Considering Data on a Patient-Reported Outcome Measure (PROM) For Chinese Patients with Thyroid Diseases Who Speak Mandarin in China

by

Miao (Christina) Tang

B.A., Simon Fraser University, 2016

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Arts

in the
Department of Psychology
Faculty of Arts and Social Sciences

SIMON FRASER UNIVERSITY
Spring 2022

Declaration of Committee

Name:	**Miao (Christina) Tang**
Degree:	**Master of Arts**
Title:	**Considering Data on a Patient-Reported Outcome Measure (PROM) For Chinese Patients with Thyroid Diseases Who Speak Mandarin in China**
Committee:	**Chair: Thomas Spalek** Professor, Psychology
	Rachel T. Fouladi Supervisor Associate Professor, Psychology
	David N. Cox Committee Member Associate Professor, Psychology
	Hui Xie Examiner Professor, Health Sciences

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

The current study examined a newly-translated Simplified Chinese version of the Thyroid-Specific Patient-Reported Outcome Short Form (SC ThyPRO-39) among 179 thyroid patients in Mainland China. This study investigated whether the ceiling/floor effect (CFE) is present in responses to the SC ThyPRO-39. The appropriateness of regression modelling strategies for data with and without CFE were considered for a variety of predictor sets, and models were compared among six distributional models. With different predictor sets, the effect of gender and mode of administration (electronic interview versus self-administration) were of particular interest. Results suggested the use of Negative Binomial or Zero-inflated Negative Binomial as modelling strategies to fit the data with significant floor effect. There were also gender and mode effects on the scale scores. Findings indicated that overall, the SC ThyPRO-39 can be used as a patient-reported outcome measure among thyroid patients who speak Mandarin or read Simplified Chinese in China.

Keywords: ThyPRO-39; PROM; Floor effect; Model selection; Gender; Mode of administration

Acknowledgements

It is a genuine pleasure to express my sincere gratitude to my supervisor, Dr. Rachel Fouladi, for her valuable support, understanding, encouragement, and guidance to my study and projects. Without her encouragement, immense knowledge, patience, and persistent help, this thesis project would not have been possible. I would also like to thank my secondary supervisor, Dr. David Cox, for his support and encouragement.

I would like to thank Dr. Torquil Watt and Dr. Huiling Liew for giving me the permission to examine the Simplified Chinese version of ThyPRO-39, and their encouragement to me for this project. I would also like to thank Dr. Hui Xie, for his valuable suggestions on the thesis during the defense.

It is also my pleasure to work with a wonderful team and fellow students in the Measurement and Modelling Lab. Especially, I would like to thank Henri Lu and Fereshteh Rashed for their encouragement and valuable advice.

Finally, I would like to express my thanks to my family, including my son, Luke Chen, for accompanying me throughout my Master's degree, and the support from my husband, Siwei Chen, and my mother, Jixia Lyu. I would also like to thank my mother-in-law, Qinzhen Ji, and my father-in-law, Hao Chen, for their support to my study and contributions to the patient recruitment in China.

Table of Contents

Chapter 1.

Introduction

People with chronic disease may experience different kinds of impairment on their quality of life (Megari, 2013). Health-related quality of life (HRQOL) measures can be used to evaluate patients' quality of life in various domains (Muragundi et al., 2012). Measures must be of good quality to be reflective of an individual's quality of life. The presence or absence of ceiling and/or floor effect (CFE) is one of the important measurement property considerations for HRQOL questionnaires (Terwee et al., 2007). CFE has been observed across various HRQOL measurements with different language versions (Hakimi et al., 2016; Lin et al., 2015; Rampazo-Lacativa et al., 2015). There are a variety of factors which can influence people's responses to questionnaires. These include personal characteristics as well as features of the questionnaire. Gender and mode of administration are important factors to be considered in the context of responses to HRQOL questionnaires (Boerma et al., 2016; Tourangeau and Smith, 1996). Other demographic and disease characteristics may also play a significant role in participants' responses to HRQOL questionnaires (Leak et al., 2013). Given the prevalence of CFE in HRQOL data, when addressing impacts of variables on HRQOL scores, it is important to select an appropriate modelling strategy that shows the best model fit to the data with CFE effects (e.g., floor effects) and which are not normally distributed (Šimkovic and Träuble, 2019).

Thyroid disease is one of the chronic diseases with high prevalence all over the world ("About Thyroid Disease", n.d.). The current thesis examines a newly translated Simplified Chinese version of the Thyroid-Specific Patient-Reported Outcome Short Form (SC ThyPRO-39, Liew, n.d.) which is based on the ThyPRO-39 (Watt et al., 2015). This thesis focuses on select measurement properties and analytic strategies for the analysis of data from the SC ThyPRO-39. Although CFE was not reported for the original version of the ThyPRO (Watt et al., 2014) or the ThyPRO-39 (Watt et al., 2015), floor effects have been reported in some other studies. Of particular relevance to this thesis are the findings of a floor effect in scores on the English version of the ThyPRO in Singaporean patients with Grave's disease (Liew, 2021), as well as in scores from thyroid patients in Hong Kong who were administered the Traditional Chinese version of

the ThyPRO-39 (TC ThyPRO-39, Wong et al, 2018). In combination with the literature on the possible influences on HRQOL scores, these findings highlight the necessity of consideration of whether CFE is present in scores on the newly developed SC ThyPRO-39, as well as the necessity of identifying optimal modelling strategies for studies of influences (e.g., gender or mode of administration effects) on these data.

The present thesis considering data on the SC ThyPRO-39 questionnaire begins with the current introductory chapter. This chapter begins by introducing the thyroid diseases, HRQOL measurements, and an overview of the ThyPRO and ThyPRO-39. This is followed with a summary of CFE among HRQOL measurements and modelling strategies which deal with CFE. In addition, there are introductions to significant factors that may influence questionnaire responses, including gender, mode of administration, and other demographic and disease-specific variables. This chapter concludes with an overview of the study and research questions for this thesis.

1.1. Thyroid diseases

Benign thyroid diseases are benign dysfunction of the thyroid, an endocrine gland in the front neck. These diseases include non-toxic goiter, hyperthyroidism, hypothyroidism, and others (International Classification of Diseases, 2018). Non-toxic goiter is shown as an enlarged thyroid gland. Hyperthyroidism is defined when the thyroid gland produces an excessive and uncontrollable amount of thyroid hormones of thyroxine (T3) and/or triiodothyronine (T4), whereas hypothyroidism occurs with limited levels of thyroxine (T3) and/or triiodothyronine (T4) (Hein and Jackson, 1990). Thyroid diseases are usually chronic and common in the general population across ages, genders, and cultures (Madariaga et al., 2014). In particular, the prevalence of thyroid dysfunction is considered as high in both North America and China. Reports indicate that Canada had a prevalence of approximately 10% in Canada (“About Thyroid Disease”, n.d.) and China had a prevalence of 11% in coastal areas of China (Wang et al., 2013). Over 50% of Canadians who have experience with thyroid diseases are un-diagnosed (“About Thyroid Disease”, n.d.). In 2013, the Chinese Health Education Centre estimated that there were 2 million patients with thyroid diseases in China, and the treatment rate was only 5% (Cui, 2013).

Patients with thyroid diseases experience various kinds of health impairments, including impairments of physical, mental, and social wellbeing, even after treatment. For example, insomnia is one of the critical diagnostic criteria in hypothyroidism (Christianson & Bender, 2011). Studies also indicated that insomnia can result from hyperthyroidism, along with energy loss and fatigue (Christianson & Bender, 2011; Demet et al., 2002). Some patients with thyroid diseases are also reported as having mental problems such as anxiety and depression (Demet et al., 2002; Sevinc & Savli, 2004). According to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5, American Psychiatric Association, 2013), hyperthyroidism can result in bipolar and related disorders, anxiety disorder, and cause panic attacks and psychotic symptoms, whereas hypothyroidism can be the primary factor in depressive disorder, cause psychotic symptoms, sleep-wake disorder, neurocognitive disorder (NCDs), and personality change. The mental and physical problems together can cause limited social life among patients with thyroid diseases. For example, patients with hyperthyroidism and high levels of psychopathology reported an inability to work and difficulties in concentration (Demet et al., 2002). These health problems altogether lead to a decreased quality of life for thyroid patients. However, many physicians only focus on the lab results and medications for diagnosis and treatment and pay little or no attention to patients' mental or social aspects of quality of life before or after treatment (Christianson and Bender, 2011). In this case, patients may not get a chance to know that instead of physical symptoms such as rapid heart rate for hyperthyroidism ("Hypothyroidism," n.d.), their anxiety may also be closely related to their thyroid dysfunction. On the other end, patients may not even view their mental problems as symptoms of thyroid dysfunction when the physical symptoms are not predominant in the clinical picture, and they may be misdiagnosed as having mental disorders. Eventually, the treatment turns out to be not valid and then leads to worse conditions (Christianson & Bender, 2011). Therefore, it seems to be necessary to raise awareness regarding patients' quality of life to both patients and physicians. To this end, the current research in quality of life in thyroid patients is of primary importance.

1.2. Health-related quality of life (HRQOL) measurements

Health-related quality of life (HRQOL) instruments measure an individual's perception of their health status in physical, mental, and social wellbeing domains

(Muragundi et al., 2012). Consequently, HRQOL instruments can be used in the evaluation of treatment of outcomes from not only a biological perspective but from a broader health perspective. The development of HRQOL instruments is critical in the measurement of patient-reported outcomes (PRO) of their quality of life. Even though HRQOL and PRO are usually used as interchangeable terms in studies (Feeny et al., 2013), PRO measures (PROMs) are often used as the measurements of the treatment outcome (Johnston et al., 2019), whereas the HRQOL instruments help with examining how the health status impact the quality of life (Yin et al., 2016). Because some patients with non-toxic goiter do not require treatment, the term HRQOL will be used more often in this thesis. There are two kinds of HRQOL instruments. Generic HRQOL instruments measure general health, and disease-specific HRQOL instruments measure specific diseases such as diabetes or thyroid diseases (Muragundi et al., 2012).

Multiple kinds of HRQOL instruments have been developed for thyroid diseases, and most of them focus on specific types of thyroid diseases, such as the GO-QOL (Quality of Life in Graves' Ophthalmopathy) questionnaire for patients with Grave's ophthalmopathy (Terwee et al., 1998), and ThyTSQ (Quality of Life and Treatment Satisfaction in Hypothyroidism) instrument for patients with hypothyroidism (McMillan et al., 2008). However, there are some overlapping symptoms across different kinds of thyroid diseases (see Table G1 for examples), such as insomnia for hyperthyroidism and hypothyroidism (Christianson & Bender, 2011; Demet et al., 2002), and enlarged thyroid gland for hyperthyroidism and non-toxic goiter (International Classification of Diseases, 2018; "Hyperthyroidism", n.d.). Also, one thyroid disease can transform into another. For example, hyperthyroidism may transform to hypothyroidism due to the treatment of radioactive iodine (Christianson & Bender, 2011), or only due to an increased number of doses of thyroid supplement (Sevinc & Savli, 2004). Thus, it is not easy to mutually exclude one thyroid disease from the other because of similarities among thyroid diseases in nature. Moreover, it is impractical for patients to respond to two different types of questionnaires because they are experiencing two thyroid diseases at the same time, or are in a state of transition between one and the other. Thus, the Thyroid-Specific Patient-Reported Outcome (ThyPRO, Watt et al., 2014) was chosen for the current study because it measured the HRQOL among patients with different kinds of benign thyroid diseases.

1.3. ThyPRO, abbreviated version of ThyPRO (ThyPRO-39) & Measurement properties

The ThyPRO (Watt et al. 2014) questionnaire as well as an abbreviated version (ThyPRO-39) of the instrument have been proposed for use. This section discusses the original version of the ThyPRO, ThPRO-39 and translated versions of the ThyPRO-39 with specific attention to translated Chinese versions of the ThyPRO-39 and their measurement properties.

1.3.1. Original version of ThyPRO & ThyPRO-39

The original version of the ThyPRO (Watt et al., 2014) measures HRQOL among patients with thyroid diseases. It consists of 84 items, with 13 scales measuring a) physical symptoms: goiter symptoms, hyperthyroid symptoms, hypothyroid symptoms, eye symptoms, b) psychological symptoms: anxiety, depressivity; c) well-being and function symptoms: emotional susceptibility, tiredness, cognitive complaints; d) participation symptoms: impaired social life, impaired daily life, impaired sex life, and cosmetic complaints. In addition to the scale scores, the original ThyPRO also provides an overall composite scale score.

The COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) taxonomy was developed by Mokkink et al. (2018) to assess the methodological quality of measurement properties of HRQOL measurements. The methodological quality addresses the risk of bias for the measurement, and it determines the trust-worthiness of the measurement properties findings (Mokkink, 2018). Specifically, the primary measurement properties considered in this framework include a) internal consistency reliability and measurement error under the broad category of reliability; b) content validity, structural validity, hypothesis testing for construct validity, cross-cultural validity, and criterion validity under the broad category of validity; and c) responsiveness.

Wong et al. (2016) conducted a systematic review of measurement properties and methodological quality on multiple HRQOL measurements related to thyroid diseases based on the COSMIN taxonomy (Mokkink et al., 2018). It was demonstrated that the original 84-item version of the ThyPRO has satisfactory measurement properties

and methodological quality, and that it is ready to be used to measure the HRQOL for patients with thyroid diseases (Wong et al., 2016). Table G2 details information about the measurement properties of the original version of the ThyPRO.

The original version of the ThyPRO was demonstrated by Wong et al. (2016) to be shown as providing the most satisfaction of measurement properties among all the existing HRQOL instruments related to thyroid diseases. Wong et al. (2016) evaluated six out of nine measurement properties in the COSMIN framework, namely the internal consistency and measurement error under reliability; the content validity, structural validity and construct validity under validity; and responsiveness. In addition, Watt et al. (2015) examined the cross-cultural validity in a separate paper. Specifically, the internal consistency was examined using Cronbach's alpha (Watt et al., 2009), and the test-retest reliability was found to be adequate via repeated measures and intra-class correlations (>0.70, Watt et al., 2000). Also, the ThyPRO was demonstrated as having strong content validity by Watt et al. (2008) using cognitive interviewing. The structural validity was examined using confirmatory factor analysis (CFA) (Watt. et al., 2014) and differential item functioning (DIF) (Watt et al. 2014). By applying CFA to the ThyPRO, 11 items were eliminated due to misfit, and the rest of the scales were shown as unidimensional (Watt. et al., 2014). Also, Watt et al. (2014) found DIF as a function of diagnosis and age, but both are small and reasonable. The known-groups validity and convergent validity were supported by a comparison between groups and comparison to another HRQOL instrument (Watt et al., 2010). The responsiveness was validated by significant changes in scores from before to after treatment, which was consistent with experts' prediction, and was more substantial than the other generic HRQOL instrument (effect size > 0.80) (Watt et al., 2014). Finally, the cross-cultural validity was high in general by results from DIF by ordinal logistic regression (Watt et al., 2015), where 12 items showed DIF as a function of country, but the impact of these items to the validity are small. Based on the COSMIN taxonomy to reviewing measurement properties (Mokkink et al., 2018), Wong et al. (2016) indicated that the ThyPRO had reliable methodological quality on internal consistency, reliability, content validity, structural validity, and moderate methodological quality on construct validity and cross-cultural validity.

Watt et al. (2015) developed the abbreviated 39-item version of the ThyPRO (ThyPRO-39) after applying an IRT model to the selection of items using a cross-

sectional and a longitudinal sample. The ThyPRO-39 was evaluated with its reliability, hypothesis testing in construct validity, and responsiveness in this study using previous data. The test-retest reliability of scores on each ThyPRO-39 scale was similar to the ThyPRO, and it was found to have high intra-class correlations (0.89 – 0.98) between ThyPRO-39 and ThyPRO scores. The expected high-score group also had significantly higher scores than the expected low-score group, which was the same as on the original ThyPRO, indicating high construct validity. Age was found to play a role in ThyPRO-39 scores when it was added to a DIF analysis, which was consistent with the ThyPRO. The results also indicated that a similar magnitude of change and responsiveness as on the ThyPRO was detected in ThyPRO-39 scores. Based on the COSMIN taxonomy (Mokkink et al., 2018), the ThyPRO-39 has satisfactory methodological quality on reliability, construct validity, and responsiveness.

1.3.2. Translated Chinese versions of ThyPRO-39

Culture plays a significant role in health in different ways. It has been demonstrated that culture not only impacts how individuals understand health but also how they manage their health conditions (Huff & Kline, 2007). Gopalkrishnan and Babacan (2015) discussed that culture affected individuals' coping styles to stressors, treatment-seeking patterns, communication, and the use of cultural and linguistic interpreters. Therefore, when the original language version of a questionnaire is translated into another language, it is common to have some problems due to cultural difference. As highlighted by the International Test Commission [ITC] (2017), even though the ITC guidelines (2017) for test translation was ready to be used for 20 years, most researchers still failed to follow the guidelines, and thus led to failing to develop a translated test that had scores with satisfactory measurement properties. Even if the researcher successfully followed the guidelines, translating the form to another language was complicated work due to cultural differences. On the other hand, the appropriateness of a translation is judged by whether it confirms the equivalence of the test structure by ITC, but at the same time, the translated version is expected to adjust to the new context of the target language version because cultural differences may lead to different functioning of items across languages (ITC, 2017). Therefore, one can argue that sometimes cultural differences are inevitable and thus lead to different responses when new language versions are developed and applied.

There are two Chinese versions of the ThyPRO-39, one is the Traditional Chinese version developed and studied in Hong Kong (TC ThyPRO-39, Wong et al., 2018), and the other is the Singapore Simplified Chinese version (SC ThyPRO-39, Liew, n.d.). It should be noted that the Traditional Chinese writing system is used in Hong Kong and Taiwan, whereas the Simplified Chinese writing system is used in Mainland China and in some countries in Southeast Asia, such as Singapore and Malaysia. Also, one considerable difference between the Chinese population in Hong Kong who use the Traditional Chinese writing system and the population in Mainland China is they use different official spoken language systems, where people in Hong Kong speak Cantonese and people in Mainland China and other countries and areas speak Mandarin as their official spoken language.

Wong et al. (2018) examined construct validity and internal consistency of the Traditional Chinese version of ThyPRO-39 (TC ThyPRO-39) scores. More specifically, the convergent validity aspect of construct validity was investigated through comparison among Traditional Chinese versions of ThyPRO-39, SF-6D, and SF-12v2 using Spearman correlations. The known-groups validity in construct validity was assessed by the comparison between different diagnosis groups of patients with thyroid diseases. The corrected item-total correlation, which measures the degree of the correlation of an item with the other items within a scale, was used to examine the “internal construct validity” (Ware & Gandek, 1998). Cronbach’s alpha was used to examine the internal consistency reliability (Nunnally, 1978). The result showed the scales in the TC ThyPRO-39 were moderately to highly correlated with SF-12v2 or SF-6D, so the convergent validity was confirmed. When comparing among different subgroups of thyroid disease, it was found that most of the scales in the TC ThyPRO-39 did not show statistically significant differences among subgroups, indicating low construct validity. Eight items in the TC ThyPRO-39 were below the threshold of good internal construct validity (item-total correlation < 0.4), pointing out problems with these items. The reliability of six scales was low (< 0.7) due to low internal construct validity with eight questionable items. The methodological qualities of the TC ThyPRO-39 are satisfactory based on the COSMIN checklist (Mokkink et al., 2018). In addition to the measurement properties listed above, it was found that the TC ThyPRO-39 had a significant floor effect (Wong et al., 2018).

1.4. Ceiling/floor effect in HRQOL instruments

HRQOL instruments are required to have a superior quality of measurement properties (Mokkink, 2018). According to Terwee et al. (2007), the absence of CFE is another essential criterion of measurement, but it was not included in the COSMIN taxonomy (Mokkink et al., 2018). Significant CFE is determined when more than 15% of the participants in the sample achieve the worst or the best score (Terwee et al., 2007). Please note that the “significant” here is not referring to the statistical significance of the difference, rather it refers to the description of a prominent level of CFE. Therefore, the terms “significant” and “prominent” CFE are used interchangeably in this study. According to Šimkovic and Träuble (2019), the CFE is problematic because the larger magnitude of CFE increases the bias and uncertainty of the statistical results based on their results from a simulation study. Explicitly, Šimkovic and Träuble (2019) stated that the presence of the CFE will cause the application of ANOVA or t-test on measures to become problematic by affecting the mean, variance, and other distributional properties, and thus influence tenability of the assumptions of normal error and homogeneity of variance. The bias caused by CFE will lead to problems with estimation of group differences, and the uncertainty will lead to wider confidence interval (CI), and thus interfere with the robustness of the results.

The CFE also affected the quality of other measurement properties (Terwee et al., 2007). With CFE, individuals who showed the lowest or the highest scores cannot be distinguished from each other. Furthermore, because the instruments with CFE could not detect the change of these individuals with CFE, the responsiveness was also negatively affected. More importantly, the presence of CFE could be an indicator of low content validity, possibly due to its failure to include extreme items that capture the features of the individuals with the lowest or highest scores.

1.4.1. Ceiling/floor effect in original language versions

A summary of the presence or absence of CFE in some HROQL measurements is shown in Table G3. Specifically, the CFE was found in scores on some original language versions of disease-specific HRQOL instruments. For example, the SWQL (Swallowing Quality of Life) instrument, which is an HRQOL measurement for patients with oculopharyngeal muscular dystrophy (Youssof et al., 2017), showed prominent CFE

on almost all the items. The PU-QOL (Pressure Ulcers Quality of Life) measurement, which measured HRQOL of patients with pressure ulcers Gorecki et al., 2013, also had significant floor effects on most of the scales. The CFE was more salient in some commonly-used generic HRQOL instruments than disease-specific HRQOL instruments for patients with different kinds of health problems. The SF-36 (36-Item Short Form Health Survey) showed significant ceiling effect on most of the dimensions among patients with total hip arthroplasty (Rampazo-Lacativa et al., 2015), and patients with brain tumours (Bunevicius, 2017). In one study about EQ-5D (Euro-QoL 5-Dimension, Hakimi et al., 2016), the CFE was present and significantly high on EQ-5D-3L (Euro-QoL 5-Dimension, 3-level) among patients with benign prostatic hyperplasia. It is reasonable that there are more remarkable CFE in generic HRQOL instruments than disease-specific HRQOL instruments because it will be harder for generic HRQOL instruments to detect the extreme condition of one specific disease compared with disease-specific HRQOL questionnaires. However, the problem is that, in most of the cases, the CFE did not raise much attention to the researchers, as it was either evaluated as not problematic or not identified as present in the original language versions of HRQOL instruments. For example, although three out of four scales in Neuro-QoL (Quality of Life in Neurological Disorders) instrument measuring social roles and activities showed significant floor effects among patients with Huntington's disease (> 15%), it was stated as not problematic by Carlozzi et al. (2018). Moreover, the percentage of CFE was not reported in previous studies of original Danish language versions of ThyPRO and ThyPRO-39 by Watt et al. (2014) and Watt et al. (2015). However, when the English version ThyPRO was validated among English-speaking Asian patients in Singapore in a recent study, it was reported that a significant floor effect was shown on 10 out of 13 symptoms and functions scales in ThyPRO (Liew et al., 2021).

1.4.2. Ceiling/floor effects in translated versions

When the original language version of HRQOL questionnaires were translated into another language, particularly into Asian language versions, there were mixed results regarding the CFE. It was found that significant CFE was absent in scores on some of the translated disease-specific HRQOL instruments, such as the GO-QOL instrument among Taiwanese patients with Graves' ophthalmopathy (Lin et al., 2013), and the CLAST (Chinese version of the Language Screening Test) among early-stroke

patients in China (Yang et al., 2018). Some of the translated versions of disease-specific HRQOL instruments showed significant CFE, such as the floor effect on the Traditional-Chinese version of ThyPRO-39 (Wong et al., 2018) and the Chinese version of QOL-RTI (Quality of Life Radiation Therapy Instrument, Chen et al., 2014). It was indicated that there was a significant ceiling effect on most of the items of QOL-RTI among patients with head and neck cancer in China (Chen et al., 2014). Moreover, for translated versions of generic HRQOL instruments, the CFE that already existed in the original language version become more salient in the new language versions, including the widely used generic HRQOL instruments of EQ-5D (Huang et al., 2017; Sun et al., 2011), SF-36 (Sararaks et al., 2005; Zhou et al., 2013) and SF-12v2 (12-Item Short Form Health Survey, Version 2, Wong et al., 2018).

In examining distributions of scores, one interesting finding from studies about response styles indicated that Asian participants tend to use more middle response styles and avoid the extreme response styles (Chen et al., 1995; Harzing et al., 2012) due to cultural values. This, however, was contradicted by findings from some other studies about HRQOL instruments, where it was shown that Asian participants were more likely to choose extreme options in their native language and report better health conditions than the Western population (Lubetkin et al., 2005; Szende & Williams, 2004). Additionally, some studies illustrated that social desirability bias played a significant role in responses to the questionnaires (Bowling, 2005; Kim & Kim, 2013). Kim and Kim (2013) argued that although the social desirability bias was present in both collectivistic (e.g., Japan and Korea) and individualistic countries (e.g., the Netherlands and the United States), the magnitude of the bias is more consistent and stronger in collectivistic countries. That is to say, the profound social desirability bias in collectivistic societies leads to CFE in East Asian culture. The acquiescence bias, which is the tendency to agree with others and report more positive responses regardless of the content of the questions, is also considered to be closely related to collectivism (Rammstedt, Danner & Bosnjak, 2017). Individuals in collectivistic cultures may experience more pressure to acquiesce than individuals in individualistic cultures (Smith & Fischer, 2008).

CFE is an important consideration in HRQOL measurements. As reported above, Wong et al. (2018) found CFE in the TC ThyPRO-39, the presence or absence of CFE in SC ThyPRO-39 is not currently described in the literature. Therefore, this study investigated the presence or absence of CFE, to see whether the same phenomenon

found in Traditional Chinese version of ThyPRO-39 (Wong et al., 2018) can also be found in Simplified Chinese version of ThyPRO-39 among participants in Mainland China.

1.5. Model comparison and selection for data with CFE

There are a variety of modelling strategies to probe influences of variables on scores. Most commonly known is the ordinary least squares multiple regression analysis, however other strategies may be more appropriate under specific data conditions. This section provides a background for the consideration of modelling strategies dealing with CFE and an overview of the six modelling strategies for the data on the Simplified Chinese ThyPRO-39 in the current study.

1.5.1. Background of modelling strategies dealing with CFE

Analyses of QOL data often involve parametric regression analyses with different distributional models. Many studies may use analytic strategies based on the normal distribution. However, if the data are not normally distributed and with a presence of CFE, other modelling strategies were also proposed by different researchers. In one study, Twisk et al. (2018) used an empirical data set of HAQ-DI scores that were not normally distributed with an excess of zeros from patients with arthritis (Claessen et al., 2009). For their analysis, the effects of treatment and time interaction were analyzed with linear mixed model and Tobit mixed model. Results indicated that the Tobit mixed model performed much better than the linear mixed model, and the linear mixed model highly underestimated the interaction effect. To better explain the Tobit model, for the Tobit regression model, the data with floor or ceiling effects for each scale are treated as left- or right-censored data at a certain point of lower and/or upper bound (Zhu & Gonzalez, 2017) and as such, scores at the floor are not considered to be “true” excess of zeros. In addition to having a lower and/or upper bound, the rest of the data still follow the linear mixed model.

