Consumer Reported Outcome Measures Consortium Task Force

Technical Report

Best Practices and Guidelines with respect to Psychometric CROM for use in Research on Tobacco and Nicotine Containing Products

March 2024

Study Coordinator and Authors:

Stacey McCaffrey*, Juul Labs, Inc., USA

Esther Afolalu* & Christelle Chrea*, Philip Morris Products S.A., Switzerland

Thomas Salzberger*, WU Wien Univ. of Economics and Business, Inst. for Statistics and Mathematics, Austria

Saul Shiffman*, Pinney Associates, USA

Lesley Giles & Daisuke Nishihara, Japan Tobacco International, Switzerland

Mohamadi Sarkar, Altria Client Services, USA

Krishna Prasad & Mandara Shetty, British American Tobacco, Southampton, UK

Ryan Black, Juul Labs, Inc., USA

**Lead Author*

Table of Contents

LIST OF ABBREVIATIONS

Abbreviation	Definition
AERA	American Educational Research Association
APA	American Psychological Association
CAT	Computerized adaptive testing
CDER	Center for Drug Evaluation and Research
CFA	Confirmatory factor analysis
CONSORT	Consolidated Standards of Reporting Trials
COPD	Chronic obstructive pulmonary disease
CORESTA	Cooperation Centre for Scientific Research Relative to Tobacco
COSMIN	COnsensus-based Standards for the selection of health Measurement INstruments
CROM	Consumer Reported Outcome Measure(s)
CTP	Center for Tobacco Products
ENDS	Electronic nicotine delivery systems
ERIQA	European Regulatory Issues on Quality of Life Assessment Group
FDA	United States Food and Drug Administration
ISOQOL	International Society for Quality of Life Research
ISPOR	The Professional Society for Health Economics and Outcomes Research
ITM	Item tracking matrix
mCEQ	Modified cigarette evaluation questionnaire
MRTP	Modified risk tobacco product
MRTPA	Modified Risk Tobacco Product Application
NCME	National Council on Measurement in Education
PRO	Patient-reported outcome
PROMIS	Patient-Reported Outcomes Measurement Information System
PMTA	Premarket Tobacco Product Application
SAP	Statistical analysis plan
SISAQOL	Setting International Standards in Analyzing Patient-Reported Outcomes and Quality of Life
SME	Subject matter expert
SPIRIT-PRO	Standard Protocol Items: Recommendations for Interventional Trials – Patient Reported Outcomes
SRNT	Society for Research on Nicotine and Tobacco
SSPT	CORESTA Smoke Science and Product Technology Conference
TF	Task Force
TNP	Tobacco and nicotine-containing product
TPPIS	Tobacco product perception and intention study
UK	United Kingdom
US	United States

DEFINITION OF TERMS

Term	Definition
Ability (sensitivity) to detect change	A psychometric property of CROM which speaks to the CROM's ability to detect change in a construct when change has actually occurred.[1] That is, we would expect to see change in a respondent's CROM scores over time in a way that is consistent with "known" change in the construct that the CROM is measuring (or a related construct). (Conversely, when scores on a CROM show change when no real change has taken place, the CROM would suffer from poor test-retest reliability). For example, if medical tests indicate a substantial decline in respiratory functioning, a CROM that is sensitive to detecting change in respiratory symptoms administered during this time would be expected to capture an increase in respiratory symptoms. This psychometric property is also referred to as responsiveness.
Backward translation	This refers to translation of the new (target) language version of a CROM back into the original (source) language.
Computerized adaptive testing (CAT)	In CAT, not all respondents complete all the constituent items of the CROM. Items are administered to each respondent depending on responses to previous items until a pre-specified level of precision (or some other stop criterion) is reached. CAT applications require a continuously updated estimation of the latent variable measure based on models of modern test theory and, therefore, need to be administered on a computer. The primary objective of CAT is the use of the smallest possible number of items administered individually. CAT applications require a large pool of pre-calibrated items and make sense only if items vary in the amount of the property being measured (e.g., items representing a low level of perceived health risk versus a high level).
Claim	In the field of tobacco and nicotine product research, a claim is usually an assertion that a particular tobacco or nicotine-containing product implies a modified or reduced risk or reduced exposure to harmful chemicals, compared to combustible cigarettes (see [1]).

[1] With Psychometric CROM, it may not be possible to know whether actual, true change in a construct has occurred. For example, even when respiratory functioning changes, a respondent's perception of the severity of their symptoms may not change or may not change in a way that is perfectly linearly correlated. And some CROM may not have an objective metric on which change could be observed. A researcher should be prepared to justify the criterion that they chose to assess a CROM's ability to detect change, and should not expect to see a 1:1 relationship in change scores due to the nature of the constructs being measured by Psychometric CROM (attitudes and perceptions).

Term	Definition
Cognitive (debriefing) interviews	A research strategy for purposes of developing a CROM and/or evaluating the content validity of a CROM. During this research session or interview, the interviewer (sometimes referred to as a "moderator" in group-based interviews) asks the participant questions about the construct being measured and/or the CROM. For example, in the case of retrospective-based interviews, the interviewer asks the participant to complete the CROM first, and then asks several questions to assess their understanding of the CROM, experience with the CROM, content coverage, etc. There are several approaches to conducting cognitive debriefing interviews, which vary based on the researcher's objective. For example, the interviews can be conducted individually or in a group setting. The interviews may be purely qualitative, or may include quantitative components, such as a survey where respondents are asked to rate the clarity or relevance of items. While cognitive interviews are typically semi-structured (the interviewer has an interviewer guide with predetermined probes to ask during the interview, and they deviate from the guide as appropriate), they may also be fully structured or unstructured.
Concept of interest, construct	The concept, or construct, of interest is the state, experience, attitude, or perception that the CROM intends to measure.
Conceptual equivalence	Conceptual equivalence refers to comparability of meaning across different languages and cultural contexts. It examines the extent to which the meaning and relevance of concept(s) of interest are consistent between a CROM's source version and translated version(s). Testing conceptual equivalence therefore means evaluating whether a concept exists in the languages and cultures of interest and whether it is constructed in the same way across those languages and cultures. Development of a clear and concise list and explanation of items and concepts in the source CROM can serve as a point of reference during the translation process to strengthen the conceptual equivalence of translations and help to avoid ambiguities.
Conceptual model (conceptual framework)	Generally depicted in the form of a figure or diagram, the conceptual model presents the key components to be measured by the CROM, as well as the theoretical structure of the concept of interest.
Construct underrepresentation	When the CROM's content does not cover all components from the ***conceptual model*** (e.g., the conceptual model includes the adverse effect of dizziness, but the CROM does not include a question pertaining to dizziness).

Term	Definition
Content validity	The relevance and completeness of what the CROM measures, in relation to the underlying construct and conceptual model. That is, the CROM content is relevant, appropriate, and comprehensively captures the construct of interest given the population of interest and the context of use. Unlike most other forms of validity, content validity is typically assessed qualitatively.
Context of use (of a CROM)	The conditions under which the CROM is used. These include the purpose and the objectives of the study in which the CROM is to be used, including the population (e.g., people who currently smoke cigarettes), the object to be assessed by the construct (e.g., the type of tobacco and/or nicotine-containing product), the mode of administration (e.g., electronic administration), the timing and number of administrations (single versus repeated administration), etc.
Convergent validity	Evidence of a statistical association (positive or negative) between the CROM and other theoretically relevant measures. For example, if a researcher observes a significant, moderate negative association between perception of product harm (new CROM being developed) and intention to try the product (previously validated CROM), consistent with the researcher's *a priori* expectations about the relationship between these constructs, this provides evidence of convergent validity of the new CROM. Convergent validity is often established by demonstrating a positive correlation between the new CROM and an existing, accepted CROM assessing a similar construct.
Cross-cultural validity, cross-cultural equivalence	Evidence that a CROM can be applied to end-users from different cultures with measurements being comparable. Cross-cultural validity requires first that rigorous translation procedures are followed, and then a quantitative evaluation showing that items do not exhibit substantial bias (often evaluated through formal testing of differential item functioning). If items do show substantial bias, appropriate corrective actions should be considered.
Discriminant validity	Evidence of validity based on the CROM being unrelated/weakly related to theoretically *un*related measures. For example, to support the discriminant validity of a new dependence[2] CROM, a researcher might evaluate the statistical association between scores on this new CROM and scores on another, theoretically unrelated CROM (e.g., a measure of health literacy). If the scores from these two theoretically unrelated CROM evidence a weak, non-significant relationship (consistent with the researcher's *a priori* expectation), this would support discriminant validity of the CROM.

[2] Within this document, "dependence CROM" refers to the individual's self-reported perception of their dependence (e.g., as opposed to a diagnosis or some other indicator of dependence).

Term	Definition
End-user(s)	The intended population of respondents (subjects, participants, consumers) to whom the CROM will be administered (e.g., people who smoke cigarettes, people who are former tobacco product users, adults, youth, etc.).
First order factor	A latent variable in a factor analytical model that causes manifest item scores and accounts for correlations among items.
Forward translation	This refers to the translation of the original language, also called the source version of a CROM into another language, often called the target language.
Item	A statement or question that is presented to an ***end-user***. An item may stand on its own (single-item CROM) or be a part of a set of items forming a multi-item scale or CROM.
Item tracking matrix (ITM)	A document that provides a record of the CROM sourcing for a study, such as whether a CROM was sourced directly from a national survey or published literature, if an existing CROM was modified, or if a CROM was developed for purposes of the study. This document would include details regarding any modifications, including addition(s) or deletion(s) of CROM components (e.g., instructions, items, response options) and rationale for such modifications.
Known-group validity	A CROM has evidence of known-group validity when it differentiates between groups of persons who are known/believed to differ with respect to the construct that the CROM is measuring. For example, people who smoke would be expected to have more favorable attitudes towards smoking than people who do not smoke. Thus, showing that a "smoking attitudes" CROM demonstrated such differences would give evidence of its known-group validity.
Latent variable	A latent variable represents a concept to be measured which cannot be observed directly. Measurements of latent variables are inferred from observable responses by end-users to a set of items that constitute a CROM. For example, "dependence" is an abstract construct, but can be estimated through a set of questions that assess behavioral expressions of perceived dependence.
Linguistic validation	Linguistic validation is a translation process to ensure that translated CROMs are as linguistically, culturally, and conceptually equivalent to their original version as possible. The process consists of a series of steps such as ***forward translation***, ***backward translation***, reconciliation/harmonization of the translations, and ***cognitive debriefing interviews***. The aim of this process is to evaluate the dependability, ***conceptual equivalence***, and accuracy of translations and produce a target language version that is equivalent to the source CROM and allows data pooling and/or comparison of data across languages and countries.

Term	Definition
Measurement equivalence, measurement invariance	Measurement equivalence ensures that scores on ***latent variables*** can be meaningfully compared across different groups of respondents (e.g., participants who smoke cigarettes and participants who use ENDS) without any bias. It requires that the latent variable is related to the observable responses in the same way for different groups of respondents (no additive bias, same strength of relationship). While full invariance of all ***items*** in a CROM is desirable, partial invariance with a subset of invariant items is sufficient to carry out mean comparisons. Bias in some items and/or different item discrimination is then corrected statistically.
Measurement model	In psychometrics, a measurement model links observable responses to CROM (observed scores) to latent variables representing the concept of interest. The observable responses are believed to be caused by the latent variable.
Modern test theory	Modern test theory (also known as item response theory or latent trait theory) comprises a family of ***measurement models*** that feature one person and, in case of dichotomous response scales (e.g., "yes"/"no"), typically one or two item parameters. The person parameter represents the person's level of the construct. Its estimation is the goal of measurement. The first item parameter indicates the item's level of the construct often referred to as item "difficulty." The greater the item difficulty, the more of the construct the item embodies and, consequently, the greater the person parameter must be for the person to agree with, or endorse, the item. The hierarchy of item difficulties in a scale provides insight into the characterization of the construct. For each item, a non-linear, s-shaped item characteristic curve describes the probability to agree for a person depending on the person parameter. Some models also feature a second item parameter describing item discrimination, which alters the slope of the item characteristic curve. The second item parameter is comparable to the factor loading in factor analysis. Its estimation requires the assumption of a normally distributed sample of persons. The Rasch model for measurement specifies only item difficulty and constrains item discrimination to be equal across all items in a scale. Multi-categorical response scales require multiple difficulty parameters (often called item thresholds) for each item indicating the transition from one response category to the next.
Predictive validity	Evidence of validity based on the extent to which scores from the CROM are statistically associated with some criteria measured later (e.g., whether behavioral intentions are related to future behavior). (Note that if criteria are measured at the same time, the evidence of validity is typically referred to as concurrent validity.)

Term	Definition
Psychometrics	The field of science that is concerned with evaluating the functioning ("psychometric properties") of self-report questionnaires, referred to here as CROM. Aspects of CROM functioning that may be evaluated include reliability, validity, invariance/bias, etc.
Rasch model for measurement	The Rasch model for measurement requires item discrimination for all items in a scale to be equal. It is set to 1 implicitly and not estimated. This feature allows for parameter separation implying that person and item parameters (see ***Modern test theory*** above) can be estimated independently of one another and that no distributional assumptions are required. Although statically part of the Modern test theory family of models, the Rasch model is unique in its role as a prescriptive model emphasizing the fit of the data to the model rather than the fit of the model to the data.
Reliability	Reliability is defined as the ratio of true variance (variance in the measurements due to genuine differences between respondents in the underlying construct) and total observed variance, which includes error variance in addition to true variance. Reliability cannot be computed but needs to be estimated. Its estimation can be based on statistical analyses of item variances and covariances (e.g., Cronbach's coefficient alpha as a measure of internal consistency, based on correlations among items) or on the correlation of repeated measurements (test-retest reliability; stability). Reliability is bound between 0 (no reliability) and 1 (perfect reliability with no random error). Since reliability is inversely related to error variance, it is often used as an indication of measurement precision. However, reliability is also a function of the true variance, which is sample dependent. Thus, reliability should be seen as the ability of a CROM to differentiate between respondents in each sample. Reliability is a key psychometric criterion but it does not inform whether the CROM measures what we want to measure, which is a matter of validity. That is, a measure can be reliable without being valid.
Response options	A set of possible responses to an item presented to the respondent. The responses are typically ordered in terms of intensity (for example agreement, or frequency). The minimum number is two (binary response scale, dichotomous scale), while response scales used in Psychometric CROM are commonly multi-categorical (polytomous scales).

