3.6	5.1
	General	33	0.89	1.9	±3.7	5.2
	Clinical	75	0.96	1.8	±3.5	5.0
	18-64 years old	59	0.95	1.8	±3.6	5.1
	64+ years old	49	0.94	1.8	±3.5	5.0
SC-GOHAI	Total	108	0.91	0.8	±1.6	2.2
	General	33	0.87	0.6	±1.2	1.7
	Clinical	75	0.91	0.9	±1.7	2.4
	18-64 years old	59	0.90	0.9	±1.8	2.6
	64+ years old	49	0.93	0.6	±1.2	1.7

4.2.3.2. Construct Validity

Four out of the six hypotheses we tested for convergent validity supported the predefined criteria. As expected, the related constructs (OHIP-14, OH-SQ) showed a strong effect size correlation with ADD-GOHAI scores, while weakly related constructs such as general health showed small correlations (EQ-5D-VAS) and medium correlations (EQ-5D-5L score). The DMFT index and the number of teeth had negligible connection with GOHAI scores. This result is consistent with international data. This may be because correctly replaced missing teeth and well-filled teeth are perceived similarly to original teeth, and OHRQoL has a much more complex underlying construct than dental status.

All ten hypotheses we used to assess known-groups validity confirmed the predefined criteria. In five cases, we found a weak to moderate effect size (d=0.2-0.5) with symptoms such as mobile teeth, mucosal lesions, bleeding gums, and specific populations. Halitosis, chewing problems, pain, DTN, and aesthetic dissatisfaction had a strong effect size (d>0.5), suggesting that these observable signs or symptoms have the strongest association with QHRQoL.

For detailed information about the results of hypotheses testing for construct validity see Table 7.

Table 7. Summary of results of hypotheses testing for convergent and known-groups validity (35)

[a]*For known-groups validity: independent t-test with one-sided p-value*

	Type of effect size	Variable	Effect size	P value[a]	Expected magnitude / sign	Hypothesis confirmed
Convergent validity	Spearman correlation	OHIP-14	0.84	<0.001	>0.50 / +	Yes
		OH-SQ	-0.52	<0 .001	>0.50 / -	Yes
		EQ-5D-5L index	0.31	<0.001	~ 0.30-0.50 / +	Yes
		EQ VAS	0.28	<0.001	~ 0.30-0.50 / +	Yes
		DMFT	-0.08	0.143	~ 0.10-0.30 / -	No
		Number of teeth	0.08	0.169	~ 0.10-0.30 / +	No
Known-groups validity	Standardized mean difference	Mobile teeth (yes / no)	-0.30	0.066	> 0.10 / -	Yes
		Mucosal lesion (yes / no)	-0.49	0.001	> 0.10 / -	Yes
		Gum bleeding (yes / no)	-0.22	0.048	> 0.10 / -	Yes
		Chewing problem (yes / no)	-1.35	<0.001	> 0.10 / -	Yes
		Pain (yes / no)	-0.72	<0.001	> 0.10 / -	Yes
		Xerostomia (yes / no)	-0.45	<0.001	> 0.10 / -	Yes
		Halitosis (yes / no)	-0.68	<0.001	> 0.10 / -	Yes
		Aesthetic satisfaction (yes / no)	1.02	<0.001	> 0.10 / +	Yes
		Dental treatment need (yes / no)	-0.94	<0.001	> 0.10 / -	Yes
		Populations (clinical / general)	-0.37	<0.001	> 0.10 / -	Yes

5. DISCUSSION

Our COSMIN-guided systematic review underscores that while GOHAI is a reliable and established tool for assessing OHRQoL, it has notable limitations that warrant further research to reinforce its psychometric foundations. GOHAI's psychometric properties appear to be stable across different age groups; however, the impact of oral health conditions on OHRQoL varies significantly with age. While internal consistency for GOHAI as a single-factor tool is well-supported, its structural validity remains uncertain due to inconsistent factor analysis results. Additionally, GOHAI's reliability is well-documented for older populations, yet it has been rarely explored in younger groups. Despite strong correlations with other OHRQoL measures, GOHAI shows only weak associations with clinical oral health indicators, highlighting a broader challenge in linking subjective quality of life measures with objective health metrics. This underscores the need for OHRQoL questionnaires, as the same clinically observable issue can have vastly different impacts on patients, reflecting their unique perceptions and experiences. Responsiveness, a key factor for tracking changes over time, is especially underexplored across age groups, limiting GOHAI's effectiveness for longitudinal monitoring. In our second study, we developed and validated the Hungarian version of GOHAI, applying COSMIN standards. The Hungarian version showed satisfactory psychometric properties across both clinical and general populations, as well as across different age groups, though it shares GOHAI's limitations—particularly in suitability for individual follow-up. For the first time in the literature, we assessed GOHAI's SEM, SDC, and measurement invariance across general and clinical populations and between age groups, providing new insights into its applicability across diverse demographic and clinical settings.