However, not all excesses of zeros are due to censoring. Some other models, which presuppose the zeros to be “true” zeros are often used to deal with count data (Liu, 2007; Ullah, Finch, & Day, 2010). For example, in Liu’s study (2007) of count data in smoking behaviour, it was discussed that the Poisson regression model can be used

for Poisson distribution that is right-skewed, with an excess of “true” zeros. Also, a zero-inflated Poisson (ZIP) regression model, which includes the probability of being an inflated zero in a binomial distribution and a Poisson regression, is also promoted by Liu (2007) for data with preponderance of zeros.

It was noted that the Poisson regression model is based on assumption of the equivalence between mean and variance, therefore they cannot be used for data with overdispersion. When the data are over-dispersed, which means the variance is larger than the mean of the distribution, the negative binomial (NB) regression model can be used to better analyze the data than Poisson regression model (Ullah, Finch, & Day, 2010).

Although Poisson and NB regression models can handle data with excesses of zeros, i.e., when there are too many zeros, two-part models were introduced to specifically deal with the preponderance of zeros. The ZIP and ZINB are both two-part models where one part follows the ordinary count models of either Poisson or NB distribution, and the second part is a binomial distribution with probability of being an inflated zero. The first part only deals with the zeros due to sampling variability; the second part can work with structural zeros that can only be zeros (Ullah, Finch, & Day, 2010). For example, for the frequency of smoking behavior in a week, zeros from people who used to smoke, but due to some reason they stopped smoking during that period, are counted as sampling zeros. In contrast, zeros from people who do not smoke at all, are counted as structural zeros (Liu, 2007). The zero-inflated Poisson (ZIP) regression model was demonstrated to show better performance in dealing with a preponderance of zeros than Poisson regression, according to Liu’s study (2007) about smoking behavior. In addition, the zero-inflated negative binomial (ZINB) regression model can be used to deal with data with both preponderance of zeros and overdispersion, based on the results from Ullah, Finch and Day’s study about falls data (2010). The zero-inflated models are also considered to show better performance than the linear mixed model and the Poisson mixed model when assessing the development of hypoglycaemic events among diabetic patients (Spriensma et al., 2013).

1.5.2. Modelling strategies for ThyPRO-39 in this study

Although scores from a questionnaire are not typically count data, in many scenarios they can be viewed as count data because the scores are almost always integers, and in HRQOL they are often a measure of frequencies of some behaviors and/or feelings. More specifically, in the ThyPRO-39, the item-level response options range from “0 = Not at all” to “4 = Very much”. As such an item-level score can be considered a measure of frequency of presence of a specific symptom, and a scale-level score as a composite measure of frequencies across a set of specific symptoms. Therefore, six models, including multiple linear (ML) regression, Tobit regression, Poisson regression, NB regression, ZIP and ZINB regression models, were applied to the data on the SC ThyPRO-39 in the current study, and various criteria for each model were computed and compared to select the best modelling strategy with best fit to the questionnaire data, especially the questionnaire data with a floor effect.

1.6. Variables impacting responses to HRQOL questionnaires

A variety of variables can impact HRQOL questionnaire data. These include personal characteristics (e.g., gender) as well as how data are collected (e.g., mode of administration). Different variables that may influence responses to HRQOL questionnaires are introduced in this section.

1.6.1. Gender

Gender has been shown to have a significant impact on responses to questionnaires related to health across the world (Boerma et al., 2016). Some studies have found that women tend to talk more about their feelings and the reasons of their diseases, and cried more with depressive symptoms, whereas men talked more about drugs and alcohol, and physical actions (Strauss et al., 1997). Izadnegahdar et al. (2014) and Cherepanov et al. (2010) indicated that women reported poorer health condition than men in every health-related domain (Boerma et al., 2016). According to Boerma et al. (2016), women were more likely to report their health problems in vision, mobility, pain, sleep, angina, arthritis, and depression, whereas men more commonly reported experiencing antisocial behavior, substance abuse, and suicide. A study

conducted in China, Gao et al. (2020) concluded that women reported more anxiety symptoms than men among college students, while men reported more experience of depression than women, which is contradictory to the findings in Boerma et al. (2016), and no gender difference shown on stress-related problems. In contrast, there were some studies in which no gender differences were found in HRQOL responses (Shafie et al., 2021; Tlusta et al., 2009). In general, in many studies, males and females may be different in their response styles. While some studies indicated no gender difference in response style (Dolnicar & Grun, 2007; Hamamura, Heine, & Paulhus, 2007). Harzing and Brown (2012) found that men tend to use more extreme response style, and middle response style was more commonly used by women.

The gender difference on responding to HRQOL questionnaires can be due to the difference in gender roles, sociodemographic and socioeconomic status (SES), and can also be due to gender differences in nature (Boerma et al., 2016; Cherepanov et al., 2010; West et al., 2015). Although disputable, according to Birk and Rieker (2008), examples of gender difference in nature include findings that women are more likely to suffer from depressive and anxiety disorder, as well as to have autoimmune disorders including thyroid diseases. Indeed, thyroid diseases are well known as more common in women than in men, especially hypothyroidism among postmenopausal women (Bauer et al., 2014; Carlé et al., 2015; Meng et al., 2015). The prevalence rate of hyperthyroidism was 8 to 5 between women and men, according to Castello and Caputo in their study published in 2019. Similar results were found for hyperthyroid dysfunction, where the prevalence for Grave's disease between women and men was 7 to 1 (Castello & Caputo, 2019). In Meng's study (2015) conducted in China, women also showed significantly higher incidence of hypothyroidism and hyperthyroidism than men.

Although gender is acknowledged as playing a significant role in HRQOL measurement and thyroid function, according to Cherepanov et al. (2010), gender was more often treated as a confounder rather than an independent variable of interest. Whatever the role, it is important to include gender into analysis when it comes to the responses to a thyroid-related quality-of-life questionnaire, such as the SC ThyPRO-39.

1.6.2. Mode of administration

The mode of administration (e.g., interview or self-administered) of the instrument can be another significant factor that influences response patterns (Bowling, 2005). Tourangeau and Smith (1996) argued that when the questionnaire related to health was administered by interviews, individuals were reported to show more positive and socially desirable responses than with self-administered questionnaires. Individuals may under-report sensitive health problems during interviews compared with self-administered questionnaires (Tourangeau & Smith, 1996), and thus result in CFE. Also, some individuals in the interview mode showed more acquiescence bias to agree with others and chose more positive responses than in the self-administered mode (Bowling, 2005). This acquiescence bias may also lead to a CFE, where individuals may choose extreme responses (Bowling, 2001). The interviewer bias could also be a factor that causes a difference between the interview and the self-administered mode, whereas the interviewer may either help respondents to choose the answers corresponding to their true feelings, or on the other hand may lead respondents to hide their true responses (Bowling, 2005). In contrast, since the self-administered mode has a weaker social presence, it may lead to a more accurate response compared to the interview mode of administration (Siemiatycki, 1979). Importantly, mode effects can also interact with individual characteristics; for example, Levin-Aspenson and Watson (2018) indicated that lower scores in depression were shown among younger and better-educated individuals when the questionnaire is administered through interviews compared to self-administration.

It is important to acknowledge that different technologies or media can be used for different modes of administration (e.g., paper-and-pencil self-administered versus electronic self-administration. Bowling (2005) concluded that there was more significant difference between interview and self-administration modes than within each mode, such that the difference between electronic self-administration and paper-and-pencil self-administration, and the difference between traditional face-to-face interviews and remote interviews (e.g., telephone, videoconference…etc.) are considered as small. Specifically, Kobak et al. (2008) conducted a study comparing the face-to-face interview mode of administration to videoconference interview and telephone interview administrations on Montgomery-Asberg Depression Rating Scale (MADRS). Results indicated that the mean scores in videoconference and telephone interview modes were not statistically

different from face-to-face interview mode. The psychometric properties of MADRS in videoconference and telephone interview modes were also comparable to face-to-face interview mode (Kobak et al., 2008). Thus, results were identified as supporting the use of videoconference and telephone interviews when assessing health-related questionnaires. Also, in Swartz et al.'s study (2007), the paper-and-pencil administration mode was compared with electronic mode on CES-D scores in a clinical sample, and results indicated that patients' scores on paper-and-pencil mode were not significantly different from scores on the electronic mode, although interaction effects were found between mode of administration and order of the modes. In another study conducted by Fouladi et al. (2002), even though statistically significant effects were found between paper-and-pencil and electronic self-administration modes on three scales, the magnitude of the mode effects was considerably small. Finally, based on Rasmussen et al.'s study (2015) of comparison between electronic and paper-and-pencil mode of administration of the original 84-item version of ThyPRO (Watt et al., 2014), there was adequate equivalence between the two modes of administration. Only scores on the cosmetic complaints scale had a significant difference between the two modes, where the mean of the scores in the electronic mode was higher than for scores in the paper-and-pencil mode. Therefore, Rasmussen et al. (2015) concluded that the electronic mode can replace the paper-and-pencil mode for ThyPRO when it is necessary.

Accordingly, interview and self-administered modes of administration were considered in the current study on analysis of responses to the SC ThyPRO-39. Due to the adequate agreement between videoconference and face-to-face interview modes, and between electronic and paper-and-pencil self-administration modes (Kobak et al., 2008; Rasmussen et al., 2015), videoconference interview and electronic self-administration may replace the traditional face-to-face interview and paper-and-pencil self-administration modes as appropriate.

1.6.3. Other demographic and disease characteristics

In addition to gender and mode of administration effects that were reviewed in the previous sections, and which are primary variables of interest in the current study, other demographic and disease characteristics, such as age, education level, duration of treatment, and level of disease control, can also be critical in predicting responses to the thyroid-related quality-of-life questionnaire.

As mentioned above in the overview on thyroid disease quality of life, in addition to gender, age plays a significant role in thyroid functioning, as more thyroid-related symptoms were reported with the increase of age, especially for women after menopause (Bauer et al., 2014; Meng et al., 2015). Leak et al. (2013) indicated that the association between age and other personal characteristics can be important for clinicians to provide different care strategies to different kinds of patients. Other examples of findings on the role of age were in analyses which indicated that age had a moderating effect on the relationship of gender and income with QOL. Watt et al. (2014) found DIF present for 8 items on ThyPRO, and results indicated that except for the item of “crying easily”, younger patients showed better QOL on the rest of the 7 items.

Regarding the effect of other personal characteristics, Leak et al. (2013) concluded that higher QOL was found in participants with higher education level, and longer time since diagnosis. It was also confirmed by Trompenaars et al. (2005) that higher education level was associated with better QOL among Dutch psychiatric outpatients. Also, longer time after treatment was associated with better HRQOL among patients with leukemia and lymphoma, according to Tacyildiz et al. (2020). However, Trompenaars et al. (2005) concluded that demographic characteristics were not as important as other factors in influencing the QOL, therefore it was not necessary to pay attention to those demographic characteristics.

In this study, in addition to considering gender, other demographic and disease characteristics (i.e., age, education level, duration of treatment, and level of disease control) were also included as predictors for the purpose of model strategy comparison (e.g., between NB and ZINB models) with various predictors. Although overall effects were reported, the main effects of these predictors were not the focus of this thesis, therefore are not reported in detail.

1.7. Study Overview and Research Questions

This study was conducted among Mandarin speakers with thyroid diseases on the Simplified Chinese form of ThyPRO-39 (SC ThyPRO-39) in China. The rationale for this study is a general lack of attention to the quality of life for patients with thyroid diseases and the fact that there is no published study about the Simplified Chinese version of ThyPRO-39 in a Mandarin-speaking population. The current study focuses on

response patterns to 12 symptoms and functions scales and to the composite scale on the Simplified Chinese version of ThyPRO-39. Moreover, because a CFE was found in a study of responses by Cantonese speaking respondents on the Traditional Chinese version of ThyPRO-39 (Wong et al., 2018), this study investigated whether CFE is also found in responses of Mandarin-speaking respondents on the Simplified Chinese version of ThyPRO-39. Other psychometric properties, such as the internal consistency reliability and internal construct validity were also reported in this study. The effects of gender and mode of questionnaire administration, particularly with respect to CFE, was also examined, and different regression modelling strategies were considered. Six distributional regression models, which are the ML, Tobit, Poisson, NB, ZIP, and ZINB models, were compared and optimal modelling strategies were selected based on series of AIC, BIC, LRTs, and Vuong tests. Results were reported separately for scales with floor effect and for scales without floor effect. In addition to gender and mode of administration, additional personal patient information, including age, education level, the duration of the thyroid diseases, and level of disease control for the thyroid diseases were also included in the regression models as predictors. Different predictor sets were included in the model comparison in order to make a comprehensive recommendation for selecting modelling strategies. The main effects of gender and of mode of administration on the 12 scales and composite scale were reported, followed by the overall effect of all six predictors. Qualitative analysis was conducted to explain some of the results on some scales. The research questions are listed below.

1.1. What are the distributional characteristics for the 12 scales, composite, and item-level scores on the Simplified Chinese ThyPRO-39 for participants overall and separated by subgroups (males versus females, and electronic interview versus electronic self-administration)?

1.2. Is there any CFE, at the scale and item level, observed in scores on the Simplified Chinese version of ThyPRO-39?

2.1. If CFE is present at the scale level, is it the same or different between self-identified males and females?

2.2. If CFE is present at the scale level, is it the same or different between the electronic interview group and the electronic self-administered group?

3.1. If there is significant CFE present at the scale level, which of the six distributional models (ML, Tobit, Poisson, NB, ZIP, ZINB) is better to use to analyze data with CFE?

*3.2. For the scales that **do not** show significant CFE, which of the six distributional models (ML, Tobit, Poisson, NB, ZIP, ZINB) is better to use to analyze the data?*

4. Do participants' scale responses vary as a function of gender, the administration mode (electronic interview versus electronic self-administered mode), age, education level, duration of the treatment, and levels of control, after adjusting analysis strategy for any CFE where appropriate?

Chapter 2. Method

2.1. Participants

This study is conducted according to university ethical guidelines and with Human Subjects Approval from the institutional review board. Data collection was conducted on 195 individuals with thyroid diseases. After excluding 16 participants with missing values on demographic variables, the current thesis is based on 179 patient-participants with various thyroid diseases in China and who are able to speak Mandarin and/or read Simplified Chinese. Among 179 patient-participants, there were 81 patient-participants in the interview administration group and 98 patient-participants in the online survey group. Among all the patient-participants in the current sample, 56 of them were self-reported male and 123 of them were self-reported female. Patient-participants were aged from 18 to 82 years, with a mean of 37 and standard deviation of 13. Other demographic information such as education level, duration of treatment, the levels of control of the diseases, were collected and were used as predictors in comparing the modelling strategies (see Table H1 for details of demographic information).

Primary analyses to address research questions include tests on proportions and analyses of different models, such as normal theory linear models. Power analysis was used to determine whether the given sample size is appropriate to have adequate power (Cohen, 1988). It was conducted using the package 'pwr' (Champely et al., 2020) in R (R Core Team, 2017). The effect size of the two-sample proportion test was set as medium (h = 0.5). The effect size for the multiple linear regression (f^2) was calculated with a R^2 = 0.3 ($f^2 = R^2/(1-R^2)$). The significance level was set as 0.05. Results indicated a power of 0.87 for the proportion test of the floor effects between two gender groups (h = 0.5, α= 0.05, n_1 = 56, n_2 = 123), a power of 0.91 for the proportion test of the floor effects between two administration groups (h = 0.5, α= 0.05, n_1 = 81, n_2 = 98), and a power > 0.99 (R^2 = 0.3 and α= 0.05, n = 179) for the multiple linear regression.

2.2. Measures

This section describes the focal PROM, namely The Simplified Chinese version of ThyPRO-39 and tools developed for the current study, as well as approaches to collect demographics and questionnaire-completion experience from patient-participants.

2.2.1. Simplified Chinese version of Thyroid-Specific Patient-Reported Outcome – 39 (Appendix A)

The Simplified Chinese version of ThyPRO-39 (SC ThyPRO-39, Liew, n.d.) was translated from the original Danish version (Watt et al., 2015) in Singapore. To date, there is no published study about the Simplified Chinese version of ThyPRO-39 to our knowledge, therefore no published evidence of its application with patients who speak Mandarin or read Simplified Chinese.

The SC ThyPRO-39 contains 39 items with 12 symptoms and functions scales. In addition to these scales, the SC ThyPRO-39 also provides an overall composite scale score. Response options are on an ordered response scale measuring frequencies of both physical and mental symptoms in the past four weeks, ranging from “0 = 完全没有”(“Not at all” in English), “1 = 有一点”(“A little” in English), “2 = 有些”(“Some” in English), “3 = 多一点”(“Quite a bit” in English), to “4 = 很多”(“Very much” in English). The original scoring of ThyPRO-39 is based on a bifactor model, with one general factor for a composite scale and sub-factors for individual scales. In a bifactor model, each item is based on both a general factor and a subfactor.

The symptoms and functions scales measure physical symptoms, psychological symptoms, well-being and function, and participation during the past four weeks (see Table G4 for details of the breakdown on items by scales). The scales of goiter symptoms, hyperthyroid symptoms, hypothyroid symptoms, and eye symptoms are labelled as physical symptoms. The anxiety and depressivity scales measure psychological symptoms. The tiredness, cognitive complaints and emotional susceptibility scales belong to the category of well-being and function. The impaired social life, impaired daily life, and cosmetic complaints scales are categorized under participation. There is one additional item about the overall quality of life, which is not included in the categories described above. There are three items for each scale except

for the hypothyroid symptom and hyperthyroid symptom scales, which have four items in each scale.

Theoretical minimum and maximum for the raw score of each scale, as well as some of the psychometric properties are shown in Table G4. The raw score of the composite scale contains 22 items selected by Watt et al. (2015) and ranges from 0 to 88. The raw scores of scales range from 0 to 12 or 16 depending on the number of items in the scale. The transformed score for each individual scale and composite scale ranges from 0 to 100 based on Orlando and Thissen IRT-based summed-score linking (Orlando et al., 2000), with higher scores indicating worse HRQOL. For comparing models to deal with scores showing CFE, this study focuses on the raw scores of the 12 symptoms and functions scales as well as the raw scores for the composite scale, instead of the transformed scoring system Watt et al. (2015) promoted in his study. All study participants in China were administered the Simplified Chinese version of ThyPRO-39 (Liew, n.d.).

2.2.2. Qualitative questionnaire on questionnaire completion experience and Post-questionnaire Interview (Appendix C)

Patient-participants experience regarding HRQOL data were collected in one of two ways. These were a) electronic qualitative questionnaire, b) a post-questionnaire interview.

The qualitative questionnaire (developed for this project by the researcher, Appendix C) measures patient-participants' interpretation of the Simplified Chinese version of ThyPRO-39. Specifically, patient-participants were asked about their interpretations of various items on the ThyPRO-39 and whether patient-participants have any concerns that are not mentioned in the HRQOL instrument. In the current study, this questionnaire was in Simplified Chinese, because participants were all Mandarin speakers who did not speak English in China. Patient-participants in the electronic self-administration mode were administered this questionnaire. The English version of this questionnaire is also shown on Appendix D as reference.

The post-questionnaire interviews used the qualitative questionnaire (developed for this project by the researcher, Appendix C) to measure questionnaire completion

experience by participants but in a remote interview mode. Only patient-participants in the interview mode of administration were administered the post-questionnaire interview.

2.2.3. Demographic Instrument (Appendix E)

In the current study, demographic instruments were administered to all participants. The demographic instrument consists of basic demographic questions such as age, sex, education level, and questions related to the disease, such as the subtype of thyroid diseases, the duration of the treatment, and whether the diseases are controlled. (See Appendix E). The English version of this instrument is provided in Appendix F as reference. More specifically, there are 7 options for education level, including “1 = lower than high school”, “2 = high school”, “3 = college diploma”, “4 = Bachelor’s degree”, “5 = Master’s degree”, “6 = PhD degree”, and “7 = Other”. Because there was no response to the option “7 = Other”, it was dropped for the further analysis. The duration of treatment variable was asked as an open-ended question, and it was recoded to a variable with 7 levels, which was “0 = no treatment at all”, “1 = less than 1 year”, “2 = 1 to 2 years”, “3 = 2 to 3 years”, “4 = 3 to 4 years”, “5 = 4 to 5 years”, and “6 = 5 + years”. The level of disease control contains 5 levels ranging from “0 = not at all”, to “4 = very much”.

2.3. Procedure

Data collection was conducted on 179 thyroid patient-participants, who speak Mandarin and read Simplified Chinese in China. Some participants were recruited through advertisements in Chinese social networks such as WeChat and Douban. Some other participants were recruited by a relative of the researcher. Thyroid patients who are resident in China, and who can speak Mandarin were invited to participate in this study. It was participants’ choice whether to do a remote interview or complete an online survey in a self-administered mode. Participants could choose whether to do a telephone interview or teleconference; all the remote interviews were recorded for analysis.

Based on the current situation of COVID-19, the original plan of using paper-and-pencil administration mode and face-to-face administration mode was replaced by electronic self-administration mode and remote interview administration mode. Questionnaires were administered through the SurveyMonkey website. Participants

completed informed consent forms first, and then completed the Simplified Chinese version of ThyPRO-39 (Liew, n.d.), followed by qualitative questionnaire or post-questionnaire interviews after completion of the Simplified Chinese version of ThyPRO-39, as well as the demographic questionnaire.

2.4. Data analysis

The data analysis section addresses the descriptive statistics procedures, the diagnostic strategies prior to further analysis, and strategies corresponding to each of the research questions.

2.4.1. Diagnostics and assumption checking

Strategies to detect and address missing values and outliers were conducted using R. Specifically, package ‘dlookr’ was used to detect missing values and outliers. Missing values of item scores are imputed by Multivariate Imputation by Chained Equations (MICE, Azur et al., 2011), which is a strategy to fill out the unobserved data predicted by observed data with multiple imputations. Missing values of the age variable were imputed using the median; cases with missing values for variables of gender, education level, duration of treatment, and levels of diseases control were excluded from the data analysis. Outliers were excluded from the data as necessary. In addition, the assumption checking for each model was conducted before each test using plots and various tests. Details about the assumption checking strategies are indicated in each research question.

As mentioned previously, because there was no data of “7 = Other” for education level, the level 7 was excluded from this variable. Age, education level, duration of treatment, and level of disease control were treated as continuous quantitative variables. Gender and mode of administration variables are treated as binary variables. In particular, the participants who self-reported as male were coded as 0 and those who self-identified as female were coded as 1; interview mode of administration was coded as 0 and self-administered mode was coded as 1.

2.4.2. Descriptive statistics and select psychometrics

The descriptive statistics of means, medians, standard deviations, skewness, and kurtosis values for scores on the Simplified Chinese ThyPRO-39 were reported at the scale level and item level. The CFE at the scale and item levels is determined by the percentage of participants' responses that are present as the top level and the bottom level of a variable, where significant CFE is defined as more than 15% of the individuals at the top or bottom levels (Terwee et al., 2007). Results were reported overall, and separately by gender and by mode of administration.

Due to the lack of investigation into the measurement properties of the Simplified Chinese form of ThyPRO-39, this study also examined the internal construct validity and internal consistency reliability of scores on language version based on the current sample in China. The internal consistency reliability of scores on the Simplified Chinese ThyPRO-39 was judged based on the Cronbach's Alpha for each scale, where a value > 0.7 indicated good internal consistency reliability (Nunnally, 1978). The corrected item-total correlation between each item and the scale was used to assess internal construct validity. Satisfactory internal construct validity was determined by correlation coefficient > 0.4 (Ware & Gandek, 1998).

2.4.3. Research questions

RQ 1.1. What are the distributional characteristics for the 12 scales, composite, and item-level scores on the Simplified Chinese ThyPRO-39 for participants overall and separated by subgroups (males versus females, and electronic interview versus electronic self-administration)?

Summary descriptive statistics for the empirical distributions of scores on the Simplified Chinese ThyPRO-39 were considered for this research question at the scale, composite, and item level. The distributional form at the scores at the scale and composite level was addressed through inspection of histograms and tests of distributional normality.

The distributions for the SC ThyPRO-39 scale and composite scores were shown in histograms for participants overall and separated by gender (males versus females) and by administration groups (electronic interview versus electronic self-administration). The distributional characteristics from the descriptive statistics (e.g., the skewness and

kurtosis) were also discussed for participants overall and separated by gender and administration groups in this section. The results of the one-sample Kolmogorov-Smirnov (K-S) test (Chakravarti, Laha, & Roy, 1967) were presented in this section to check whether the data on the SC ThyPRO-39 scores in this study are normally distributed. Statistical significance was determined after the adjustment of Bonferroni correction (α = 0.05/13 = 0.0038).

RQ 1.2. Is there any CFE observed in scale and item level scores on the Simplified Chinese version of ThyPRO-39?

The proportion of the participants' scores at the bottom or top level for each of the 12 scales, composite scale, and individual items were presented in the descriptive analysis of SC ThyPRO-39 data overall, by gender and by each mode of administration. Interpretation of the presence or absence of significant CFE was based on criteria recommended by Terwee et al. (2007).

RQ 2.1. If CFE is present at the scale level, is it the same or different between self-identified males and females?

The proportion of floor and/or ceiling effects at the SC ThyPRO-39 scale level were compared between two gender groups using a set of two-sample proportion tests. The null hypothesis is that there is an equal proportion of either floor or ceiling effect for male (n = 56) and female (n = 123) groups. Regarding the assumption checking protocol, the two gender groups are assumed to be mutually exclusive and independent, and there are at least 5 cases in each group that reach the lowest or highest scores, and 5 cases in each group that do not reach the lowest or highest scores. The Bonferroni correction was applied to determine statistical significance.

RQ 2.2. If CFE is present at the scale level, is it the same or different between the electronic interview group and the electronic self-administered group?

The proportion of floor and/or ceiling effects at the scale level of the SC ThyPRO-39 were compared between two administration groups using a set of two-sample proportion tests. The null hypothesis is that there is an equal proportion of either floor or ceiling effect for interview (n = 81) and self-administration (n = 98) groups. The assumption checking procedures is the same as the proportional test between gender, where the proportional test for two administration groups is assumed to be mutually

exclusive and independent, and there are at least 5 cases in each group that reach the lowest or highest scores, and 5 cases in each group that do not reach the lowest or highest scores. Similar to the previous research question, the Bonferroni correction was applied to determine statistical significance.

RQ 3.1. If there is significant CFE present in the scale scores, which of the six distributional models (ML, Tobit, Poisson, NB, ZIP, ZINB) is better to use to analyze data with CFE?