Term	Definition
Probing	In the context of qualitative interviews (e.g., ***cognitive debriefing interviews***), "probing" means that the interviewer asks the participant questions to understand how the participant interpreted the CROM, to gather feedback on the CROM, etc. In "retrospective" probing, such questions are asked *after* the CROM has been administered (as opposed to *during* CROM administration).
Social desirability	Occurs when the respondent's answers do not solely reflect their actual beliefs, opinions, states, or traits, but responding is influenced by the desire to come across in a positive (or socially desirable) manner. It is further distinguished between impression management (looking good in the eyes of others) and self-deception (where the participants are deluding themselves). Social desirability, if present, implies a bias in measurement.
Translatability assessment	TA of CROM is the review and evaluation of its source text ideally during its development stage, prior to its use, in order to determine the extent to which it can be suitably and meaningfully translated into another language.
Validation	The process to establish that the performance of a CROM is acceptable for its intended purpose.
Validity	In general, validity refers to whether a CROM measures what it is intended to measure (the ***concept of interest***) in a specific ***context of use*** (purpose of measurement). Validity is a continuous criterion that comes in degrees of accuracy. Empirical evidence of validity is multifaceted and may comprise qualitative aspects (***content validity)*** as well as quantitative, or statistical, aspects (e.g., convergent, discriminant, predictive, know-group validity).

1. INTRODUCTION

CORESTA is an organization developed with the purpose of promoting international cooperation in scientific research relative to tobacco and its derived products. Its vision is "to be recognized by our members and relevant external bodies as an authoritative source of publicly available, credible science and best practices related to tobacco and its derived products."
(https://www.coresta.org/who-we-are-29290.html)

In 2018, CORESTA approved the formation of a new TF to establish best practices and guidelines for the development and use of CROM[3] in research on TNPs[4]. This TF defined a CROM as: a measurement instrument where data are collected by self-report from the subject of research[5].

The CROM TF consists of members from seven contributing manufacturers. Its primary objectives are 1) to provide guidance on the development, modification, and application of CROM, and 2) to facilitate the identification and access to recommended CROM. The CORESTA Scientific Commission provides oversight of the consortium to ensure conformity of the work with CORESTA standards. A consortium approach, with contributions from manufacturers and industry partners, has been taken to develop a scientific framework based on the following shared vision:

- To work together to create a paradigm shift in the way CROM are conceptualized and implemented in research on TNPs,
- To work with SMEs to establish guidance for developing and validating new measures,
- To establish consensus on existing measures and research methods,
- To use a core set of concepts and tools to facilitate sharing, comparing, and replicating findings, and integrating data from multiple sources.

The CROM TF distinguishes between "Psychometric" CROM, which are intended to measure underlying (unobservable) attributes of an individual, and "Descriptive" CROM, which are intended to measure observable characteristics and behavior. To achieve its primary objective, the CROM TF created several working groups (see Figure 1 for an overview of the purpose of each working group). Two separate best practices and guidelines were developed by the working groups:

A. "Consumer-Reported Outcome Measure (CROM) Best Practices and Guidelines with Respect to Psychometric CROM for Use in Research on TNPs"
B. "Consumer-Reported Outcome Measure (CROM) Best Practices and Guidelines with Respect to Descriptive CROM for Research on TNPs"

[3] Within this document, "CROM" can refer to "measure" (singular) or "measures" (plural), which can be inferred through context.

[4] Within these guidelines, "TNPs" refer to tobacco products, as well as nicotine-containing products that do not contain tobacco.

[5] Although not common practice in the field of TNP research, in theory, a CROM could also be completed by someone other than the subject of research. For example, a parent could be asked about their child's use of tobacco products.

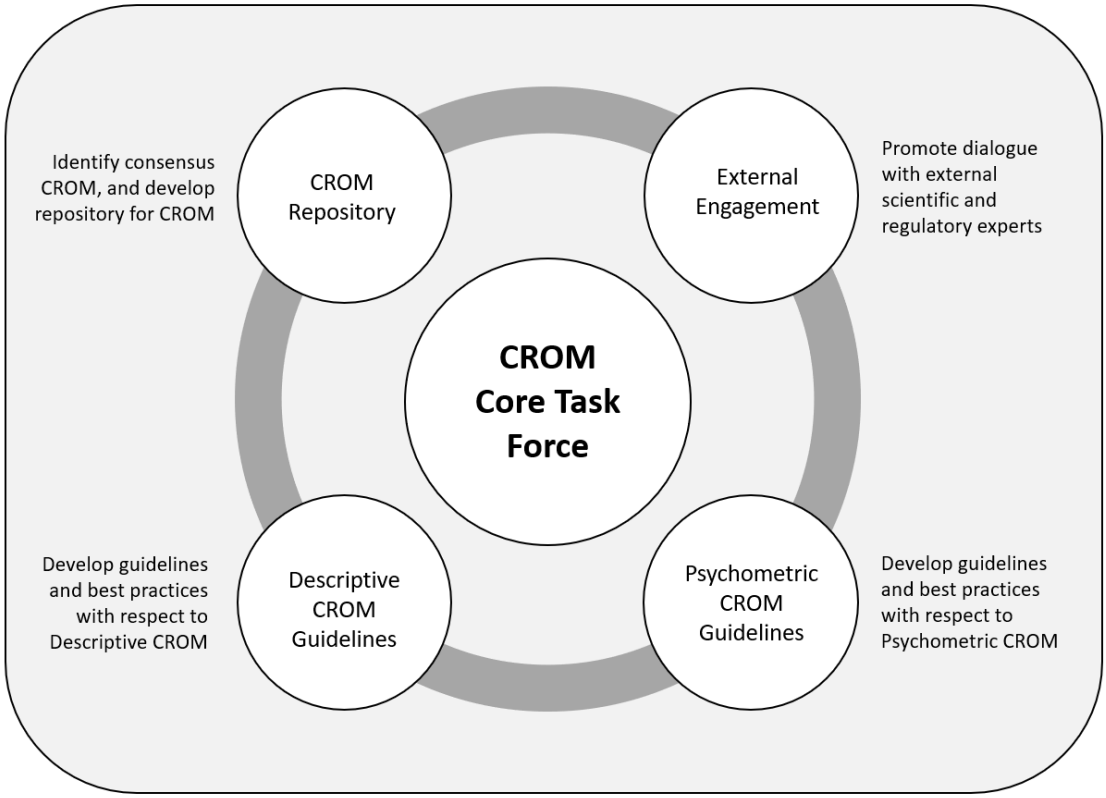

Figure 1 - Overview of the CROM Task Force

This document represents the final deliverable of the working group focused on Psychometric CROM (i.e., guidelines articulating best practices for the selection, development and ***validation***, modification, and implementation of Psychometric CROM for use in research on TNPs).

1.1 Psychometric CROM

"Psychometric CROM" refers to CROM intended to measure underlying individual psychological attributes/unobservable latent constructs. Examples of Psychometric CROM commonly used in TNP regulatory research include but are not limited to the following: product perceptions (e.g., relative and absolute risk perceptions), behavioral intentions (e.g., susceptibility to smoke, likelihood of trying the product, intention to use the product, intention to quit smoking cigarettes), responses to the product/reactions to product use (e.g., dependence[6], craving, withdrawal symptoms, reinforcing effects, taste/sensory effects, liking/satisfaction), ***claim*** perceptions (e.g., believability of a MRTP claim)[7], and impact on health and functioning (e.g., quality of life)[8].

[6] While some items intended to assess dependence ask participants about their (directly observable) behaviors, such as time until first cigarette, these items are not considered Descriptive CROM because responses to these items are intended to reflect an underlying latent construct of dependence and are often combined with other items to estimate the underlying construct. That is, the intention of such items is to extrapolate beyond the self-reported behavior (time until first cigarette) to dependence.

[7] Items assessing comprehension of an MRTP claim are generally descriptive in nature (Descriptive CROM). For example, if a multiple-choice item is developed to determine the percentage of people who smoke cigarettes who correctly understand that a claim is communicating a reduction (as opposed to an elimination) in exposure to harmful chemicals for people who smoke who switch completely to the proposed modified risk product, and results from the item are interpreted in this way, as an expression of a specific understanding (e.g., 80 % of people who smoke selected the correct response to this item), then this item would be considered Descriptive CROM.

[8] Health and functioning CROM may be either Descriptive or Psychometric CROM. See Figure 2 and the discussion below.

Psychometric CROM should be differentiated from "Descriptive CROM," which are CROM intended to measure observable characteristics and behavior. For example, items pertaining to TNP consumption, such as the average number of cigarettes smoked per day, are observable behaviors[9] and therefore, would be considered Descriptive CROM, as would demographic characteristics. Additional information about Descriptive CROM can be found in "Consumer-Reported Outcome Measure (CROM) Guidelines with Respect to Descriptive CROM for Research on Tobacco and Nicotine Containing Products".

A Venn diagram is used in Figure 2 to illustrate how CROM related to concepts such as health and functioning status may include items that could be Descriptive or Psychometric. For example, a question asking whether the participant has been diagnosed with COPD by a doctor or other health professional would be a Descriptive CROM because it relates specifically to an observable objective event. Conversely, asking a participant a series of questions about symptomatology to generate an estimate of that participant's current respiratory symptom severity would be a Psychometric CROM as it attempts to estimate an underlying construct through a combination of items thought to reflect the construct. Importantly, the way that an item/CROM is being applied and interpreted can dictate whether it would be classified as Psychometric or Descriptive. For instance, an item asking about the presence or absence of morning cough in the past 30 days could be Psychometric or Descriptive; if the intention in asking the item is to simply determine the presence or absence of cough, this item would be a Descriptive CROM. Conversely, if the item is administered because the presence or absence of cough is believed to be indicative of an underlying latent construct being measured, such as severity of respiratory symptomatology, then this item would be a Psychometric CROM. As this example illustrates, the distinction is not based on the content of the assessment, but rather on whether there is a step of inference from the literal content of the item to an underlying latent construct.

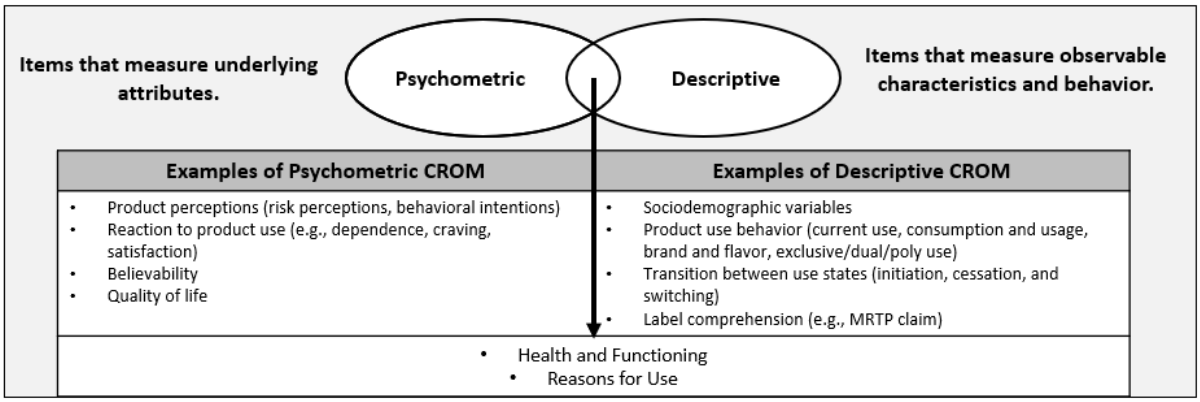

Figure 2 - Venn diagram illustrating the potential overlap in content between Psychometric and Descriptive CROM

[9] The core concept is that they be, in principle, capable of being directly observed, even if in practice we rely on the respondents' self-reports.

1.2 Context of The Psychometric CROM Guidelines

The objective of this working group of the CROM TF is to generate best practices and guidelines for the selection, development, modification, and implementation of Psychometric CROM in research on TNPs, including tobacco harm reduction and public health research, behavioral research, and regulatory research. Currently, aside from the guidelines for Descriptive CROM developed by one of the other CROM TF working groups, only one guidance document exists pertaining to CROM in the tobacco product space; this is the US FDA CTP TPPIS guidance [2]. The scope of the TPPIS guidance is somewhat different from this document in that, while it provides general recommendations related to the development, adaptation, and use of measures of perception, intentions, and understanding within the context of tobacco product perception studies, it does not address other Psychometric CROM, such as measures of dependence or craving, and it does not address the use of CROM in research other than TPPIS, or outside of the regulatory context. Therefore, additional guidance is needed.

The lack of additional guidance around CROM in the TNP regulatory space in the US and globally is somewhat surprising given the need for CROM data to support regulatory filings, such as PMTAs and MRTPAs in the US. Of note, while there are several regulatory guidance documents from the FDA related to PRO measures [3, 4], these guidance documents may not be directly translatable to CROM in the TNP space. First, US expectations of FDA CDER, which regulated drug products but not tobacco products, regarding the development, modification, and implementation of PRO measures may differ from the FDA CTP expectations for CROM. Indeed, there are substantial differences in FDA PRO guidance vs. FDA TPPIS guidance regarding standards for measure development and the robustness of ***validity*** evidence that is recommended. Such differences may be due to the fact that no FDA guidance for CROM in the TNP space existed until very recently while FDA guidance for PROs have existed for decades.

Second, there are important and fundamental differences in the regulatory frameworks under which CDER operates and those under which CTP operates. As described later in this document, some CROM are developed or applied with the express purpose of meeting a regulatory need or requirement (e.g., addressing a US CTP PMTA or MRTPA requirement). In such instances, it is important to keep regulatory requirements central to CROM development. This can be at odds with a consumer perspective, which is sometimes given priority in PRO measures developed for drug evaluation by CDER or may be more relevant for TNP public health research outside of regulatory purview. For example, the risks and risk perceptions that regulators may be most concerned with are risks associated with TNPs that impose considerable burdens on morbidity and mortality. Conversely, people who use TNPs themselves may emphasize social perceptions or aesthetic consequences, such as dirt, odor, or cosmetic effects (e.g., wrinkles, yellowed fingers) that may not be given much weight by regulators but could still be relevant and contribute to the overall scientific evaluation of and evidence base for TNPs and may also be relevant to developing interventions for consumers. Thus, it is important to define and focus on the intended use and purpose of the CROM.

1.3 Overview of the Psychometric CROM Guidelines

This document is intended to provide an overview of best practices and general guidelines for the selection, development, modification, and implementation of Psychometric CROM in research pertaining to TNPs. These guidelines contain five chapters. Following this introduction to the guidelines (Chapter 1), the chapters delve into particulars.

Chapter 2: Defining the Construct to be Measured and Identifying the Ideal CROM Characteristics based on the Objective of the Study, is intended to facilitate a researcher's decision as to whether an existing "off the shelf" CROM is appropriate for their study (without

any additional modifications or testing), an existing CROM might be modified to meet the researcher's needs, or whether it is necessary to develop a new CROM for purposes of the study.