5.1. Internal Structure (Structural Validity, Internal Consistency)

The structural validity of GOHAI raises questions about whether it functions as a purely reflective model or has formative characteristics. In a formative framework, each item uniquely contributes to the construct of OHRQoL, capturing distinct facets like pain, functional limitation, or psychological impact. In this model, items are not necessarily correlated, meaning that low scores in one domain (e.g., pain) do not imply low scores in another (e.g., psychosocial impact), thereby enabling GOHAI to capture a

multidimensional OHRQoL profile that collectively assesses oral health impact. The EQ-5D-5L is a widely used formative measure of HRQoL, featuring a descriptive system that captures health status across five dimensions. Instruments like the EQ-5D and similar tools often develop preference-based scores, which reflect the relative importance of specific health problems in the context of overall quality of life (51). In contrast, without preference weights, the scoring system of the GOHAI assumes that every symptom, regardless of its severity or frequency, holds equal importance. For instance, a person experiencing aesthetic issues either constantly or occasionally would receive the same GOHAI score difference as someone experiencing swallowing difficulties with the same frequency. This represents a strong assumption, as it implies that the impact of different health issues is considered equal, regardless of their nature.

Conversely, a reflective model assumes that all items are manifestations of a single underlying construct. For GOHAI to operate as a reflective tool, each item would need to equally represent aspects of OHRQoL, implying high internal consistency, with differences arising only from measurement error.. Our findings support a single-factor model in line with previous studies, indicated by a Cronbach's α of 0.83 for the Hungarian ADD-GOHAI (31), consistent with meta-analytic findings ($\alpha = 0.81$). This strong internal consistency suggests an interrelatedness among GOHAI items, though the variability seen in international factor analyses implies that GOHAI may also tap into distinct aspects of oral health that contribute independently to OHRQoL.

In our Hungarian sample, a strong loading was observed on functional items like eating difficulties, with secondary factors related to psychosocial concerns. Pain or orofacial appearance did not emerge as salient dimensions, suggesting that while GOHAI effectively captures physical and social impacts, it may not fully address other emerging OHRQoL dimensions, such as aesthetic concerns (90). These findings suggest that GOHAI could be interpreted as a mixed reflective-formative measure, potentially explaining observed inconsistencies in its internal structure and the challenges in establishing measurement invariance across populations. This dual nature should be carefully considered when interpreting GOHAI scores across diverse groups.

5.2. Measurement Invariance

In our study, we were the first to assess GOHAI's measurement invariance properties across subgroups. We observed configural invariance for both ADD-GOHAI and SC-GOHAI, suggesting that the factor structure is comparable across groups. However, strong measurement invariance was not achieved, and ADD-GOHAI did not meet metric invariance requirements between general and clinical populations. This suggests that OHRQoL assessments may vary due to subjective and demographic factors such as cultural background, age, and clinical status. For example, aesthetic concerns may influence younger individuals' OHRQoL more than older adults, who may accept such issues as part of aging. Additionally, clinical and general populations may experience and interpret oral health impacts differently, as reflected in our findings. SC-GOHAI demonstrated more stability across subgroups, yet we must interpret these findings with caution, considering the potential heterogeneity within groups. For instance, variations within age groups or among clinical and general samples may influence measurement invariance testing outcomes. Thus, while our results provide preliminary insights into the measurement invariance properties of GOHAI, they underscore the need for cautious interpretation and further research into subgroup-specific measurement invariance.

5.3. Reliability and Measurement Error

Our meta-analysis of the literature found that the test-retest reliability of ADD-GOHAI was robust, with an overall estimate of r=0.84 and values ranging from 0.79 to 0.87 across four age groups without significant differences between them. While this evidence is moderate for older age groups, it remains limited for younger populations, indicating a need for further validation studies in these groups. In our assessment of the Hungarian GOHAI, test-retest reliability was similarly satisfactory for both scoring methods, ADD-GOHAI and SC-GOHAI, across the entire sample and in all subgroups, with ICC values from 0.87 to 0.96. These ICC values indicate that the Hungarian GOHAI demonstrates high reliability across all scoring methods and subgroups, suggesting consistent measurement of OHRQoL across diverse groups. However, understanding measurement error is crucial for interpreting these scores accurately. Measurement error includes both random and systematic errors in a patient's score that do not reflect true changes in OHRQoL. To quantify this, we calculated the SEM and SDC for the Hungarian GOHAI,

marking the first time these values have been assessed for this instrument. The SDC is particularly useful in clinical settings: for the ADD-GOHAI, a score change of about 5 points, and for the SC-GOHAI, a change of 2-3 points, can be considered beyond the threshold of measurement error and indicative of true change. However, whether such changes are meaningful to patients (i.e., meet the minimal important difference, MID) remains an interpretative issue. The responsiveness of GOHAI scores to clinically relevant changes requires further research, but our findings on SDC provide an important benchmark for evaluating individual patient progress.

5.4. Construct Validity

Our systematic review found that the GOHAI effectively measures OHRQoL and aligns with Locker's comprehensive model for this construct (10). Although Locker initially questioned the association between GOHAI and GHRQoL (12), the literature indicates a clear link, with 23 out of 24 studies demonstrating at least a weak correlation (31). This supports the idea that OHRQoL contributes to GHRQoL, a finding further supported by the results of the Hungarian validation process (35). Due to the subjective nature of dental health perception, however, our systematic review also observed that approximately 28% of hypotheses related to dental status showed no correlation with ADD-GOHAI scores (31). This limited alignment with objective indicators underscores the value of PROMs like GOHAI in capturing patient experiences that may not be fully captured by objective measures.