For the SC ThyPRO-39 scales that presented CFE, the effect of CFE was taken into consideration due to the fact that a CFE can significantly interfere with the robustness of analytic results (Šimkovic & Träuble, 2019; Terwee et al., 2007; Zhu & Gonzalez, 2017). Therefore, based on the results from Research Questions 1 and 2, scales with significant CFE were analyzed using the six regression modelling strategies proposed, suggestions and recommendations of which model is best to be used to analyze data with significant CFE were based on the results of model comparison. For each of the six distributional models at the scale level, the response variable is the raw test score of each scale, and the predictors are all between-subject variables, including gender, mode of administration, age, education level, duration of treatment, and level of disease control for the thyroid disease. In order to get a comprehensive result to be able to be generalized to other test scores with significant CFE, different sets of predictors were tested. Within each distributional model, the regression analysis started with the simplest model where only a single predictor is present. More specifically, the predictor of gender or of mode of administration was selected, and model selection was considered to determine the optimal analytic strategy (e.g., ML versus NB versus ZINB). The model selection to determine the optimal analytic strategy for the full predictor set with all six predictors was also performed. Also, other sets of predictors choosing from the six predictors were included in the analyses to check whether the results of model selection are consistent across different predictor sets, but the details of those results are not reported.

For the performance of the models, different goodness-of-fit criteria were used to compare and select the best models, including the Akaike's information criterion (AIC) and the Bayesian information criterion (BIC) (Burnham & Anderson, 1998; Warton, 2005). A lower value of AIC and BIC indicates better performance of goodness-of-fit.

Also, different tests were used to compare the models, including sets of likelihood ratio tests that were used to select from models that are nested (Lewis et al., 2010), and Vuong tests for non-nested models (Vuong, 1989). Specifically, because the NB model has an extra parameter that estimates the over-dispersion compared to the Poisson model, the Poisson model is nested within the NB model; and the ZIP is also nested within the ZINB for the same rationale. Thus, the likelihood ratio tests were used to compare between Poisson and NB models, as well as between ZIP and ZINB models. The null hypothesis for the likelihood ratio test is that the first model is better than the second model. The Vuong test was used to measure the goodness-of-fit between non-nested models with or without the zero-inflated part (Wilson, 2015), therefore sets of Vuong tests were used to compare Poisson to ZIP models and compare NB to ZINB models. According to Desmarais and Harden (2013), the results from the original uncorrected Vuong test showed a significant bias to support the models with zero-inflated part, thus they proposed the AIC- and BIC-corrected Vuong tests to increase the power. Moreover, they suggested that the BIC-corrected Vuong test performed better at rejecting the zero-inflation than the AIC-corrected Vuong test when there is an absence of zero-inflation. Therefore, this study only focused on the results from BIC-corrected Vuong tests. The null hypothesis for the BIC-corrected Vuong test is that there is no difference between the two models, and the alternative hypothesis is one model is better than the other.

The assumptions and diagnostics for each distributional model with different sets of predictors were checked prior to the application of the model. The distributional diagnostics for the response variables were checked first, where the diagnosis of normal distributions were checked by one-sample K-S test (Chakravarti, Laha, & Roy, 1967); the diagnosis of Poisson distributions were checked by the goodness-of-fit tests for Poisson; and the diagnosis of negative binomial distributions were checked by goodness-of-fit tests for negative binomial (Friendly, 2000).

For the ML regression model, the assumption of linearity, specification of the predictors, and independence of errors were checked using scatterplots between fitted values and regular residuals. The homogeneity of variance assumption was checked by scatterplots between fitted values and squared root of standardized residuals. All scatterplots are required to follow the loess line of 0. The Normal Q-Q plot helped with checking the normality of errors assumption, where all the points should follow a roughly

positive-sloped, straight line. Finally, for predictor sets with more than one predictor, the diagnostic for lack of multicollinearity (where lack of multicollinearity is the desired state) was determined using the variance inflation factor (VIF), where VIF lower than 10 and tolerance larger than 0.1 were identified as low multicollinearity (Cohen et al., 2003).

For other models, the assumption checking procedure was similar to the ML model. Assumptions including the linearity, specification of predictors, independence of errors were checked by scatterplots, and the normality of errors was checked by normal Q-Q plot. The lack of multicollinearity was checked by VIF and tolerance.

For the Poisson and NB regression models, the assumption of the homogeneity of variance was checked by scatterplot between fitted value and standardized residuals, and the normality of errors was checked by a normal Q-Q plot. The lack of collinearity was checked by VIF. In addition, the Poisson regression assumes the population mean equals the variance (Cameron & Trivedi, 1999), therefore the dispersion test was also performed to check whether the dispersion parameter equals 1 or is larger than 1 (H_0: Dispersion parameter = 1, H_1: Dispersion parameter > 1). More specifically, if the null hypothesis was not rejected, the diagnosis of equal mean and variance was determined as satisfied for the Poisson regression.

Compared to the Poisson regression model that assumes the equivalence between mean and variance, the NB model allows for overdispersion (Johnson, Kemp, and Kotz, 2005). The assumption of normality of errors was checked by the normal Q-Q plots. Lack of collinearity was checked by VIF for sets of predictors larger than one. Whether there was overdispersion for the NB model was also checked by the dispersion test, and if the null hypothesis for the dispersion test was rejected, the presence of overdispersion was determined as satisfied for the NB regression.

For the assumption checking/diagnostic procedures of ZIP and ZINB models, the dispersion tests were conducted to check the equivalence between mean and variance. Also, because the zero-inflated models are not simply linear models, the assumption of normality of errors and/or the homogeneity of variance cannot be checked. Instead, the assumption of zero-inflation was checked by zero-inflation tests. For the zero-inflation tests, if the observed zeros are larger than the predicted zeros, the zero-inflated models are recommended to be used (Lüdecke et al., 2021).

RQ 3.2. For the scales that do not show significant CFE, which of the six distributional models (ML, Tobit, Poisson, NB, ZIP, ZINB) is better to use to analyze the data?

For the SC ThyPRO-39 scales with an absence of significant CFE, six distributional models were also compared using the same criteria and tests. Models were compared to determine whether there was an optimal modelling strategy for these type of data for SC ThyPRO-39. The response variables are the test scores from each scale, and the predictors are the six predictors mentioned previously (gender, mode of administration, age, education level, duration of treatment, and level of disease control for thyroid disease). For the purpose of presenting the main effect of gender and mode of administration, the models with only one predictor (gender or mode of administration) were also compared among six distributional models. Similar to scales with floor effect, other predictors sets were also analyzed to make a better conclusion across different predictors sets but are not reported in detail. The assumption checking procedures were the same as mentioned in RQ 3.1. for scales with floor effect. The results from the model comparison strategies were reported for each scale.

In addition to the 12 scales in the ThyPRO-39, the composite scale, which contains the total scores of 22 items from the questionnaire selected by Watt et al. (2015), was also fitted using six distributional models, and the results from goodness-of-fit criteria and tests were also reported.

RQ 4. Do participants' composite and scale responses vary as a function of the administration mode (interview versus self-administered mode), gender, age, education level, duration of the treatment, and levels of control, after adjusting for any CFE where appropriate?

The effects of the six predictors on SC ThyPRO-39 scores on 12 individual scales and on the composite scale were shown on the table presented for models with best fit (e.g., ZINB) according to the results from RQ 3. Several models were presented in cases where no optimal modelling strategy was identified. The assumption checking/diagnostic procedures were conducted in RQ 3, therefore no further diagnostics were included in this part. The main effects of gender and mode of administration in the single predictor analyses were presented first, followed by the effects of the full predictor set. The results were separately reported for scales with and without floor effect. The results included the estimated coefficients, estimated standard

error (SE), the significance level (p), and 95% confidence interval for the estimated coefficient of each predictor in the model. In addition, in order to see whether the whole model was statistically significant, the results from chi-squared tests comparing the full model to the null model without any predictors were presented in tables.

Chapter 3. Results

In this section, results pertaining to diagnostics, descriptive and inferential analysis are presented. The results from data cleaning, including detecting and handling missing values and outliers, are reported. Descriptive statistics including select psychometric results are detailed. Detailed results from each research question are also presented.

3.1. Missing value and outlier strategies

The percentage of missing values for each of the demographic variables (gender, age, education level, duration of treatment, levels of disease control) are presented in Table I2. The missing values from the item scores were imputed using MICE (Azur et al., 2011) through R. The missing values for the age variable were replaced by the median. The cases that contain missing values with other 4 demographic/medical history variables were deleted, and as a result there were 179 participants in the current sample. In terms of the outliers, there were no outliers detected in this study according to the package ‘dlookr’ (Ryu, 2021) in R, therefore no cases were excluded.

3.2. Descriptive statistics and select psychometrics for scores at the scale, composite, and item-level on the SC ThyPRO-39

This section includes summary statistics and select psychometrics of scores on the SC ThyPRO-39. Distributional descriptive statistics for the mean, median, standard deviation, skewness, and kurtosis, and the percentage of floor and ceiling for 12 scales and the composite scale, are presented in Table I1. It shows that the mean of 12 scale scores ranged from 2.02 to 5.31, and standard deviations were from 2.50 to 3.29. All of the distributions for the 12 scale and composite scores were positively skewed. Nine out of 13 scales showed positive kurtosis, and four scales showed negative kurtosis. The percentage of responses at the floor at the scale level ranged from 0.56% on the tiredness scale and composite scale to 37.43% of impaired social life scale. The percentage of responses at the ceiling were rarely observed from the data, where it was ranged from 0% to 2.23%.

Table I3 presents the descriptive statistics (mean, median, standard deviation, skewness, kurtosis, and percentage of floor and ceiling) for each item. It should be noted that there are three items that are positively worded thus require reverse coding, which are tq3b, tq6g, and tq7h. The tables for descriptive statistics at the item level only present the statistics before reverse coding. Please note that because the ThyPRO-39 does not include the sexual life scale that was on the original ThyPRO, item level tables do not include item labels starting with "tq10" – which are the items related to the impaired sex life.

All items showed positive skewness and most of the items showed positive kurtosis. The percentage at the floor at the item level was ranged from 7.26% to 65.92%, which was more salient than at the scale level. The percentage at the ceiling was very low.

The descriptive statistics for males and females for the 12 scales and composite scale are presented in Table I4. It can be observed that there were more female participants than male participants at the floor at the scale level. Also, both female and males showed negative kurtosis on most of the scales. Item-level descriptive statistics for gender are shown in Table I5 and Table I6.

Table I7 shows the descriptive statistics for two mode of administration groups for the 12 scales and composite scale. Nine out of 13 scales in electronic self-administered group showed negative kurtosis, where there were only four scales in the electronic interview group that showed negative kurtosis, which means the distribution of the responses in the electronic self-administered group was more varied than the distribution of responses in the electronic interview group. The percentage at the floor was much higher in the electronic interview group for the impaired social life scale and the impaired daily life scale, but it was lower for the goiter symptoms scale. Tables I8 and I9 show the descriptive statistics based on the mode of administration at the item level.

Regarding the internal consistency reliability, the Cronbach’s Alphas for five scales were lower than the criterion of 0.7 (Nunnally, 1978). These scales were the hyperthyroid symptoms, hypothyroid symptoms, tiredness, depressivity, and emotional susceptibility scales. In particular, the emotional susceptibility showed the lowest internal consistency reliability among all scales, where the Cronbach’s α was 0.518. In terms of

the internal construct validity, 5 items do not reach the standard of corrected item-total correlation (r > 0.4, Ware & Gandek, 1998). These items are tq1t “您有肚子不舒服吗” (“Had an upset stomach”, r = 0.337), in hyperthyroid symptoms, tq1q “您有对冷较敏感吗” (“Been sensitive to cold”, r = 0.372) in hypothyroid symptoms, tq3b “您有感到精力充沛吗” (“Felt energetic”, r = 0.341) in Tiredness, tq6g “您有感到自信吗” (“Had self-confidence”, r = 0.271) in depressivity, and tq7h “您有感到生活在自己的掌控之中” (“Felt in control of your life”, r = 0.112) in emotional susceptibility, which corresponded to the low internal consistency reliability of the five scales indicated above. Noted that for the 5 items which showed low corrected item-total correlation, the item-total correlations were computed after reverse coding where appropriate; yet three of the 5 items were reverse coded items. Details for the Cronbach’s Alpha and corrected item-total correlation are presented in Table I12.

3.3. RQ 1.1. Distributional characteristics for the 12 scales, composite, and item level scores on the SC ThyPRO-39 for participants overall and separated by subgroups

Distributional characteristics of scores include summary statistics of empirical distributions. Summary statistics characterizing the distributions of the scores on the ThyPRO-39 at the scale, composite, and item level were detailed in Section 3.1. This section focuses on the illustration and tests on the distributional form of the scores on 12 separate scales and on the composite scale of the SC ThyPRO-39, contextualizing the data visualization with the summary statistics, and followed by summary of results from tests of distributional normality. The distributional forms for these sets of scores are shown in the histograms in Figure N1 (Figures N1.1 to N1.13) for the participants overall, in Figure N2 (Figure N2.1 to N2.13) by gender, and from the Figure N3 (Figure N3.1 to N3.13) by mode of administration.

Based on inspection of the histograms, for the participants overall, only the scores on the tiredness and emotional susceptibility scales looked like the normal distribution; for the other scales and the composite scale, the distributions are all right-skewed, and some also come along with an excess of zeros. The skewness from the descriptive statistics in Table I1 are consistent with the observations from the histograms.

When the histograms are displayed by gender in Figure N2 (Figure N2.1 to N2.13), it is clear that females had their scores concentrated on the lower end of the distribution for most of the scales (goiter symptoms, hyperthyroid symptoms, cognitive complaints, anxiety, depressivity, impaired social life, impaired daily life, and cosmetic complaints scales).

According to the histograms displayed by mode of administration groups in Figure N3 (Figure N3.1 to N3.13), a difference in patterns can be found on the goiter symptoms scale, where there are more participants in the electronic self-administered group at the floor than participants in the electronic interview group. In contrast, for the scale scores of hyperthyroid symptoms, tiredness, cognitive complaints, anxiety, depressivity, emotional susceptibility, impaired daily life, cosmetic complaints, and composite scales, it seems that there are more participants in the electronic self-administered groups with scores in the middle or higher end of the distribution than the participants in the electronic interview group.

The results of the sets of one-sample K-S tests for all the scales are shown on Table I10. The null hypothesis of the data being normally distributed was rejected for 11 scales after adjusting for the Bonferroni correction (α = 0.05/13 = 0.0038), except for the emotional susceptibility scale (d = 0.09, p = 0.08) composite scale (d= 0.09, p=0.09). It indicated that except the emotional susceptibility and composite scales, the scores of the 11 scales are not normally distributed.

3.4. RQ 1.2. The CFE in scale and item level scores of Simplified Chinese version of ThyPRO-39

Based on the descriptive statistics shown in Table I1 and the Terwee's definition of significant floor or ceiling effect of 15% (2007), four scales present significant floor effect, and there was no ceiling effect presented at the scale level in this Simplified Chinese ThyPRO-39. Specifically, the four scales with significant floor effects are goiter symptoms scale, social life scale, impaired daily life scale, and cosmetic complaints scale. The impaired social life scale showed the most significant floor effect, where 37.43% of participants obtained the lowest score for this scale.

In terms of the item level, 35 out of 39 items showed significant floor effect (see details in Table I3). Among these items, the floor effect was the highest for item tq9c (65.92%), which was the item of “不能参与日常的活动” (“not be able to participate in life around you”) on the impaired daily life scale. Ceiling effects were also observed for 3 items, excluding the three items with reverse coding.

In terms of the floor effects presented in each administration group, the floor effects presented among participants who did the online survey were on the four scales that showed significant floor effects for overall participants, with an additional scale of eye symptoms. Also, the floor effects were shown on more scales in interview administration mode, which were goiter symptoms scale, cognitive complaints scale, anxiety scale, impaired social life scale, impaired daily life scale, and cosmetic complaints scale (Table I4). When looking at the floor effect based on two gender groups on Table I7, it was shown that the floor effect only presented for males with impaired social life and impaired daily life scales, and the floor effect presented for females on five scales of goiter symptoms, eye symptoms, impaired social life, impaired daily life, and cosmetic complaints. The details about the CFE presented at item level by gender are shown on Table I5 and I6, and the CFE presented at item level by mode of administration is shown on Tables I8 and I9.

3.5. RQ 2.1. The CFE difference at the scale level and the composite between self-identified males and females

Relevant assumption checking procedures were conducted before the set of two-sample proportion tests. First, the gender groups were mutually exclusive because the gender was forced-choice and self-identified by participants while completing the questionnaires. Also, the tiredness, depressivity, emotional susceptibility, and composite scales were excluded from analyses because there were less than 5 cases in both gender groups. The number of cases among males for the hypothyroid symptom scales was also less than 5, but the results are still reported. Further study should consider increased sample size before conducting the two-sample proportion tests (see Table J1 for details).

The proportion tests on the ceiling effect were not required for this part. Table J2 summarizes the results of the two-sample proportion tests for the 9 scales with exclusion

of 3 scales that violated the assumptions. Descriptively, it can be observed that except for the cognitive complaints and anxiety scales, males showed more proportion at the floor than females on the other seven scales. After adjusting for the Bonferroni correction, where α = 0.05/9 = 0.045, there was no gender difference found on the proportion of floor for any of the symptoms and functions scales and the composite scale.

3.6. RQ 2.2. The CFE difference at the scale level and the composite between electronic interview and electronic self-administered group

Because the floor effect between two modes of administration was also tested through two-sample proportion tests, the assumption checking procedure is the same as RQ 2.1. for the gender difference on proportion of floor. The two administration groups were considered to be mutually exclusive and independent during the patient recruitment, because participants can choose to do either interview or an online survey. Next, for the requirement that 5 cases of successes and 5 cases of failure, the tiredness and emotional susceptibility scales have less than 5 cases with the lowest score in both administration groups; therefore, the above two scales were excluded from the two-sample proportion tests. In addition, the number of cases that reached the lowest score was less than 5 for the hyperthyroid symptoms scale in the electronic interview group and depression scale in the electronic self-administered group; the results for these two scales are still reported, but further study should consider increasing the sample size. Details about the results of this assumption checking are shown in Table J3.

Because there is no ceiling effect shown on any scales, the proportion of ceiling effect was not required in the analyses. Analyses focus on the proportion of floor on 10 scales (excluding the tiredness, emotional susceptibility scales, and composite scale with assumption violation), and results indicated that mode of administration difference was detected on the anxiety scale (χ_1^2= 6.43, p = 0.01) and the impaired social life scale (χ_1^2= 8.12, p = 0.004). However, after adjusting for the Bonferroni correction, the mode difference on the proportion of floor for the anxiety scale became insignificant (p = 0.01 > α = 0.05/10 = 0.005). The impaired social life scale also showed the most profound floor effect in electronic interview group (49%) compared to the floor effect in the electronic self-administered group (28%). The other nine scales did not show any significant

difference between administration group for the proportion of floor. Details about the sample estimates, degrees of freedom, and p-value are presented in Table J4.

3.7. RQ 3.1. Comparing modelling strategies for scales with prominent floor effect

Based on the results from RQ 1 and 2 which determined which scales had a prominent floor effect, candidate modelling strategies were first compared for four scales that showed prominent floor effect, which are the goiter symptoms scale, social scale, daily life scale, and cosmetic scale.

The candidate modelling strategies were the six distributional models of ML, Tobit, Poisson, NB, ZIP, and ZINB. The following sections are divided based on the procedures to compare models. Specifically, part 3.7.1 focuses on the assumption checking/diagnostic procedures prior to the model comparison, and part 3.7.2 presents the results of model comparison based on AIC and BIC. In addition, the results based on the comparing of models using LRT for nested models are presented in part 3.7.3, and the results of model comparison using the Vuong test for non-nested models are presented in part 3.7.4. Within each part, results are presented in subsequent order based on the predictor sets: one predictor of gender; one predictor of mode of administration; and full predictor set of all six variables. Other predictor sets, including the three predictors set of the duration of treatment, education, and level of disease control; three predictors set of administration group, gender, and age; and four predictors set of administration group, duration of treatment, control, and gender, which are the predictors sets for checking the consistency of the results of model comparison, were also tested but not reported in the following sections. The overall summary for the section indicates the optimal modelling strategies for these data.

3.7.1. Assumption checking/diagnostic procedures

The assumption checking/diagnostic procedures were conducted within each predictor set individually but were integrated and are summarized together in the next paragraphs. Specifically, the predictor sets are one predictor set of gender, one predictor set of mode of administration, and full predictor set of all six predictors (mode of administration, gender, age, education level, duration of treatment, and level of disease

control). The assumption checking/diagnostic procedures for other predictors sets (the three predictors set of the duration of treatment, education, and level of disease control; three predictors set of administration group, gender, and age; and four predictors set of administration group, duration of treatment, control, and gender) were also conducted but not reported in detail in this section. The assumption checking results for the one predictor model of gender are summarized in Table K5, for the one predictor model of mode of administration in Table K8, for the full predictor model of six predictors in Table K11. The details of the assumption checking/diagnostics are discussed below.

The distributional diagnostics of normal distribution and Poisson distribution were violated for all scale scores with floor effect, and the assumption of NB distribution was violated for the impaired social scale only. The summary of results is shown in Table K1. Specifically, the test results for the K-S tests of the 4 scale scores are shown in Table K2, and the test results of goodness-of-fit tests for Poisson distribution are shown in Table K3. In addition, the goiter symptoms scale, impaired daily life scale, and cosmetic complain scale support the use of NB regression and the scores of impaired social life scale does not support the NB distributional assumption (see Table K4 for test results).

Regarding other assumption checking/diagnostic results, in sum, the assumption checking for the ML and Tobit models were violated in some cases with one predictor sets but were accepted for the full predictor set and other different predictor sets. The assumptions of equal mean and variance were all violated for Poisson and ZIP models, no matter which predictor set it is applied to. This also means that overdispersion was present for all four scales with NB and ZINB models with all predictor sets. Finally, the assumptions of zero-inflation were accepted for all scores of four scales with ZIP models regardless of the predictor sets and were violated in some cases with ZINB models.

The assumption checking/diagnostic results for the six distributional models (ML, Tobit, Poisson, NB, ZIP, and ZINB) under one predictor set of gender are presented in figures from Figure O1 to Figure O4, and in tables from Table K6 to Table K7. For the one predictor set of mode of administration, the assumption checking/diagnostic results for the six distributional models are shown from Figure O14 to Figure O17, and on Table K9 and K10. In terms of the results for the full predictor set, the details are presented from Figure O27 to Figure O30, and from Table K12 to Table K16. Note that, for the

figures of assumption checking results, the figures are presented in tables, where each table contains all assumption checking figures with one predictor set on one scale.

3.7.2. Model comparison results for AIC and BIC for analyzing data with prominent floor effects

Overall, according to the AIC and BIC for all predictors sets, the best model for scales with prominent floor effect was either ZINB or NB regression models, and the worst model was either ML or Poisson regression models.

Specifically, the results of AIC and BIC for the single predictor models of gender for the raw scores of each of the four scales (goiter, impaired social life, impaired daily life, cosmetic complaints) are shown on Table L1 and Table L2. The results of AIC and BIC for the one predictor models of mode of administration are presented in Tables L3 and L4. The AIC and BIC of the full model with all six predictors (gender, mode of administration, age, education level, duration of treatment, level of disease control) are presented in Tables L5 and L6. All six predictor sets (including the three predictors sets that are not discussed in detail) yielded the same results, where the ZINB and NB regression models were recommended to be used to analyze the four scales, and ML and Poisson regression models were the worst models.

3.7.3. Model comparison for nested models using LRTs for scales with prominent floor effect

Overall, NB models (including NB regression and ZINB regression models) perform better than Poisson models (Poisson and ZIP regression models) according to the results from LRTs with every set of predictors across different scales with prominent floor effects. Details of the test results are shown from Table L7 to Table L9.

3.7.4. Model comparison for non-nested models using Vuong tests for scales with prominent floor effect

The BIC-corrected Vuong tests were used to compare models between ordinary count models (Poisson and NB) and zero-inflated models (ZIP and ZINB). In contrast with the results from the LRTs, where all of the tests with different sets of predictors and varying scales showed consistent results, the results from Vuong tests did not show

consistent results of one model being better than the other. Specifically, most of the cases indicated the advantage of ZIP over Poisson regression models, and some cases showed no difference between ZIP and Poisson regression models. Regarding the comparison between the NB and ZINB models, NB showed better performance in some cases, and there were no differences between NB and ZINB in some other cases, and ZINB only showed better performance on one occasion for impaired daily life scale based on the results of Vuong tests. Details of the test results are shown in Table L10 to Table L12.

3.7.5. Summary of the model comparison based on different criteria and tests for scales with prominent floor effects

For the analysis of data on SC ThyPRO-39 scales with significant floor effect, according to the results from AIC, BIC, LRTs and Vuong tests with various predictor sets and different scales, the conclusion can be made that Poisson and ML regression models had the worst performance compared to other models, NB and ZINB had the best performance than the other models, and Tobit regression and ZIP models were the ones in between. The summary of results including different comparison sets are shown in Table L13 to Table L15.

3.8. RQ 3.2. Comparing modelling strategies for scales without prominent floor effect and the composite scale

The strategies for comparing models for the eight SC ThyPRO-39 symptoms and functions scales without prominent floor effects and for the composite scale were the same as the four scales with prominent floor effect. Models using three sets of predictors were fit to the six distributional models (ML, Tobit, Poisson, NB, ZIP, and ZINB). The predictors sets are the one predictor of gender, one predictor of mode, and full model of all six predictors. Part 3.8.1 presents the assumption checking/diagnostic procedures for these scales and composite scale, part 3.8.2 shows the results of model comparison according to AIC and BIC. Part 3.8.3 summarizes the results from LRTs, part 3.8.4 presents the results from Vuong tests, and part 3.4.5 shows the summary of results.

3.8.1. Assumption checking/diagnostic procedures

The assumption checking/diagnostic procedures are the same for SC ThyPRO-39 scales without significant floor effect, including the composite scale. They were also conducted within each predictor sets (one predictor set of gender, one predictor set of mode of administration, and full predictor set of all six predictors) separately, but results are integrated in this section. The assumptions/diagnostics for other predictors sets were also checked, but not reported in detail in this section. The summary of results for the eight scales without floor effect can be found in Table K21 for the one predictor set of gender, and Table K24 for the one predictor set of mode of administration, and Table K27 for the full predictor set. The summary of results for the composite scale are shown in Table K33. The assumption checking/diagnostic results are reported below.

First, the results of distributional diagnostics for the eight scales without floor effect and composite scale are summarized in Table K17. The results from one-sample K-S tests shown in Table K18 indicated the violation of normal distribution for most scales without floor effect, but not for the emotional susceptibility scale and composite scale. The goodness-of-fit tests of Poisson distribution showed that only the raw score of emotional susceptibility can be fit to the Poisson distribution, and other scales and composite scale do not support the distributional assumption of Poisson (see Table K19 for details). Also, the results from goodness-of-fit tests of NB distribution indicated that five scales supported the assumption of NB distribution (hyperthyroid symptom scale, tiredness scale, depressivity scale, and emotional susceptibility scale), whereas the other scales did not support the diagnosis of NB distribution (the hyperthyroid symptoms scale, eye symptoms scales, cognitive complaints scale, and composite scale). Details are shown in Table K20.