Chapter 3: Modifying an Existing Psychometric CROM presents best practices for modifying/adapting an existing Psychometric CROM. This includes: (1) examples of the type and extent of modifications that a researcher might make, and (2) qualitative and quantitative strategies that could be used to gather evidence to support the modification, and factors that influence the type and extent of evidence recommended to support the modifications.

Chapter 4: Developing and Validating a New Psychometric CROM provides an overview of the general stages of Psychometric CROM development, including best practices for executing each stage.

Finally, *Chapter 5: Application, Implementation, and Interpretation of a Psychometric CROM* discusses considerations for the application/implementation of Psychometric CROM in a research study as well as considerations pertaining to scoring and interpretation of the scores and findings generated from the CROM.

The following are important points concerning this guideline document:

- These guidelines and the CROM TF do not pretend to represent authoritatively the views of regulatory bodies and any guidance they may publish. This document is intended to represent basic principles of psychometric measurement science, and to serve as a guide for those conducting TNP research using or considering the use of Psychometric CROM.
- This document describes the current thinking of this CROM working group and should be viewed only as recommendations. The use of the word "should" simply means that something is suggested or recommended. The recommendations in this document are grounded in scientific rationale and aim to provide an overview of foundational principles grounded in ***psychometrics*** and what may currently be considered best practices regarding the use of Psychometric CROM in research on TNPs. However, best practices may also evolve over time with advances in research on TNPs, psychometrics, and measurement science.
- The guidelines are not intended to reflect unattainable standards; that said, researchers in the field of TNPs should be knowledgeable about these guidelines, and then make an informed decision as to what extent they are applicable or necessary for a particular study. The researcher is ultimately responsible for defending their research.
- These guidelines are intended to represent a set of principles and recommendations, forming a foundation of best practices related to Psychometric CROM; they do not (and cannot) provide recommendations around which specific CROM a researcher should use, as the CROM that are most appropriate for a particular study depend on several study-specific factors, a topic which is discussed further in later sections of these guidelines. These guidelines are an important precursor to the development of a CROM repository (addressed by a different working group from the CROM TF), as the information contained within Chapter 2 of these guidelines will help researchers use such a CROM repository most effectively (see also Chapter 2).
- The intended audience of these guidelines are individuals who not only have appropriate knowledge of behaviors associated with the use of TNPs, but who also have basic familiarity or experience with how CROM (or similarly, PRO measures) can be used as endpoints in research studies. That is, the guidelines assume some basic knowledge of relevant concepts and terminology, although a table at the beginning of this document (see Definition of Terms) is provided to aid readers unfamiliar with more advanced

concepts and terminology. Words or phrases found in the Definition of Terms appear in ***bold italics*** at first mention. Additionally, recommended reading/references are incorporated throughout the guidelines for readers interested in learning more.

- The recommendations presented in these guidelines do not reflect the views of individual companies whose members are part of the CROM TF and are not intended to have binding implications on past, current, or future research conducted by individual companies or research that may be conducted in support of TNP regulatory applications or scientific publications.
- Based on consensus within the research community [5, 6], throughout the guidelines, we adopt the use of person-first language (e.g., "people who smoke") rather than commonly used labels (e.g., "smokers") to promote greater respect and convey dignity for people who use TNPs. It has been suggested that the use of precise and bias-free language to describe people who use TNPs has the potential to reduce tobacco-related stigma and may enhance the precision of scientific communication [6].
- We recognize that there may be instances where a researcher is interested in using Psychometric CROM with those underage to purchase TNP, such as youth. This document does not provide recommendations regarding whether a researcher should or should not collect data from those underage to purchase. However, if a researcher does intend to use a Psychometric CROM to collect data from those underage to purchase, the recommendations and best practices contained within this document (e.g., such as how to identify an appropriate CROM, modify an existing CROM, develop a new CROM, etc.) are applicable. Because Psychometric CROM may be used with youth, there are examples throughout this document which reference youth.

1.4 Methodology for Guideline Development

This CROM TF working group includes 11 researchers representing companies participating in the CROM TF. Members have diverse expertise and backgrounds, with experience in psychometrics, PRO measures, survey methodology, and behavior associated with use of TNPs. Since its inception in 2020, this working group has met regularly to discuss topics related to the guidelines, external dissemination strategies, and to outline and draft these guidelines.

As a starting point, working group members reviewed around 50 different documents, including peer-reviewed publications and publicly available guidelines and best practices published by other prominent organizations from related fields. These include but are not limited to guidance and best practices authored by US FDA, CONSORT PRO Group, ISOQOL, COSMIN, SISAQOL Consortium, ISPOR, PRO Harmonization Group, PROMIS, ERIQA, SPIRIT-PRO Group, and AERA/APA/NCME.

The working group adopted a consensus-based approach for drafting the guidelines, inspired by approaches taken by prominent outcomes research organizations, such as ISPOR. Development of the guidelines was a collaborative, iterative process, and the working group sought active collaboration from SMEs with diverse perspectives and expertise representing public health, academia, and the tobacco industry throughout the guideline development process. These guidelines were collaboratively drafted by several working group members in conjunction with three external SMEs based on their review of relevant literature and expertise.

The working group authors presented outlines and proposed content of the draft guidelines at international conferences, including the SRNT, ISPOR and ISOQOL conferences, and the CORESTA Smoke Science and Product Technology Conference SSPT. The authors used the conferences as opportunities to reach out and actively seek feedback and collaborations from individuals with relevant expertise who were interested in assisting with guideline development or providing feedback on draft content.

2. DEFINING THE CONSTRUCT TO BE MEASURED AND IDENTIFYING THE IDEAL CROM CHARACTERISTICS BASED ON THE OBJECTIVE OF THE STUDY

When planning a research study that will require Psychometric CROM, researchers should start by defining the construct to be measured and considering the qualities of a CROM that are most important for purposes of the study. That is, within the context of the study and its purpose, *what is (are) the construct(s) that need(s) to be measured?* Careful definition and elaboration of the construct, and consideration of the study context (e.g., the population being studied, the mode of administration) can then lead to a further question: *what would the ideal CROM look like?*

Going through this exercise will allow the researcher to concretely assess the match between the needs of the study and existing Psychometric CROM[10], facilitating determination of whether an existing CROM would be an appropriate choice for the study and to justify this CROM selection.

For example, a researcher may want to evaluate the impact on respiratory symptoms when people who smoke who are not diagnosed with pulmonary disease switch completely from combustible cigarettes to the candidate product, as an indication that the change in behavior has resulted in clinically meaningful changes in respiratory health status. That is, the researcher is looking for a CROM appropriate to assess respiratory symptoms within a particular timeframe under consideration. In this hypothetical example, the researcher hopes to use the CROM to generate supportive evidence for a regulatory application claiming that a reduction in the likelihood of developing respiratory disease is reasonably likely among people who smoke who switch from cigarettes to the candidate product.

For this purpose, the researcher may need a CROM of respiratory symptoms that (1) is appropriate for use with a non-diseased population of people who smoke and (2) is sensitive to detect change in respiratory symptoms expected to occur within the designated time after switching. Further, if the researcher would like to assess changes in respiratory symptoms monthly during the study[11], then the ideal respiratory symptom CROM for this study might have a 30-day recall period (i.e., ask participants to report their symptoms over the past 30 days). Selection of a recall period for a CROM also needs to consider the limits of autobiographical memory to ensure that respondents can reasonably recall the relevant content over the period designated.

As another example, a researcher may need to measure satisfaction obtained from use of a product repeatedly (e.g., every 5 or 10 minutes) during a 30-minute period of *ad libitum* product use, where the candidate product will be tested against a comparator product (combustible cigarettes). As before, the researcher would need to start by operationalizing the construct to be measured (i.e., *how is "satisfaction" defined?* and if satisfaction is conceptualized as a

[10] Within the context of these guidelines, "existing CROM" refers to any existing CROM, regardless of whether it is published in peer-reviewed literature, extensively validated, etc. CROM may be sourced from any number of places, including but not limited to published literature, publicly available regulatory filings, national surveys, and existing databases (e.g., PhenX Toolkit, PROQOLID, etc.). That said, FDA TPPIS guidance recommends using CROM that have demonstrated "some type of validity" in peer reviewed literature when possible; when that is not possible, the researcher should consider CROM that are "widely used in peer-reviewed literature".

[11] One important factor to consider when determining the frequency of assessment is the extent to which the outcome is expected to fluctuate over time. That is, if a construct is expected to be fairly stable over time, a researcher may choose longer intervals between assessments compared to a construct which is expected to be more responsive to the independent variable/fluctuate over time (e.g., nicotine withdrawal symptoms).

multidimensional construct, *what are the components of satisfaction?*). Because the researcher intends to compare satisfaction of the candidate product against a comparator product, it is important that the CROM includes content that assesses aspects of satisfaction that are applicable and relevant for both products[12]. Within the context of this study, the researcher requires a measure that (1) is sensitive to detect immediate changes in satisfaction within participants over brief periods of time, (2) is extremely brief, and (3) can be administered repeatedly. Further, the CROM's content would need to be applicable and appropriate for both product categories (i.e., for the candidate and comparator products), and scores on the measure should be comparable across product categories (e.g., does a score of "5" reflect the same amount of satisfaction for both product categories?).

This exercise of identifying the optimal characteristics of a Psychometric CROM within the context of the particular study will help the researcher determine whether (1) an existing CROM may be appropriate (with or without any additional testing[13]), (2) an existing CROM might be modified to meet the study's needs, or (3) it may be necessary to develop a new CROM. Later sections of this document provide recommendations for modifying an existing CROM and developing a new CROM.

Following on the first hypothetical example from above (i.e., identifying a CROM appropriate to assess respiratory symptoms that may be amenable to change within a particular timeframe): The researcher conducts a review of peer-reviewed literature, national surveys, and CROM databases but is unable to identify an existing respiratory symptom CROM appropriate for use with a non-diseased population of people who smoke. That is, existing respiratory symptom CROM were developed and validated specifically for use with clinical populations, such as those with asthma or COPD, and their content is inappropriate (reflects severe respiratory symptoms) or not applicable for people who smoke who do not have such diagnoses (and are likely experiencing more mild respiratory symptoms.[14] Therefore, because existing CROM do not meet the researcher's needs, the researcher decides to develop a new CROM. Without going through this exercise, the researcher may have elected to go with a commonly used, well-validated respiratory symptom PRO endorsed by the FDA (e.g., St. George's Respiratory Symptom Questionnaire), without realizing that it would be an inappropriate choice for purposes of their specific study (as it was developed and validated to be used in clinical COPD populations).

Table 1 includes considerations when determining the optimal Psychometric CROM characteristics for a particular study. This list is not intended to be comprehensive but touches upon some key factors for consideration. With the exception of the first consideration (definition of the construct to be measured), the considerations are not listed in order of importance, as order of importance will be dictated by the study. Starting this exercise with a clear definition of the construct to be measured is a critical initial step. Even a commonly used, validated CROM may not be the correct choice for a study if it does not align with what the researcher needs to measure. See Section 4.1 for various approaches that can be used to facilitate the process of defining the construct and developing a conceptual model.

[12] See Section 4.1 for various approaches that can be used to facilitate the process of defining the construct and developing a conceptual model.

[13] Even if the researcher decides to use an existing Psychometric CROM for their study, the researcher may still conclude that additional testing would be beneficial to evaluate a critical psychometric property (e.g., ability to detect change over time).

[14] When there is a mismatch between the "difficulty" of a CROM's items and the persons being assessed (such as when using a CROM with items assessing severe respiratory symptoms to evaluate respiratory symptoms among people who smoke with mild or moderate respiratory symptoms), this can lead to reduced measurement precision and a floor effect, inhibiting our ability to detect reductions in respiratory symptoms over time.

When evaluating the suitability of an existing Psychometric CROM to be used in a study, the most pivotal criterion is that the CROM must measure what the researcher intends to measure in the study. It should be noted though that each CROM has been validated for a specific *context of use*. The context might include a reference to, for example, a particular, or possibly several, TNPs (e.g., combustible cigarettes versus ENDS), to a specific population (e.g., people who use currently versus people who never used; adults who use versus adolescents who use; people of a certain nationality or ethnicity), or to a method of administration (e.g., online versus paper and pencil; self-report versus interviewer-administered). If these contextual factors of the study in question differ from those the Psychometric CROM was validated for, the existing CROM may not be appropriate in its current form and modifications may be necessary (see Chapter 3).

Table 1 - Considerations when Defining the Psychometric CROM Characteristics of Greatest Importance for a Particular Study

Consideration	Description/Examples
Definition of the construct to be measured	What is the concept to be measured, and how is it defined? What are the components/aspects of the construct that should be represented in the CROM?[15] Is the construct likely stable (trait) or unstable (state) over time?[16]
Score interpretation	What should the score reflect? Is the researcher looking for a single total score (e.g., subjective effects of a product), or separate scores to reflect the different aspects of the construct (e.g., negative reinforcing effects and positive reinforcing effects)?
Defining context of use within the study	Who are the participants in the study? What "target population" do they represent (e.g., adults who smoke cigarettes)? An ideal CROM would have psychometric evidence supporting its use with participants representing the target population. Will the CROM need to be appropriate for any population(s) of interest (e.g., those with limited health literacy, youth)? What is the study type (e.g., clinical, real-world evidence, TPPIS, etc.) and study design (e.g., cross-sectional vs longitudinal; if longitudinal, over what duration)? Will the CROM be applied to different products (candidate and comparator products)?
Psychometric functioning	What are the psychometric properties of greatest importance within the context of the study (e.g., ***ability to detect change***, ***known-group validity***, equivalence of scores across product categories, ***predictive validity***, etc.)?

[15] For example, if the researcher intends to measure a multi-faceted construct such as quality of life, a psychometric CROM that assesses a single component of quality of life would be insufficient to meet the researcher's needs.

[16] Some CROM reference specific timeframes (e.g., urge to smoke "over the past 7 days" vs. "right now"). The researcher will need to decide whether such timeframes are appropriate given the construct to be measured, the objective of the study, and the demands on memory.