5.5. Responsiveness

Responsiveness is a critical attribute of a PROM, reflecting its ability to detect meaningful changes over time in the construct being measured. Like construct validity, assessing responsiveness involves hypothesis testing; however, while validity pertains to the accuracy of a single score, responsiveness addresses the validity of a change score (36). Although GOHAI has been widely studied as a dental PROM, evidence regarding its responsiveness remains limited. Responsiveness is inherently challenging to evaluate for most PROMs, as it requires well-defined interventions and stable populations over time. Our systematic review highlighted this gap, revealing limited data on GOHAI's responsiveness and a lack of information for younger subgroups (<60 and ≤45 years) (31).

The Hungarian validation of GOHAI did not assess responsiveness either, underscoring the need for further research to establish the tool's sensitivity to change in diverse patient populations.

5.6. Scoring Variants and Reference Period of GOHAI

The standard timeframe for GOHAI items is the past three months, using response options across 3-, 5-, or 6-point Likert-type frequency scales. Locker, for comparability, extended the reference period to one year in his study of both GOHAI and OHIP-14 (30). Interestingly, research by Sutinen et al. with OHIP-14 shows that extending the reference period does not necessarily lead to a decline in scores, as might be expected, highlighting the need to further investigate how reference periods affect score interpretation (91). Additionally, the number of response options warrants further study. Some researchers caution that a high number of response choices (5 or 6) may pose difficulties for patients with lower educational levels, potentially reducing the reliability of their responses (81, 89). However, scales with more response categories generally show improved reliability and validity in psychometric literature (82). This suggests a nuanced balance between accessibility and psychometric robustness that should be explored in future studies.

5.7. Floor and Ceiling Effect

Hungarian ADD-GOHAI showed a notable ceiling effect, while SC-GOHAI exhibited a floor effect. This finding contrasts with prior studies, such as Hassel's (83) and Locker's (30), where the rates of highest scores were significantly lower. Our higher rate of ceiling effects (18.8%) may reflect the participants' favorable socio-demographic backgrounds, possibly indicating better-than-average OHRQoL. This limited range for higher scores suggests GOHAI is less effective for distinguishing among individuals with high OHRQoL, while it remains sensitive for lower scores. Similar limitations in the score range have been seen in other GHRQoL measures, like EQ-5D (92). Minimizing floor and ceiling effects is essential for improving responsiveness, an aspect that remains underexplored for GOHAI, particularly in populations with above-average OHRQoL, where the tool's suitability may be in question based on these results.

6. CONCLUSIONS

6.1. Answers to Research Questions and Key Findings

1. What insights does a systematic COSMIN review provide about the measurement properties of GOHAI as well as the quality and strength of the supporting evidence?

Internal structure:
GOHAI both follows the theoretical dimensions of OHRQoL and encompasses a broad range of oral health issues. This dual nature has led to divergent findings in previous studies regarding its factor structure, with some authors questioning whether it can truly be considered an OHRQoL questionnaire. The COSMIN review confirms that, despite these concerns, GOHAI can be regarded as a valid single-factor OHRQoL instrument, supported by high-quality evidence of internal consistency.
Construct validity:
The GOHAI demonstrates a strong correlation with OHRQoL instruments, a moderate correlation with GHRQoL instruments, and a weak correlation with objective oral health symptoms, which also varies by age. Based on this, GOHAI can be considered a valid OHRQoL measure across all age groups, but it highlights that patients' subjective experiences may differ significantly from what is reflected by the objective clinical picture.
Responsiveness:
An important feature in the clinical follow-up of an individual is the ability of the questionnaire to measure changes over time. However, the COSMIN review revealed that responsiveness remains a scarcely investigated area for GOHAI, with the few studies conducted being methodologically inadequate. This gap in research is evident across all age groups.

2. Are the psychometric properties of the GOHAI, including structural validity, construct validity, and reliability, uniformly established across different age groups to support its use as a general OHRQoL instrument?

According to our COSMIN review, the measurement properties of GOHAI have been examined across all age groups, and its structural validity, construct validity, and reliability are adequate in all age groups, although the quality of evidence varies. GOHAI can be used as an OHRQoL measure in all adult age groups, though the relationship between specific physical symptoms and quality of life may vary with age.

3. Is the construct validity of the Hungarian GOHAI supported by at least 75% of the predefined hypotheses, as suggested by the COSMIN guidelines?

Construct validity of the Hungarian GOHAI is supported by at least 75% of the predefined hypotheses, in line with the COSMIN guidelines. Specifically, four out of six hypotheses tested for convergent validity met the predefined criteria, and all ten hypotheses assessed for known-groups validity were confirmed. These results indicate that the GOHAI-HU effectively reflects the underlying concept of OHRQoL.

4. Does a single-factor structure of Hungarian GOHAI with secondary dimensions of physical function, psycho-social function, and pain and discomfort demonstrate adequate fit in confirmatory factor analysis (CFA)?