Regarding other assumption checking/diagnostic results, similar to the scales with floor effect, the assumption checking for the ML and Tobit models for symptoms and functions scales and composite scale without floor effect were violated in some cases with one predictor sets but was accepted for the full predictor set and other different predictors sets. The assumption of equal mean and variance was only accepted for emotional susceptibility scale and was violated for other scales and composite scale with Poisson and ZIP models. The test of overdispersion also indicated that the existence of overdispersion for scores of most of the scales except for the emotional susceptibility

scale with NB and ZINB models regardless of the predictors sets. Finally, the assumptions/diagnostics of zero-inflation were accepted for most of the scales and composite scale except for the tiredness scale with ZIP models and ZINB models across various predictors sets.

The assumption checking/diagnostic results of the eight symptoms and functions scales and composite scale without significant floor effect for the six distributional models (ML, Tobit, Poisson, NB, ZIP, and ZINB) under one predictor set of gender are presented from Figure O5 to Figure O13, and in Table K22 and Table K23. Regarding the one predictor set of mode of administration, the assumption checking/diagnostic results for the five distributional models are shown from Figure O18 to Figure O26, and in Table K25 and K26. In terms of the results for the full predictor set, the details are presented from Figure O31 to Figure O39, and from Table K28 to Table K32.

3.8.2. Model comparison results for AIC and BIC for analyzing data without prominent floor effect and the composite scale

The results of AIC and BIC for the eight scales without prominent floor effect and composite scale are presented on Table L16 and L17 with one predictor set of gender, Table L18 and L19 with one predictor set of mode of administration, and Table L20 and L21 with full predictor set. Overall, the NB was most frequently shown as the best model, followed by ZINB. The emotional susceptibility scale can be best fit with Poisson models in most cases, which is different from other scales. Regarding the worst model, the results were also varied, as the Poisson regression models were shown as the worst models most frequently based on the result from AIC and BIC, followed by ZIP, and ML regression models. The ZINB model was the worst model for emotional susceptibility in all cases.

3.8.3. Model comparison for nested models using LRTs for scales without prominent floor effects and the composite scale

According to the results from LRTs on the nine SC ThyPRO-39 scales without prominent floor effects (eight symptoms and functions scales and one composite scale) shown in Tables L22, L23, and L24, the NB regression model performed better than the Poisson regression model on most of the scales, except for the emotional susceptibility

scale and sometimes the tiredness scale. The ZINB model also performed better than ZIP on most of the scales except for the tiredness and emotional susceptibility scales.

3.8.4. Model comparison for non-nested models using Vuong tests for scales without prominent floor effects and the composite scale

When comparing the ordinary count models (Poisson and NB regression models) to the zero-inflated models (ZIP and ZINB) using BIC-corrected Vuong tests, unlike the scales with significant floor effect, the zero-inflated models did not show many advantages over ordinary count models for SC ThyPRO-39 scales without significant floor effect and for the composite scale. Specifically, the ZIP had better performance in some cases, but worse performance in some other cases, and there were also some cases that there was no difference between Poisson and ZIP. In terms of the comparison between NB and ZINB, NB models showed some advantages in most of the cases, and in some other cases, there was no difference between NB and ZINB models. It was also noted that the ZINB did not perform better than NB models in all cases. Table L25 to L27 provide details.

3.8.5. Summary of the results comparing models based on different criteria and tests for scales without prominent floor effects and the composite scale

Based on the results from AIC, BIC, LRTs and Vuong tests on the eight SC ThyPRO-39 scales without significant floor effect and composite scale, there were mixed results, and it was summarized based on the predictor sets shown in Tables L28 to L30. However, NB still had a major advantage on most of the scales, followed by ZINB models, except for the analysis of the emotional susceptibility scale, where Poisson was most well fitted.

3.9. Effects of predictors

For these models, based on the results from the model comparison reported on RQ 3.1. & RQ 3.2., preferred modelling strategies were identified for each SC ThyPRO-39 scale with different models. Results are first summarized for the symptoms and functions scales, then separately for the composite scale.

The main effect of gender on SC ThyPRO-39 test scores of four scales with floor effect is shown in Table M1, the main effect of the mode of administration on the raw scores of the four scales (goiter symptoms, impaired social life, impaired daily life, and cosmetic complaints scales) with prominent floor effects are shown on Table M2, and the effects of the six predictors on the raw scores of the four scales are shown on Table M3 to M6. In addition, Table M7 presents the main effect of gender on test scores of eight scales without prominent floor effect; Table M8 shows the main effect of mode of administration; and Table M9 to M16 indicates the main effects of predictors in the full model. Lastly, the main effect of gender, mode of administration, and the full model on test scores of the composite scale are shown in Table M17. Specifically, each table presents the estimated coefficient, estimated SE, z score, p-value, and lower level and upper level of 95% confidence interval of the NCP estimate when the data were fit to the models of NB and ZINB selected by RQ 3.1. Moreover, each table presents the results of the chi-squared test for the effect of the overall model compared to the null model, which also indicates whether the effect of the predictor(s) is/are statistically significant.

3.9.1. Effects of predictors on scales with prominent floor effect

The main effect of gender in single predictor model

According to the results shown in Table M1, there was a gender difference on two SC ThyPRO-39 scales with prominent floor effect (χ_1^2 = 20.38, p = < 0.001 for the goiter symptoms scale with NB model, and χ_1^2 = 23.65, p < 0.001 for the goiter symptoms scale with ZINB model; χ_1^2 = 11.59, p < 0.001 for the impaired daily life scale with NB model, and χ_1^2 = 13.16, p < 0.001 for the impaired daily life scale with ZINB model). The other two scales did not show any difference between males and females with both NB models (χ_1^2 = 2.55, p = 0.11 for the impaired social life scale with NB model, χ_1^2 = 3.06, p = 0.08 for the impaired social life scale with ZINB model; χ_1^2 = 3.80, p = 0.05 for the cosmetic complaints scale with NB model, χ_1^2 = 3.81, p = 0.05 for the cosmetic complaints scale with ZINB model). Moreover, females showed lower scores than males on all scales with significant results. When looking at the zero-inflation part of the ZINB models, gender did not significantly predict the zero-inflation on any of the scales without prominent floor effects.

The main effect of mode of administration in single predictor model

Regarding the main effect of mode of administration on SC ThyPRO-39 scales with prominent floor effect presented in Table M2, three out of four scales showed mode effects, except the goiter symptoms scale with NB model (χ_1^2 = 0.10, p = 0.76). More specifically, for scales with significant mode effects, participants in the online self-administered survey group showed higher scores than the participants in electronic interview group. In addition, mode of administration was not a predictor to the zero-inflation in the ZINB models for all four scales.

The effects in the full model

The results of the effects for the full model with all six predictors on four SC ThyPRO-39 scales with prominent floor effect are shown on Table M3 to Table M7. Based on the results of model comparison for the full model on scales with prominent floor effect from RQ 3.1, the results were all presented in both NB and ZINB models for the four scales with floor effect.

Overall, according to the chi-squared tests comparing the full model to the null model, there were significant results for all scales with significant floor effect when fitted the data into NB and ZINB models.

In terms of the significance level of the predictors on the ThyPRO-39 test scores, the results showed that gender and mode of administration significantly predicted the test scores in the full model in most of the cases with both NB and ZINB models. Similar to the direction in the single predictor models, females showed lower scores than males when gender had a significant effect on test scores with floor effect. Participants in the electronic interview group showed lower scores than participants in the electronic self-administered group in all cases where the mode of administration significantly predicted the test scores among scales with floor effect. Other predictors, including age, education level, duration of treatment and level of disease control can also significantly predict some of the test scores with floor effect. In particular, it was noted that the raw scores tend to increase as age decreases, as education level decreases, as the duration of treatment increases, and as the level of disease control decreases. In addition, the mode of administration, age, education level and level of disease control also significantly predict the zero-inflation of test scores of some scales with prominent floor effects.

3.9.2. Effects of predictors on scales without prominent floor effect

The main effect of gender in single predictor model

In terms of the gender effect on SC ThyPRO-39 scales without floor effect, most of the scales did not show any gender difference according to the chi-squared tests results shown on Table M8, where only the hypothyroid symptoms scale with NB and ZINB models (χ_1^2 = 3.94, p = 0.047 with NB model and χ_1^2 = 26.05, p = 0.03 with ZINB model) and depressivity scale with ZINB model (χ_1^2 = 41.89, p = 0.047) showed significant results. The raw scores of hypothyroid symptoms scale was higher in females than in males, and the raw score of depressivity scale was higher in males than in females. The estimated coefficient of the zero-inflated part of ZINB model for the hypothyroid scale was not significant (estimate = 6.83, p = 0.0.90 for the zero-inflated part). And the estimated coefficients of both the count and zero-inflated parts of the ZINB model for depressivity scale were not significant (estimate = -0.01, p = 0.90 for the count part; estimate = -2.03, p = 0.32 for the zero-inflated part).

The main effect of mode of administration in single predictor model

In contrast with the non-significant effect of gender on most of the ThyPRO-39 scales without floor effect, the mode effect between self-administration and interview showed greater difference on most of the scales without floor effect, where only eye symptoms scale with NB model did not show any mode effect (χ_1^2 = 2.43, p = 0.12). In addition, for the effects of the six predictors set, all scales showed significant results for the overall models, where the participants in the electronic self-administered group showed higher scores compared to participants in the electronic interview group across scales with significant results. Table M9 provides details.

The effects of all six predictors in the full model

The results of the effects of gender, mode, and all six predictors on eight SC ThyPRO-39 scales without prominent floor effect are presented in Tables M10 to M16. The tables contain the same statistics as the tables for scales with prominent floor effect. However, based on results of comparing models from RQ 3.2, for the emotional susceptibility scale, because the Poisson regression model were reported as the best models among all six distributional models, followed by ZIP model, the effects of

predictors were presented under Poisson regression and ZIP models for this particular scale, and the rest of the scales, effects were still reported under NB and ZINB models.

Overall, the full model had a significant effect on most SC ThyPRO-39 scales without significant floor effect, except for the eye symptoms scale with NB model. The significance levels of the partial effects of different predictors are varied across different scales. Similar to the results shown in single predictor sets, gender did not have significant impact on test scores of most scales without floor effect in the full model with both NB and ZINB models (with Poisson and ZIP for the emotional susceptibility scale), but mode of administration significantly predicted the test scores in most of the scales in the full model. Among scales without floor effect with significant coefficients for mode of administration, participants in the electronic self-administered group tend to have higher scores than participants in the electronic interview group. In addition, other predictors, including age, education level, and level of disease control, can also significantly predict the test scores without prominent floor effects. More specifically, controlling for the effect of other variables, all of the age, education level, and level of disease control had negative relationship with test scores in analyses cases that showed significant effects. That said, the results indicated that, controlling for the other variables in the model, none of the six predictors can predict the zero-inflation of the test scores in all scales without prominent floor effects.

In addition to the significant results that were mentioned previously, it was also noted that regarding the eye symptoms and tiredness scales, the statistics of the partial coefficients in the six predictor set with the ZINB models cannot be shown. The possible reason for the missing results and solutions to this problem are discussed in the following Discussion section of 4.4.3.

3.9.3. Main effects on composite scale

The results of the main effects of gender, mode, and overall effect of all six predictors on the SC ThyPRO-39 composite scale are presented in Table M17 for the mian effect of gender, Table M18 for the main effect of mode of administration, and Table M19 for the overall effects of all six predictors. Specifically, there was no gender effect on the composite scale with NB model (χ_1^2 = 1.52, p = 0.22) as well as the ZINB model (χ_1^2 = 55.00, p = 0.15). There was a significant effect of mode of administration on

composite scale with both NB model (χ_1^2 = 22.68, p < 0.001), and ZINB model (χ_1^2 = 26.49, p < 0.001). In both NB and ZINB models, participants in the electronic self-administered group showed higher scores than participants in the electronic interview group. For the overall effects of all six predictors, the model was statistically significant with both NB model (χ_6^2 = 45.38, p < 0.001) and ZINB model (χ_6^2 = 55.00, p < 0.001). In particular, the partial coefficients were significant for mode of administration (estimate = 0.39, p < 0.001), age (estimate = -0.01, p = 0.04), and level of disease control (estimate = -0.11, p < 0.001) with NB model. This indicates that participants in the electronic interview group show lower composite scores than participants in the electronic self-administered group, and the composite scores increase as age decreases, and as the level of disease control decreases. In addition, similar to the eye symptoms scale, the statistics of the coefficients for the predictors on the composite scale with ZINB model are not shown. The reasons and possible solutions are also discussed in the Discussion section 4.4.3.

Chapter 4. Discussion

The current study examined a newly-translated Simplified Chinese version of the Thyroid-Specific Patient-Reported Outcome Short Form (SC ThyPRO-39) among 179 patients with thyroid diseases who speak Mandarin in Mainland China. The study focused on select measurement properties and analytic strategies for the analysis of data from the SC ThyPRO-39. The discussions in this chapter are based on the results from the research questions pertaining to scores and analysis of data on the SC ThyPRO-39. The distributional characteristics of the data and the presence of floor effect and other measurement properties are discussed first in section 4.1. The effect of gender and mode of administration on the proportion of floor effects are discussed in section 4.2. The best model(s) for the scales with prominent floor effects are discussed, as well as the best model(s) for the scales without prominent floor effects and for the composite scale are discussed in section 4.3. The findings regarding main effects of gender and mode of administration are reviewed in single as well as in multiple predictor models, and can be found in section 4.4. Other considerations for the SC ThyPRO-39 are indicated in section 4.5. Implications of this study are discussed in section 4.6. Finally, limitations and future directions are explored in section 4.7 and 4.8.

4.1. Distributional characteristics, floor effects, and other measurement properties

Regarding the distributional features of the SC ThyPRO-39 data for participants overall, it was found that most distributions of the SC ThyPRO-39 scores at the item and scale levels were not normally distributed according to the histograms and descriptive statistics. Additionally, only the emotional susceptibility and composite scores can be fit to a normal distribution based on the K-S tests after Bonferroni correction for the statistical significance.

A significant floor effect was present in scores on most of the items in SC ThyPRO-39 (35 out of 39 items), and some of the scales (goiter symptoms scale, impaired social life scale, impaired daily life scale, and cosmetic complaints scale), which was consistent with the results in Wong et al.'s study (2018) conducted in Hong Kong for the TC ThyPRO-39. However, as mentioned in the introduction, a discussion regarding

the presence of the floor effects was not found in Watt et al.'s publications of the original Danish version of ThyPRO (Watt et al., 2009) and ThyPRO-39 (Watt et al., 2015). The presence of floor effect on the English version of ThyPRO was only reported among Asian thyroid patients in Singapore (Liew et al., 2021). It is unknown whether the floor effect was not an issue to be reported in the original Danish version of ThyPRO and ThyPRO-39, or it is just that the researchers did not focus on the possibility of a floor effect. As indicated earlier, the presence of floor effects may depend on the population of the study participants. It is possible that the floor effects were more salient among Chinese or Asian populations than among Western populations (Szende & Williams, 2004; Lubetkin et al., 2005). However, if it was an issue that was present but was not reported, the floor effects cannot be ignored as it is a significant aspect of measurement properties, as indicated by Terwee et al. (2007). The importance of reporting the presence of floor effect was also confirmed in this study, where the presence of floor effects can be an indicator of the importance of applying different modelling strategies other than models based on a normal distribution.

Seven out of 12 SC ThyPRO-39 scales as well as the composite scale showed satisfactory internal consistency reliability. As indicated in the results section, the five SC ThyPRO-39 scales with low internal consistency reliability also contained five items with low internal construct validity. It is also the same problem shown on TC ThyPRO-39 according to Wong et al. (2018). The low construct validity of the five items may in part due to the reason that indicated by Wong et al. (2018) in their study about TC ThyPRO-39, as they stated that sometimes concepts in the ThyPRO-39 may be culturally specific and not translatable for Asian populations, such that those items may only be applied to European populations where this questionnaire is originally developed.

The cause of the low internal construct validity can also be explored by the post-questionnaire interviews and qualitative questionnaire in this study. Content analysis of the qualitative data showed that 35 out of 179 participants reported they have difficulty in understanding some of the items, 22 participants indicated that the reason why they did not understand those items was that they felt like those items were not related to their thyroid diseases. Moreover, some participants explained that they already had those symptoms before they had thyroid disease, therefore showing symptoms of some items was not an indication of their quality of life that related to thyroid diseases. Items that were mentioned by participants included some items that showed low internal construct

validity such as “had an upset stomach?”, “had self-confidence”, and “felt in control of your life”. Also, according to the post-questionnaire interview, 92 out of 179 participants reported that they did not relate self-confidence to their thyroid diseases in the qualitative questionnaire. Therefore, the low construct validity of those items and related low internal consistency of the scales that contain those items may also be due to the fact that participants did not relate the symptoms of those items to their thyroid diseases.

In addition, among the five items which showed low internal construct validity, three of them are items with reverse coding. It may point out the problems with the three reverse worded items. Previous studies (Zhang et al., 2016) compared the model with both positively worded and reverse worded items to the models with all positively worded items and all negatively worded items; results indicated that the model with both positively worded and reverse worded items performs worse than the other two models. The results further suggested that the use of reverse worded items can be a disadvantage, because the reverse worded items can negatively affect the factor structure, and more complicated models were necessary to be used in order to reach a better fit.

In terms of the results of floor effects for the participants overall, the Cronbach’s Alpha, the item-total correlation, and number of SC ThyPRO-39 scales or items that showed significant floor effects were compared to the results from Wong et al.’s study for the TC ThyPRO-39 (2018), and it showed that the Simplified Chinese version in this study performs better in all three domains overall. The differences comparing the results from the current study with those from Wong et al. (2018) are summarized in Table I12. Details are shown in Table I13. More specifically, where there were 35 items on the SC ThyPRO-39 that showed significant floor effect in the current study, all of the items showed significant floor effect on the TC ThyPRO-39 in Wong et al.’s study (2018). Also, in addition to the SC ThyPRO-39 scales with prominent floor effects in the current study, there were also four more scales which showed significant floor effect on the TC ThyPRO-39 (Wong et al., 2018), namely the hypothyroid symptoms, eye symptoms cognitive complaints, and anxiety scales.

In addition to the five scales with low internal consistency reliability in the Simplified Chinese version present in the current study, the eye symptoms scale also showed low internal consistency reliability (Cronbach’s $\alpha = 0.680$) in the TC ThyPRO-39

(Wong et al., 2018). In addition, there were three more items in the TC ThyPRO-39 that showed low internal construct validity, which are tq1m “您有出很多汗的倾向吗” (“Had a tendency to sweat a lot”, r = 0.352), tq1ee “您有皮肤发痒吗” (“Had itchy skin”, r = 0.385), and tq1w “您有眼睛干燥或眼睛里有异物的感觉吗” (“Had the sensation of dryness or ‘grittiness’ in the eyes”, r = 0.351) (Wong et al., 2018).

Regarding the better performance of floor effect, internal construct validity, and internal consistency reliability of SC ThyPRO-39 in this study compared to the TC ThyPRO-39 in Wong et al.’s study (2018), one possible explanation was that the sample size is relatively small in this study compared to Wong et al.’s study (2018), where there were 179 patient-participants in this study, and there were 308 participants in Wong et al.’s study (2018). In addition, the distributions of the thyroid diseases are different between two samples in Hong Kong and Mainland China (see Table I11). Whereas there were 93.8% of participants who were having thyroid nodules and non-toxic goiter in Wong et al.’s sample (2018), there were only 44.0% participants who had the same thyroid diseases in this sample. As mentioned by Wong et al. (2018), patients with thyroid nodules in their sample had milder symptoms than other thyroid diseases that may not impair patients’ HRQOL. It also corresponded to the patients in this study, where there were more than half of the patients with thyroid nodules and nontoxic goiter (59.3%) stated that they did not require any treatment.

In sum, this study considered the distribution of the SC ThyPRO-39 scores and in terms of satisfaction of HRQOL questionnaire measurement properties of CFE, internal construct validity and internal consistency reliability. Overall, the current study showed a better performance of SC ThyPRO-39 scores on these measurement properties compared to the TC ThyPRO-39 in Wong et al.’s study (2018). That said, importantly, with the exception of there being documented nonnormality on many scales, there was a presence of significant floor effect on some scale scores in this study, which means participants showed good health conditions in some health domains related to thyroid diseases. Although internal consistency reliability was acceptable on many scales, the low construct validity and low internal consistency reliability on some scales pointed out problems of some items; improvement to items on those scales may be necessary.

4.2. Gender and mode of administration differences on the CFE

For the difference of floor effects between males and females, it was found that females tend to show significantly more floor effects on goiter symptoms and impaired daily life symptoms scales in the SC ThyPRO-39. After adjusting for the Bonferroni correction, no gender difference was found on the proportion of floor. The results were in part consistent with previous studies, where it was shown that there was no gender difference among patients with Grave's disease (Delfino et al., 2017). Also, another study only indicated a significant association between gender and cosmetic complaints scale scores in ThyPRO (Bukvic et al., 2015), which was not found in this study. It is unknown that whether the gender effect on floor was influenced by other factors. Studies conducted in China have found that other demographic variables, such as income (Lee et al., 2020), regional difference (rural versus urban areas) (Rong et al., 2020), and the number of comorbidities (Dong et al., 2020) can moderate the gender effects on scores of HROQL measurements. Furthermore, results from chi-squared test and two-samples Welch's t tests of the gender effect on other predictors in this study showed that gender was significantly associated with mode of administration (χ_1^2 = 4.65, p = 0.04), and there were significant gender differences on age ($t_{113.8}$ = 17.30, p < 0.001), and education level ($t_{96.60}$ = 7.214, p = 0.006). In particular, in this sample, men were significantly younger than women, and men were significantly less educated than women. This is an indication that gender is significantly related and may interact with other variables that are included or not included in this study, in such a way that may affect its effect on the proportion of floor of QOL scores. Future studies can investigate other variables and the effects of interaction terms in order to better understand the gender difference on floor effects of SC ThyPRO-39.

In terms of the mode of administration effects on the proportion of floor, the chi-squared tests comparing the proportions of floor effects between the two mode of administration groups showed non-significant results for most of the SC ThyPRO-39 scales. It can be inferred that the mode did not have an impact on the presence of the floor effect on the SC ThyPRO-39 in most of the cases, which means participants' choice of extreme options were not affected by whether they completed the SC ThyPRO-39 in interview or self-administered mode in most of the cases.

Regarding the two SC ThyPRO-39 scales showing a significant mode of administration effect after Bonferroni adjustment, which was the impaired social life scales, participants showed a more salient floor effect in the electronic interview group rather than the electronic self-administered group. These results indicated that there may be a presence of extreme response bias (Cook, 2010) for the social related health domains among participants in the electronic interview group than in the electronic self-administered group. In other words, it may be that participants in this study tended to select the extreme options to emphasize their best health condition related to their interaction with others. One study investigating the mode effect also found the mode effect present for items related to psychosocial health (Hoebel et al., 2014). It is reasonable that participants in a social interaction context, in this case that is the interview, tried to show better social-related health than people who only completed the self-report questionnaire. These effects may have been salient especially for items like “felt you were a burden to other people”, and “had difficulty being together with other people” in the impaired social life scale in the SC ThyPRO-39.

4.3. Comparing modelling strategies

In this section, discussions are made based on the results sections of comparing modelling strategies for scales with and without prominent floor effects.

4.3.1. Comparing models for scales with prominent floor effect

Based on the results from RQ 3.1 of comparing the model fit for the ThyPRO-39 scale scores with prominent floor effects, the NB and ZINB are shown to have the best performance on model fit in the analysis of data of all four scales with floor effect, regardless of the predictor sets that were considered. In contrast, the ML regression model was the worst model across different predictor sets. Therefore, the conclusion can be made that the NB and ZINB regression models are the best models to analyze data with prominent floor effects, and the ML model is the worst. There are several good examples for using NB and ZINB regression models in the published literature. Previous studies that applied NB and ZINB regression models mostly dealt with count outcome with health-related data, such as the number of falls (Ullah, Finch, & Day, 2010), difficulty in activities of daily living (Zaninotto & Falaschetti, 2011), and the number of

days with unhealthy QOL (Lyu & Wolinsky, 2017). There are also some studies which fit data into NB and/or ZINB models for HRQOL scores. For example, Alemu et al. (2020) in their study examined the effect of the QOL domain outcome directly using the NB model because the data were right-skewed with a significant floor effect. In another study, the frequency responses on the QOL scale were also modelled using the NB model with data that is over-dispersed and without excess zeros (Schneider & Stone, 2016). The effect of age on the QOL of kidney function was modelled through NB regression by adding one additional variable of QOL deficits (Canney et al., 2018).

Many studies investigating the discriminative/known-group validity and responsiveness of the HRQOL measurements involved the comparison between groups. For the scales with CFE, some studies provided alternative ways to analyze data instead of the tests such as ML regression or t-tests that are based on the assumption of normality. For example, Zhu and Gonzalez (2017), Huang et al. (2017), and Wang and Zhang (2009) in their studies used Tobit regression models to analyze HRQOL scores due to the presence of floor effect. Bunevicius (2017) and Kashkouli et al. (2017) used the Mann-Whitney test, which is a non-parametric test that avoids the assumption of normal distribution, to evaluate the known-group validity of some scales in SF-36 and GO-QOL. Ferreira et al. (2016) also fit the data of EQ-5D into non-parametric models to measure the known-group validity. However, there were still some studies that used the tests based on normal distribution such as the ANOVA, linear-mixed model, and t-tests with the presence of the CFE of the measurement (Bharmal & Thomas, 2006; Chen et al., 2014; Ponto et al., 2011; Youssof et al., 2017; Zhou et al., 2013).

Wong et al. (2015) also used the t-tests on means for known-group comparison for construct validity where significant floor effects were present on most of the scales of the Traditional Chinese version of ThyPRO-39. Although the known-group comparison in Wong et al.'s study was based on the transformed scores rather than the raw scores, it is still possible that the transformed scores were still highly-skewed. It may require careful consideration of the choice of tests that are based on the normal distribution or not. According to the results from this study, the models based on the assumption of normal distribution (i.e., ML models) yielded the worst fit for the data with floor effect and right-skew. Similar to the ML models, the t-tests on means are also based on the assumption of normal distribution, therefore more consideration should be taken based on the model fit.

Overall, the results of the current study suggest that the NB and ZINB regression models can be used for questionnaire data with floor effect in the future, and the models that are based on the assumption of normal distribution, such as the ML regression model may need to be carefully avoided.