Consideration	Description/Examples
Administration considerations	Mode/method of administration: • Does the study require electronic administration? ○ If so, on devices with different sized screens? • Does the CROM require electronic administration/scoring (e.g., CAT, display of digital images)? • Does administration of the CROM involve conditions that might affect participant responses (e.g., will a research assistant ask CROM items out loud while the participant is getting their blood drawn as part of a clinical study?) • Does the CROM need to be administered over the phone? • If administered repeatedly, how frequently will the CROM be completed? • Are there study restrictions regarding the length (time required for administration) of the CROM?
Accessibility of the CROM	Licensing fees, permission to use, copyright clearance

Once the researcher has identified the Psychometric CROM characteristics of greatest importance for purposes of the study, the next step is to compare these characteristics against existing Psychometric CROM to determine whether an existing Psychometric CROM may be appropriate. The researcher may be concerned about the match between the ideal and actual Psychometric CROM characteristics depending on whether the CROM addresses a primary, secondary, tertiary, or exploratory study objective. For example, the researcher may place greater importance on the rigor of the validity evidence for the Psychometric CROM if the CROM will be used to address a primary study objective or a regulatory requirement vs. an exploratory study objective.

If the researcher decides that an existing Psychometric CROM does not meet the study's needs, they may decide to proceed with either (1) modifying/adapting an existing Psychometric CROM or (2) developing a new Psychometric CROM. Guidelines and best practices for these activities are described below. Of note, at this stage, the researcher may also determine that an existing Psychometric CROM may be an appropriate fit for the study, but that additional research would be beneficial to evaluate a critical psychometric property. For instance, a researcher may want to evaluate whether an existing measure of behavioral intentions predicts actual behavior in the future (predictive validity), or whether a measure of satisfaction is sensitive to detect changes in nicotine consumption. Similarly, if data related to a critical psychometric property is inconclusive or contradictory across studies (especially if the Psychometric CROM will be used to support a primary study objective or needed to address a regulatory requirement), or if consultation with the relevant regulator indicates that the currently-available data supporting the Psychometric CROM are not adequate, the researcher may decide to collect data and evaluate the psychometric property before using the existing Psychometric CROM (see Chapter 4 for a discussion around the collection of data to evaluate psychometric functioning).

Some key aspects particularly relevant in TNP research warrant further comment. First, existing CROM are typically validated in one specific language. If the language does not match the language to be used in the application study, a translation is required unless the respondents'

language skills can be considered sufficient, or adequate, for responding to items in a foreign language. A suitable translation needs to consider linguistic as well as cultural aspects (see Section 3.4.1) to provide a psychometrically equivalent version of the CROM. Sometimes the languages of different countries are broadly speaking the same but represent well-defined variants, e.g., English for the UK vs. English for the US. An adaptation of a CROM developed, in the US, for example, but to be applied in the UK, may be dispensable if the CROM developers have been careful to avoid Americanisms as opposed to using US-specific spellings, terms, and expressions. When scales developed in languages other than English are published in English-language academic journals, the authors often provide English working translations, which must not be confused with properly validated linguistic versions of the CROM. Another issue related to a CROM's language is the level of language used. A CROM using sophisticated language should not be used in populations with limited language skills and low literacy. However, as a rule, CROM should use simple language that is unambiguous and easy to understand.

Second, many CROM in the field of TNP research are developed for a particular age group, typically either adults or adolescents. A CROM solely validated for adults should not be straightforwardly administered to adolescents without qualitative research investigating whether the items are properly understood and meaningful. As an example, in the measurement of dependence, items may reflect manifestations of dependence logically applicable only to adults (e.g., product use at the workplace). Conversely, CROM developed and validated for youths may require adaptations when applied to adults. CROM validated for a particular group of adults (i.e., individuals 20 to 30 years old), can generally be used for other adult age brackets (i.e., 40 to 70 years old), unless there is a reasonable basis for believing that the new group may be materially different from the group for whom the CROM has been validated.

Third, CROM are typically developed and validated for a specific type of person (e.g., adults who smoke cigarettes). For instance, a CROM designed to measure the perceived risk of smoking cigarettes in people who currently smoke should not be applied to people who never smoked without adaptations unless qualitative evidence supports its applicability. In contrast, the application to people who used formerly might be argued as people who used formerly necessarily were previously people who used currently. That said, adaptations of sentence-stems or instructions may still be required.

Fifth, a frequent feature of concepts of interest in research on TNPs, such as perceived risk, self-reported dependence, subjective product evaluation, or intent to use, is their reference to a particular tobacco or nicotine product. A CROM developed to measure dependence on cigarettes may focus on phenomena indicating dependence on this specific tobacco or nicotine product and therefore, may not be applicable to ENDS. Studies aiming to compare different products must use CROM for which there is evidence supporting their applicability/appropriateness to all products included in the application study. While the same set of self-report questions may be applicable to different products indicating the same concept of interest (e.g., dependence), some manifest questions may stand in a different relationship to the latent concept to be measured. For example, the item "time to first use after awakening" may prove to be an excellent indicator of dependence on cigarettes but perform poorly for other TNPs [7]. If such peculiarities are not taken into account, comparisons between products can be biased. Thus, in addition to evidence of applicability to different TNPs, evidence of comparability across different products (supported by evidence of ***measurement equivalence***, or invariance) is particularly helpful.

Finally, it is worth mentioning that evidence of a Psychometric CROM's applicability is often found outside the original publication introducing the CROM. The same or other authors might have conducted further empirical research that allows for a broader application of the CROM. After all, a CROM's validation is an on-going process that is never completely finished.

3. MODIFYING AN EXISTING PSYCHOMETRIC CROM

There are numerous modifications that could be made to an existing Psychometric CROM, and these modifications vary in terms of the *type* of modification (i.e., changes to content, administration, and/or application) and the *extent* of the modification (i.e., minor, moderate, substantial). In this section, we present definitions and examples of the type and extent of modifications that a researcher might make to an existing CROM. Then, we discuss qualitative and quantitative strategies that can be used to gather evidence to support the modification, as well as the factors that influence the type and extent of evidence recommended to support the modifications.

3.1 Types of CROM Modifications

Illustrative (non-exhaustive) examples of the three types of Psychometric CROM modifications are presented in Table 2. It is not uncommon that a CROM modification may impact multiple areas, such as content and application, such that modifying a CROM to a new population would likely include both content and application modifications.

Table 2 - Types of Psychometric CROM Modifications

Type of Modification	Illustrative Examples (Non-Exhaustive)
Content: Modifying the instructions, items, and/or response options	• Removing or introducing a ***response option*** of "I don't know" • Adding response labels so that a scale is fully labeled • Changing the number of response categories • Changing response category labels • Changing instructions and/or item content to reference a different product category (e.g., "ENDS" instead of "cigarettes," updating language/terminology) • Adding item(s) • Removing item(s)/only administering a subset of items • Adding images to items to improve clarity/comprehension • Changing the recall period (e.g., "in the past 30 days" to "in the past 7 days")
Administration: Changing the mode, method, and/or format of administration	• Administering a CROM developed for paper-and-pencil electronically • Changing the method of administration from self-report to interviewer administered • Modifying a CROM to fit a small screen device (smartphone) by administering one item per screen instead of the items together as a grid • Changing a rating task (asking the participant to respond to each item by selecting a value on a numerical rating scale) to a drag-and-drop task • Changing the order of item administration (fixed order vs. randomized)

Type of Modification	Illustrative Examples (Non-Exhaustive)
Application: Applying the CROM in a new way, such as to a new population or product (from which it was originally developed/ validated)	• A measure of smoking susceptibility developed to assess susceptibility among youth is used to assess smoking susceptibility in adults • A measure of cigarette dependence developed for use with adults who smoke cigarettes is administered to people who use ENDS to assess dependence on ENDS • Translating a CROM into a different language and administering it to a new population (i.e., individuals whose primary language differs from languages the CROM has been validated for) • Administering a CROM to individuals from another culture (i.e., individuals whose cultural background differs from the background of individuals for whom the CROM was originally validated for)

3.2 Extent of CROM Modifications

In principle, the extent of modifications can be mapped onto a continuum ranging from very minor to very substantial (Figure 3). The key question though is whether the modification requires some sort of evidence based on qualitative and/or quantitative research. This implies two broadly defined classes of modifications, which we name "Minor" and "Substantial," respectively. Table 3 includes the definitions, which are driven by the need for evidence to support the modification, and illustrative, non-exhaustive examples of modifications which likely fall into the Minor[17] and Substantial classifications. Mapping a continuum (of modifications) onto two categories necessarily represents a simplification. Accordingly, it may not be readily apparent whether a given modification, possibly considered to be "Moderate", ultimately is to be classified as Minor or Substantial as defined in Table 3. This decision will also depend on the specific circumstances. As a rule, a modification should be classified as Minor only if it is minimal or merely consists of unambiguous clarifications added (see examples below). If in doubt, collecting qualitative and/or quantitative evidence can be helpful to support modifications even when it may not be considered necessary.

[17] The examples of Minor modifications listed in Table 3 might be considered Moderate or even Substantial depending on the likely impact on end-users' interpretation and response to the CROM. See discussion below.

Figure 3 - Extent of CROM Modifications

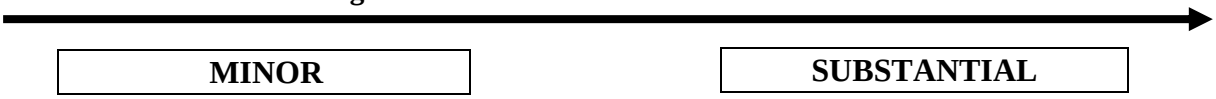

Table 3 - Recommendations pertaining to CROM Modifications

Modification	Minor	Substantial
Definition	Modifications that are **not** reasonably likely to impact end-users' interpretation of CROM content and response to the CROM, above and beyond changes to interpretation and response that are a result of improving clarity/reducing measurement error.[a]	Modifications that could reasonably change end-users' interpretation of the CROM content and response to the CROM items.
Examples	• Making the text bold and underlining the recall period in the instructions ("In the **past 7 days"**) for visibility and emphasis • Changing font size or font style • Adding additional clarifying language to an item or instruction • Adding an image of the product being referenced • Adding an "I don't know" response option • Administering a paper-and-pencil CROM electronically, without changing the presentation of the CROM • Administering items forming a single dimension from a multi-dimensional CROM[b]	• Administering a subset of items from a unidimensional CROM (developing a "short form") • Changing the type of task (e.g., a numerical rating task is changed to a drag-and-drop task) • Changing the type of response scale (e.g., from a 5-category fully labeled scale to a visual analog scale, from 5-point descriptive response scale to 11-point numerical rating scale) • Changing the content of the response scale (e.g., replacing a frequency scale with an intensity scale) • Adding items to a CROM • Administering the CROM to a population for which it was not developed (e.g., a measure of cigarette dependence developed for use with adults who smoke cigarettes is administered to adolescents) • Applying the CROM to TNPs for which it was not developed (e.g., a measure of cigarette dependence is administered to individuals who use ENDS to assess dependence on ENDS) • Translating a CROM into a new language and administering it to this new cultural population

Modification	Minor	Substantial
Recommended Approach(es) to Support Modification	• Generally, no evidence is needed • In certain circumstances, qualitative evidence may be helpful (e.g., to ensure that new clarifying language added to instructions is clear) • Usability testing may be helpful when modifying a paper-and-pencil CROM for electronic administration	Qualitative and/or quantitative evidence is always recommended • Quantitative evidence is recommended to support development of a short form[c] • If CROM content is substantially changed (e.g., changing response task or response scale, adding new items), either (or both) qualitative and quantitative evidence could be used to support the modification[d] • If scores from two versions of a CROM are being directly compared in a study (e.g., ENDS dependence vs. cigarette dependence), quantitative evidence is recommended • Quantitative evidence is needed when administering a CROM to a new population (e.g., youth vs. adults) or product • Qualitative and in some cases quantitative evidence is recommended when translating a CROM into a new language

[a] Often, minor modifications are made with the explicit intention of correcting inaccurate interpretation or misunderstanding (reducing measurement error), which may subsequently *correct* interpretation.

[b] This modification would be considered Minor, assuming that the items from that dimension are scored and interpreted separately from the remaining items comprising the full CROM. If the researcher is dropping items of a unidimensional CROM to create a short form, impacting scoring, this would generally constitute a Substantial modification.

[c] Qualitative strategies may also be helpful, such as asking SMEs to review the items comprising the new short form to ensure that no critical content from the long-form of the CROM is missing.

[d] Depending on the modification, qualitative evidence is generally helpful to ensure that participants understand the new content, such as the new response task (e.g., drag-and-drop task), type of rating scale (e.g., participants perceive that the new response categories reflecting intensity make sense given the construct being measured and are the appropriate level of granularity), or new item(s). In many cases, quantitative evidence is recommended to verify adequate psychometric functioning of the modified CROM (e.g., that the response categories are ordered, that new/modified items are internally consistent with other items and loading onto factors as anticipated, etc.).

As can be seen from Table 3, the criteria differentiating the modification classifications have to do with the likelihood that end-users' interpretation of the CROM content and response to the CROM is impacted as a result of the change. In general, Minor modifications reflect changes that a researcher might make with the intention of improving clarity/reducing ambiguity, ultimately reducing measurement error. For instance, a researcher may choose to modify the question "In your opinion, how harmful is smoking to your health?" by adding the word "cigarettes" (i.e., "In your opinion, how harmful is smoking cigarettes to your health?") to reduce potential confusion that the item refers to ENDS or another product. Because this modification improves clarity and reduces measurement error, this modification would be considered Minor. Of note, by *correcting misinterpretation*, Minor modifications may indeed

have an impact on participants' responses to the CROM by reducing random noise/response or systematic error (e.g., originally misunderstanding an item to be referencing ENDS, the participant would have selected "moderately harmful", but now understanding that the item is asking about perception of combusted cigarettes, they select "extremely harmful"). Another example might be adding clarifying language to an instruction; "Please think about ALL of the tobacco/nicotine products you use" could be modified to "Please think about ALL of the tobacco/nicotine products you use. Some examples of tobacco and/or nicotine products include cigarettes, e-cigarettes, heated tobacco products, or smokeless tobacco products (e.g., chewing tobacco, snus, snuff, dissolvable)." In this case, the modification is intended to facilitate understanding among those participants who may be unfamiliar with the range of TNPs that they are expected to consider, presumably reducing measurement error previously caused by random guessing or misunderstanding. A similar example of a Minor modification would be adding an image of the product being referenced to the CROM to reduce random guessing or misunderstanding. Indeed, it is well-established that the general public may misreport types of ENDS products used (e.g., [8]) and images can improve accuracy of responding.