Yes, the single-factor structure of the Hungarian GOHAI demonstrates an adequate fit in the CFA (Add-GOHAI CFI = 0.98, SC-GOHAI CFI = 0.99), with additional item correlations explained by eating-related factors.

5. When tested for measurement invariance, does the Hungarian GOHAI exhibit at least metric invariance between the general and clinical populations, as well as different age groups?

Configural invariance was achieved for both ADD-GOHAI and SC-GOHAI, indicating that the basic factor structure is comparable across groups. Metric invariance was demonstrated for SC-GOHAI across both clinical and general populations, as well as across different age groups. However, for ADD-GOHAI, metric invariance was observed only across age groups and not between the general and clinical populations. As

a result, strong measurement invariance could not be established for either version, which limits the comparability of scores across certain populations.

6. What are the test-retest reliability, standard error of measurement (SEM), and the smallest detectable change (SDC) of the Hungarian GOHAI after repeated administrations?

The test-retest reliability of the Hungarian GOHAI was found to be adequate for both ADD-GOHAI and SC-GOHAI across the total sample and all subgroups. The standard error of measurement was approximately 2 points for ADD-GOHAI and 1 point for SC-GOHAI, consistent across clinical and general populations, as well as different age groups. The smallest detectable change was approximately 5 points for ADD-GOHAI and 2-3 points for SC-GOHAI.

6.2. Implementation for Practice and Research

1. Responsiveness of the GOHAI, a critical factor for tracking changes over time, remains underexplored. This limitation reduces its usefulness for longitudinal monitoring.

2. The GOHAI shows limited suitability for individual-level follow-up, primarily due to relatively high SDC thresholds. While ADD-GOHAI appears somewhat more appropriate for personal monitoring, its use is more reliable for assessing study populations rather than tracking individual changes over time. Any observed changes at the individual level must be interpreted cautiously, considering the measurement error associated with GOHAI. In individual follow-ups, a difference of 5 points for ADD-GOHAI and 2 points for SC-GOHAI can be reliably considered as not resulting from measurement error.

3. Since strong measurement invariance was not established, the same GOHAI scores across different populations do not necessarily reflect the same OHRQoL value, making comparisons between groups unfeasible. The lack of metric invariance between clinical and general populations highlights the potential for different groups to interpret the

impact of various oral health conditions on OHRQoL differently, emphasizing the need for caution when comparing these populations.

SC-GOHAI, however, demonstrated greater stability across subgroups, making it more suitable for comparisons between different populations. The metric invariance observed between age groups indicates that, while different items may carry varying weights in different age groups, overall, GOHAI evaluates OHRQoL similarly for both younger and older individuals. However, due to the absence of strong invariance, comparisons of mean values are only valid when populations are highly homogeneous.

4. The ceiling effect, indicating a limited range for higher scores, suggests that GOHAI is less effective at distinguishing between individuals with high OHRQoL.However, it remains sensitive and reliable for detecting lower OHRQoL scores.

7. SUMMARY

Objectives: Our research examines the GOHAI, a widely used tool for assessing OHRQoL. Using the COSMIN methodology, we systematically review its psychometric properties across age groups, validate the Hungarian version, and assess its SEM, SDC, measurement invariance across general and clinical populations, as well as age groups.

Methods: We conducted a systematic review of English peer-reviewed articles on the development, translation, or validation of the GOHAI, using data from PubMed, Web of Science, and EMBASE (1990–2023). Methodological evaluation followed the COSMIN guidelines, analyzing results across four age groups (≥60, all ages, <60, ≤45). Structural validity was summarized qualitatively, while internal consistency and reliability were assessed via random-effects meta-analysis of T-transformed Cronbach's α and Fisher's Z-transformed correlation coefficients. Construct validity and responsiveness were evaluated using effect sizes. GOHAI-HU was translated using a forward-backward process. 306 participants (45.1% general, 54.9% clinical) from two age groups were recruited in Budapest, with 108 receiving two administrations of GOHAI. Structural validity and measurement invariance were assessed using CFA; internal consistency was evaluated with Cronbach's α, and test-retest reliability was measured using ICC. Construct validity was tested against predefined hypotheses.

Results: Our systematic review included 60 studies from 497 records. Structural validity varied across age groups, while internal consistency was sufficient (α=0.81). Test-retest reliability was strong (r=0.84). Construct validity supported 361 hypotheses. Responsiveness was not assessed for younger groups, leading to low evidence quality. For GOHAI-HU, a single-factor model was confirmed for structural validity. The instrument showed strong internal consistency (α=0.76–0.85) and test-retest reliability (ICC: 0.87–0.96). Construct validity was robust, but measurement invariance could not be confirmed between subpopulations. SDC was 5 points for ADD-GOHAI.

Conclusion: Our systematic review confirms that GOHAI has sufficient psychometric properties across age groups, though its responsiveness remains underexplored. The Hungarian version demonstrates satisfactory performance in both general and clinical populations. While SC-GOHAI is more stable across populations, ADD-GOHAI is better suited for tracking individual changes, though observed changes should be interpreted with caution, accounting for the inherent SEM.