4.3.2. Comparing models for scales without prominent floor effect

According to results from RQ 3.2 of comparing models for SC ThyPRO-39 scale scores without prominent floor effect, although the results were not consistent among all scales, it was noticed that NB still showed the best performance on most of the scales, followed by ZINB models. In contrast, the ML model was also shown as having the worst model fit among all six distributional models in many cases, which was similar to the results among scales with prominent floor effect. The only scale that was significantly different from other scales was the emotional susceptibility scale, where the best model was the Poisson regression model, and the worst model was the ZINB model. The tiredness scale also showed slightly different from other scales, where in the full predictor model, the NB and ZINB models did not show significant advantages over Poisson and ZIP models. It was consistent with the results from distributional characteristics, in which the distribution of the raw scores on the tiredness and emotional susceptibility scales looked like the normal distribution, and little floor effects were observed on these two scales (0.56% for tiredness scale and 1.68% for emotional susceptibility scale).

In addition to identifying optimal modelling strategies, the results also highlighted the importance of assumption checking/diagnostic procedures. In reviewing the literature, it was noted that the assumption checking procedures were rarely mentioned in published papers, but in this study, it was determined as a critical procedure that cannot be ignored. To be more specific, in this study, the results of assumption checking/diagnostics were closely related to the results of model comparison. In most of the cases, for the models in which the assumptions could not be met, the model performance was not satisfactory as well, such as the assumption of equivalence between mean and variance for the Poisson models. For this specific assumption, only the data for the emotional susceptibility scale met this assumption, and it also showed that the Poisson regression model was the best model for this scale. Also, the reason why in some cases that the NB and ZINB are indistinguishable regarding the model fit

was also reflected on the assumption checking/diagnostic procedures, where the zero-inflation was shown in some cases but not the others that varied across different scales and predictors sets.

Therefore, a conclusion can be made that NB is still recommended to be used, even for the scales without significant floor effect, and it was not wise to choose models based on normal distribution to analyze data like these. Distributional figures and assumption checking can help decide the best model. Every step in data analysis cannot be disregarded and skipped.

4.4. Effect of predictors

Among several predictors examined in this study, the discussion of the effects focuses on the gender and mode of administration. The effect of the overall model with all six predictors were also investigated but are not discussed in detail because it is not the focus of this study.

4.4.1. The main effect of gender

Overall, the gender difference was shown on goiter symptoms and impaired daily life for scales that showed significant floor effects and hypothyroid symptoms, and on depressivity scales that did not show significant floor effects. Among the four scales showing significant gender differences, females showed better quality of life in most of the scales except for the hypothyroid symptoms scale. This result was inconsistent with some previous studies that females showed poorer HRQOL than males (Boerma et al., 2016; Cherepanov et al., 2010; Izadnegahdar et al., 2014). The slightly but not significantly higher QOL for females on cosmetic complaints in this study (χ_1^2 = 3.80, p = 0.05) were also not consistent with results from a previous study for ThyPRO conducted among thyroid patients in Serbia (Bukvic et al., 2015), where Bukvic et al (2015) found that females showed significantly lower QOL than males regarding the cosmetic complaints. In addition, the results were not consistent with some of the previous studies that females tend to show lower QOL on depression and anxiety (Shafie et al., 2021; Tlusta et al., 2009). The results are mostly consistent with the RQ 2.2. of the gender difference on the floor effect, where gender had a significant effect on the floor effects of goiter symptoms and impaired daily life scales. One interesting finding is that, although

females showed worse QOL than males on the hypothyroid symptoms scale, there was still more proportion of floor for females than males (7% versus 5%), although the difference of proportion was not significant. It means that both males and females show some symptoms related to the hypothyroid function to some extent, and females showed more severe symptoms that was not related to their choices to show their full health or not.

The results in the current thesis may have been affected by the imbalanced distribution of gender in this particular sample, where there were more than twice of the females than males (123:53). Further study can include more males in this study to see whether the results are the same. The gender role of the female researcher in the current thesis may have also impacted participants' responses to the questionnaire. In addition, the opposite direction of gender effect in this study compared to previous studies can also be due to cultural differences. Researchers have indicated that culture is a critical factor influencing the QOL (Shek, 2010). While the Western societies emphasize happiness and satisfaction for the quality of life, the quality of life among the Chinese population focuses more on forbearance, endurance and humility that are derived from Confucianism, Buddhism, and Taoism.

4.4.2. The main effect of mode of administration

Regarding the main effect of mode of administration, although there was no difference of the mode of administration on the choices of the lowest scores for most SC ThyPRO-39 scales based on the results from RQ 2.2, participants in the electronic self-administered group showed poorer QOL than participants in the electronic interview group on all scales, including the composite scales. It indicated that almost all scales showed significant mode effects, no matter which scale showed significant floor effect or not, which may explain why the proportion of floor did not differ by mode of administration in most cases.

The results confirmed the critical role of mode of administration on responding to the HRQOL measurements, where participants in interviews reported to have better QOL compared to participants in the self-administered mode. The results were consistent with Bowling's theory about the acquiescence bias (2005) and Tourangeau and Smith's theory about social desirability bias (1996). In other words, participants may

be more drawn to show better health condition while completing the HRQOL in interviews that are possibly influenced by acquiescence bias and social desirability bias.

The existence of mode effects may raise the discussion about whether research should be conducted with only one of the modes of administration in order to control for the variability and get a consistent result. However, both modes of administration had their strengths and drawbacks. Apparently, the response rate was one of the most important issues for self-administration survey. In this study, all of the missing values from demographic variables of gender, age, duration of treatment, and education level are from electronic self-administered group. In contrast, studies also indicated the advantage of self-administered mode, where it resulted in more reports of morbidity, disability, and socially undesirable behaviors compared to interview mode (Bergner et al., 1981; Unruh et al., 2003).

In terms of the interview mode of administration, although responses from participants in the interview mode may be influenced by acquiescence bias and social desirability bias, studies also indicated the importance of interview mode. For example, Unruh et al. (2003) in their study pointed out that in some cases where participants have problems with vision, and who have comorbidity with other diseases that stopped them from completing questionnaires in self-administered mode, the interview modes may be particularly important. The prevalence of thyroid dysfunction was 25% among older adults in the U.S. population according to Diab et al., (2019), which means there may be a considerable proportion of thyroid patients who are older adults with different levels of disabilities. Therefore, it is highly possible that for patients with thyroid diseases, if some patients who are not able to complete the self-administered survey and the interview mode is not provided, those patients will be excluded from the study, and it will influence the generalization of the results (Unruh et al., 2003). In the current study, although there were only 3 participants who have problems with vision and cannot read the questionnaire by themselves, considering the source of the recruitment (i.e., mostly from researcher's relative, and some from social network), and the online form of mode of administration (electronic interviews and online survey), it is likely that most of the older people who have disabilities were not reached. Therefore, despite the mode effects on the responses to the questionnaire, it is still worth conducting a study using both interview and self-administered modes so that results better generalize to the target population by ensuring the sample is not restricted to individuals with specific technology

access, communication and literacy abilities. Most importantly, at the same time, the mode of administration can be a critical factor when examining the questionnaire responses that cannot be ignored.

As indicated from the previous sections about gender differences on floor effects as well as the test scores of some scales, it was noted that there was a significant association between mode of administration and gender (χ_1^2 = 4.65, p = 0.03), i.e., there was a differential proportion in the mode of administration that was chosen by women and men. When looking at the mode of administration effect on other demographic/disease-specific variables using Welch's t-tests, there were no statistically significant mean differences by mode of administration on these variables (age: $t_{169.93}$ = -1.381, p = 0.169; education level: $t_{165.14}$ = -1.025, p = 0.307; duration of treatment: $t_{176.46}$ = -0.639, p = 0.524; level of disease control: $t_{176.92}$ = 1.367, p = 0.173). Although, with the exception of gender, there was no significant association between mode of administration and the above variables, because in the current study, participants chose the mode of administrations, the mode of administration may be related to and interact with other variables that were not included in this study (e.g., occupation, subtype of thyroid diseases). Future studies can consider these variables and add interaction terms between mode and other variables to better understand the results.

4.4.3. The effects in the full model with all six predictors

When looking at the overall effect of all six predictors (gender, mode of administration, age, education level, duration of treatment, and level of disease control), it was shown that the overall model had significant results on most of the scales with both NB and ZINB models, except for the eye symptoms scale with NB models. Beyond the effects of gender and mode of administration discussed above, the results suggested that the demographic and disease-related may be predictors to the responses to HRQOL questionnaires. It further highlighted the critical role of these demographic and disease-related variables. The results were consistent with the previous study investigating the HRQOL among thyroid cancer survivors in China, where various demographic and disease-related variables such as sex, education, employment status, income, and type of surgeries are demonstrated to have an impact on the responses to the SF-36, and it was further indicated that professionals should raise their awareness to these factors in order to provide better care to patients with thyroid cancer in China (Wong et al., 2018).

Specifically, this study found that age can negatively predict nine out of 12 symptoms and functions scales as well as the composite scale, which are goiter symptoms, impaired social life, impaired daily life, cosmetic complaints, hyperthyroid symptoms, anxiety, depressivity, emotional susceptibility, and composite scores, where older people seemed to show better QOL on all scales mentioned above. It was also observed from the think-aloud process, in which younger participants showed more anxiety than older participants on their change of appearance and social activity. For example, one participant aged 24 thought she became someone else's burden because others had to consider her diet when having dinner with her, and she thought her mother cannot take that she had thyroid disease at this young age, and the family had to ask friends and relatives for help to look for a specialist in thyroid disease (quote: "因为就是像我刚刚说的，你跟别人出去吃饭他们就得迁就我以及...额...我感觉我妈觉得我得病对她来说是一件很...难以接受的事情，然后...并且我们就是要拜托到...就是...家里的一些朋友啊，医生朋友啊，去帮我找一些更厉害的甲状腺科的一些医生"). In contrast, one participant aged 49 stated that the thyroid surgery did not affect her appearance and it was just a "little scar" that did not need to pay much attention to (quote: "额…这些都没有，就是做了一个手术嘛，有一个小疤。"). The results are contradicted with the findings from Watt et al. (2014) that younger people tend to show better QOL with most of the items that showed DIF. Indeed, from a biological perspective, for thyroid patients, the endocrine system and rate of metabolism of hormones may be changed as people getting older (Aggarwal & Razvi, 2013). However, one longitudinal study found that the quality of life increases from 50 years to 68 years after controlling for other socioeconomic and disease specific variables (Netuveli et al., 2006). The findings from Netuveli et al. (2006) indicated that older age can also be associated with better QOL in some cases. In addition, the results from this study were in part consistent with previous studies conducted among thyroid patients, where these studies reported older age had negative influence on the physical functioning and cognitive functioning but had a positive influence on the emotional functioning (Husson et al., 2013; Lee et al., 2010).

In the current study, education level was negatively associated with scores on goiter symptoms, impaired social life, impaired daily life, and cosmetic complaints scales, which means that the higher the education, the better self-reported goiter symptoms, and less of a negative impact on their daily life and less of concern with respect to appearance change due to thyroid diseases. These findings were consistent with a

previous study where lower education level was an indication of worse functioning and more severe symptoms among thyroid patients (Husson et al., 2013). Another interesting finding is that, the education level also predicted the likelihood of having structured zeros or sampling zeros for scores on goiter symptoms scale, which means that the education level may determine whether people do not have goiter symptoms at all or people who have goiter symptoms but chose not to report for some reasons.

Regarding the disease-specific predictors, it was found that the duration of treatment can be a predictor of impaired social life, cosmetic complaints, and hyperthyroid symptoms scores, and the longer the treatment, the worse QOL shown on impaired social life, cosmetic complaints and hyperthyroid symptoms. It may be all indication that having thyroid diseases that cannot be cured during a longer period of time may lead someone to experience more impaired interaction with others, more serious impact to their appearance, and more severe hyperthyroid symptoms. Health professionals may need to pay more attention to patients with chronic diseases with regard to their physical symptoms and social interaction functioning. Also, the level of disease control to thyroid diseases was negatively associated with the scores on 11 out of 12 symptoms and functions scales and the composite scale, except for the eye symptoms scale. It can be concluded that the better control of thyroid diseases reported by patients was associated with better QOL on most of the scales in SC ThyPRO-39. Future study can also examine the interaction effect of duration of treatment and level of disease control to make more confident conclusions of whether the improvements result from treatment. In addition, in terms of the level of disease control, it was found that the level of disease control predicted the excess zeros of scores on goiter symptoms and impaired social life scales, indicating that participants may actually show a full health status on their goiter symptoms and social interaction problems based on what extent they thought their diseases were controlled.

As provided in the discussion section of gender difference on floor effects, it was shown that gender had a significant association with other predictors (mode of administration, age, and education level). Therefore, the other predictors may significantly interact with the gender effects on the scale scores, including some other variables that were not included in this study, such as occupations, income, regional difference, and others. Future study can add these variables as well as interaction terms to the analysis to see whether the gender effect on QOL is influenced by other variables.

Regarding the issues of missing results for the overall effect on eye symptoms, tiredness, and composite scales, it was observed that the estimated coefficients for the intercepts of the three scales are large (estimate = -130.08 for the eye symptoms scale, estimate = -150.71 for the tiredness scale, and estimate = -150.71 for the composite scale), and some estimated coefficients for other predictors in those three models are also large (e.g., estimate = -23.48 for the "level of disease control" in the zero-inflated part of eye symptoms scale). Because the coefficients are estimated on the logit link function for the zero-inflated part of ZINB model (Cameron & Trivedi, 1999), the zero-inflated probabilities for the baseline level of three scales are close to zero (e.g., the probability at the baseline level for the eye symptoms scale is $3.21 * 10^{-57}$). In other words, the likelihood of the zero-inflated part of the ZINB models with the three scales are extremely close to zero, which makes it impossible for any program to generate estimated coefficients. In addition, complete separation can be another reason for the missing results based on the error message from the program output, whereby complete separation occurs when the outcome variable completely separates one or more predictors. The complete separation causes the maximum likelihood estimation to become infinite, and as a result, the estimate cannot be appropriately estimated (Rainey, 2016). Further analysis was conducted with elimination of one of the predictors from the full predictor sets. It was shown that there were no errors present with the elimination of the variable of level of disease control for eye symptoms scales, tiredness scale and composite scale, and the results from the new models are shown on Table M20, M21 and M22 for the three scales. After excluding the "level of disease control" from the zero-inflated model, there is a significant effect of level of disease control on the negative binomial part of the eye symptoms scale (see Table M20); significant effects of mode of administration and level of disease control on the negative binomial part of the tiredness scale (see Table M21); and significant effects of gender, age, and level of disease control on the negative binomial part of the composite scale (see Table M22). The directions of the effects of gender, mode of administration, age, and level of diseases control from the simplified models are the same as directions from other scales discussed previously. In sum, it was suggested that the predictors in the zero-inflated part of the zero-inflated models should be carefully considered and selected, and researchers should be alert when the estimated coefficients are large. Tests for detecting the complete separation can also be conducted as appropriate (Konis, 2007).

It was noticed that the eye symptoms scale was the only scale that showed insignificant results in most of the cases. The insignificant effects of predictors on this particular scale compared to others may be a result in part of participants' difficulty in interpreting or considering the relevance of the three items under the eye symptoms scales. According to the further content analysis of the think-aloud process and post-questionnaire interview questions for the electronic interview group, 21 out of 98 participants indicated that they were confused by the questions related to the eye function, as they did not know whether their eye symptoms related to their thyroid diseases. Also, there were only two participants who indicated that they have eye symptoms related to their thyroid diseases. Moreover, the most frequently asked question was "Is my nearsightedness/farsightedness can be counted as 'impaired vision'?" ("那我近视/老花算是视力受损吗?").

Previous research has indicated that prior knowledge has a significant effect on questionnaire responses (Langer & Nicolich, 1981); it can be the case in this study. Because ThyPRO-39 is a comprehensive HRQOL measurement that includes items asking about the symptoms of different thyroid diseases, the eye symptoms may only be relevant to patients with Grave's disease but no other thyroid diseases. Therefore, it is reasonable that participants without Grave's diseases could not understand what those questions are supposed to ask, and therefore chose the options that did not represent their true health conditions related to thyroid diseases. It is also possible that the lack of effects is due to the few participants with Grave's diseases, insofar as there were only five participants who reported that they had Grave's diseases in this study.

4.5. Other consideration for the analysis of ThyPRO-39 data

In addition to the Tobit model and series of models related to Poisson regression, the log-transformation of the scores to the normal distribution were also considered in initially developing this study, because the log-transformation was widely used for skewed data (Feng et al., 2014). However, during the course of data analysis, it was found that even after different types of transformation (log-transformation, squared-transformation, squared-root transformation, box-cox transformation…etc.), the data were not normally distributed. This was consistent with previous research which suggested that it is not always the case that the log-transformation can make the data

more normally distributed (Feng et al., 2014). More specifically, it depends on whether the original data follow a log-normal distribution. In other words, only if the original data was log-normally distributed, the log-transformed data can be normal. Moreover, it became harder to interpret the scores after transformation. Therefore, using transformation was not recommended to be used according to Feng et al. (2014). This study then excluded the log-transformation of the data in modelling strategy comparisons, and only regression modelling with Tobit and Poisson-related distributions were compared with the multiple linear regression model.

It was also noted that there was a linear transformation presented for the scores of the original scales on the ThyPRO-39 (Watt et al., 2014). The raw scores of each scale were transformed to scales ranging from 0 to 100, using the Orlando and Thissen IRT-based summed-score linking (Orlando et al., 2000). Although it is reasonable to use this transformation to make it to be easier for patients and physicians to interpret quality of life scores, there also can be a potential problem with this linear transformation. Feng et al. (2014) pointed out that each value under different scales is transformed to different values after rescaling. For example, a raw score of 0 was rescaled as 2 on goiter symptom scale and was rescaled as 0 on the tiredness scale. Although it seemed easier to understand the level of QOL after linear transformation, because the same value under different scales was transformed to different values, it also became harder to interpret the meaning of scores before and after the transformation, as stated by Feng et al. (2014). Thus, the scores after linear transformation were not used in this study, instead, this study only focused on the raw scores of each scale, as well as on the composite scale.

In addition to the recommended parametric regression modelling strategies (i.e., NB and ZINB regression models), there is another method proposed by Liu and Wang in their most recent study (2021) to deal with two independent sample comparisons of data with prominent floor effects using t-test based on truncated normal distribution. This approach assumes the normality of the distribution, and it also assumes that the data are censored, which means the data at zero are not “true” zeros, and true levels may be lower than zero (Liu & Wang. 2021). In other words, this method is similar to the Tobit regression model that is based on the assumption of censored data, and it is different from the NB and ZINB regression models considered in this study that assume the data at zero to be “true” zeros. Specifically, Liu and Wang’s approach (2021) involves

generating a new sample mean and variance adjusting for the ceiling and floor effects, and using the new mean and variance to compute a t statistic. They compared their new method to three other methods. The first one is the method that ignores the CFE and treats the data as normally distributed, which is similar to the multiple linear regression model in this study; the second one is removing the data at floor and/or ceiling and treating the data as truncated data; and the third one is the censored regression method, which is the Tobit regression that was used in this study. Results from their simulated study indicated that when the assumption of homogeneity of variance is met, both the Tobit regression model and the new method perform well, but when the homogeneity of variance is not met, the new method was preferred over the Tobit regression model when dealing with data with CFE.

When applying the proposed method using t-tests proposed by Liu and Wang (2021) to examine gender differences and mode of administration effects on the SC ThyPRO-39 data from the current study, the results are mostly consistent with the results from NB and ZINB models. Specifically, females showed significantly better QOL than males on goiter symptoms scales and impaired daily life scales, but not the other scales, and participants in the electronic interview group showed significantly better QOL than participants in the electronic self-administered group for most of the scales except goiter symptoms scales, hypothyroid symptoms scales, and eye symptoms scales (details are shown on Table M23 and M24). Although the conclusions in single predictor gender and mode of administration models were most consistent, it is still unknown whether Liu and Wang's method (2021) has an advantage over the NB and ZINB models when considering between group differences in the current study. This uncertainty is largely because the Liu-Wang method assumes the zeros at floor are not "true" zeros. But given the nature of the ThyPRO-39, the zeros on each item of this study are the equivalent of "not at all", and as such may be "true zeros. In many understandings of the state-of-the-world, it is impossible to be less than "not at all"; as such, there seems no room left for data to be below the option of "not at all". Therefore, the assumption of censored data may not be held in this study. Also, if the zeros at floor are "true" zeros, the assumption of the normality may not be met as well, which is also a key assumption for Liu and Wang's proposed method. On the other hand, the same condition may not be applicable to questionnaires using other response options at the bottom of each item, such as the "0 = rarely" from the Center for Epidemiologic Studies Depressive Symptomology Scale

(CES-D, Radloff, 1977), where “rarely” does not mean nothing at all. Finally, as presently implemented, the Liu-Wang method can only be used for group comparisons with models with categorical predictors as in independent samples t-tests on means or in one-way ANOVA; there is no evidence that this method can be applied to modelling data with more than one or more continuous or quasi-continuous quantitative predictor, such as was considered in the present thesis. Future studies focused on models with appropriately coded categorical variable predictors may also apply the method from Liu and Wang (2021) to analyze data with CFE when comparing group means, with the consideration of the assumption of censored data and normality.

4.6. Implications

In general, this study highlighted the importance of thorough psychometric evaluation when a measurement is translated and adapted in a different language/cultural context, because there may be unique considerations to the new version regarding the cultural and linguistic features. More specifically, it is important to check distributional features before conducting analysis, including distributional plots and tests, because the results from the plots and tests may provide significant insights to the following modelling strategies. Assumption checking/diagnostics are also critical, because if the assumption of one model was not met, the model fit of that model may not be satisfactory.

This study also points out the importance of floor effect in questionnaire responses. First, the level of floor or ceiling effect should be reported for studies related to questionnaire responses. The results regarding comparing distributional models for the data with and without prominent floor effects also indicated that one should be careful in analyzing this kind of data. For data with significant floor effects, the NB and ZINB models are recommended for use rather than models based on normal distribution. Also, even if the data do not show significant floor effect, if they are skewed, the tests from models based on normal distribution (e.g., ML models) may not be appropriate.

This is the first study considering the Simplified Chinese version of ThyPRO-39 in Mainland China in a moderately large sample. The SC ThyPRO-39 was translated and developed by Liew (n.d.) in Singapore, and was only validated in a pilot study with 5 Mandarin-speaking thyroid patients in Singapore. Therefore, the results from this current

study have implications for the use of SC ThyPRO-39 among Chinese population. The application of this newly-translated Simplified Chinese version of ThyPRO-39 in Mainland China may be a starting point to raise awareness of the quality of life with and without treatment to patients with thyroid diseases, as some patient-participants in this study reported that they failed to relate their symptoms to their thyroid diseases before completing the questionnaires.

When looking at the SC ThyPRO-39 specifically, the presence of low internal construct validity and low internal consistency reliability in both SC and TC ThyPRO-39 versions compared to the original Danish version may be an indication of poor-quality translation, but it may also be an indication of a fundamental difference between Chinese culture and European culture where the ThyPRO-39 is originally developed. In other words, some items in the original version of ThyPRO-39 may not be compatible when translated and adapted in the Chinese culture. In addition, it was noted that based on different distributional characteristics and results from assumption checking/diagnostics, not all scales in the SC ThyPRO-39 are best analyzed with the same modelling strategy. Because there is no over-dispersion present in the emotional susceptibility scale, results from performance of model fits indicated it is better to use Poisson or Zero-inflated Poisson regression to analyze data on the emotional susceptibility scale, rather than NB and ZINB regression models that were preferred for analysis of data on other scales with prominent floor effects in the SC ThyPRO-39.

4.7. Limitations

All studies have limitations, with many limitations being tied to the context of data collection. The current study on data considerations for measurement of HRQOL with a patient-reported outcome measure focused on Chinese patients with thyroid diseases administered the Simplified Chinese ThyPRO-39 (Liew, n.d.). This section addresses limitations pertaining to the generalizability of findings from the current study.

Data collection for the current project took place in the context of the first year of the onset of the Covid-19 pandemic. This required data collected using online platforms with electronic self-administered questionnaires and online technology-mediated interviews. Findings, particularly with regard to the distributions of scores, may not be

similar to findings from paper-and-pencil administrations and traditional face-to-face interviews (Tourangeau & Smith, 1996) due to potentially different impacts of medium of administration. Additionally, findings may not be generalizable to people with limited access to technology.

Data collection on the SC ThyPRO-39 was conducted on study participants in Mainland China, and therefore may not be generalizable to Chinese thyroid patients elsewhere. Additionally, although random sampling from the population of interest is an ideal, many studies make use of convenience samples and word-of-mouth recruitment. Findings based on one recruitment strategy may be different from findings based on other recruitment strategies, that is recruitment strategies can be a source of variation contributing to different responses. Although social media recruitment as well as word-of-mouth recruitment were conducted, there was a considerable proportion of the participants, approximately half of the participants, recruited by the researcher's relative. As observed from interview transcriptions, people who were recruited directly by the relative of the researcher provided richer reflection and elaboration than participants who were recruited from the social network site and who were strangers to the researchers in both the think-aloud process and post-questionnaire interview. However, because the questions about the recruitment source were not asked in the demographic questionnaire, the two groups of people could not be distinguished from each other among participants who were in the electronic self-administered group completing online surveys. The researcher's relative lives in a small town named Huai'an in the east coast area of China, and is a teacher. In this study, it is known that 60 participants are from Huai'an, and among these participants, 52 of them are teachers. This means that participants in this study were highly concentrated within a specific region and specific occupation; it is unlikely that the study sample is representative of the population with thyroid diseases in China.

Regarding the criteria used for comparing modelling strategies, it was noted that because the ML and Tobit regression models treat the data as continuous, whereas the series of Poisson models (Poisson, NB, ZIP, and ZINB models) treat the data as discrete (Cameron & Trivedi, 1999), problems may be raised if only using the AIC and BIC criteria for the model performance. Therefore, other ways to evaluate the performance may be necessary to be considered, such as the scatterplots between observed values and fitted values from each of the six models (Carty et al., 2015). In another study, to

compare the Tobit, linear, and Poisson-gamma models, Brown and Dunn (2011) divided the data into estimation and validation data sets, and then use the validation set to generate various simulations. The means, medians, 75th quantiles, percentage zeros and 95% CI were generated for both estimation and simulated data sets. Finally, the performance of the model is determined by the chance of capturing the true values by the simulated data sets from the estimation data set. Future studies can use scatterplots as well as validation sets to better evaluate the model performance.

Another limitation is with regard to concern for the generalizability of the findings on the Simplified Chinese ThyPRO-39 (Liew, n.d.) and Mandarin speaking people with thyroid disease to Chinese with different dialects. Although Simplified Chinese in Mainland China and Traditional-Chinese in Hong Kong share a similar writing system, there is a considerable difference between the spoken language system of Mandarin and Cantonese. The difference between the spoken language system may lead to different results on the responses on the SC ThyPRO-39 versus those on the TC ThyPRO-39 (Wong et al. 2018) which are discussed in this discussion chapter.