Conversely, if a CROM with 4-point descriptive response scale (from "not at all likely" to "extremely likely") is modified so that it now uses an 11-point numerical rating scale (from 0 % to 100 % likely), this is a Substantial modification, as the change to the response options may impact how participants think about and respond to the CROM content. For example, a participant who previously said "not at all likely" may select any number of responses when provided with a more granular numeric rating scale (e.g., 0 %, 10 %, 20 %), as the granular scale may allow them to express their perception of likelihood more effectively. However, a numeric rating scale that is too granular may lead to greater measurement error if participants are not able to effectively use the scale or those who have low numeracy find it difficult to express their perceptions on a numerical scale. Therefore, as with other Substantial modifications, such a modification warrants additional testing.

Of note, when modifying content of an existing CROM, it is recommended to clearly document such changes using a table (or in some other organized fashion) with columns showing the original CROM content and the modified CROM content, as well as rationale justifying the modification(s) (See Chapter 4 for additional information on item tracking matrices).

3.3 Types of Evidence that can be Gathered to Support the Modification

A researcher may choose to gather qualitative and/or quantitative evidence to support the modification, i.e., that integrity of the CROM has been maintained (the modification has not led to end-users interpreting the CROM differently and/or responding differently, aside from reducing measurement error [see definition of Minor modification in Table 3]).

For many instances of CROM modifications, especially when content is modified, conducting individual cognitive debriefing interviews (see Chapter 4 for more information about cognitive debriefing interviews) to qualitatively assess understanding and interpretation of the modified CROM would be helpful and is recommended. Focus groups are another qualitative strategy that can be used for this purpose. In conducting cognitive interviews or focus groups, the researcher may choose to present content from both the original and modified CROM to participants to determine if the modification resulted in differential interpretation and/or response (again, this depends on the purpose of the modification). Usability testing may also be useful in certain circumstances, such as when modifying CROM formatting to fit a small screen electronic device (smartphone).

Quantitative strategies may sometimes be necessary to support CROM modifications and are specifically recommended when the researcher is making direct comparisons between scores from the original and modified CROM (e.g., leveraging the example from Table 2, comparing level of product dependence for people who smoke cigarettes and people who use ENDS). In this case, demonstrating measurement equivalence quantitatively across populations is of primary importance[18]. Quantitative strategies may also be useful in situations where a researcher would like to evaluate a modification retrospectively and has access to quantitative data from both the original and modified CROM. As an example, CFA is one appropriate statistical approach for evaluating whether a CROM's internal structure (i.e., number of factors, factor loadings) is invariant across groups; interested readers are referred elsewhere [9, 10].

Of note, researchers may choose to leverage both qualitative and quantitative strategies, especially in cases where Substantial modifications have been made.

3.4 Type and Extent of Evidence Recommended to Support Modifications

The type of evidence (e.g., qualitative, quantitative) and amount of the evidence which may be useful to support the modification depends on two factors. The first of these factors is the extent of the modification (Minor or Substantial, see Table 3). All else being equal, Minor modifications generally do not necessitate additional testing, although cognitive testing may still be helpful to provide evidence that the modification did not negatively interfere with accurate interpretation of the CROM[19]. Conversely, additional evidence supporting the modification is typically recommended for Substantial modifications. Certainly, in some circumstances, modifications to a CROM may be so substantial that the CROM might reasonably be considered a "new" CROM, as opposed to being "modified" from the original; in such instances, the researcher should consider following recommendations outlined in Chapter 4, as this approach is generally more thorough and may provide greater assurances that the Psychometric CROM will ultimately function well.

The second factor impacting the type and extent of the evidence needed to support a modification is the way in which the modified CROM will be used and interpreted. There are specific circumstances where it is important for the original and modified CROMs to be psychometrically equivalent ("parallel forms"); as an example, if a CROM assessing cigarette dependence is modified to reference ENDS dependence and both versions of the CROM are being used in a study to directly assess differences in cigarette and ENDS dependence as a primary study objective or to address a regulatory requirement, quantitative evidence of measurement equivalence is recommended. As another example, if a paper-and-pencil CROM is modified to electronic format (requiring substantial modifications to the formatting) and both the paper-and-pencil and electronic administrations are used within the same study, both qualitative and quantitative evidence is recommended to support equivalence prior to study

[18] Measurement equivalence ensures that scores on latent variables can be meaningfully compared across different groups of respondents (scores mean the same; no bias). It requires that the latent variable is related to the observable responses in the same way for different groups of respondents. This includes invariance of item discrimination and absence of additive bias. While full invariance of all items in a CROM is desirable, partial invariance with a subset of invariant items is sufficient to carry out mean comparisons. Then bias in some items and/or different item discrimination is corrected statistically.

[19] As previously noted, the burden to justify any CROM modification falls on the researcher, and when in doubt about the potential impact of a modification on end-users' interpretation and response to the CROM, the more conservative approach is to gather evidence to support the modification. If applicable, it can also be important to consult the regulatory body receiving the data to understand their views of what is required.

implementation. Alternatively, leveraging this same example, if modifications to the paper-and-pencil CROM during migration were very minor and only electronic administration was used within the study, additional evidence is likely not necessary to support the modification.

3.4.1 Linguistic/Cultural Adaptations

Further discussion about one type of modification, linguistic/cultural adaptation, is warranted due to the unique recommendations around establishing evidence to support this type of modification. While an overview of the recommended process is provided here, interested readers are referred to other source documents detailing the processes for ensuring linguistic and cultural equivalence for outcome measures (see [11-16]). It is generally recommended that the researcher work closely in collaboration with an expert or organization specialized in linguistic services to determine and execute the most appropriate linguistic and cultural validation strategy for developing or modifying an existing CROM. Additionally, as with other types of CROM modifications, the researcher should determine whether there are any restrictions related to existing translated versions of the CROM, access, and licensing from copyright owner.

With this type of CROM modification, it is important that the CROM measures the same concepts in a comparable way across different languages and cultures and that data from multiple languages and countries can be compared, if necessary (i.e., the CROM are conceptually equivalent). The assessment of linguistic and ***cross-cultural equivalence*** generally requires qualitative, and in some cases, quantitative, evidence. Typically, the process begins with qualitative investigations to help increase the likelihood that the content is conceptually equivalent, and quantitative strategies are used after the measure has been translated to compare data from the two (original CROM and translated CROM) versions.

Linguistic validation is a comprehensive translation process to ensure that translated CROMs are as linguistically, culturally, and conceptually equivalent to their original version as possible. The process typically consists of all or some of the following phases to increase the likelihood of conceptual equivalence. First, a document with a clear explanation of the different items/concepts present in the CROM and translation tips is developed to help to avoid any ambiguities and misinterpretation of the items/concepts during the translation process and to ensure harmonization of translation process across different languages If a ***Translatability Assessment*** was conducted during the development stage of a new CROM (see Chapter 4), this assessment can be used to support this phase. Next, the translation phase involves forward translation (translation from source language to the target language) and back translation (target-language translation translated back into the source language) by professional translators or individuals who are fluent or native speakers of the target and source languages. Additional consolidation and reviews by the translators would follow, a SME and speaker of the target language or developer of the CROM may also be involved to review the translations and give additional recommendations.

An additional cognitive debriefing interview phase with speakers of the target language who also represent the target population (e.g., people who smoke cigarettes) could also be conducted at this time. The purpose of these cognitive interviews is to check the cognitive equivalence, comprehensibility, interpretation, and cultural relevance of the translation in the target language. It also provides an opportunity to test any translation alternatives that may not have been resolved by the translators during the forward and backward translation process, to highlight any items that may be conceptually inappropriate or may cause respondents to misunderstand or misinterpret items in the target languages. Respondents included during this process should ideally represent a cross-section within the target population, for example, both men and women, adults of different ages and people from different socioeconomic groups. Revisions to the modified CROM may be implemented based on learnings from these interviews.

Next, all input is consolidated and the translated CROM is finalized. A certificate of translation is typically produced as certified documentation of the translated version, and a final report may also be included to provide a description of all translation and cultural adaptation processes and decisions.

While qualitative evidence to support the linguistic/cultural adaptation of Psychometric CROM is always recommended, the type and extent of the evidence needed to support such a modification may be dependent on several factors. For example, if the CROM is to be used as a primary study objective, to address a regulatory requirement, or if an objective of the study is to make cross-cultural comparisons, a thorough linguistic validation process (including cognitive debriefing interviews) may be desired. The complexity of the construct being measured by the CROM and the complexity/simplicity of the CROM language is also relevant. Another factor influencing the extent of the evidence needed to support a linguistic/cultural adaptation is the closeness of the source and target language/culture; for instance, the English language in the UK and English in Ireland could be considered more linguistically similar than English in the UK when compared to Spanish in Spain. Only a minor language adaptation and review would likely be needed to update the CROM content from English in the UK to English in Ireland, whereas a linguistic validation process may be preferred to translate the CROM from English in the UK to Spanish in Spain.

While the qualitative approaches described above can help ensure conceptual equivalence between the original and translated CROM, the researcher may also decide to leverage more robust, quantitative approaches to establish measurement equivalence (also known as measurement invariance). Achieving invariance provides evidence that scores from the CROM are comparable, allowing the researcher to directly compare scores between the original source CROM and a translated CROMs, different translated CROMs languages, and use of same language CROM in different countries. This can be evaluated using different statistical and psychometric analyses (e.g., multi-group factor analysis, tests for differential item functioning in modern test theory). See Chapter 5 for an overview of quantitative methods to evaluate measurement equivalence and psychometric properties.

Without quantitative assessment of invariance, the researcher is not able to determine whether differences in means between the CROM reflect true differences in the construct being measured or biases in item scores. While following the translation procedure described above will help reduce the likelihood of biased measurement, it does not necessarily guarantee comparability without statistical evaluation (and corrections, if applicable) for bias. If quantitative testing reveals substantial non-invariance, it is the responsibility of the researcher to develop practical solutions to resolve issues and strengthen measurement equivalence and overall cross-cultural equivalence across different translated versions of a CROM. Additional qualitative inquiries and quantitative sensitivity analysis may be required to determine the sources of differences, assess if there are any potential explanations for the differences, and understand the impact of the differences on interpretation of the CROM data.

4. DEVELOPING AND VALIDATING A NEW PSYCHOMETRIC CROM

TNPs are undergoing substantial evolution, and the TNP regulatory and research landscape is also fast developing. Accordingly, there will be circumstances in which a measure that fits the researcher's objectives does not exist. If an existing Psychometric CROM does not satisfy the researcher's needs and modifying an existing CROM would not be sufficient, it may be necessary to develop a new Psychometric CROM[20]. This section of the guidelines provides an overview of the general stages of Psychometric CROM development, including best practices for executing each stage. We recognize that the specifics of a CROM's development and validation will (and should) be idiosyncratic. That is, the specific process that a researcher chooses to follow to develop a measure would depend on the construct to be measured and the qualities of the Psychometric CROM that are most important (see Chapter 2). The Psychometric CROM development and validation process is often iterative, and a researcher may decide to repeat or skip certain steps. That said, the researcher is always recommended to work in close collaboration with a measurement expert to determine the most appropriate development and validation strategy for their CROM.

As described in Chapter 2, prior to making the determination that a new Psychometric CROM needs to be developed, the researcher should first identify and carefully elaborate the target construct and consider the CROM characteristics that are of greatest importance in the context of the research. As part of this process, the researcher would have, for example, defined the concept to be measured, defined the scientific and regulatory need that it is intended to fill (if applicable), determined the context of use, defined the end-users of the CROM, identified the psychometric properties of greatest importance, and searched the literature for existing measures of relevance. Engaging in the exercise described in Chapter 2 is a prerequisite to the CROM development process described in this section.

4.1 Conceptual Model Development

While completing the exercise described in Chapter 2, the researcher would have already started the conceptual model development process. That is, when defining the construct to be measured and articulating how the CROM would be used to address a particular regulatory need or research question, the researcher would have considered content of the CROM, including identifying various components of the construct that need to be represented in the CROM.

This chapter begins by introducing the reader to a conceptual model through hypothetical examples. Next, the reader is provided with a high-level overview of various strategies that could be used to develop a conceptual model. These sections only apply to multi-item CROM, as a single-item Psychometric CROM would not require a conceptual model.

[20] This determination would be made based on the outcome of the exercise described in Chapter 2, i.e., defining the construct to be measured and identifying the qualities of the Psychometric CROM that are of upmost importance to achieve the study objective. This is a critical step prior to engaging in CROM development, as results from this exercise will help guide the development process. For example, if establishing predictive validity of a new CROM is of critical importance, then the development process would likely include a longitudinal component as part of the quantitative validation to prospectively assess the CROM's association with future outcomes.

4.1.1 General Principles

Building upon the definition of a construct, a conceptual model, generally depicted in the form of a figure or diagram, presents the key content/components to be measured by the CROM, as well as the theoretical structure of the concept of interest. For example, a conceptual model of dependence might include craving, withdrawal, tolerance, and perceived loss of control. Having a conceptual model can help to ensure adequate construct representation, that is, the content of the new CROM adequately and comprehensively reflects all critical parts of the construct (in the aforementioned example, this would mean that the new dependence CROM includes items pertaining to craving, withdrawal, tolerance, and perceived loss of control). A conceptual model should also include information about the theorized structure, which should be empirically evaluated during later stages of the CROM development process. If confirmed quantitatively, this structure will inform scoring and interpretation.

If the new CROM's content covers all components from the conceptual model (craving, withdrawal, tolerance, and perceived loss of control, from the example of a dependence CROM above), this can help to provide evidence of adequate construct representation and content validity of the new CROM. Conversely, if the new CROM's content does not cover all components from the conceptual model (e.g., does not include one or more questions pertaining to withdrawal, which was identified by the researcher as a fundamental feature of dependence), the CROM may have construct underrepresentation which may threaten the CROM's validity. That is, the CROM measures part of the construct, but not the whole construct, keeping it from being regarded as an adequate measure of the construct.

4.1.2 Methods To Develop A Conceptual Model

Qualitative, quantitative, and mixed-methods approaches can be used to develop a conceptual model. For example, a researcher might choose to use one or more of the following: surveys, individual interviews with SMEs or individuals representing the target population, focus groups, a card sorting task, social media analysis, literature review, etc. The purpose of these approaches is to identify and gain in-depth information about relevant aspects, domains, and facets of the concept(s) of interest, such as the experience of dependence among people who use TNPs. Regulatory guidance documents may also inform the development of the conceptual model, if applicable.

For illustration, two hypothetical examples of approaches to developing a conceptual model are provided below. These are intended to show how different approaches can be used in combination, and that there is no "right" way to develop a conceptual model that fits all circumstances.

The approach that the researcher chooses to pursue will likely depend on various factors, including the construct and complexity of the construct to be measured, extent of peer-reviewed literature published on the topic, which could be leveraged to inform conceptual model development, the regulatory need that the CROM is being used to address (if applicable), etc.