8. REFERENCES

1. Preamble to the Constitution of the World Health Organization, (22 June 1946, entered into force on 7 April 1948).
2. Skevington SM, Lotfy M, O'Connell KA, Group W. The World Health Organization's WHOQOL-BREF quality of life assessment: psychometric properties and results of the international field trial. A report from the WHOQOL group. Qual Life Res. 2004;13(2):299-310.
3. Logan HL. The patient and the shifting health-care paradigm. J Am Coll Dent. 1997;64(1):16-8.
4. Alauddin MS, Baharuddin AS, Mohd Ghazali MI. The Modern and Digital Transformation of Oral Health Care: A Mini Review. Healthcare (Basel). 2021;9(2).
5. Crossing the Quality Chasm: A New Health System for the 21st Century. Washington (DC)2001.
6. Scambler S, Delgado M, Asimakopoulou K. Defining patient-centred care in dentistry? A systematic review of the dental literature. Br Dent J. 2016;221(8):477-84.
7. Sischo L, Broder HL. Oral health-related quality of life: what, why, how, and future implications. J Dent Res. 2011;90(11):1264-70.
8. Kleinman A. The illness narratives: suffering, healing, and the human condition. . New York: Basic Books; 1989.
9. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life. A conceptual model of patient outcomes. JAMA. 1995;273(1):59-65.
10. Locker D. Measuring oral health: a conceptual framework. Community Dent Health. 1988;5(1):3-18.
11. Measuring Oral Health and Quality of Life1997; University of North Carolina-Chapel Hill: Department of Dental Ecology, School of Dentistry, University of North Carolina.
12. Locker D, Allen F. What do measures of 'oral health-related quality of life' measure? Community Dentistry and Oral Epidemiology. 2007;35(6):401-11.
13. John MT, Hujoel P, Miglioretti DL, LeResche L, Koepsell TD, Micheelis W. Dimensions of oral-health-related quality of life. J Dent Res. 2004;83(12):956-60.

14. Tapsoba H, Deschamps JP, Leclercq MH. Factor analytic study of two questionnaires measuring oral health-related quality of life among children and adults in New Zealand, Germany and Poland. Qual Life Res. 2000;9(5):559-69.
15. Su N, van Wijk A, Visscher CM. Psychosocial oral health-related quality of life impact: A systematic review. J Oral Rehabil. 2021;48(3):282-92.
16. Sischo L, Wilson-Genderson M, Broder HL. Quality-of-Life in Children with Orofacial Clefts and Caregiver Well-being. J Dent Res. 2017;96(13):1474-81.
17. Lee GH, McGrath C, Yiu CK, King NM. A comparison of a generic and oral health-specific measure in assessing the impact of early childhood caries on quality of life. Community Dent Oral Epidemiol. 2010;38(4):333-9.
18. Slade GD, Spencer AJ. Development and evaluation of the Oral Health Impact Profile. Community Dent Health. 1994;11(1):3-11.
19. Atchison KA, Dolan TA. Development of the Geriatric Oral Health Assessment Index. J Dent Educ. 1990;54(11):680-7.
20. Liu J, Wong MCM, Lo ECM. Development of Oral Health Impacts on Daily Living Questionnaire Items - a Qualitative Study. Chin J Dent Res. 2017;20(2):79-88.
21. Deana NF, Pardo Y, Ferrer M, Espinoza-Espinoza G, Garin O, Muñoz-Millán P, et al. Evaluating conceptual model measurement and psychometric properties of Oral health-related quality of life instruments available for older adults: a systematic review. Health and Quality of Life Outcomes. 2024;22(1).
22. Slade GD. Derivation and validation of a short-form oral health impact profile. Community Dent Oral Epidemiol. 1997;25(4):284-90.
23. S. Adulyanon AS, editor Oral Impacts on Daily Performances. Measuring Oral Health and Quality of Life; 1996; University of North Carolina-Chapel Hill, North Carolina: epartment of Dental Ecology, School of Dentistry, University of North Carolina.
24. Riva F, Seoane M, Reichenheim ME, Tsakos G, Celeste RK. Adult oral health-related quality of life instruments: A systematic review. Community Dent Oral Epidemiol. 2022;50(5):333-8.
25. Slade GD, editor The Oral Health Impact Profile. Measuring Oral Health and Quality of Life; 1996; University of North Carolina-Chapel Hill, North Carolina: Department of Dental Ecology, School of Dentistry, University of North Carolina.