4.8. Future directions

The current thesis on the Simplified Chinese ThyPRO-39 (Liew, n.d.) used the original scoring guidelines for the ThyPRO-39 as adopted in translated versions, e.g., TC ThyPRO-39. Although a Confirmatory Factor Analysis (CFA) was conducted by Watt et al. (2014) for the original long form of ThyPRO in Danish, and unidimensionality across items for each scale was demonstrated, a CFA was not conducted with ThyPRO-39, nor with the SC ThyPRO-39. Although not necessary if appropriate strategies are used, authors often indicate that the univariate or multivariate normality of the observed variables must be satisfied before conducting the CFA (Mahmoud & Khalifa, 2015; Meyers et al., 2016). Therefore, unless appropriately addressed, the lack of normality in the current data may raise some problems when conducting CFA. One may be cautious when conducting a CFA and use some corrections or other estimator instead of the standard maximum likelihood estimator, such as the diagonally weighted least squares (WLSMV) to deal with the non-normality of the data (Li, 2016). Future studies can conduct CFA to measure the measurement structural validity of SC ThyPRO-39 with the caution of appropriate handling of non-normality of the data.

Overall, the Simplified Chinese ThyPRO-39 (Liew, n.d.) had better performance than the Traditional Chinese ThyPRO-39 (Wong et al., 2018) on measurement properties. However, problems still existed for some SC ThyPRO-39 items and scales in terms of low internal consistency reliability and low internal construct validity. Therefore, some revisions are suggested to be made for the future application of the SC ThyPRO-39. For example, where all three reverse worded items ("Felt energetic", r = 0.341; "Had self-confidence", r = 271; "Felt in control of your life", r = 0.112) showed low corrected item-total correlations, the three reverse worded items should be considered to be rephrased as the positive worded items and the item-total correlations can be compared between positive worded items and reverse worded items to see if the results improve.

A problem was also shown on the SC ThyPRO-39 eye symptoms scale. Participants reported they had difficulty responding to items related to eye symptoms. Although they felt like they had those symptoms, those symptoms were not related to their thyroid diseases. To deal with this problem, instructions can be added as a subheading to the questionnaire, such as "Please note that this questionnaire is for a variety of thyroid function problems, you may be asked about some symptoms that may not be relevant to your thyroid disease." ("请注意此问卷是针对于不同情况的甲状腺患者，所以您可能会被问到一些与您甲状腺疾病症状无关的问题。"). Also, on the current version of the SC ThyPRO-39, conditions about "have your thyroid disease caused you to" are only presented for scales with impaired social life and impaired daily life scales, but not the other scales; such phrasing may be useful for other items. Based on the results from this study, clarifying conditions about the influence of thyroid diseases may be necessary to be added to the five items with low item-total correlation on the five scales, and items in the eye symptoms scale. For example, the item "had an upset stomach" ("您有肚子不适吗?", r = 0.337) can be rephrased as "have you had an upset stomach due to your thyroid diseases?" ("您有因为甲状腺疾病而肚子不适吗?") and the item of "have you had impaired vision?" in eye symptoms function can be rephrased as "have you had impaired vision because of the thyroid diseases?" ("您有因为甲状腺疾病而视力受损吗?").

In addition, after completing the SC ThyPRO-39, 10 out of 123 women indicated that women-specific questions, such as the influence of thyroid diseases on pregnancy, menstruation, and sexual life were not mentioned in the SC ThyPRO-39. Although the

scale about sexual life was included in the original long-form of ThyPRO, it was removed for the ThyPRO-39 because of a higher rate of missing responses than for other scales (Watt et al., 2014). It may be necessary to reintroduce the items about sexual life and other questions specific to women in the questionnaire, even if there may be missing values for those items; based on comments on the SC ThyPRO-39, it is still important to the QOL of some women who have thyroid diseases.

A major priority of this project was on identifying optimal modelling strategies for analyzing test scores of the SC ThyPRO-39 with floor effect in different predictors sets. However, this study only focuses on parametric models, and semi-parametric and non-parametric models can also be used to analyze data with prominent floor effects that are not normally distributed. For example, Qian and Xie (2021) proposed a semi-parametric odds ratio (SOR) model to deal with data with non-normality that are not collected through random sampling. In the future, this model can be applied to the data in this study because of the selective sampling of the sample and the non-normality of the data. In addition, as discussed in the previous sections, some non-parametric models (Bunevicius, 2017; Kashkouli et al., 2017) can also be applied to the current data and compared with other models.

In addition to looking at the distributional form of scores, the primary main effects considered were for gender and mode of administration. Other effects, such as the main effect of age and duration of treatment on each scale were not investigated in detail. Also, there may be other important relevant variables that were not included in the current study. For example, future studies can focus on the interaction terms such as age and gender interaction. Some additional predictors, such as the variable of occupation and region of the participants can also be included in the analysis when more data are collected for participants with more varied occupations and from more varied areas in Mainland China. In addition, the source of recruitment can also be asked in the demographic questionnaire to see whether there is a source of variation in questionnaire responses.

Finally, future studies should also be conducted among other Mandarin speaking, Simplified Chinese reading participants, not just in other parts of China, but also in other countries. The original plan for this study involved a comparison between English and Simplified Chinese versions of ThyPRO-39, with an additional variable of acculturation

measuring the influence of culture on the responses to ThyPRO-39 among Mandarin-English bilinguals and Mandarin monolinguals in Canada. However, due to the impact of COVID-19, the recruitment rate in Canada was very poor, as there were only 20 participants in the Canadian sample to date. Therefore, data on those participants were not included and the variables of language version and acculturation were removed from the analysis. Future studies can continue collecting data after the COVID-19 pandemic, and after enough data are collected, the language version and acculturation can be included to investigate the influence of language and culture.

4.9. Conclusion

This is the first study investigating the SC ThyPRO-39 among thyroid patients in China to our knowledge. In sum, 11 out of 13 scales (including composite scale) in the SC ThyPRO-39 were not normally distributed. A floor effect presented for 4 out of 13 scales and 35 out of 39 items. In addition, five items showed low internal construct validity, which corresponds to the low internal consistency reliability of the five scales. Comparing results from the current study to those reported on the TC ThyPRO-39 by Wong and colleagues (2018), the three measurement properties (presence or absence of CFE, internal construct validity and internal consistency reliability) were better on the SC ThyPRO-39 than on the TC ThyPRO-39. The proportion of floor did not differ by gender or mode of administration on most scales. NB and ZINB models are recommended to be used for scales with prominent floor effect; and for scales without prominent floor effect, the NB model is still suggested to be used, except for the emotional susceptibility scale without overdispersion and zero-inflation.

Although there was generally little impact of gender and mode of administration on CFE, this study also illustrated the significant effect of gender and mode of administration on responses to the SC ThyPRO-39 in general; other variables, including age, education level, duration of treatment, and level of disease control can also contribute to different response patterns to the SC ThyPRO-39. Importantly, findings depended on appropriate analysis of data, highlighting that careful analysis of data should be contingent upon model assumptions and presence or absence of CFE. Future studies should be cautious when choosing the model for the analysis of ThyPRO-39 scores. Based on the preliminary results on select measurement properties of the SC ThyPRO-39, we conclude that overall, the SC ThyPRO-39 has strong potential for use

as a Patient-Reported Outcome Measure among thyroid patients in Mainland China and who read Simplified Chinese worldwide. When analyzing ThyPRO-39 data, whether on the SC ThyPRO-39 or on another language form, the potential floor effects cannot be ignored; explicit statements regarding the presence or absence of CFE are recommended; and analytic strategies should be carefully chosen based on the distributional features of the data. Researchers should also pay attention to the variables regarding the mode of administration and personal characteristics (e.g., gender, age, and level of disease control) on the responses to the SC ThyPRO-39.

References

About Thyroid Disease. Thyroid Foundation of Canada. Retrieved from: https://thyroid.ca/thyroid-disease/.

Acrscommunity. (2017, August 1). *What is Culturally Competent Mental Health Care?* Retrieved from https://acrs.org/culturally-competent-mental-health-care/.

Aggarwal, N., & Razvi, S. (2013). Thyroid and Aging or the Aging Thyroid? An Evidence-Based Analysis of the Literature. *Journal of Thyroid Research*, *2013*, e481287. https://doi.org/10.1155/2013/481287

Alemu, S., Herklots, T., Almansa, J., Mbarouk, S., Sulkers, E., Stekelenburg, J., de Zeeuw, J., Jacod, B., & Biesma, R. (2020). Mental Health and Quality of Life of Women One Year after Maternal Near-Miss in Low and Middle-Income Countries: The Case of Zanzibar, Tanzania. *International Journal of Environmental Research and Public Health*, *17*(23), 9034. https://doi.org/10.3390/ijerph17239034

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Bowling, A. (2009). *Research Methods in Health: Investigating Health and Health Services. Open University Press, Buckingham, UK. ISBN 0335206433.*

Bauer, M., Glenn, T., Pilhatsch, M., Pfennig, A., & Whybrow, P. C. (2014). Gender differences in thyroid system function: Relevance to bipolar disorder and its treatment. *Bipolar Disorders*, *16*(1), 58–71. https://doi.org/10.1111/bdi.12150

Bergner, M., Bobbitt, R. A., Carter, W. B., & Gilson, B. S. (1981). The Sickness Impact Profile: Development and final revision of a health status measure. *Medical Care, 19*(8), 787–805. https://doi.org/10.1097/00005650-198108000-00001

Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health, 27*(3), 281-291.

Bharmal, M., & Thomas, J. (2006). Comparing the EQ-5D and the SF-6D Descriptive Systems to Assess Their Ceiling Effects in the US General Population. *Value in Health, 9*(4), 262–271. https://doi.org/10.1111/j.1524-4733.2006.00108.x

Bird, C. E., & Rieker, P. P. (2008). *Gender and Health: The Effects of Constrained Choices and Social Policies*. Cambridge University Press. https://www.rand.org/pubs/commercial_books/CB412.html

Boerma, T., Hosseinpoor, A. R., Verdes, E., & Chatterji, S. (2016). A global assessment of the gender gap in self-reported health with survey data from 59 countries. *BMC Public Health, 16*(1), 675. https://doi.org/10.1186/s12889-016-3352-y

Bradley, E. A., Sloan, J. A., Novotny, P. J., Garrity, J. A., Woog, J. J., & West, S. K. (2006). Evaluation of the National Eye Institute Visual Function Questionnaire in Graves' Ophthalmopathy. *Ophthalmology, 113*(8), 1450–1454. https://doi.org/10.1016/j.ophtha.2006.02.060

Brazier, J. E., Harper, R., Jones, N. M., O'Cathain, A., Thomas, K. J., Usherwood, T., & Westlake, L. (1992). Validating the SF-36 health survey questionnaire: New outcome measure for primary care. *BMJ (Clinical Research Ed.), 305*(6846), 160–164. https://doi.org/10.1136/bmj.305.6846.160

Brown, J. E., & Dunn, P. K. (2011). Comparisons of Tobit, Linear, and Poisson-Gamma Regression Models: An Application of Time Use Data. *Sociological Methods & Research*, *40*(3), 511–535. https://doi.org/10.1177/0049124111415370

Bukvic, B., Zivaljevic, V., Sipetic, S., Diklic, A., Tausanovic, K., & Paunovic, I. (2015). Validation and cross-cultural adaptation of the questionnaire ThyPRO in thyroid patients in Serbia. *Vojnosanitetski Pregled, 72*(7), 583–588. https://doi.org/10.2298/VSP131112035B

Bunevicius, A. (2017). Reliability and validity of the SF-36 Health Survey Questionnaire in patients with brain tumors: A cross-sectional study. *Health and Quality of Life Outcomes, 15*(1), 92. https://doi.org/10.1186/s12955-017-0665-1

Burnham, K. P., & Anderson, D. R. (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*. Springer-Verlag. https://doi.org/10.1007/978-1-4757-2917-7

Cameron, A., & Trivedi, P. (1999). Regression analysis of count data. 2nd ed. In *Technometrics* (Vol. 41). https://doi.org/10.1017/CBO9780511814365

Campinha-Bacote, J. (1994). Cultural competence in psychiatric mental health nursing. A conceptual model. *The Nursing Clinics of North America, 29*(1), 1–8.

Canney, M., Sexton, E., Tobin, K., Kenny, R. A., Little, M. A., & O'Seaghdha, C. M. (2018). The relationship between kidney function and quality of life among community-dwelling adults varies by age and filtration marker. *Clinical Kidney Journal, 11*(2), 259–264. https://doi.org/10.1093/ckj/sfx084

Carlé, A., Bülow Pedersen, I., Knudsen, N., Perrild, H., Ovesen, L., & Laurberg, P. (2015). Gender differences in symptoms of hypothyroidism: A population-based DanThyr study. *Clinical Endocrinology, 83*(5), 717–725. https://doi.org/10.1111/cen.12787

Carlozzi, N. E., Hahn, E. A., Goodnight, S. M., Kratz, A. L., Paulsen, J. S., Stout, J. C., Frank, S., Miner, J. A., Cella, D., Gershon, R. C., Schilling, S. G., & Ready, R. E. (2018). Patient-reported outcome measures in Huntington disease: Quality of life in neurological disorders (Neuro-QoL) social functioning measures. *Psychological Assessment, 30*(4), 450–458. https://doi.org/10.1037/pas0000479

Carty, D. M., Young, T. M., Zaretzki, R. L., Guess, F. M., & Petutschnigg, A. (2015). Predicting and Correlating the Strength Properties of Wood Composite Process Parameters by Use of Boosted Regression Tree Models. *Forest Products Journal*, *65*(7–8), 365–371. https://doi.org/10.13073/FPJ-D-12-00085

Castello, R., & Caputo, M. (2019). Thyroid diseases and gender. *Italian Journal of Gender-Specific Medicine, 5*(3), 136–141.

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Amandkumar, A., Ford, C., Volcic, R., De Rosario, H. (2020). *Package.* Retrived from https://cran.r-project.org/web/packages/pwr/pwr.pdf

Chen, C., Lee, S.-Y., & Stevenson, H. W. (1995). Response Style and Cross-Cultural Comparisons of Rating Scales among East Asian and North American Students. *Psychological Science*, *6*(3), 170–175.

Chen, X., Qiu, Z., Gu, M., Su, Y., Liu, L., Liu, Y., Mo, C., Xu, Q., Sun, J., & Li, D. (2014). Translation and validation of the Chinese version of the quality of life radiation therapy instrument and the head & neck module (QOL-RTI/H&N). *Health and Quality of Life Outcomes, 12*(1), 51. https://doi.org/10.1186/1477-7525-12-51

Cherepanov, D., Palta, M., Fryback, D. G., & Robert, S. A. (2010). Gender differences in health-related quality-of-life are partly explained by sociodemographic and socioeconomic variation between adult men and women in the US: Evidence from four US nationally representative data sets. *Quality of Life Research, 19*(8), 1115–1124. https://doi.org/10.1007/s11136-010-9673-x

Christianson, D. A., & Bender, H. (2011). *The Complete Idiot's Guide to Thyroid Disease* (1st edition). Alpha.

Claessen, S. J. J., Hazes, J. M. W., Huisman, M. A. M., van Zeben, D., Luime, J. J., & Weel, A. E. A. M. (2009). Use of risk stratification to target therapies in patients with recent onset arthritis; design of a prospective randomized multicenter controlled trial. *BMC Musculoskeletal Disorders, 10*, 71. https://doi.org/10.1186/1471-2474-10-71

Cohen, J. (1988*). Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Collins, K., Hughes, D., Doty, M., Ives, B., Edwards, J., & Tenney, K. (2002). *Diverse Communities, Common Concerns – Assessing Health Care Quality for Minority Americans: Findings from the Commonwealth Fund 2001 Health Care Quality Survey.* New York: Commonwealth Fund.

Cook, C. (2010). Mode of administration bias. *Journal of Manual & Manipulative Therapy, 18*(2), 61–63. https://doi.org/10.1179/106698110X12640740712617

Towards a Culturally Competent System of Care: *A Monograph on Effective Services for Minority Children who are Severely Emotionally Disturbed. (Volume I).* (1989). CASSP Technical Assistance Center. http://archive.org/details/towardsculturall00un

Cui, F. (2013) *Patients with thyroid diseases may exceed 2 million and the treatment rate is only 5%.* Retrieved from: http://jiankang.cntv.cn/2013/05/21/ARTI1369098074937803.shtml

Demet, M.M., Ozmen, B., Deveci, A., Boyvada, S., Adiguzel, H., Aydemir, O. (2002). Depression and anxiety in hyperthyroidism. *Arch Med Res, 33*(6):552-556.

Desmarais, B. A., & Harden, J. J. (2013). Testing for Zero Inflation in Count Models: Bias Correction for the Vuong Test. *The Stata Journal, 13*(4), 810–835. https://doi.org/10.1177/1536867X1301300408

Diab, N., Daya, N. R., Juraschek, S. P., Martin, S. S., McEvoy, J. W., Schultheiß, U. T., Köttgen, A., & Selvin, E. (2019). Prevalence and Risk Factors of Thyroid Dysfunction in Older Adults in the Community. *Scientific Reports, 9*(1), 13156. https://doi.org/10.1038/s41598-019-49540-z

Dolnicar, S., & Grün, B. (2007). Cross-cultural differences in survey response patterns. *International Marketing Review, 24*(2), 127–143. https://doi.org/10.1108/02651330710741785

Feeny, D. H., Eckstrom, E., Whitlock, E. P., & Perdue, L. A. (2013). Patient-Reported Outcomes, Health-Related Quality of Life, and Function: An Overview of Measurement Properties. In *A Primer for Systematic Reviewers on the Measurement of Functional Status and Health-Related Quality of Life in Older Adults [Internet]*. Agency for Healthcare Research and Quality (US). https://www.ncbi.nlm.nih.gov/books/NBK169156/

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. M. (2014). Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry*, *26*(2), 105–109. https://doi.org/10.3969/j.issn.1002-0829.2014.02.009

Ferreira, L. N., Ferreira, P. L., Ribeiro, F. P., & Pereira, L. N. (2016). Comparing the performance of the EQ-5D-3L and the EQ-5D-5L in young Portuguese adults. *Health and Quality of Life Outcomes, 14*(1), 89. https://doi.org/10.1186/s12955-016-0491-x

Friendly, M. (2000). *Visualizing categorical data*. SAS Institute.

Fouladi, R. T., Mccarthy, C. J., & Moller, Naomip. (2002). Paper-and-Pencil Or Online?: Evaluating Mode Effects on Measures of Emotional Functioning and Attachment. *Assessment*, *9*(2), 204–215. https://doi.org/10.1177/10791102009002011

Gao, W., Ping, S., & Liu, X. (2020). Gender differences in depression, anxiety, and stress among college students: A longitudinal study from China. *Journal of Affective Disorders, 263*, 292–300. https://doi.org/10.1016/j.jad.2019.11.121

Gomez, S. L., Kelsey, J. L., Glaser, S. L., Lee, M. M., & Sidney, S. (2004). Immigration and acculturation in relation to health and health-related risk factors among specific Asian subgroups in a health maintenance organization. *American Journal of Public Health, 94*(11), 1977-1984.

Gopalkrishnan, N., & Babacan, H. (2015). Cultural diversity and mental health. *Australasian Psychiatry, 23*(6_suppl), 6–8. https://doi.org/10.1177/1039856215609769

Gorecki, C., Brown, J. M., Cano, S., Lamping, D. L., Briggs, M., Coleman, S., Dealey, C., McGinnis, E., Nelson, A. E., Stubbs, N., Wilson, L., & Nixon, J. (2013). Development and validation of a new patient-reported outcome measure for patients with pressure ulcers: The PU-QOL instrument. *Health and Quality of Life Outcomes, 11*, 95. https://doi.org/10.1186/1477-7525-11-95

Hakimi, Z., Herdman, M., Pavesi, M., Devlin, N., Nazir, J., Hoyle, C., & Odeyemi, I. A. O. (2017). Using EQ-5D-3L and OAB-5D to assess changes in the health-related quality of life of men with lower urinary tract symptoms associated with benign prostatic hyperplasia. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *26*(5), 1187–1195. https://doi.org/10.1007/s11136-016-1460-x

Hamamura, T., Heine, S. J., & Paulhus, D. L. (2008). Cultural differences in response styles: The role of dialectical thinking. *Personality and Individual Differences, 44*(4), 932–942. https://doi.org/10.1016/j.paid.2007.10.034

Harzing, A.-W. (2006). Response Styles in Cross-national Survey Research: A 26-country Study. *International Journal of Cross Cultural Management*, *6*(2), 243–266. https://doi.org/10.1177/1470595806066332

Hein, M. D., & Jackson, I. M. (1990). Review: Thyroid function in psychiatric illness. *General Hospital Psychiatry, 12*(4), 232–244. https://doi.org/10.1016/0163-8343(90)90060-p

Hoebel, J., von der Lippe, E., Lange, C. *et al.* Mode differences in a mixed-mode health interview survey among adults. *Arch Public Health* 72, 46 (2014). https://doi.org/10.1186/2049-3258-72-46

Hocking, R. R. (1976). A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression. *Biometrics, 32*(1), 1–49. https://doi.org/10.2307/2529336

Honkavaara, N., Al-Ani, A., Campenfeldt, P., Ekström, W., & Hedström, M. (2016). Good responsiveness with EuroQol 5-Dimension questionnaire and Short Form (36) Health Survey in 20-69 years old patients with a femoral neck fracture: A 2-year prospective follow-up study in 182 patients. *Injury, 47*(8), 1692-1697.

Huang, W., Yu, H., Liu, C., Liu, G., Wu, Q., Zhou, J., Zhang, X., Zhao, X., Shi, L., & Xu, X. (2017). Assessing Health-Related Quality of Life of Chinese Adults in Heilongjiang Using EQ-5D-3L. *International Journal of Environmental Research and Public Health*, *14*(3), 224. https://doi.org/10.3390/ijerph14030224

Husson, O., Haak, H. R., Buffart, L. M., Nieuwlaat, W.-A., Oranje, W. A., Mols, F., Kuijpens, J. L., Coebergh, J. W., & van de Poll-Franse, L. V. (2013). Health-related quality of life and disease specific symptoms in long-term thyroid cancer survivors: A study from the population-based PROFILES registry. *Acta Oncologica, 52*(2), 249–258. https://doi.org/10.3109/0284186X.2012.741326

Hypothyroidism | Thyroid Foundation Of Canada. (n.d.). Retrieved November 28, 2021, from https://thyroid.ca/resource-material/information-on-thyroid-disease/hypothyroidism/

Ihara, E. (2004). *Cultural Competence in Health Care: Is it important for people with chronic conditions?* Retrieved from https://hpi.georgetown.edu/cultural/.

Izadnegahdar, M., Norris, C., Kaul, P., Pilote, L., & Humphries, K. H. (2014). Basis for sex-dependent outcomes in acute coronary syndrome. *The Canadian Journal of Cardiology, 30*(7), 713–720. https://doi.org/10.1016/j.cjca.2013.08.020

Johnston, B. C., Patrick, D. L., Devji, T., Maxwell, L. J., Bingham III, C. O., Beaton, D. E., Boers, M., Briel, M., Busse, J. W., Carrasco-Labra, A., Christensen, R., da Costa, B. R., El Dib, R., Lyddiatt, A., Ostelo, R. W., Shea, B., Singh, J., Terwee, C. B., Williamson, P. R., … Guyatt, G. H. (2019). Patient-reported outcomes. In *Cochrane Handbook for Systematic Reviews of Interventions* (pp. 479–492). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119536604.ch18

Johnson, Kemp, Kotz, Kemp, Adrienne W, & Kotz, Samuel. (2005). *Univariate discrete distributions / Norman L. Johnson, Adrienne W. Kemp, Samuel Kotz*. (3rd ed. / Norman L. Johnson, Adrienne W. Kemp, Samuel Kotz. ed.). Wiley.

Karson, M. (1968). Handbook of Methods of Applied Statistics. Volume I: Techniques of Computation Descriptive Methods, and Statistical Inference. Volume II: Planning of Surveys and Experiments. I. M. Chakravarti, R. G. Laha, and J. Roy, New York, John Wiley; 1967, *Journal of the American Statistical Association, 63*(323), 1047–1049. https://doi.org/10.1080/01621459.1968.11009335

Kashkouli, M. B., Karimi, N., Aghamirsalim, M., Abtahi, M. B., Nojomi, M., Shahrad-Bejestani, H., & Salehi, M. (2017). Measurement Properties of the Persian Translated Version of Graves Orbitopathy Quality of Life Questionnaire: A Validation Study. *Ophthalmic Epidemiology, 24*(1), 3–10. https://doi.org/10.1080/09286586.2016.1255974

Kemmelmeier, M. (2016). Cultural differences in survey responding: Issues and insights in the study of response biases. *International Journal of Psychology, 51*. https://doi.org/10.1002/ijop.12386

Kim, Seung Hyun, & Kim, Sangmook. (2016). National Culture and Social Desirability Bias in Measuring Public Service Motivation. *Administration & Society, 48*(4), 444-476. https://doi.org/10.1177/0095399713498749

Konis, K. (2007). *Linear programming algorithms for detecting separated data in binary logistic regression models*. https://ora.ox.ac.uk/objects/uuid:8f9ee0d0-d78e-4101-9ab4-f9cbceed2a2a

Noguchi, K., Gel, Y., Brunner, E., & Konietschke, F. (2012). nparLD: An R Software Package for the Nonparametric Analysis of Longitudinal Data in Factorial Experiments. *Journal of Statistical Software, 50.* https://doi.org/10.18637/jss.v050.i12

Huff, R., & Kline, M. (2007). Health Promotion in the Context of Culture. *Health Promotion in Multicultural Populations*.

Kobak, K. A., Williams, J. B. W., Jeglic, E., Salvucci, D., & Sharp, I. R. (2008). Face-to-face versus remote administration of the Montgomery-Asberg Depression Rating Scale using videoconference and telephone. *Depression and Anxiety, 25*(11), 913–919. https://doi.org/10.1002/da.20392

Kolker, C. (2004, January 5). *Familiar Faces Bring Health Care to Latinos*. Retrieved from https://www.washingtonpost.com/archive/politics/2004/01/05/familiar-faces-bring-health-care-to-latinos/d2293a96-c74f-4d33-814d-94255b34fd22/.

Kong, H., & Hsieh, E. (2011). The social meanings of traditional Chinese medicine: Elderly Chinese immigrants' health practice in the United States. *Journal of Immigrant and Minority Health*. https://doi.org/10.1007/s10903-011-9558-2

Langer, J., & Nicolich, M. (1981). Prior Knowledge and Its Relationship to Comprehension. *Journal of Literacy Research - J LIT RES, 13*, 373–379. https://doi.org/10.1080/10862968109547426

Leak, A., Smith, S. K., Crandell, J., Jenerette, C., Bailey, D. E., Zimmerman, S., & Mayer, D. K. (2013a). Demographic and Disease Characteristics Associated with Non-Hodgkin Lymphoma Survivors' Quality of Life: Does Age Matter? *Oncology Nursing Forum, 40*(2), 157–162. https://doi.org/10.1188/13.ONF.157-162

Lee, J. I., Kim, S. H., Tan, A. H., Kim, H. K., Jang, H. W., Hur, K. Y., Kim, J. H., Kim, K.-W., Chung, J. H., & Kim, S. W. (2010). Decreased health-related quality of life in disease-free survivors of differentiated thyroid cancer in Korea. *Health and Quality of Life Outcomes*, *8*(1), 101. https://doi.org/10.1186/1477-7525-8-101

Lewis, F., Butler, A., & Gilbert, L. (2010). A unified approach to model selection using the likelihood ratio test. *Methods in Ecology and Evolution, 2*, 155–162. https://doi.org/10.1111/j.2041-210X.2010.00063.x

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7

Li, L., Wang, H. M., & Shen, Y. (2003). Chinese SF-36 Health Survey: Translation, cultural adaptation, validation, and normalisation. *Journal of Epidemiology and Community Health*, *57*(4), 259–263. https://doi.org/10.1136/jech.57.4.259

Liew, H., Watt, T., Nan, L., Tan, A. W. K., Chan, Y. H., Chew, D. E. K., & Dalan, R. (2021). Psychometric properties of the thyroid-specific quality of life questionnaire ThyPRO in Singaporean patients with Graves' disease. *Journal of Patient-Reported Outcomes, 5*(1), 54. https://doi.org/10.1186/s41687-021-00309-x

Liew, H., Simplified Chinese Thyroid-Specific Patient-Reported Outcome short-form (ThyPRO-39). *Unpublished Manuscript*.