> ***Hypothetical Example 1.*** A researcher has determined that no existing CROM is appropriate to assess respiratory symptoms in people who smoke but who have not been diagnosed with clinical pulmonary disease and has, therefore, decided to develop a new CROM. To develop a conceptual model, the researcher begins by reviewing relevant literature, PRO measures relevant to pulmonary disease in published literature and national/international surveys and consulting several SMEs (pulmonologists and researchers working in the area) to identify respiratory symptoms that are likely experienced by people who smoke but who have not yet developed pulmonary disease, and to understand how respiratory symptoms change over time following smoking cessation.

Results from the aforementioned approaches lead the researcher to draft a conceptual model that includes content related to morning cough with phlegm, cough throughout the day, shortness of breath interfering with normal (non-strenuous) daily activities, becoming easily winded during normal daily activities, and wheezing during normal daily activities. Notably, results from these approaches also informed the researcher as to what should *not* be part of the conceptual model. For example, general symptoms that may be related to respiratory disease but are also common to many other disorders (e.g., fatigue, difficulty sleeping) were not included because they do not discriminate well between pulmonary disease and other conditions. Similarly, respiratory symptoms that reflect more severe forms of respiratory diseases were not included (e.g., difficulty getting out of bed because of respiratory symptoms), as they are less relevant for individuals without pulmonary disease (these items are too severe to be experienced by a population with mild respiratory symptoms, and therefore these items would not help with measurement precision).

Hypothetical Example 2. A researcher is looking for a CROM to assess withdrawal symptoms among people who use ENDS exclusively. An initial review of the relevant literature reveals that while a range of CROM to assess tobacco/nicotine withdrawal have been developed [17], these existing CROM are not appropriate to fit the researcher's needs (e.g., most refer to the use of cigarettes and a population of people who smoke as opposed to those who use ENDS, or they were developed specifically as diagnostic assessments of "Tobacco Use Disorder" or "Nicotine Dependence").

It is probable that withdrawal symptoms among people who use ENDS exclusively may entail signs and symptoms beyond those of tobacco and nicotine withdrawal generally described in existing literature [18]. In addition to a thorough review of existing literature and consultation with relevant SMEs, a researcher may decide to conduct concept elicitation individual interviews or focus groups and social media analysis to identify and confirm experiences of withdrawal symptoms in people who smoke cigarettes exclusively compared to the target population of those who use ENDS exclusively. As a result, the researcher may note that while some signs and symptoms of withdrawal are similar between the two groups, other aspects appear to be specific and unique to the target population of people who use ENDS exclusively. With these findings, the researcher is now able to generate a draft conceptual model representing a broader conceptualization of the subjective, physiological, or behavioral indices of withdrawal symptoms specifically related to exclusive ENDS use.

It is recommended that the researcher consult an expert in measure development to determine the most appropriate approach to developing a conceptual model for their Psychometric CROM. Interested readers are referred to other source documents for additional information [19-21]. For an example of conceptual model development in the TNP space, see [22].

4.2 Item Generation and CROM Drafting

Drafting the CROM includes developing any instructions, items, and response options (referred to here as "CROM components"). Item generation should aim to develop an adequate range of items to cover the breadth of content within the concepts of interest defined in the draft conceptual model. If prior qualitative research was conducted with the target population, items can be constructed using as many of the respondents' own words and descriptions of the concepts of interest as possible and appropriate to strengthen relevance and content validity of the CROM. At this stage, the researcher should also begin to consider the intended mode and method of administration. For instance, if the researcher intends to allow the new CROM to be administered on the participant's preferred device (i.e., permitting all screen sizes), it is recommended that the researcher take screen size into consideration when drafting the CROM to facilitate maintenance of integrity of the CROM across screen sizes[21]. Some response formats (e.g., response grids) and tasks (e.g., drag-and-drop) are not easily administered on a small screen. Challenges may also arise with items that have many response categories (e.g., 10 or more), and instances where multiple items need to be administered on the same screen. In brief, considering plans for CROM administration early in the CROM drafting process can help prevent having to modify the CROM at a later date.

Using a table or tracking matrix (often referred to as an ***item tracking matrix***, or ITM) is recommended to document CROM component sourcing (if applicable), and to track revisions, additions, and removal of items and rationale for changes as the researcher moves through the remaining phases of the CROM development process.

At this stage, a researcher may also wish to start proactively preparing and planning for linguistic/cultural translation as part of the development process by conducting a TA as part of the item generation process. The TA is conducted preferably during the development stage prior to the use of the CROM use in order to determine the CROM contents' suitability for future translations The goal of TA is to facilitate future translations and use of the CROM in global studies by 1) identifying and categorizing potential translation issues in the source text (e.g., potential difficulties to translate idiomatic expressions or colloquialism) and 2) providing alternative wordings on which translations can be based and/or recommendations of how to modify the source text so that future translations are conceptually and culturally appropriate for the target populations [12]. The researcher may also consider conducting cognitive testing and evaluation of psychometric properties across all target populations (all relevant languages/cultures).

[21] In general, participants should be given the opportunity to complete the CROM on their preferred device. Today, fewer participants complete CROMs on desktop or laptop computers but use tablets and mobile phones instead with the latter potentially becoming the most prevalent device depending on the target population, sampling and data collection strategy. It may be advisable to design CROMs specifically for mobile phones as the smallest common denominator. This means avoiding long item text, large grids or presenting many items at once.

General best practices when drafting a new CROM include the following[22]:

Global recommendations

- Having an initial pool of items with content that adequately covers the conceptual model[23]. It can be useful to start with multiple items for each element of the model, to be winnowed down later as needed.
- Use simple language (be cognizant of reading level[24] [25]) and avoid technical terminology, slang, idiomatic expressions, or colloquialisms (if possible)
- Use direct, unambiguous language
- Avoid leading questions and biasing language
- Use of images can be helpful to aid comprehension/reduce confusion

With respect to CROM instructions/item content

- Each item should communicate a single concept (e.g., not asking about severity of cough and shortness of breath within the same item, or asking about severity and duration in the same item)
- Avoid hypothetical questions, especially "double hypotheticals" (e.g., asking a person who does not use tobacco if they would adopt a TNP and then switch to a more harmful product, such as cigarettes)
- Recall period should be relevant and appropriate

With respect to response options

- Response option labels should relate to the construct being measured (e.g., severity, frequency, agreement, etc.)
- Response options should cover the full range of potential responses; avoid building in assumptions about the minimum or maximum level of the construct
- Consider whether response options would result in ceiling/floor effects, keeping the target population in mind
- Response categories should be distinguishable (e.g., participants can articulate the differences between them)[26] and, if ordered, ordering is perceived as intended
- Avoid response option labels that may bias the direction of the responses
- Bipolar scales (i.e., rating scales with a continuum between two opposing end points) should generally be symmetrical
- Use of "not applicable" should be avoided when possible (items should be applicable for participants, and skip patterns can be used to avoid administering items to participants for whom they are truly not applicable)

---

[22] These recommendations pertain to Psychometric CROM broadly. FDA and others have provided recommendations specific to certain categories of CROM (e.g., comprehension, risk perception, behavioral intentions) [23] [2]

[23] This speaks to construct representation, discussed in the earlier section Conceptual Model Development.

[24] The researcher can assess reading level (Flesch-Kincaid grade level) using a feature in Microsoft Word or other program.

[25] FDA TPPIS Guidance (2022) recommends that the reading level be "appropriate for those with less than a high school education" (p. 14).

[26] Later in the development process it is recommended that researchers collect qualitative and/or quantitative evidence of this.

- If appropriate, use similar response option scales across items (at a minimum, maintain polarity of response options between questions) to minimize measurement error due to respondent confusion, and to facilitate combining items into a composite score (if applicable)
- "I don't know" (or other similar response options) should be visually distinct from the other response options, and should be placed last in the response set

Of note, these recommendations pertain to the drafting of a single CROM (e.g., a risk perception CROM); additional recommendations and considerations when developing a survey (combining multiple CROMs) for a research study (as are often used in TPPIS) can be found in Chapter 5.

4.3 A Note about CROM Content, Length, and Measurement Precision

When drafting content of a CROM, ***measurement precision*** is an important aspect to be considered[27]. Measurement precision (i.e., ability to discriminate between participants with similar levels of the construct being measured) is largely a function of the number of items and their relative position to participants[28], although other item properties, such as the number of response options or item discrimination, and participant properties, such as random/careless responding or acquiescent responding ("yea-saying" or "nay-saying"), are relevant, too. That is, generally speaking, longer CROMs with more granular response options *can*[29] have higher measurement precision than shorter CROMs. While this potential benefit may make longer CROMs seem preferrable, shorter CROMs help alleviate response burden and can increase acceptability on the part of respondents, which in turn may increase attentive response behavior. Thus, rather than striving for maximization of precision with a lengthy CROM, measurement precision can be optimized by using a shorter CROM that is well targeted to the population of interest; such a CROM discriminates better between participants and allows for more precise measurement than a longer CROM that poorly matches the target population. Reliability also tends to be higher for well-targeted CROM. In other words, longer scales can sometimes compensate for weaker items, but that can be a very unfavorable trade-off.

With good item-selection in mind, at this stage of the measurement development process, the researcher may consider including *more* items in the draft CROM than less, even intentionally including items with very similar content which seemingly measure the same aspect of the construct. It is easy to remove items at later stages of the CROM development process based on learnings from the quantitative evaluation stage, and more difficult to add items, as new items may need to be cognitively tested.

---

[27] Having adequate content representation, which speaks to the content validity of a CROM, is always recommended. That is, the CROM should have items that cover all relevant content, as defined in the conceptual model. Once the researcher has items that cover all aspects of the construct of interest, measurement precision, which is influenced by the number of items, the number of response options, and the targeting of the items' "difficulty" to the target population, should be considered.

[28] The role of the number of items is well-recognized regardless of the measurement model applied, while the full appreciation of the impact of targeting (how well the items' "difficulty" matches the target population) is largely confined to modern test theory applications. Items occupying similar positions on the construct to be measured as the majority of the respondents provide more information to be used in the estimation of participant measures.

In traditional test theory, strong floor or ceiling effects (i.e., many participants have extreme responses) is an indication of poor matching of the items and the population.

[29] This is not always the case. First, overly granular response options can also lead to poor measurement precision if participants are not able to adequately distinguish between response categories. Second, longer CROMs may not lead to increased measurement precision if the items are at the same location (have the same item difficulty), do not adequately target the sample, are irrelevant to the construct being measured, or fatigue respondents to the point of undermining attentiveness, etc.

4.4 Refine the Draft CROM through Cognitive Testing

4.4.1 General Principles

For new Psychometric CROM, conducting individual cognitive debriefing interviews with participants who represent target population of CROM (e.g., adults who smoke cigarettes) is recommended. Participants in such interviews should include individuals who represent the full range of the target population (e.g., people who smoke currently and people who never smoked), being sure to include socio-demographically diverse participants and those with limited health literacy, as appropriate. In brief, cognitive interviews allow the researcher the opportunity to refine the CROM based on participant feedback – reducing measurement error and/or bias – before using the CROM.

Through cognitive testing, the researcher can determine whether:

- components of the CROM (e.g., instructions, item stem, items, response options) are understood and interpreted as intended
- content and recall period are appropriate
- response categories are perceived as meaningfully different and appropriately granular
- items are perceived as applicable/relevant and not redundant
- whether any important content is missing (if applicable)

While qualitative feedback from cognitive testing often provides the researcher with direction on how to modify the CROM to enhance content validity, it can also provide clarification and context regarding *why* participants respond to items in a particular way[30]. This can help the researcher interpret their findings if similar response patterns are observed in quantitative studies using the CROM in the future.

4.4.2 Measurement Challenges Addressed with Cognitive Testing

There are specific measurement challenges often faced in TNP research; researchers should attempt to mitigate these measurement challenges to the extent possible[31] through appropriate CROM development, and cognitive testing can be an effective tool to address many of these challenges. Examples of such challenges related to Psychometric CROM include:

- the potential for ***social desirability*** to bias responses if participants perceive that a particular response is socially acceptable, or conversely, that a particular response is socially disapproved (e.g., participants may feel that it is not acceptable to express no intention to quit smoking); in this context response options extending beyond the range of actual responses may prevent participants from considering their suitable response as

[30] As a hypothetical example: a researcher notices that some people who smoke report lower intention to try a smokeless tobacco product after being exposed to an MRTP claim. Upon probing, the researcher learns that these people who smoke disbelieved the claim, making them more skeptical than before about the health effects of the product (subsequently reducing their intention to try the product). This qualitative finding leads the researcher to include a CROM of claim believability in the research study to account for this phenomenon.

[31] We acknowledge that several of these measurement challenges cannot and will not be eliminated through the use of cognitive testing or other measure development strategies. For example, as articulated by FDA in their TPPIS Guidance (2022), "participants may have limited ability to forecast their future patterns of use behavior without having tried the product." (p. 12). Similarly, response biases such as social desirability will exist to some extent despite appropriate CROM development processes. That said, cognitive testing and other qualitative research strategies can help the researcher better understand and potentially begin to address some of these challenges.

too extreme and socially unacceptable (e.g., when asking about how many cigarettes are smoked per day, the response categories should not top out at, say, "20 or more" as that category might then appear extreme and dissuade respondents from endorsing it even though it applies objectively)[32]

- using the most appropriate/current language for various products and behaviors (e.g., "e-cigarettes" or "ENDS" vs. "e-vapor")
- minimizing confusion of terminology (e.g., participants confusing "smoking" to refer to ENDS use; confusing different product categories)
- being appropriately specific when asking about heterogeneous categories of behaviors such as "dual use" so that participants are interpreting these items uniformly
- assessing participants' perception of products that they may not be familiar with/have no experience using (e.g., asking people who do not use tobacco to rate their perception of the specific health effects of a product); this may include products that are not yet on the market (premarket TNPs) or are new to the market (e.g., heated tobacco products); appropriate product descriptions must convey an accurate understanding of the product even for participants with low-literacy.

4.4.3 Methodological Considerations for the Conduct of Cognitive Testing

Here we summarize general best practices for cognitive testing when used to develop and refine CROM for use in TNP research. Many books, articles, and guidance documents exist to provide interested readers with greater detail regarding the conduct of these interviews and analysis of cognitive interviewing data (see [25-27]).