26. Ohrn K, Jönsson B. A comparison of two questionnaires measuring oral health-related quality of life before and after dental hygiene treatment in patients with periodontal disease. Int J Dent Hyg. 2012;10(1):9-14.
27. Gokturk O, Yarkac FU. Comparison of two measures to determine the oral health-related quality of life in elders with periodontal disease. Community Dent Health. 2019;36(2):143-9.
28. El Osta N, Tubert-Jeannin S, Hennequin M, Bou Abboud Naaman N, El Osta L, Geahchan N. Comparison of the OHIP-14 and GOHAI as measures of oral health among elderly in Lebanon. Health and Quality of Life Outcomes. 2012;10.
29. Ikebe K, Hazeyama T, Enoki K, Murai S, Okada T, Kagawa R, et al. Comparison of GOHAI and OHIP-14 measures in relation to objective values of oral function in elderly Japanese. Community Dent Oral Epidemiol. 2012;40(5):406-14.
30. Locker D, Matear D, Stephens M, Lawrence H, Payne B. Comparison of the GOHAI and OHIP-14 as measures of the oral health-related quality of life of the elderly. Community Dent Oral Epidemiol. 2001;29(5):373-81.
31. Oszlanszky J, Gulacsi L, Pentek M, Hermann P, Zrubka Z. Psychometric Properties of General Oral Health Assessment Index Across Ages: COSMIN Systematic Review. Value Health. 2024.
32. Locker D, Allen F. What do measures of 'oral health-related quality of life' measure? Community Dent Oral Epidemiol. 2007;35(6):401-11.
33. Szentpetery A, Szabo G, Marada G, Szanto I, John MT. The Hungarian version of the Oral Health Impact Profile. Eur J Oral Sci. 2006;114(3):197-203.
34. Cseh Károly SG, Marada Gyula, Szentpétery András. Szájegészség által meghatározott életminőség: Két rövid magyar OHIP változat kifejlesztése és értékelése. Mentalhigiéné és Pszichoszomatika. 2008;9 (2008) 1,:81-96.
35. Oszlanszky J, Mensch K, Hermann P, Zrubka Z. Validation of the Hungarian version of the General Oral Health Assessment Index (GOHAI) in clinical and general populations. BMC Oral Health. 2024;24(1):1402.
36. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63(7):737-45.

37. Mokkink LB, de Vet HCW, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. Qual Life Res. 2018;27(5):1171-9.

38. Prinsen CAC, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HCW, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018;27(5):1147-57.

39. 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):1159-70.

40. Vogt L, Pretzl B, Eickholz P, Ramich T, Nickles K, Petsos H. Oral health-related quality of life and patient-reported outcome measures after 10 years of supportive periodontal care. Clin Oral Investig. 2023;27(6):2851-64.

41. Borges GA, Barbin T, Dini C, Maia LC, Magno MB, Barao VAR, et al. Patient-reported outcome measures and clinical assessment of implant-supported overdentures and fixed prostheses in mandibular edentulous patients: A systematic review and meta-analysis. J Prosthet Dent. 2022;127(4):565-77.

42. Terwee CB, Jansma EP, Riphagen, II, de Vet HC. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual Life Res. 2009;18(8):1115-23.

43. Judit Oszlánszky MP, László Gulácsi, Zsombor Zrubka, Péter Hermann. Psychometric properties of Geriatric/General Oral Health Assessment Index in different age groups: COSMIN Systematic Review

[Available from: https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=384132.

44. Atchison KA. Chapter 7: The General Oral Health Assessment Index. In: Slade GD, editor. Measuring Oral Health and Quality of Life; Chapel Hill: University of North Carolina: Dental Ecology; 1997. p. 71-80.

45. Campos J, Zucoloto ML, Bonafé FSS, Maroco J. General Oral Health Assessment Index: A new evaluation proposal. Gerodontology. 2017;34(3):334-42.

46. H. C. W. de Vet CBT, L. B. Mokkink and D. L. Knol. Measurement in Medicine, A Practical Guide: Cambridge University Press, New York; 2011.

47. Gerke O. Reporting Standards for a Bland-Altman Agreement Analysis: A Review of Methodological Reviews. Diagnostics (Basel). 2020;10(5).
48. COOPER H, LARRY V. HEDGES, and JEFFREY C. VALENTINE. Handbook of Research Synthesis and Meta-Analysis, The.: Russell Sage Foundation; 2009.
49. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests: Danmarks Paedagogiske Institut; 1960.
50. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine (Phila Pa 1976). 2000;25(24):3186-91.
51. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res. 2011;20(10):1727-36.
52. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res. 1995;4(4):293-307.
53. Kressin NR, Atchison KA, Miller DR. Comparing the impact of oral disease in two populations of older adults: application of the geriatric oral health assessment index. J Public Health Dent. 1997;57(4):224-32.
54. David W. Gerbing JCA. On the Meaning of Within-Factor Correlated Measurement Errors. Journal of Consumer Research. June 1984;Volume 11(Issue 1):Pages 572–80.
55. J. Micah Roos SB. Confirmatory Factor Analysis (Quantitative Applications in the Social Sciences): SAGE Publications, Inc; 2021.
56. Satorra A, Bentler PM. Corrections to test statistics and standard errors in covariance structure analysis. In: Clogg AvECC, editor. Latent variables analysis: Applications for developmental research. Thousand Oaks, CA: Sage Publications Inc.; 1994. p. 399-419.
57. Sörbom D. Model modification. Psychometrika. 1989;54(3):371-84.
58. Floyd FJ, Widaman KF. Factor analysis in the development and refinement of clinical assessment instruments. Psychol Assessment. 1995;7(3):286-99.
59. Costello ABJO. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. Practical Assessment Research & Evaluation. 2005;10(7).