Lin, I.-C., Lee, C.-C., & Liao, S.-L. (2015). Assessing quality of life in Taiwanese patients with Graves' ophthalmopathy. *Journal of the Formosan Medical Association*, *114*(11), 1047–1054. https://doi.org/10.1016/j.jfma.2013.12.002

Liu, H. (2007). Growth Curve Models for Zero-Inflated Count Data: An Application to Smoking Behavior. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(2), 247–279. https://doi.org/10.1080/10705510709336746

Liu, Q., & Wang, L. (2021). T-Test and ANOVA for data with ceiling and/or floor effects. Behavior Research Methods, *53*(1), 264–277. https://doi.org/10.3758/s13428-020-01407-2

Lubetkin, E. I., Jia, H., Franks, P., & Gold, M. R. (2005). Relationship among sociodemographic factors, clinical conditions, and health-related quality of life: Examining the EQ-5D in the U.S. general population. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *14*(10), 2187–2196. https://doi.org/10.1007/s11136-005-8028-5

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R Package for Assessment, Comparison and Testing of Statistical Models. Journal of *Open Source Software, 6*(60), 3139. https://doi.org/10.21105/joss.03139

Lyu, W., & Wolinsky, F. D. (2017). The Onset of ADL Difficulties and Changes in Health-Related Quality of Life. *Health and Quality of Life Outcomes, 15*(1), 217. https://doi.org/10.1186/s12955-017-0792-8

Madariaga, A. G., Palacios, S. S., Guillén-Grima, F., & Galofré, J. C. (2014). The Incidence and Prevalence of Thyroid Dysfunction in Europe: A Meta-Analysis. *The Journal of Clinical Endocrinology & Metabolism, 99*(3), 923–931. https://doi.org/10.1210/jc.2013-2409

Marshall, R. D., Collins, A., Escolar, M. L., Jinnah, H. A., Klopstock, T., Kruer, M. C., Videnovic, A., Robichaux-Viehoever, A., Swett, L., Revicki, D. A., Bender, R. H., & Lenderking, W. R. (2019). A Scale to Assess Activities of Daily Living in Pantothenate Kinase-Associated Neurodegeneration. *Movement Disorders Clinical Practice*, *6*(2), 139–149. https://doi.org/10.1002/mdc3.12716

Mahmoud, A., & Khalifa, B. (2015). A Confirmatory Factor Analysis for SERVPERF Instrument based on a Sample of Students from Syrian Universities. *Education and Training, 57*, 343–359. https://doi.org/10.1108/ET-04-2014-0038

Megari, K. (2013). Quality of Life in Chronic Disease Patients. *Health Psychology Research*, *1*(3), e27. https://doi.org/10.4081/hpr.2013.e27

Meng, Z., Liu, M., Zhang, Q., Liu, L., Song, K., Tan, J., Jia, Q., Zhang, G., Wang, R., He, Y., Ren, X., Zhu, M., He, Q., Wang, S., Li, X., Hu, T., Liu, N., Upadhyaya, A., Zhou, P., & Zhang, J. (2015). Gender and Age Impacts on the Association Between Thyroid Function and Metabolic Syndrome in Chinese. *Medicine, 94(*50), e2193. https://doi.org/10.1097/MD.0000000000002193

Meyers, L. S., Gamst, G. C., & Guarino, A. J. (2016). *Applied Multivariate Research: Design and Interpretation (Third edition).* SAGE Publications, Inc.

Mice.pdf. (n.d.). Retrieved June 24, 2021, from https://cran.r-project.org/web/packages/mice/mice.pdf

McMillan, C. V., Bradley, C., Woodcock, A., Razvi, S., & Weaver, J. U. (2004). Design of new questionnaires to measure quality of life and treatment satisfaction in hypothyroidism. *Thyroid: Official Journal of the American Thyroid Association, 14*(11), 916–925. https://doi.org/10.1089/thy.2004.14.916

Mokkink, L. B., de Vet, H. C. W., Prinsen, C. a. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *27*(5), 1171–1179. https://doi.org/10.1007/s11136-017-1765-4

Muñiz, J., Elosua, P., Hambleton, R. K., & International Test Commission. (2013). [International Test Commission Guidelines for test translation and adaptation: Second edition]. *Psicothema*, *25*(2), 151–157. https://doi.org/10.7334/psicothema2013.24

Muragundi, P., Tumkur, A., Shetty, R., & Naik, A. (2012). Health-related Quality of Life Measurement. *Journal of Young Pharmacists: JYP*, *4*(1), 54. https://doi.org/10.4103/0975-1483.93568

Netuveli, G., Wiggins, R. D., Hildon, Z., Montgomery, S. M., & Blane, D. (2006). Quality of life at older ages: Evidence from the English longitudinal study of aging (wave 1). *Journal of Epidemiology and Community Health*, *60*(4), 357–363. https://doi.org/10.1136/jech.2005.040071

Norris, C. M., Murray, J. W., Triplett, L. S., & Hegadoren, K. M. (2010). Gender roles in persistent sex differences in health-related quality-of-life outcomes of patients with coronary artery disease. *Gender Medicine, 7*(4), 330–339. https://doi.org/10.1016/j.genm.2010.07.005

Nunnally, J. C., & Nunnaly, J. C. (1978). *Psychometric Theory*. McGraw-Hill.

Orlando, M., Sherbourne, C. D., & Thissen, D. (2000). Summed-score linking using item response theory: Application to depression measurement. *Psychological Assessment, 12*(3), 354–359. https://doi.org/10.1037/1040-3590.12.3.354

Palmer, R. C., Fernandez, M. E., Tortolero-Luna, G., Gonzales, A., & Dolan Mullen, P. (2005). Acculturation and mammography screening among Hispanic women living in farmworker communities. *Cancer Control: Journal of the Moffitt Cancer Center*, *12 Suppl 2*, 21–27. https://doi.org/10.1177/10732748050120004S04

Papou, A., Hussain, S., McWilliams, D., Zhang, W., & Doherty, M. (2017). Responsiveness of SF-36 Health Survey and Patient Generated Index in people with chronic knee pain commenced on oral analgesia: Analysis of data from a randomised controlled clinical trial. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *26*(3), 761–766. https://doi.org/10.1007/s11136-016-1484-2

Pérez-Escamilla, R., & Putnik, P. (2007). The role of acculturation in nutrition, lifestyle, and incidence of type 2 diabetes among Latinos. *The Journal of Nutrition*, *137*(4), 860–870. https://doi.org/10.1093/jn/137.4.860

Ponto, K. A., Hommel, G., Pitz, S., Elflein, H., Pfeiffer, N., & Kahaly, G. J. (2011). Quality of Life in a German Graves Orbitopathy Population. *American Journal of Ophthalmology, 152*(3), 483-490.e1. https://doi.org/10.1016/j.ajo.2011.02.018

Qian, Y., & Xie, H. (2021). *Simplifying Bias Correction for Selective Sampling: A Unified Distribution-Free Approach to Handling Endogenously Selected Samples* (Working Paper No. 28801; Working Paper Series). National Bureau of Economic Research. https://doi.org/10.3386/w28801

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Radloff, L. S. (1977). The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. *Applied Psychological Measurement, 1*(3), 385–401. https://doi.org/10.1177/014662167700100306

Rainey, C. (2016). Dealing with Separation in Logistic Regression Models. *Political Analysis*, *24*(3), 339–355. https://doi.org/10.1093/pan/mpw014

Rammstedt, B., Danner, D., & Bosnjak, M. (2017). Acquiescence response styles: A multilevel model explaining individual-level and country-level differences. *Personality and Individual Differences*, *107*, 190–194. https://doi.org/10.1016/j.paid.2016.11.038

Rampazo-Lacativa, M. K., Santos, A. A. dos, Coimbra, A. M. V., & D'Elboux, M. J. (2015). WOMAC and SF-36: Instruments for evaluating the health-related quality of life of elderly people with total hip arthroplasty. A descriptive study. *Sao Paulo Medical Journal = Revista Paulista De Medicina*, *133*(4), 290–297. https://doi.org/10.1590/1516-3180.2014.8381508

Rasmussen, S. L., Rejnmark, L., Ebbehøj, E., Feldt-Rasmussen, U., Rasmussen, Å. K., Bjorner, J. B., & Watt, T. (2016). High Level of Agreement between Electronic and Paper Mode of Administration of a Thyroid-Specific Patient-Reported Outcome, ThyPRO. *European Thyroid Journal*, *5*(1), 65–72. https://doi.org/10.1159/000443609

Ryu, C. (2021). dlookr: Tools for Data Diagnosis, Exploration, Transformation. R package version 0.4.5. https://CRAN.R-project.org/package=dlookr.

Sararaks, S., Azman, A. B., Low, L. L., Rugayah, B., Aziah, A. M., Hooi, L. N., Abdul Razak, M., Norhaya, M. R., Lim, K. B., Azian, A. A., & Geeta, S. (2005). Validity and reliability of the SF-36: The Malaysian context. *The Medical Journal of Malaysia*, *60*(2), 163–179.

Schneider, S., & Stone, A. A. (2016). The meaning of vaguely quantified frequency response options on a quality of life scale depends on respondents' medical status and age. *Quality of Life Research, 25*(10), 2511–2521. https://doi.org/10.1007/s11136-016-1293-7

Sevinc, A., & Savli, H. (2004). Hypothyroidism masquerading as depression: The role of noncompliance. *Journal of the National Medical Association*, *96*(3), 379–382.

Shafie, S., Samari, E., Jeyagurunathan, A., Abdin, E., Chang, S., Chong, S. A., & Subramaniam, M. (2021). Gender difference in quality of life (QoL) among outpatients with schizophrenia in a tertiary care setting. *BMC Psychiatry, 21*(1), 61. https://doi.org/10.1186/s12888-021-03051-2

Shek, D. T. L. (2010). Introduction: Quality of Life of Chinese People in a Changing World. *Social Indicators Research, 95*(3), 357–361. https://doi.org/10.1007/s11205-009-9534-6

Siemiatycki, J. (1979). A comparison of mail, telephone, and home interview strategies for household health surveys. *American Journal of Public Health*, *69*(3), 238–245. https://doi.org/10.2105/ajph.69.3.238

Šimkovic, M., & Träuble, B. (2019). Robustness of statistical methods when measure is affected by ceiling and/or floor effect. *PLOS ONE*, *14*(8), e0220889. https://doi.org/10.1371/journal.pone.0220889

Smith, P. B., & Fischer, R. (2008). Acquiescence, extreme response bias and culture: A multilevel analysis. In *Multilevel analysis of individuals and cultures* (pp. 285–314). Taylor & Francis Group/Lawrence Erlbaum Associates.

Spriensma, A. S., Eekhout, I., Boer, M. R. D., Luime, J. J., Jong, P. H. D., Bahçecitapar, M. K., Heymans, M. W., & Twisk, J. W. R. (2018). Analysing outcome variables with floor effects due to censoring: A simulation study with longitudinal trial data. *Epidemiology Biostatistics and Public Health, 15*(2), 1–9. https://doi.org/10.2427/12850

Spriensma, A. S., Hajos, T. R. S., de Boer, M. R., Heymans, M. W., & Twisk, J. W. R. (2013). A new approach to analyse longitudinal epidemiological data with an excess of zeros. *BMC Medical Research Methodology*, *13*, 27. https://doi.org/10.1186/1471-2288-13-27

Strauss, J., Muday, T., McNall, K., & Wong, M. (1997). Response Style Theory Revisited: Gender Differences and Stereotypes in Rumination and Distraction. *Sex Roles*, *36*(11), 771–792. https://doi.org/10.1023/A:1025679223514

Sun, S., Chen, J., Johannesson, M., Kind, P., Xu, L., Zhang, Y., & Burström, K. (2011). Population health status in China: EQ-5D results, by age, sex and socio-economic status, from the National Health Services Survey 2008. *Quality of Life Research*, *20*(3), 309–320. https://doi.org/10.1007/s11136-010-9762-x

Swartz, R. J., de Moor, C., Cook, K. F., Fouladi, R. T., Basen-Engquist, K., Eng, C., & Carmack Taylor, C. L. (2007). Mode effects in the center for epidemiologic studies depression (CES-D) scale: Personal digital assistant vs. paper and pencil administration. *Quality of Life Research*, *16(*5), 803–813. https://doi.org/10.1007/s11136-006-9158-0

Szende, A., Janssen, B., & Cabases, J. (Eds.). (2014). *Self-Reported Population Health: An International Perspective based on EQ-5D*. Springer. http://www.ncbi.nlm.nih.gov/books/NBK500356/

Tacyildiz, N., Karakose, T., Unal, E. C., Dincaslan, H., Tanyildiz, G., & Çakmak, H. M. (2020). Evaluation of the quality-of-life (QOL) and socio-demographic characteristics of patients with leukemia and lymphoma: Comparison with sibling and control group. *Journal of Clinical Oncology, 38*(15_suppl), e22524–e22524. https://doi.org/10.1200/JCO.2020.38.15_suppl.e22524

Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., Bouter, L. M., & de Vet, H. C. W. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, *60*(1), 34–42. https://doi.org/10.1016/j.jclinepi.2006.03.012

Terwee, C. B., Gerding, M. N., Dekker, F. W., Prummel, M. F., & Wiersinga, W. M. (1998). Development of a disease specific quality of life questionnaire for patients with Graves' ophthalmopathy: The GO-QOL. *The British Journal of Ophthalmology, 82*(7), 773–779. https://doi.org/10.1136/bjo.82.7.773

Tlusta, E., Zarubova, J., Simko, J., Hojdikova, H., Salek, S., & Vlcek, J. (2009). Clinical and demographic characteristics predicting QOL in patients with epilepsy in the Czech Republic: How this can influence practice. *Seizure, 18*(2), 85–89. https://doi.org/10.1016/j.seizure.2008.06.006

Tourangeau, R., & Smith, T. W. (1996). Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Context. *The Public Opinion Quarterly*, *60*(2), 275–304.

Treanor, C., & Donnelly, M. (2015). A methodological review of the Short Form Health Survey 36 (SF-36) and its derivatives among breast cancer survivors. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *24*(2), 339–362. https://doi.org/10.1007/s11136-014-0785-6

Trompenaars, F. J., Masthoff, E. D., Van Heck, G. L., Hodiamont, P. P., & De Vries, J. (2005). Relationships between demographic variables and quality of life in a population of Dutch adult psychiatric outpatients. Social Psychiatry and Psychiatric Epidemiology, *40*(7), 588–594. https://doi.org/10.1007/s00127-005-0946-0

Trotti, A. (1998). *Development of a head and neck companion module for the quality of life-radiation therapy instrument (QOL-RTI).* 42(2), 5.

Twiss, J., McKenna, S., Ganderton, L., Jenkins, S., Ben-L'amri, M., Gain, K., Fowler, R., & Gabbay, E. (2013). Psychometric performance of the CAMPHOR and SF-36 in pulmonary hypertension. *BMC Pulmonary Medicine, 13*(1), 45. https://doi.org/10.1186/1471-2466-13-45

Ullah, S., Finch, C. F., & Day, L. (2010). Statistical modelling for falls count data. *Accident Analysis & Prevention, 42*(2), 384–392. https://doi.org/10.1016/j.aap.2009.08.018

Unruh, M., Yan, G., Radeva, M., Hays, R. D., Benz, R., Athienites, N. V., Kusek, J., Levey, A. S., & Meyer, K. B. (2003). Bias in Assessment of Health-Related Quality of Life in a Hemodialysis Population: A Comparison of Self-Administered and Interviewer-Administered Surveys in the HEMO Study. *Journal of the American Society of Nephrology, 14*(8), 2132–2141. https://doi.org/10.1097/01.ASN.0000076076.88336.B1

Vuong, Q. (1989). Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica, 57*, 307–333. https://doi.org/10.2307/1912557

Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2008). Investigating Ceiling Effects in Longitudinal Data Analysis. *Multivariate Behavioral Research, 43*(3), 476–496. https://doi.org/10.1080/00273170802285941

Wang, J., Ma, X., Qu, S., Li, Y., Han, L., Sun, X., Li, P., Liu, X., & Xu, J. (2013). High prevalence of subclinical thyroid dysfunction and the relationship between thyrotropin levels and cardiovascular risk factors in residents of the coastal area of China. *Experimental & Clinical Cardiology*, *18*(1), e16–e20.

Wang, T., Jiang, M., Ren, Y., Liu, Q., Zhao, G., Cao, C., & Wang, H. (2018). Health-Related Quality of Life of Community Thyroid Cancer Survivors in Hangzhou, China. *Thyroid: Official Journal of the American Thyroid Association, 28*(8), 1013–1023. https://doi.org/10.1089/thy.2017.0213

Ware, J., Snoww, K., MA, K., & BG, G. (1993). SF36 Health Survey: Manual and Interpretation Guide. *Lincoln, RI: Quality Metric, Inc, 1993*, *30*.

Warton, D. I. (2005). Many zeros does not mean zero inflation: Comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics, 16*(3), 275–289. https://doi.org/10.1002/env.702

Watt, T., English Thyroid-Specific Patient-Reported Outcome short-form (ThyPRO-39). *Unpublished Manuscript*.

Watt, T., Barbesino, G., Bjorner, J. B., Bonnema, S. J., Bukvic, B., Drummond, R., Groenvold, M., Hegedüs, L., Kantzer, V., Lasch, K. E., Marcocci, C., Mishra, A., Netea-Maier, R., Ekker, M., Paunovic, I., Quinn, T. J., Rasmussen, Å. K., Russell, A., Sabaretnam, M., … Feldt-Rasmussen, U. (2015). Cross-cultural validity of the thyroid-specific quality-of-life patient-reported outcome measure, ThyPRO. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *24*(3), 769–780. https://doi.org/10.1007/s11136-014-0798-1

Watt, T., Bjorner, J. B., Groenvold, M., Cramon, P., Winther, K. H., Hegedüs, L., Bonnema, S. J., Rasmussen, Å. K., Ware, J. E., & Feldt-Rasmussen, U. (2015). Development of a Short Version of the Thyroid-Related Patient-Reported Outcome ThyPRO. *Thyroid: Official Journal of the American Thyroid Association*, *25*(10), 1069–1079. https://doi.org/10.1089/thy.2015.0209

Watt, T., Bjorner, J. B., Groenvold, M., Rasmussen, A. K., Bonnema, S. J., Hegedüs, L., & Feldt-Rasmussen, U. (2009). Establishing construct validity for the thyroid-specific patient reported outcome measure (ThyPRO): An initial examination. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *18*(4), 483–496. https://doi.org/10.1007/s11136-009-9460-8

Watt, T., Cramon, P., Hegedüs, L., Bjorner, J. B., Bonnema, S. J., Rasmussen, Å. K., Feldt-Rasmussen, U., & Groenvold, M. (2014). The Thyroid-Related Quality of Life Measure ThyPRO Has Good Responsiveness and Ability to Detect Relevant Treatment Effects. *The Journal of Clinical Endocrinology & Metabolism*, *99*(10), 3708–3717. https://doi.org/10.1210/jc.2014-1322

Watt, T., Hegedüs, L., Groenvold, M., Bjorner, J. B., Rasmussen, A. K., Bonnema, S. J., & Feldt-Rasmussen, U. (2010). Validity and reliability of the novel thyroid-specific quality of life questionnaire, ThyPRO. *European Journal of Endocrinology*, *162*(1), 161–167. https://doi.org/10.1530/EJE-09-0521

Watt, T., Groenvold, M., Deng, N., Gandek, B., Feldt-Rasmussen, U., Rasmussen, Å. K., Hegedüs, L., Bonnema, S. J., & Bjorner, J. B. (2014). Confirmatory factor analysis of the thyroid-related quality of life questionnaire ThyPRO. *Health and Quality of Life Outcomes*, *12*(1), 126. https://doi.org/10.1186/s12955-014-0126-z

Watt, T., Groenvold, M., Hegedüs, L., Bonnema, S. J., Rasmussen, Å. K., Feldt-Rasmussen, U., & Bjorner, J. B. (2014). Few items in the thyroid-related quality of life instrument ThyPRO exhibited differential item functioning. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *23*(1), 327–338. https://doi.org/10.1007/s11136-013-0462-1

Watt, T., Rasmussen, A. K., Groenvold, M., Bjorner, J. B., Watt, S. H., Bonnema, S. J., Hegedüs, L., & Feldt-Rasmussen, U. (2008). Improving a newly developed patient-reported outcome for thyroid patients, using cognitive interviewing. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *17*(7), 1009–1017. https://doi.org/10.1007/s11136-008-9364-z

Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2009). Investigating Ceiling Effects in Longitudinal Data Analysis. *Multivariate Behavioral Research*, *43*(3), 476–496. https://doi.org/10.1080/00273170802285941

West, C., Paul, S., Dunn, L., Dhruva, A., Merriman, J., & Miaskowski, C. (2015). Gender Differences in Predictors of Quality of Life at the Initiation of Radiation Therapy. *Oncology Nursing Forum, 42*(5), 507–516. https://doi.org/10.1188/15.ONF.507-516

Wilson, P. (2015). The misuse of the Vuong test for non-nested models to test for zero-inflation. *Economics Letters*, *127*, 51–53. https://doi.org/10.1016/j.econlet.2014.12.029

Winkelmann, R. (2008). *Econometric Analysis of Count Data (5th ed.).* Springer-Verlag. https://doi.org/10.1007/978-3-540-78389-3

Ware, J. E., & Gandek, B. (1998). Methods for testing data quality, scaling assumptions, and reliability: The IQOLA Project approach. International Quality of Life Assessment. *Journal of Clinical Epidemiology, 51*(11), 945–952. https://doi.org/10.1016/s0895-4356(98)00085-7

Wong, C. K. H., Choi, E. P. H., Woo, Y. C., & Lang, B. H. H. (2018). Measurement properties of ThyPRO short-form (ThyPRO-39) for use in Chinese patients with benign thyroid diseases. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *27*(8), 2177–2187. https://doi.org/10.1007/s11136-018-1857-9

Wong, C. K. H., Lang, B. H. H., & Lam, C. L. K. (2016). A systematic review of quality of thyroid-specific health-related quality-of-life instruments recommends ThyPRO for patients with benign thyroid diseases. *Journal of Clinical Epidemiology*, *78*, 63–72. https://doi.org/10.1016/j.jclinepi.2016.03.006

World Health Organization. (2018). *International classification of diseases for mortality and morbidity statistics (11th Revision).* Retrieved from https://icd.who.int/browse11/l-m/en

Wu, Q., Ge, T., Emond, A., Foster, K., Gatt, J. M., Hadfield, K., Mason-Jones, A. J., Reid, S., Theron, L., Ungar, M., & Wouldes, T. A. (2018). Acculturation, resilience, and the mental health of migrant youth: A cross-country comparative study. *Public Health*, *162*, 63–70. https://doi.org/10.1016/j.puhe.2018.05.006

Yang, H., Tian, S., Flamand-Roze, C., Gao, L., Zhang, W., Li, Y., Wang, J., Sun, Z., Su, Y., Zhao, L., & Liang, Z. (2018). A Chinese version of the Language Screening Test (CLAST) for early-stage stroke patients. *PLoS ONE*, *13*(5), e0196646. https://doi.org/10.1371/journal.pone.0196646

Yarlas, A., Bayliss, M., Cappelleri, J. C., Maher, S., Bushmakin, A. G., Chen, L. A., Manuchehri, A., & Healey, P. (2018). Psychometric validation of the SF-36® Health Survey in ulcerative colitis: Results from a systematic literature review. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation*, *27*(2), 273–290. https://doi.org/10.1007/s11136-017-1690-6

Yin, S., Njai, R., Barker, L., Siegel, P. Z., & Liao, Y. (2016). Summarizing health-related quality of life (HRQOL): Development and testing of a one-factor model. *Population Health Metrics*, *14*, 22. https://doi.org/10.1186/s12963-016-0091-3

Youssof, S., Romero-Clark, C., Warner, T., & Plowman, E. (2017). Dysphagia-related quality of life in oculopharyngeal muscular dystrophy: Psychometric properties of the SWAL-QOL instrument. *Muscle & Nerve*, *56*(1), 28–35. https://doi.org/10.1002/mus.25441

Zaninotto, P., & Falaschetti, E. (2011). Comparison of methods for modelling a count outcome with excess zeros: Application to Activities of Daily Living (ADL-s). *Journal of Epidemiology & Community Health, 65*(3), 205–210. https://doi.org/10.1136/jech.2008.079640

Zhang, X., Noor, R., & Savalei, V. (2016). Examining the Effect of Reverse Worded Items on the Factor Structure of the Need for Cognition Scale. *PLOS ONE, 11*(6), e0157795. https://doi.org/10.1371/journal.pone.0157795

Zhang, Y., Zhou, F., & Sun, Y. (2015). Assessment of health-related quality of life using the SF-36 in Chinese cervical spondylotic myelopathy patients after surgery and its consistency with neurological function assessment: A cohort study. *Health and Quality of Life Outcomes*, *13*, 39. https://doi.org/10.1186/s12955-015-0237-1

Zhou, K., Zhuang, G., Zhang, H., Liang, P., Yin, J., Kou, L., Hao, M., & You, L. (2013). Psychometrics of the Short Form 36 Health Survey Version 2 (SF-36v2) and the Quality of Life Scale for Drug Addicts (QOL-DAv2.0) in Chinese Mainland Patients with Methadone Maintenance Treatment. *PLoS ONE, 8*(11), e79828. https://doi.org/10.1371/journal.pone.0079828

Zhu, L., & Gonzalez, J. (2017). Modeling Floor Effects in Standardized Vocabulary Test Scores in a Sample of Low SES Hispanic Preschool Children under the Multilevel Structural Equation Modeling Framework. *Frontiers in Psychology, 8,* 2146. https://doi.org/10.3389/fpsyg.2017.02146

Appendix A. Simplified Chinese ThyPRO-39

Liew, H., Simplified Chinese Thyroid-Specific Patient-Reported Outcome short-form (ThyPRO-39). *Unpublished Manuscript*.