In general, cognitive testing should be conducted individually with persons representing the target population for which the CROM will be administered. The sample should be appropriately diverse with respect to potentially relevant variables (e.g., TNP use, demographics, such as age, sex, and race, etc.). For instance, if the CROM is being developed for use with adults who use TNPs currently and adults who do not use TNPs, then cognitive testing of the new CROM should include adults who represent both of these groups. If there is reason to believe that the CROM could function differently across different populations (e.g., individuals who do not use TNPs currently but who used previously vs. individuals who never used TNPs, individuals with limited health literacy vs. individuals with adequate health literacy), these individuals should also be represented in the sample. Cognitive interviews should be conducted until the point of saturation has been reached[33], which is defined as the point at which additional interviews seem unlikely to yield new or useful information [4]. This is typically determined by testing several participants at a time (e.g., 6-8) and tracking participant feedback using an informal saturation tracking table to track themes in feedback as well as the emergence of new relevant feedback (saturation). Although the number of cognitive interviews needed to reach saturation depends on various factors, such as the heterogeneity of the end-users, and cannot be determined *a priori*, it is not uncommon to reach saturation after approximately 25-30 participants. Ideally, interviews would be conducted in waves, so that modifications made to the CROM can be tested through additional interviews. All modifications, including reasons for modifications, should be documented using the ITM.

[32] The impact of response scales on the response options chosen goes beyond social desirability. There is some indication that respondents assume that the options offered reflect the known or assumed distribution by the researcher in the population, which is then taken into account when forming a judgement [24].

[33] This recommendation is consistent with guidance documents in the PRO space, e.g., ISPOR guidance ([28]).

Cognitive interviews can be conducted in-person or virtually. If conducted virtually, it is ideal for the interviewer to observe the participant complete the CROM using video (to watch the participant's reaction) and screenshare (to watch them answer the questions) features simultaneously to allow for behavioral observations (e.g., changing an answer multiple times, taking an unusual amount of time to answer a question, clicking back and forth in the CROM, facial expressions suggesting negative or positive feelings, or confusion), if feasible. Interviewers should be experienced with cognitive interviewing techniques and should follow a semi-structured interviewing guide, which allows for deviation from the guide to fully understand the participant's experience with the CROM and to identify opportunities for improvement. Although there are several different approaches to ***probing***, it is generally recommended that at least during final waves of testing that retrospective probing (probing retrospectively, *after* the participant has completed the CROM; see [25, 27]) be used to enhance realism and generalizability of findings. Similarly, mode and method of survey administration should mimic the intended mode and method by which the CROM will be administered in the study.

4.5 Quantitative Methods to Evaluate Key Psychometric Properties

After cognitive testing, the researcher should conduct a quantitative study to gather information about relevant psychometric properties of the new CROM. As previously indicated, the exercise described in Chapter 2, which includes identifying the psychometric properties of greatest importance, will function to drive the design and analysis plan of the quantitative study. For example, if test-retest reliability, ability to detect change over time, or predictive validity are of great importance, then the quantitative study may be a prospective longitudinal study where the CROM is administered multiple times.[34] In addition to a foundational CROM validation study, other sources of quantitative data stemming from the use of the CROM in observation studies, post-market surveys, clinical studies etc. can also be used to generate information about the CROM's psychometric functioning.

The sample for a quantitative psychometric evaluation typically includes individuals representing the target population in which the CROM will be administered, in order to demonstrate reliability and validity of the CROM when used with this population. However, psychometric studies may also include other groups of individuals for purposes of establishing key psychometric properties (e.g., known-group validity). Further, in contrast to other quantitative studies (e.g., TNP use prevalence studies) where the purpose is to generate a population-level estimate, the researcher typically need not aim to have a sample that is *representative* of the population as a whole (e.g., imposing demographic target quotas to reflect the population of adults who smoke cigarettes in the US), but instead may target specific populations for purposes of facilitating the psychometric evaluation (e.g., imposing a soft-target for the minimum number of women who complete the study to permit evaluation of item functioning across gender).

While it is beyond the scope of this document to provide recommendations regarding the appropriate conduct of a psychometric evaluation and the psychometric analyses most appropriate to evaluate key psychometric properties, some examples of psychometric properties evaluation methods include Classical Test Theory, Item Response Theory, and ***Rasch Measurement Model,*** and interested researchers may consult other publications and documents on the topic (e.g., [4]). Recent publications presenting the development and validation of

[34] As previously articulated in earlier sections of these guidelines, the number of administrations and length between assessment periods would depend on the construct being measured and anticipated variability in the construct over time.

CROM in the TNP space may also be helpful reference documents [29-33]. It is always recommended that researchers work with appropriately qualified experts in designing and executing quantitative psychometric evaluations.

Output from the psychometric evaluation generally includes a report detailing the psychometric analyses conducted and the results, as well as the following: the final CROM with administration instructions[35], the completed ITM with final CROM components[36], the quantitatively verified conceptual model, and empirically-based recommendations for scoring and interpretation. Portions of this output can reside in a User's Guide, which is essentially a brief document summarizing recommendations for CROM administration, scoring, and interpretation intended to facilitate appropriate implementation of the CROM in future studies. The User's Guide should include a copy of the CROM and instructions for administration, including any important programming notes for electronic administration (i.e., instructions if each item should be administered on a separate screen with the instructions repeated on the top of each screen, whether a "back" button that allows participants to change their responses to items is permitted, inclusion of skip logic, whether responses are "forced" [or items can be skipped/left blank], etc.). Additionally, the User's Guide should indicate the intended users of the CROM (e.g., adults who smoke cigarettes), recommendations for mode and method of administration, handling of missing data (if applicable), scoring, and interpretation. Importantly, all information in the User's Guide should be science-based, derived from the CROM development and validation studies. Finally, if the CROM is intended to be distributed externally, the User's Guide should include any relevant licensing information, including fees and permission for use, as well as all languages that the CROM is available in, or the process required for new translations (see Section 3.4.1).

[35] If the CROM is intended to be administered electronically, a screenshot of the CROM from the electronic survey should be provided to facilitate maintenance of integrity with respect to formatting and CROM presentation in future studies. If the CROM was administered in a "bring-your-own-device" format (on the participant's preferred device) and the CROM appeared differently on the different screen sizes (e.g. mobile phones, tablets), including screenshots from each device is recommended. Translated screenshots should also ideally be checked against English source screenshots and available paper translation to ensure all screenshots are confirmed as accurately displaying and reflecting what is intended.

[36] Modifications to the CROM may occur during the quantitative psychometric evaluation process; documenting any CROM changes using the ITM (e.g., removing an item) along with a rationale for the modification is recommended.

5. APPLICATION, IMPLEMENTATION, AND INTERPRETATION OF A PSYCHOMETRIC CROM

The ultimate purpose of a CROM is as a measurement tool in a research study. In this section, we describe what to consider when implementing an existing, adapted, or newly developed CROM and when interpreting the resulting measurements. In this chapter, unless otherwise specified, the term *study* refers to the study in which the CROM is applied (application study), not the study/studies conducted to develop it (validation study).

Before implementing a CROM in a study, the exercise described in Chapter 2 is strongly recommended. By clearly defining the construct to be measured and carefully considering the CROM characteristics that are of greatest importance for purposes of the study, this exercise facilitates identification of an appropriate existing CROM (if one exists), guides modifications (or psychometric testing) that need to be made to an existing CROM to make it an appropriate fit for the study (see Chapter 3), or guides the CROM development process (see Chapter 4). Therefore, if the reader skipped over these earlier sections of the guidelines, they should go back and read these chapters before proceeding.

5.1 Application of a CROM

Previous sections of these guidelines provide recommendations intended to assist the researcher in identifying a suitable CROM. For example, a suitable CROM must adequately measure what is intended to be measured in the study, and it should be validated for a context of use (e.g., for use with particular TNPs, for use with specific populations, for a specific mode/method of administration, etc.) that is consistent with the researcher's study. Readers are referred to Chapter 2 for further discussion of determining CROM applicability for a particular study.

5.2 Comments Regarding the Sequence of the CROM Validation and Application Studies

For obvious reasons, any CROM must exist prior to its implementation in a study regardless of whether it has been developed specifically for this purpose, adapted, or taken from the body of existing CROMs. Ideally, the CROM has been validated in a separate study focusing on, and confirming, its psychometric properties. In practice, time and financial constraints may not always allow for a separate validation study. Rather, the data collected in the application study itself is used to assess the CROM's psychometric properties. This approach is suboptimal and generally discouraged unless the adaptations to a previously validated CROM are minimal. The reasons for this are varied and also interrelated. First, validation studies may require different sampling designs. For example, when developing a CROM that measures dependence on cigarettes, a proper validation sample would include a wide range of participants ranging from people who are only slightly dependent to those who are heavily dependent. In contrast, the application study may focus on a particular population of interest (i.e., people who are moderately-to-heavily dependent), which would preclude proper evaluation and calibration of items targeted to assess light dependence. Second, the validation study might ideally include additional data (e.g., administering other CROM alongside the new CROM for purposes of assessing convergent or ***discriminant validity***) that would not be collected as part of the application study. Third, the assessment of psychometric properties may support modifications to the CROM (e.g., reducing the number of items, modifying a response scale, assigning items to other dimensions). Even if the data do not suggest the need for modifications, using the same data to evaluate the CROM and to estimate measurements of participants runs the risk of capitalizing on chance and overfitting a CROM to a specific sample. As a consequence, the

study findings might not be generalizable. Finally, there is always the chance that validation of the CROM reveals fundamental problems, which would require substantial changes to the CROM and collection of new data.

In practice, the ideal is not always achievable as funds and time are limited. Researchers must be pragmatic and find a suitable and defendable spot on the continuum from best practice with minimal limitations to acceptable approaches with some qualifications while avoiding poor practice with considerable, if not fatal, limitations. If a combined validation and application study is the only possible option, it should include all variables necessary for the validation of the CROM, while its design should be optimized by considering sampling requirements (size and structure of the sample) for both the validation and the application study objectives. This may involve inclusion of other measures that are useful for psychometric evaluation (e.g., an existing measure of a similar construct to assess ***convergent validity***, or a measure of a very different construct to assess discriminant validity), even if they are not needed for the study's primary objectives. A large enough sample size allows for setting aside a subsample to be used for quantitative assessment of the validity of the CROM and calibration of its parameters, while the remaining participants (application sample) are used in the analyses addressing the objectives of the application study. Such an approach would mimic a sequential process of CROM validation and CROM application but managing on a single study and data collection. Complementary analyses of the application sample could check whether any changes to the instrument suggested by the validation sample can be replicated in terms of a cross-validation.

5.3 Implementation of a CROM

The mode of administration of the CROM (e.g., online, offline using an electronic device, paper-and-pencil administered, self-completed versus interviewer-administered) and the timing (e.g., single versus repeated measurement) is largely determined by the study objectives and the role of the construct measured by the CROM. Once again, consistency with the intended use of the CROM is important. If multiple modes of administration are implemented (e.g., online data collection for the general population, mail-administered data collection using the CROM on paper for participants without online access), instructions in the CROM manual and empirical evidence in the literature should be sought. If such guidelines are not available and a separate validation study is not feasible, as much empirical evidence as possible that supports the validity of measurements and their comparability across different modes, based on data from the study itself, should be provided. As a starting point, the visual appearance of the CROM on the screen and on paper should match as closely as possible (e.g., number of items presented at once). Most online data collection tools record the type of device used by the participant, allowing for empirical checks comparing responses in terms of frequency of missing values, use of extreme categories, straight-lining (repeating the same response across many items), etc. If a CROM is interviewer-administered, either face-to-face or over the telephone, consistency between interviewers and adherence to interviewer-guidelines (e.g., instruction to read all questions in full, how to probe, etc.) is crucial. In cases where the interviewer is making a judgment that shapes the data, such as coding a response to a category, the consistency or coding between interviewers becomes an important part of the psychometric performance of the CROM, and must be empirically established (e.g., by having other raters code the responses from recordings of the interview).

Related to the mode of administration is the context of administration, i.e., the setting in which data collection takes place. Administering a CROM in a clinical study shortly before or after a blood sample is drawn, may make the participants feel uncomfortable, which may compromise the validity of their responses to the CROM. Being observed while answering sensitive

questions may also impact participants' responses to a CROM, which may have been developed for conditions ensuring complete anonymity. Contexts that do not provide privacy or anonymity may be particularly prone to social desirability bias.

A frequent feature of applied studies is the administration of multiple CROMs at once, resulting in lengthy surveys. This practice raises two questions. First, does the order in which CROMs are presented impact the responses? Some CROMs may be more susceptible to order effects than others depending on what they assess. The study's data collection design should consider the possibility of order effects and the researcher will need to be prepared to defend and provide a rationale for the design implemented. One option is to implement a unique order that is assumed to have the least impact on responses. Previous studies reported in the literature may help define a reasonable design. Alternatively, different orderings may be randomly implemented, allowing for empirical analyses of order effects. However, such designs may quickly become very complex and shift the focus from the actual study objectives to methodological research questions.

Second, does the length of the data collection compromise the quality of the data? Lengthy surveys are prone to provoke response burden and fatigue, with data quality decreasing over time.[37] While it is unclear what the maximum length of a survey should be as it likely varies by target population, type of questions, mode of administration, etc., an average duration of no more than about 15 minutes is generally recommended. If the flow of the survey and possible order effects allow it, it is generally advisable to place the CROM most relevant to the study objectives at the beginning when participants may be most alert. While incentives may help improve participant retention during the survey, they do not protect against "speeding" (completing the survey very fast, responding carelessly without attending to the CROM content). Indeed, inattentive respondents are a key threat to validity of self-reported measurement and there are many approaches to determining the extent of this problem. In online data collection, time stamps usually allow for computation of the time needed to complete the survey, which would allow for the identification of "speeders." Another option is to include validity checks in the form of one or more interspersed questions that instruct the participant (hence called "instructed response items") to select a particular response (e.g., "Please select strongly disagree for this item") or skip the item ("Please skip this item to show you carefully read the questions"; [34] p. 84). Failing such items is indicative of inattentive responding. The researcher may also consider including checks for consistency [35]. For example, in a TNP research context, the reported duration of smoking in years should generally match the time span between the stated starting age of smoking and the participant's current age.[38] Finally, very unusual responses that strongly deviate from the mean response of other participants (outliers) should be scrutinized.

These strategies should not be considered a solution to the problem of response burden; they rather provide diagnostic indicators to determine to what extent successful mitigation or avoidance of adverse effects of response burden have been achieved. If inattentive response behavior occurs frequently in a study, a key question remains: Should data from participants who demonstrated careless response behavior be eliminated from the dataset? The potential advantage of cleaner data could be offset by reduced statistical power due to the decreased sample size. Data cleaning may also result in differential removal of participants representing

[37] Gamification, the use of game technology outside the context of games, is another trend in some fields to enhance the entertaining aspect of data collection and counteract response fatigue. It remains to be seen whether gamification will contribute to data quality in tobacco research also.