60. Putnick DL, Bornstein MH. Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research. Dev Rev. 2016;41:71-90.
61. Mokkink LB, Boers M, van der Vleuten CPM, Bouter LM, Alonso J, Patrick DL, et al. COSMIN Risk of Bias tool to assess the quality of studies on reliability or measurement error of outcome measurement instruments: a Delphi study. BMC Med Res Methodol. 2020;20(1):293.
62. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol. 2006;59(10):1033-9.
63. Oszlanszky J, Gulacsi L, Pentek M, Hermann P, Zrubka Z. Psychometric Properties of General Oral Health Assessment Index Across Ages: COSMIN Systematic Review. Value Health. 2024;27(6):805-14.
64. Ellis PD. The essential guide to effect sizes : statistical power, meta-analysis, and the interpretation of research results United Kingdom: Cambridge University Press; 2010.
65. Mokkink LB, Prinsen CA, Patrick DL, Alonso J, Bouter LM, de Vet HC, et al. COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs). Amsterdam: Vrije University Medical Center, Amsterdam; 2017.
66. Campos JA, Zucoloto ML, Geremias RF, Nogueira SS, Maroco J. Validation of the Geriatric Oral Health Assessment Index in complete denture wearers. J Oral Rehabil. 2015;42(7):512-20.
67. Campos JA, Carrascosa AC, Zucoloto ML, Maroco J. Validation of a measuring instrument for the perception of oral health in women. Braz Oral Res. 2014;28.
68. Vettore MV, Rebelo MAB, Rebelo Vieira JM, Cardoso EM, Birman D, Leão ATT. Psychometric Properties of the Brazilian Version of GOHAI among Community-Dwelling Elderly People. Int J Environ Res Public Health. 2022;19(22).
69. Othman WN, Muttalib KA, Bakri R, Doss JG, Jaafar N, Salleh NC, et al. Validation of the Geriatric Oral Health Assessment Index (GOHAI) in the Malay language. J Public Health Dent. 2006;66(3):199-204.
70. Naito M, Suzukamo Y, Nakayama T, Hamajima N, Fukuhara S. Linguistic adaptation and validation of the General Oral Health Assessment Index (GOHAI) in an elderly Japanese population. J Public Health Dent. 2006;66(4):273-5.

71. Hassel AJ, Rolko C, Koke U, Leisen J, Rammelsberg P. A German version of the GOHAI. Community Dent Oral Epidemiol. 2008;36(1):34-42.
72. Atieh MA. Arabic version of the Geriatric Oral Health Assessment Index. Gerodontology. 2008;25(1):34-41.
73. Gutiérrez Quiceno B, Calzada Gutiérrez MT, Fandiño-Losada A. Cultural adaptation and validation of the Geriatric Oral Health Assessment Index - GOHAI - Colombian version. Colomb Med (Cali). 2019;50(2):102-14.
74. Sánchez-García S, Heredia-Ponce E, Juárez-Cedillo T, Gallegos-Carrillo K, Espinel-Bermúdez C, de la Fuente-Hernández J, et al. Psychometric properties of the General Oral Health Assessment Index (GOHAI) and dental status of an elderly Mexican population. J Public Health Dent. 2010;70(4):300-7.
75. Atchison KA, Der-Martirosian C, Gift HC. Components of self-reported oral health and general health in racial and ethnic groups. J Public Health Dent. 1998;58(4):301-8.
76. Franchignoni M, Giordano A, Levrini L, Ferriero G, Franchignoni F. Rasch analysis of the Geriatric Oral Health Assessment Index. Eur J Oral Sci. 2010;118(3):278-83.
77. Denis F, Bizien P, Tubert-Jeannin S, Hamad M, Trojak B, Rude N, et al. A Rasch Analysis between Schizophrenic Patients and the General Population. Transl Neurosci. 2017;8:139-46.
78. Denis F, Hamad M, Trojak B, Tubert-Jeannin S, Rat C, Pelletier JF, et al. Psychometric characteristics of the "General Oral Health Assessment Index (GOHAI) » in a French representative sample of patients with schizophrenia. BMC Oral Health. 2017;17(1):75.
79. Tubert-Jeannin S, Riordan PJ, Morel-Papernot A, Porcheray S, Saby-Collet S. Validation of an oral health quality of life index (GOHAI) in France. Community Dent Oral Epidemiol. 2003;31(4):275-84.
80. Shyama M, Honkala S, Al-Mutawa SA, Honkala E. Oral health-related quality of life among parents and teachers of disabled schoolchildren in Kuwait. Med Princ Pract. 2013;22(3):285-90.
81. Mathur VP, Jain V, Pillai RS, Kalra S. Translation and validation of Hindi version of Geriatric Oral Health Assessment Index. Gerodontology. 2016;33(1):89-96.

82. Luis M. Lozano EGa-C, and José Muñiz. Effect of the Number of Response Categories on the Reliability and Validity of Rating Scales. Methodology. 2008;4(2):73–79.

83. Hassel AJ, Steuker B, Rolko C, Keller L, Rammelsberg P, Nitschke I. Oral health-related quality of life of elderly Germans--comparison of GOHAI and OHIP-14. Community Dent Health. 2010;27(4):242-7.

84. Niesten D, Witter D, Bronkhorst E, Creugers N. Validation of a Dutch version of the Geriatric Oral Health Assessment Index (GOHAI-NL) in care-dependent and care-independent older people. BMC Geriatr. 2016;16:53.

85. Rodakowska E, Mierzyńska K, Bagińska J, Jamiołkowski J. Quality of life measured by OHIP-14 and GOHAI in elderly people from Bialystok, north-east Poland. BMC Oral Health. 2014;14:106.

86. W AD, Jun-Qi L. Factors associated with the oral health-related quality of life in elderly persons in dental clinic: validation of a Mandarin Chinese version of GOHAI. Gerodontology. 2011;28(3):184-91.

87. Daradkeh S, Khader YS. Translation and validation of the Arabic version of the Geriatric Oral Health Assessment Index (GOHAI). J Oral Sci. 2008;50(4):453-9.

88. Santucci D, Camilleri L, Kobayashi Y, Attard N. Development of a Maltese version of oral health-associated questionnaires: OHIP-14, GOHAI, and the Denture Satisfaction Questionnaire. Int J Prosthodont. 2014;27(1):44-9.

89. Hägglin C, Berggren U, Lundgren J. A Swedish version of the GOHAI index. Psychometric properties and validation. Swed Dent J. 2005;29(3):113-24.

90. John MT. Foundations of oral health-related quality of life. J Oral Rehabil. 2020.

91. Sutinen S, Lahti S, Nuttall NM, Sanders AE, Steele JG, Allen PF, et al. Effect of a 1-month vs. a 12-month reference period on responses to the 14-item Oral Health Impact Profile. Eur J Oral Sci. 2007;115(3):246-9.

92. Parkin D, Devlin N, Feng Y. What Determines the Shape of an EQ-5D Index Distribution? Med Decis Making. 2016;36(8):941-51.

9. BIBLIOGRAPHY OF PUBLICATIONS

Publications related to the thesis:

Oszlánszky J, Gulácsi L, Péntek M, Hermann P, Zrubka Z. Psychometric Properties of General Oral Health Assessment Index Across Ages: COSMIN Systematic Review. **Value in Health**. 2024 Jun;27(6):805-814. doi: 10.1016/j.jval.2024.02.022. Epub 2024 Mar 14. PMID: 38492926.

D1, IF: 4.9

Oszlánszky J, Mensch K, Hermann P, Zrubka Z. Validation of the Hungarian version of the General Oral Health Assessment Index (GOHAI) in clinical and general populations. **BMC Oral Health**. 2024 Nov 19;24(1):1402. doi: 10.1186/s12903-024-05198-2. PMID: 39563321; PMCID: PMC11575072.

Q1, IF: 2.6

∑IF: 7.5

Publications not related to the thesis:

Oszlánszky J, Kádár L, Hermann P, Schmidt P, Gyulai Gaál Szabolcs; Implantációs fogpótlások protetikai aspektusai kivehető teljes alsó fogpótlások esetén [Removable dentures with implants for edentulous lower jaw]; FOGORVOSI SZEMLE 106 : 3 pp. 91-95. , 5 p. (2013)

Oszlánszky J, Mikulás K, Déri T; Implantátumon elhorgonyzott kivehető fogpótlások In: Hermann, P; Kispélyi, B (szerk.) Fogpótlástan 1-2; Budapest, Magyarország: Semmelweis Kiadó és Multimédia Stúdió (2022) 1,248 p. pp. 904-919. , 16 p.

Oszlánszky Judit, Gyulai-Gaál Szabolcs, Kádár László, Schmidt Péter, Hermann Péter Tízéves követéses vizsgálat fogászati kezeléstől kórosan félő beteg összetett ellátása során: FOGORVOSI SZEMLE 116 : 3 pp. 127-135. , 9 p. (2023)

Márta, Péntek ; Áron, Hölgyesi ; Zsombor, Zrubka ; Petra, Baji ; **Judit, Oszlánszky** ; Péter, Hermann ; Levente, Kovács ; László, Gulácsi., "Subjective Expectations on Living with Innovative Digital Implantable Medical Devices at Older Ages," 2024 IEEE 28th International Conference on Intelligent Engineering Systems (INES), Gammarth, Tunisia, 2024, pp. 000043-000048, doi: 10.1109/INES63318.2024.10629145.

10. ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to my supervisors for their invaluable support. I am deeply thankful to Professor Hermann, who provided unwavering encouragement and always strived to address any emerging needs as quickly as possible. I am especially indebted to Dr. Zsombor Zrubka, whose profound influence over the past years has shaped not only my professional development but also my research, medical, and ethical perspectives. The countless hours spent working together, solving challenges, and achieving shared goals have been a defining part of my recent years. I sincerely hope that this dissertation is just the beginning of our collaboration and not its conclusion.

I am profoundly grateful to my family, as this work demanded a great deal of time away from them. Without the support of my husband and my mother, this accomplishment would not have been possible. My husband, Balázs Sziklai, deserves special recognition, not only for keeping our family grounded but also for helping me find my professional path.

I would also like to thank everyone who contributed to the success of my research, even if I cannot mention them all by name, as they are too numerous to list. My gratitude extends to the patients who participated in the studies, the assistants, our librarians, and my colleagues—each of you played a vital role in making this work possible.

Finally, I would also like to thank my loyal four-legged companion, Lancelot, who faithfully stayed by my side no matter how long I sat in front of the computer, ensuring I never worked alone.

Thank you all for your support and encouragement.