[38] In practice, periods of abstinence in between should be taken into account, provided these are also queried.

certain sociodemographic backgrounds or other potentially relevant subgroups (e.g., those with limited health literacy). As data deletion can be controversial, the study protocol or the SAP should prescribe in advance any procedure arranging for data editing. A sensitivity analysis can also be conducted comparing the results based on the complete and the reduced data.

5.4 Repeated Measurements

In longitudinal studies, CROM may be administered repeatedly to the same study participants. Such repeated measurements pose their own challenges. On the one hand, the item properties can change over time, on the other hand, measurement error considered to be random and uncorrelated over different timepoints can be related across time introducing artificial dependencies. Particularly in the assessment of test-retest reliability of CROM assessing trait-like constructs, where no intervention is involved, the time interval should be long enough to avoid excessive dependency of measurements (replication of the same response pattern). Any recommendation provided in this regard in the User's Guide or from existing studies should be followed, if possible. A time span of 7 to 14 days between successive administrations of most CROMs may avoid dependency over time [36, 37]. However, if measurements are to be taken on a daily basis (e.g., in a diary study), the CROM should be appropriate for such an application. Also, in case of interventions (e.g., using a product and then completing a CROM intended to assess impact on craving), the same CROM measuring a state rather than a trait might need to be administered repeatedly within a short time span. In such cases, the CROM's sensitivity to change (whether the CROM identifies true changes) is important, while dependencies over repeated administrations can still be investigated statistically.

5.5 Measurement Precision

As discussed earlier in these guidelines, measurement precision is optimized when the items' difficulty ("severity" of the construct being measured) matches the level of the construct in the participants in the study. In some cases, the researcher may not know *a priori* the extent to which the operational range of the CROM (where it provides adequately precise measurements) matches the study target population. This may most commonly be the case when a researcher is using an existing CROM, which was developed for a population that may differ on (potentially) relevant characteristics from the target population. In practice, this may not always be straightforward, since the distribution of measurements of participants is only known after the application of the CROM. Nevertheless, efforts should be undertaken to judge the expected distribution of the participant measures with sufficient accuracy. If this is difficult to achieve, the characteristics of the populations in the validation study and in the application study should correspond reasonably closely.

5.6 Interpretation of the CROM

The interpretation of scores/measurements based on a Psychometric CROM should be in line with the guidelines provided by the developers of the CROM. Instructions as to the scoring of item responses, sum score formation and, if applicable, transformation of scores to measurements can be found in manuals and/or other publications. If a CROM features multiple domains, it should be clarified whether individual domain scores are to be formed and/or whether a sum score across all domains is permissible (i.e., a total or composite score). The mCEQ [38] is an example of a multi-domain CROM consisting of 12 items that are assigned to three multi-item domains (Satisfaction, Psychological Reward, Aversion) and two single-item domains. The mCEQ provides five domain scores, whereas a composite score across domains is not justified, as it combines unrelated and even opposing constructs.

CROMs that have been developed based on the ***Modern test theory*** approach (e.g., Item response theory, Rasch measurement theory) may include tables or computerized scoring that convert simple sum scores to linear measurements. If provided and recommended for use by the CROM developer, these conversions should be applied, as, strictly speaking, only the converted measurements allow for the use of parametric statistics that require interval-scaled data. However, except for cases with many extreme or near-extreme scores, the use of raw scores is not uncommon. Missing data should be handled as recommended by the CROM developers. Should such guidelines be unavailable, suitable imputation techniques may be considered.

Deviations from recommended procedures in the implementation of a CROM constitute a modification (see Chapter 3) and require appropriate consideration in the interpretation of measurements.

5.7 Documentation

The selected CROM needs to be included and described in the study protocol and referenced in the SAP (if applicable). These documents should also provide the background and the rationale supporting the selection and implementation of the CROM (e.g., include references to User's Guide and published literature). They should stipulate how the CROM is to be implemented, how responses are to be scored, how missing values are to be handled, and how measurements are to be derived and subject to statistical analysis, as per the study endpoints and objectives. An overview can be included in the protocol, and details should be provided in the SAP. Any deviation from the User's Guide of a CROM should be considered very carefully and justified.

If the application and implementation of a well-established and validated psychometric CROM is inconsistent with the instrument's intended use, additional analyses undertaken to address the modification (see Chapter 3) should be documented.

Finally, the analysis and study report need to mention all deviations from the protocol that have occurred during data collection or analysis. If deviations in the application and implementation bear on the interpretation of measurements, potential limitations should be mentioned in the report and any study publications.

As most studies involve their peculiarities, a perfect consistency of a CROM's intended and actual application and implementation may not be achievable. However, the goal should always be the best possible implementation.

6. SUMMARY AND CONCLUSIONS

Consistent with other areas of science, behavioral science requires objective measurement that is precise, replicable, and measures what it is intended to measure. When measuring behavioral constructs relevant to TNPs, such as dependence or risk perceptions, it is critical that the researcher has reliable and valid measurement tools, i.e. CROM. The purpose of these guidelines is to provide researchers with recommendations and best practices for the use of Psychometric CROM (CROM which measure underlying attributes [latent constructs], which cannot be directly observed but are estimated by participants' responses to a set of items) in the TNP space.

As emphasized in these guidelines, CROM selection needs to be an informed decision. At first, the researcher must conceptually define the construct and explicate the role of the construct in the study (i.e., the purpose of measurement). Chapter 2 outlines how a researcher might determine whether an existing CROM would be an appropriate fit for their study by comparing the psychometric properties and context of use of existing CROM against the needs of their study (i.e., do(es) the CROM(s) really measure what the researcher wants to measure). Through this exercise, it will become apparent whether an existing CROM is sufficient, whether modifications to an existing CROM are needed, or whether a new CROM needs to be developed.

Should modifications need to be made, Chapter 3 walks the reader through the recommendations for when and how to collect evidence to support a modified Psychometric CROM, based on the type and extent of CROM modifications. As described in this chapter, unless minor, collecting evidence to support the adequacy of the modifications is generally recommended. If successful, modifications add to the body of evidence for a CROM's validity and applicability, therefore, they are a valuable contribution to the literature. If an existing CROM does not lend itself to modifications to satisfy the current study's objectives, it would then be necessary to develop a new CROM.

Chapter 4 provides recommendations for developing and validating a new Psychometric CROM, including both qualitative and quantitative approaches to support its development and validation. The chapter elaborates on general principles to support the development of a new CROM's conceptual model, items generation, cognitive testing, and evaluation of its psychometric properties.

Regardless of whether the researcher is using an existing, modified, or new CROM, caution should be taken when implementing the CROM in the study and interpreting data from the CROM.

Chapter 5 of these guidelines presents recommendations for the application of CROM, including considerations when multiple CROM are combined into a survey for purposes of the study and the interpretation of measurements (i.e., what does a particular score mean).

The recommendations in this document are grounded in scientific rationale and aim to provide an overview of foundational principles grounded in psychometrics and what may currently be considered best practices regarding the use of Psychometric CROM in research on TNPs. However, best practices may also evolve over time with advances in research on TNPs, psychometrics, and measurement science.

7. BIBLIOGRAPHY

[1] Center for Tobacco Products, *Modified Risk Tobacco Product Applications: Draft Guidance for Industry*. 2012.

[2] Center for Tobacco Products, *Tobacco Products: Principles for Designing and Conducting Tobacco Product Perception and Intention Studies*. 2022.

[3] United States Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), and C.f.B.E.a.R. (CBER), *Patient-Focused Drug Development: Collecting Comprehensive and Representative Input: Guidance for Industry, Food and Drug Administration Staff, and Other Stakeholders*. 2020.

[4] United States Food and Drug Administration, *Patient-reported outcome measures: Use in medical product development to support labeling claims*. 2009.

[5] Volkow, N.G., JA; Koob, GF., *Choosing appropriate language to reduce the stigma around mental illness and substance use disorders.* Neuropsychopharmacology, 2021. **46**(13): p. 2230-2.

[6] Williamson, T.R., KE; Carter-Harris, L; Ostroff, JS., *Changing the language of how we measure and report smoking status: implications for reducing stigma, restoring dignity, and improving the precision of scientific communication.* Nicotine and Tobacco Research, 2020. **22**(12): p. 2280-2.

[7] Strong, D.R., et al., *Indicators of dependence for different types of tobacco product users: Descriptive findings from Wave 1 (2013-2014) of the Population Assessment of Tobacco and Health (PATH) study.* Drug Alcohol Depend, 2017. **178**: p. 257-266.

[8] Kaplan, B., E. Crespi, J.J. Hardesty, and J.E. Cohen, *Assessing Electronic Nicotine Delivery Systems Device Type Accurately in Surveys.* Nicotine & Tobacco Research, 2023.

[9] Brown, T.A., *Confirmatory Factor Analysis for Applied Research*. Second ed. 2015.

[10] Hair, J.F., W.C. Black, B.J. Babin, and R.E. Anderson, *Multivariate data analysis*. 8th ed. 2019, Boston: Cengage.

[11] Acquadro, C., K. Conway, G. Christelle, and M. I, *Linguistic Validation Manual for Health Outcome Assessments*. 2012.

[12] Acquadro, C., et al., *Emerging good practices for Translatability Assessment (TA) of Patient-Reported Outcome (PRO) measures.* J Patient Rep Outcomes, 2017. **2**(1): p. 8.

[13] Anfray, C., et al., *Reflection paper on copyright, patient-reported outcome instruments and their translations.* Health Qual Life Outcomes, 2018. **16**(1): p. 224.

[14] Regnault, A. and M. Herdman, *Using quantitative methods within the Universalist model framework to explore the cross-cultural equivalence of patient-reported outcome instruments.* Qual Life Res, 2015. **24**(1): p. 115-24.

[15] Wild, D., et al., *Multinational trials-recommendations on the translations required, approaches to using the same language in different countries, and the approaches to support pooling the data: the ISPOR Patient-Reported Outcomes Translation and Linguistic Validation Good Research Practices Task Force report.* Value Health, 2009. **12**(4): p. 430-40.

[16] Wild, D., et al., *Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: report of the ISPOR Task Force for Translation and Cultural Adaptation.* Value Health, 2005. **8**(2): p. 94-104.

[17] Patten, C.A. and J.E. Martin, *Measuring tobacco withdrawal: a review of self-report questionnaires.* J Subst Abuse, 1996. **8**(1): p. 93-113.

[18] Barakat, M., et al., *The Era of E-Cigarettes: A Cross-Sectional Study of Vaping Preferences, Reasons for Use and Withdrawal Symptoms Among Current E-Cigarette Users in the United Arab Emirates.* J Community Health, 2021. **46**(5): p. 876-886.

[19] Center for Drug Evaluation and Research, *Patient-Focused Drug Development: Collecting Comprehensive and Representative Input*. 2020.

[20] Patrick, D.L., et al., *Content validity--establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part 1--eliciting concepts for a new PRO instrument.* Value Health, 2011. **14**(8): p. 967-77.

[21] Center for Drug Evaluation and Research, *Patient-Focused Drug Development: Selecting, Developing, or Modifying Fit-for-Purpose Clinical Outcome Assessments*. 2022.

[22] Afolalu, E.F., et al., *Impact of tobacco and/or nicotine products on health and functioning: a scoping review and findings from the preparatory phase of the development of a new self-report measure.* Harm Reduct J, 2021. **18**(1): p. 79.

[23] Kaufman, A.R., et al., *Measuring Cigarette Smoking Risk Perceptions.* Nicotine Tob Res, 2020. **22**(11): p. 1937-1945.

[24] Schwarz, N. and H.-J. Hippler, *Response alternatives: the impact of their choice and presentation order.* (ZUMA-Arbeitsbericht, 1990/08). Mannheim: Zentrum für Umfragen, Methoden und Analysen 1990.

[25] Beatty, P.C., *Cognitive interviewing: the use of cognitive interviews to evaluate ePRO instruments.* , in *Pro: Electronic Solutions for Patient-Reported Data* B. Byrom and B. Tiplady, Editors. 2010, Gower: United Kingdom. p. 2-48.

[26] Office of Management and Budget, *Statistical Policy Directive No. 2: Standards and Guidelines for Statistical Surveys; Addendum: Standards and Guidelines for Cognitive Interviews*. 2016.

[27] Willis, G.B., *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. 2004: Sage Publications, Inc. 352.

[28] Rothman, M., et al., *Use of existing patient-reported outcome (PRO) instruments and their modification: the ISPOR Good Research Practices for Evaluating and Documenting Content Validity for the Use of Existing Instruments and Their Modification PRO Task Force Report.* Value Health, 2009. **12**(8): p. 1075-83.

[29] Cano, S., et al., *Development and validation of a new instrument to measure perceived risks associated with the use of tobacco and nicotine-containing products.* Health Qual Life Outcomes, 2018. **16**(1): p. 192.

[30] McCaffrey, S.A., et al., *Development and validation of behavioral intention measures of an E-vapor product: intention to try, use, dual use, and switch.* Health Qual Life Outcomes, 2021. **19**(1): p. 123.

[31] Morean, M.E. and K.W. Bold, *The Modified E-Cigarette Evaluation Questionnaire: Psychometric Evaluation of an Adapted Version of the Modified Cigarette Evaluation Questionnaire for Use With Adults Who Use Electronic Nicotine Delivery Systems.* Nicotine Tob Res, 2022. **24**(9): p. 1396-1404.

[32] Morean, M.E., et al., *Development and psychometric validation of a novel measure of sensory expectancies associated with E-cigarette use.* Addict Behav, 2019. **91**: p. 208-215.

[33] Shiffman, S., et al., *A New Questionnaire to Assess Respiratory Symptoms (The Respiratory Symptom Experience Scale): Quantitative Psychometric Assessment and Validation Study.* JMIR Form Res, 2023. **7**: p. e44036.

[34] Kam, C.C.S. and G.H. Chan, *Examination of the validity of instructed response items in identifying careless respondents.* Personality and Individual Differences, 2018. **129**: p. 83-87.

[35] Bauer, U.E. and T.M. Johnson, *Editing data: what difference do consistency checks make?* Am J Epidemiol, 2000. **151**(9): p. 921-6.

[36] Deyo, R.A., P. Diehr, and D.L. Patrick, *Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation.* Control Clin Trials, 1991. **12**(4 Suppl): p. 142s-158s.

[37] Polit, D.F., *Getting serious about test-retest reliability: a critique of retest research and some recommendations.* Qual Life Res, 2014. **23**(6): p. 1713-20.

[38] Cappelleri, J.C., et al., *Confirmatory factor analyses and reliability of the modified cigarette evaluation questionnaire.* Addict Behav, 2007. **32**(5): p. 912-23.