Please note that the copy of the Simplified Chinese version of ThyPRO-39 is not included in this thesis due to copyright. Please contact Dr. Torquil Watt for the complete form (https://ikm.ku.dk/english/ansatte/?pure=en/persons/72342).

Appendix B. English Version of ThyPRO-39

Watt, T., English Thyroid-Specific Patient-Reported Outcome short-form (ThyPRO-39). *Unpublished Manuscript*.

Watt, T., Bjorner, J. B., Groenvold, M., Cramon, P., Winther, K. H., Hegedüs, L., Bonnema, S. J., Rasmussen, Å. K., Ware, J. E., & Feldt-Rasmussen, U. (2015). Development of a Short Version of the Thyroid-Related Patient-Reported Outcome ThyPRO. *Thyroid: Official Journal of the American Thyroid Association*, *25*(10), 1069–1079. https://doi.org/10.1089/thy.2015.0209

Please note that the copy of the English version of ThyPRO-39 is not included in this thesis due to copyright. Please contact Dr. Torquil Watt for the complete form (https://ikm.ku.dk/english/ansatte/?pure=en/persons/72342).

Appendix C. Qualitative Questionnaire in Simplified Chinese

对于 ThyPRO-39 理解的问卷

这份问卷是调查您对于上一份问卷的理解。

1. 对于第 21 题“感到害怕或焦虑”，请问您觉得“害怕”和“焦虑”的意思相同或者不同？请阐述具体原因

_____

2. 对于第 26 题 “您有感到自信吗？”，请问您在回答这一题的时候有没有把自信和您的甲状腺疾病联系在一起，还是说您只是回答了大体上的“自信”，与甲状腺疾病无关？

_____

3. 对于第 29 题 “您有感到自己的生活在掌控之中吗？” 和 33 题 “有困难处理日常生活吗？”， 请问您觉得“生活在掌控之中”和“处理日常生活”的意思相同或者不同？请阐述具体原因

_____

4. 对于第 34 题“您的甲状腺疾病是否造成您不能参与日常的活动？”，不考虑甲状腺疾病对您的影响，请问在什么情况下您才会觉得您的健康状况使您“不能参与日常的活动”？（考虑到严重程度，例如：发高烧时，重感冒时，胃疼时，失眠等）

_____

5. 如果您需要从以下五中选择中回答“甲状腺疾病对我的生活质量没有负面影响。”，请问您会选哪一个？

强烈同意　　同意　　中立　　不同意　　强烈不同意

6.1. 如果将 37 题 “您会在意其他人对您的眼光吗？”改成 ”您会在意其他人看着您吗？” 您会做出什么样的回答？

完全没有　　有一点　　有些　　多一点　　很多

6.2. 请问您觉得 “眼光” 和 “看着您” 的意思相同或者不同？请您解释具体原因。

_____

7. 请问在以上的甲状腺生活质量调查问卷中，有哪些问题您觉得特别的疑惑或者难以理解？可以请您解释为什么会觉得疑惑或者难以理解吗？

_____

8. 请问您觉得有什么您关心的健康相关的问题在这份问卷中没有被提及的吗？

_____

Appendix D. Qualitative Questionnaire in Simplified Chinese

Qualitative Questionnaire to ThyPRO-39

This questionnaire contains questions about responding to the Thyroid questionnaire in different languages. Please choose and/or type your response under each question.

1. For item 21 “Felt afraid or anxious”, Do you think “afraid” and “anxious” are the same or different in meaning? And why?
2. For item 26, “Had self-confidence.”, 1) when you answered this question, did you relate the “self-confidence” to your thyroid diseases? Or you just answered in general and did not relate to your thyroid diseases?
3. For item 29 “Felt in control of your life.” and item 33 “Had difficulty managing your daily life”. 1) Do you think “in control of life” and “manage daily life” are the same or different in meaning? Please provide reasons.
4. For item 34, “your thyroid disease caused you to not be able to participate in life around you”. In what situation do you think you are “not able to participate in life around you” because of your health condition? (For example, were you unable to participate in life around you because you had a high fever, serious cold, serious stomachache, insomnia…etc.)
5. If you are asked to rate the item “The thyroid disease did not have a negative effect on my quality of life.” Which one will you choose?

Strongly agree Neutral Disagree Strongly disagree

6. If the question in item 37 of “**您会在意其他人**对您的眼光吗？” changed to “**您会在意其他人看着您**吗？”, how you would answer?

Not at all A little Some Quite a bit Very much

7. Were there any of items in the questionnaire that you found particularly confusing or difficult to process? Can you describe why you feel confusing or difficult to process?
8. Do you have any health-related concerns that you think are not mentioned in the thyroid questionnaire?

Appendix E. Demographic Instrument in Simplified Chinese

1. 请问您来自哪个城市/地区?

_____

2. 请问您的年龄是?

_____

3. 请问您的性别是?

_____

4. 请问您已完成的最高学历是?

A. 高中以下学历

B. 高中或同等学历

C. 大专学历

D. 本科学历

E. 研究生/硕士或同等学历

F. 博士生

G. 其他 _____ (请具体说明)

5. 请问您现在的工作情况是?

A. 全职工作

B. 兼职工作

C. 在校学生

D. 无工作, 并没有在找工作

E. 正在寻找工作

F. 已退休

G. 无法工作

6. 如果您正在工作，请具体写明您的职业:

_____

7. 请问您属于下列哪一种甲状腺疾病?

A. 甲状腺功能减退症 (甲减)

B. 甲状腺功能亢进症 (甲亢)

C. 非毒性甲状腺肿大

D. 其他 _____ (请具体说明)

8. 请问您的甲状腺疾病的具体名称是 (如果您知道，例: 桥本氏甲状腺炎，葛瑞夫兹氏病，甲状腺结节)

_____

9. 请问您接受治疗多久了?

_____

10. 请问您觉得您的甲状腺疾病已经控制住了吗 (从 0 到4， 0 = ”完全没有”, 4 = “控制得非常好”)?

0 1 2 3 4

Appendix F. Demographic Instrument in English

1. Which area/region are you from?

_____

2. What is your age?

_____

3. What gender do you identify as?

_____

4. What is the highest degree or level of education you have completed?

A. Less than a high school diploma
B. High school degree or equivalent
C. Diploma
D. Bachelor's degree (e.g. BA. BS)
E. Master's degree (e.g. MA, MS, MEd)
F. Doctorate (e.g. PhD, EdD)
G. Other _____ (Please specify)

5. What is your current employment status?

A. Employed Full-Time
B. Employed Part-Time
C. Student
D. Unemployment (not currently looking for work)
E. Unemployment (currently looking for work)
F. Retired
G. Unable to work

6. If you are employed, please specify your occupation.

_____

7. Please indicate your type of thyroid disease:

A. Hyperthyroidism
B. Hypothyroidism
C. non-toxic goiter
H. Other _____ (Please specify)

8. Please specify your subtype of thyroid disease (if known, e.g.: Hashimoto's disease, Graves' disease, nodular goiter):

_____

9. How long have you been treated?

_____

10. Do you think your disease is under control (from 0 to 4, 0 = "Not at all", 4 = "Very much")?

0 1 2 3 4

Appendix G. Tables for Background Information

Content of the tables in Appendix G

Table G1. Summary of Symptoms of Common Thyroid Diseases [1]

Type of thyroid diseases	Hypothyroidism	Hyperthyroidism
Common Symptoms	Enlarged neck	Thyroid growths, called goiters
	Frequent exhaustion/Fatigue	Fatigue
	Menstrual problems	Menstrual problems
	Insomnia	Insomnia
	Thinning or dry hair	Thinning hair
	Weak muscles	Weak muscles
	Gaining weight for no apparent reason	Losing weight for no apparent reason (or gaining weight due to an abrupt increase in appetite)
Reversed Symptoms	Slowed thinking	A racing mind that makes it difficult to focus
	Slowed heart rate	Heart pounding in chest at over 90 beats per minute while at rest
	Depression	Severe anxiety and panic attacks
	Lowered interest in sex	A sex drive that's in overdrive
	Constipation	Frequent bowel movements and/or a loose stool
	Sluggishness	Irritability
	Rough, itchy, and/or thinning skin	Tingling in the hands and feet
	Feeling unusually cold	Feeling overheated
	Sweating too little	Over-sweating (especially in the head, hands, and feet)
	Numbness in the hands and feet	Shakiness, including hand tremors
Disease-specific symptoms	Dry, brittle nails	Eye sensitive to light
	Hair loss	A dry, gritty feeling in the eyes
	Infertility	Enlarged, protruding eyes (creating a "bug-eyed" look)
	Memory problems	
	Acne	
	Puffy skin	
	Cold skin	
	Hoarse voice	

Note: [1] Table is summarized based on Christianson & Bender (2011).

Table G2. Measurement Properties for Original Version of ThyPRO

Measurement properties	Analyses	Results	Methodological qualities [8]
Internal consistency [1]	Cronbach's alpha	α > 0.70 except for except for hypothyroid symptoms	Satisfactory
Reliability [2]	Repeated measures in two weeks and intra-class correlations	> 0.70 for all scales	Satisfactory
Content validity [3]	Cognitive interviewing	Strong	Satisfactory
Structural validity	Confirmatory factor analysis (CFA) [4.1]	11 misfit items are eliminated and rest show unidimensionality	Satisfactory
	Differential item functioning (DIF) [4.2]	DIF with diagnosis and age, but small	
Construct validity [5]	Known-group validity is analyzed by comparison between groups with expected low or high scores	Results confirm the known-group validity	Moderate
	Convergent validity is analyzed by a comparison to another scale (HADS)	Small correlation between ThyPRO and HADS but reasonable	
Cross-cultural validity [6]	DIF	12 items show DIF with countries, but small	Moderate
Responsiveness [7]	Comparison before and after treatment, and relative validity to another scale (SF-36)	Effect size > 0.80 for change after treatment, higher responsiveness than SF-36	Satisfactory

Note: [1] was examined by Watt et al. (2009). [2] was examined by Watt et al. (2010).
[3] was examined by Watt et al. (2008). [4.1] was examined by Watt et al. (2014).
[4.2] was examined by Watt et al. (2014). [5] was examined by Watt et al. (2010).
[6] was examined by Watt et al. (2015). [7] was examined by Watt et al. (2014).
[8] was examined by Wong et al. (2018).

Table G3. Summary of CFE on HRQOL measurements

Measurement	Type of HRQOL	Language	CFE	Details on CFE
SWAL-QOL[1]	Disease-specific	Original English	Yes	Significant ceiling effect on 40/44 items
PU-QOL[2]	Disease-specific	Original English	Yes	Significant floor effects on 4/10 scales
NEI VFQ-25[3]	Disease-specific	Original English	Yes	Substantial ceiling effect on 4/12 subscales; significant ceiling effect on 3/12 subscales
PKAN-ADL[4]	Disease-specific	Original English	Yes	Substantial floor effect on 5/12 items; substantial ceiling effect on 3/12 items
SF-36	Generic	Original English[5]	Yes	Significant floor effect on 2/8 scales; significant ceiling effect on 5/8 scales
		Simplified Chinese[6]	Yes	Significant ceiling effect on 5/8 scales
EQ-5D	Generic	Original English[7]	Yes	Significant ceiling effect on total scores among type 2 diabetes patients (16.1%), benign prostatic hyperplasia patients (45%), and general populations (47%)
		Simplified Chinese[8]	Yes	All scales showed significant ceiling effects
Neuro-QoLs	Disease-specific	Original English[9]	Yes	3 out of 4 scales showed significant floor effect, but it was stated as reliable
GO-QOL	Disease-specific	Original Dutch[10]	Not specified	n/a
		Traditional Chinese[11]	No	No significant CFE on subscales, no data provided with individual item
LAST	Disease-specific	Simplified Chinese[12]	No	No significant CFE in all domains
QOL-RTI	Disease-specific	Original English[13]	Not specified	n/a
		Simplified Chinese[14]	Yes	Significant ceiling effect on 16/20 items
ThyPRO-39	Disease-specific	Original Danish[15]	Not specified	n/a
		Traditional Chinese[16]	Yes	Significant ceiling effect on 9/13 scales

Note: n/a = not applicable.
[1] Youssof et al., 2017, [2] Gorecki et al., 2013, [3] Bradley et al., 2006, [4] Marshall et al., 2019, [5] Bunevicius, 2017, [6] Huang et al., 2017, [7] Bharmal & Thomas, 2006, [8] Ferreira et al., 2016, [9] Carlozzi et al., 2018, [10] Terwee et al., 1998, [11] Lin et al., 2015, [12] ,[12] Yang et al., 2018, [13] Trotti, 1998, [14] Chen et al., 2014, [15] Watt et al., 2015, [16] Wong at al., 2018.

Table G4. Details for Scales in ThyPRO-39

Categories	Scale	Number of items	Min & maximum raw scores of scales	Test-retest reliability English form [1]	Cronbach's α original English form (84 items) [2]	Cronbach's α Traditional Chinese form [3]
Physical symptoms	Goiter symptoms	3	[0, 12]	0.83	0.89	0.80
	Hyperthyroid symptoms	4	[0, 16]	0.89	0.82	0.60
	Hypothyroid symptoms	4	[0, 16]	n/a	0.70	0.66
	Eye symptoms	3	[0, 12]	0.78	0.84	0.68
Psychological symptoms	Anxiety	3	[0, 12]	0.75	0.90	0.90
	Depressivity	3	[0, 12]	0.86	0.92	0.65
Well-being and function	Tiredness	3	[0, 12]	0.84	0.94	0.65
	Cognitive complaints	3	[0, 12]	0.84	0.93	0.88
	Emotional Susceptibility	3	[0, 12]	0.82	0.93	0.63
Participation	Impaired social life	3	[0, 12]	0.80	0.82	0.74
	Impaired daily life	3	[0, 12]	0.82	0.94	0.84
	Cosmetic complaints	3	[0, 12]	0.75	0.94	0.83
Single Item	Overall quality of life	1	[0, 4]	n/a	n/a	n/a
Composite	Composite	22	[0, 88]	0.90	n/a	0.929

Note: [1] Watt et al. (2015), [2] Watt et al. (2009), [3] Wong et al. (2018) ; n/a= not applicable.

Appendix H. Participants Information

Table H1. Demographic characteristics (N = 179)

Characteristic	%	Mean	SD
Age (in years)		37	13
Gender			
Male	32.09		
Female	67.91		
Mode of administration			
Interview	43.08		
Self-administration	57.92		
Education level			
Lower than high school	4.81		
High school	13.90		
College Diploma	17.11		
Bachelor's degree	48.66		
Master's degree	14.43		
PhD	1.07		
Duration of treatment [1]			
No treatment at all	26.23		
Less than 1 year	38.80		
1 to 2 years	9.29		
2 to 3 years	6.01		
3 to 4 years	4.37		
4 to 5 years	7.10		
5 + years	8.20		
Levels of control			
0 = not at all	6.42		
1	9.09		
2	13.90		
3	27.81		
4 = very much	42.78		

Note: [1]: The duration of treatment was an open-ended question in the demographic questionnaire, but participants provided different units, some were based on days, some were based on months. and most of them cannot provide a specific amount of time based on days or months. Therefore, the duration of treatment variable was categorized into 7 levels shown above.

Appendix I. Tables for Descriptive Statistics and for RQ 1

Content of the tables in Appendix I

Table I1. Descriptive Statistics at the Scale Level

Scale	Mean	SD	Skewness	Kurtosis	Floor (%)	Ceiling (%)
Goiter symptoms	2.86	2.60	1.08	1.06	**22.35%**	0.56%
Hyperthyroid symptom	4.08	3.11	0.77	0.29	12.29%	0.00%
Hypothyroid symptom	4.54	3.29	0.78	-0.20	6.15%	0.00%
Eye symptoms	2.99	2.57	0.97	0.29	14.53%	0.00%
Tiredness	5.31	2.68	0.45	-0.37	0.56%	2.23%
Cognitive complaints	3.54	2.76	0.90	0.50	12.29%	1.12%
Anxiety	3.63	2.91	0.85	0.01	12.29%	1.12%
Depressivity	4.62	2.78	0.55	-0.25	4.47%	1.12%
Emotional susceptibility	5.51	2.50	0.14	-0.28	1.68%	1.68%
Impaired social life	2.21	2.64	1.41	1.57	**37.43%**	0.56%
Impaired daily life	2.02	2.59	1.66	2.37	**36.87%**	0.00%
Cosmetic complaints	2.94	2.88	1.25	1.05	**17.88%**	1.68%
Composite	25.95	13.89	0.69	0.09	0.56%	0.00%

Note: **Bolding texts** indicate significant CFE (>15%)

Table I2. Percentage of Missing Data on Demographic variables

Variables	Missing (%)
Gender	4.10
Age	3.07
Duration of Treatment	5.64
Education Level	2.56

Table I3. Descriptive Statistics at the Item Level

Scale	Item	Mean	SD	Skewness	Kurtosis	Floor (%)	Ceiling (%)
Goiter	tq1a	0.99	0.96	0.89	0.44	**35.20%**	6.15%
symptoms	tq1c	1.02	1.00	1.03	0.81	**34.08%**	6.15%
	tq1h	0.85	1.08	1.22	0.65	**50.84%**	7.82%
Hyperthyroid	tq1l	0.60	0.90	1.53	1.89	**62.01%**	3.35%
symptoms	tq1m	1.22	1.23	0.82	-0.30	**35.75%**	9.50%
	tq1n	1.34	1.20	0.68	-0.44	**28.49%**	11.73%
	tq1t	0.92	1.08	1.22	0.91	**44.13%**	6.15%
Hypothyroid	tq1q	1.49	1.26	0.57	-0.75	**24.58%**	**15.08%**
Symptoms	tq1cc	0.78	1.07	1.37	1.04	**55.31%**	7.26%
	tq1dd	1.22	1.21	0.92	-0.06	**32.40%**	9.50%
	tq1ee	1.06	1.11	0.88	-0.02	**39.66%**	8.94%
Eye	tq1w	1.09	1.03	0.88	0.32	**31.84%**	7.82%
symptoms	tq1x	0.92	1.05	1.07	0.40	**44.69%**	8.38%
	tq1bb	0.98	1.05	1.06	0.62	**40.22%**	6.15%
Tiredness	tq2a	1.85	1.16	0.49	-0.79	7.26%	**16.76%**
	tq2c	1.17	1.11	0.99	0.49	**30.73%**	5.59%
	tq3b	1.71	1.28	0.40	-0.85	**18.99%**	12.29%
Cognitive	tq4a	1.27	1.04	0.76	0.20	**23.46%**	8.38%
Complaints	tq4b	1.13	0.99	0.85	0.45	**28.49%**	6.70%
	tq4f	1.14	1.13	0.96	0.32	**33.52%**	6.70%
Anxiety	tq5b	1.38	1.14	0.79	-0.02	**21.79%**	8.94%
	tq5c	1.19	1.06	0.91	0.42	**27.37%**	7.26%
	tq5e	1.06	1.15	1.02	0.22	**39.66%**	8.38%
Depressivity	tq6a	1.28	1.14	0.78	-0.11	**27.37%**	10.06%
	tq6e	1.37	1.12	0.83	0.07	**20.67%**	9.50%
	tq6g	2.03	1.29	0.14	-1.17	10.61%	**21.23%**
Emotional	tq7c	1.68	1.16	0.57	-0.59	11.73%	**15.64%**
susceptibility	tq7d	1.77	1.11	0.41	-0.63	9.50%	**17.88%**
	tq7h	1.94	1.23	0.08	-1.07	12.29%	**26.26%**
Impaired	tq8a	0.78	1.12	1.46	1.42	**56.98%**	3.35%
social life	tq8b	0.73	1.11	1.47	1.19	**61.45%**	6.70%
	tq8c	0.70	0.94	1.46	1.81	**54.19%**	4.47%
Impaired daily	tq9a	0.55	0.91	1.94	3.76	**64.80%**	2.23%
life	tq9c	0.64	1.09	1.83	2.46	**65.92%**	5.03%
	tq9e	0.84	1.02	1.22	0.89	**48.04%**	6.70%
Cosmetic	tq11a	1.07	1.19	1.06	0.20	**40.22%**	8.94%
Complaints	tq11d	1.12	1.20	1.03	0.22	**37.99%**	6.70%
	tq11e	0.75	1.14	1.44	1.04	**60.89%**	7.26%
	tq12	1.03	1.09	1.06	0.59	**37.99%**	6.15%

Note: **Bolding texts** indicate significant CFE (>15%).
Items starting with tq10 were removed for the ThyPRO-39 from ThyPRO, which were items related to impaired sex life.

Table I4. Descriptive Statistics at the Scale Level Based on Gender

Male (N = 56)						
Scale	Mean	SD	Skewness	Kurtosis	Floor (%)	Ceiling (%)
Goiter symptoms	4.25	3.06	0.73	-0.21	8.93%	1.79%
Hyperthyroid symptom	4.5	3.49	0.98	0.50	8.93%	0.00%
Hypothyroid symptom	3.82	3.12	1.22	0.75	5.36%	0.00%
Eye symptoms	2.59	2.40	1.45	1.77	12.50%	0.00%
Tiredness	5.3	2.55	0.69	-0.33	0.00%	1.79%
Cognitive complaints	3.46	2.99	1.03	0.54	14.29%	1.79%
Anxiety	3.64	2.96	0.92	0.17	12.50%	1.79%
Depressivity	4.96	3.03	0.29	-0.78	7.14%	1.79%
Emotional susceptibility	5.43	2.49	-0.21	-0.66	3.57%	0.00%
Impaired social life	2.73	3.01	1.22	0.86	**32.14%**	1.79%
Impaired daily life	3.07	3.14	1.08	0.11	**21.43%**	0.00%
Cosmetic complaints	3.59	2.98	0.95	0.20	10.71%	1.79%
Composite	27.93	15.56	0.65	-0.14	0.00%	0.00%
Female (N = 123)						
Scale	Mean	SD	Skewness	Kurtosis	Floor (%)	Ceiling (%)
Goiter symptoms	2.23	2.09	0.80	-0.04	**28.46%**	0.00%
Hyperthyroid symptom	3.89	2.91	0.48	-0.64	13.82%	0.00%
Hypothyroid symptom	4.86	3.33	0.59	-0.51	6.50%	0.00%
Eye symptoms	3.17	2.64	0.75	-0.24	**15.45%**	0.00%
Tiredness	5.31	2.74	0.35	-0.50	0.81%	2.44%
Cognitive complaints	3.57	2.66	0.78	0.21	11.38%	0.81%
Anxiety	3.63	2.9	0.78	-0.22	12.20%	0.81%
Depressivity	4.46	2.66	0.66	-0.03	3.25%	0.81%
Emotional susceptibility	5.54	2.51	0.29	-0.27	0.81%	2.44%
Impaired social life	1.97	2.44	1.38	1.28	**39.84%**	0.00%
Impaired Daily life	1.54	2.15	1.85	3.53	**43.90%**	0.00%
Cosmetic complaints	2.65	2.79	1.38	1.44	**21.14%**	1.63%
Composite	25.05	13.03	0.61	-0.22	0.81%	0.00%

Note: **Bolding texts** indicate significant CFE (>15%)

Table I5. Descriptive Statistics at the Item Level for Males (N = 56)

Item	Mean	SD	Skewness	Kurtosis	Floor (%)	Ceiling (%)
tq1a	1.48	1.03	0.44	-0.38	**16.07%**	12.50%
tq1c	1.38	1.09	0.65	-0.16	**21.43%**	8.93%
tq1h	1.39	1.34	0.56	-1.01	**33.93%**	**16.07%**
tq1l	0.80	1.13	1.33	0.73	**55.36%**	8.93%
tq1m	1.54	1.19	0.36	-0.74	**23.21%**	12.50%
tq1n	1.39	1.23	0.56	-0.72	**28.57%**	12.50%
tq1t	0.77	0.93	0.99	-0.06	**50.00%**	7.14%
tq1q	1.16	1.14	0.77	-0.38	**33.93%**	12.50%
tq1cc	0.66	1.07	1.66	1.95	**62.50%**	5.36%
tq1dd	1.07	1.01	0.91	0.19	**30.36%**	10.71%
tq1ee	0.93	1.04	0.71	-0.83	**46.43%**	10.71%
tq1w	0.95	0.96	0.71	-0.54	**39.29%**	8.93%
tq1x	0.8	0.96	1.23	1.13	**46.43%**	5.36%
tq1bb	0.84	1.04	1.17	0.52	**48.21%**	8.93%
tq2a	1.77	1.03	0.46	-0.36	7.14%	14.29%
tq2c	1.18	1.13	1.14	0.60	**26.79%**	7.14%
tq3b	1.64	1.27	0.36	-0.93	**21.43%**	14.29%
tq4a	1.07	0.99	0.75	0.01	**32.14%**	7.14%
tq4b	1.18	1.10	0.87	0.22	**30.36%**	5.36%
tq4f	1.21	1.20	0.95	-0.08	**30.36%**	10.71%
tq5b	1.25	1.12	0.97	0.39	**25.00%**	5.36%
tq5c	1.16	1.06	1.04	0.72	**26.79%**	5.36%
tq5e	1.23	1.28	0.90	-0.35	**33.93%**	10.71%
tq6a	1.27	1.21	0.69	-0.63	**32.14%**	14.29%
tq6e	1.48	1.24	0.69	-0.49	**21.43%**	8.93%
tq6g	1.79	1.28	0.45	-1.03	12.50%	**16.07%**
tq7c	1.46	1.08	0.65	-0.27	**16.07%**	12.50%
tq7d	1.68	1.13	0.42	-0.73	12.50%	**17.86%**
tq7h	1.71	1.25	0.26	-1.07	17.86%	**21.43%**
tq8a	1.02	1.17	0.98	0.08	**44.64%**	5.36%
tq8b	0.84	1.23	1.27	0.36	**58.93%**	8.93%
tq8c	0.88	1.11	1.17	0.49	**50.00%**	7.14%
tq9a	0.86	1.09	1.37	1.39	**48.21%**	1.79%
tq9c	0.93	1.23	1.10	0.01	**53.57%**	8.93%
tq9e	1.29	1.19	0.54	-0.85	**32.14%**	**16.07%**
tq11a	1.30	1.19	0.62	-0.61	**30.36%**	12.50%
tq11d	1.30	1.14	0.69	-0.34	**26.79%**	10.71%
tq11e	0.98	1.31	0.93	-0.59	**57.14%**	12.50%
tq12	1.32	1.16	0.80	-0.15	**25.00%**	8.93%

Note: **Bolding texts** indicate significant CFE (>15%)
Items starting with tq10 were removed for the ThyPRO-39 from ThyPRO, which were items related to impaired sex life.

Table I6. Descriptive Statistics at the Item Level for Females (N = 123)

Item	Mean	SD	Skewness	Kurtosis	Floor (%)	Ceiling (%)
tq1a	0.77	0.85	1.09	1.10	**43.90%**	3.25%
tq1c	0.85	0.91	1.19	1.38	**39.84%**	4.88%
tq1h	0.60	0.84	1.26	0.73	**58.54%**	4.07%
tq1l	0.50	0.76	1.20	0.16	**65.04%**	0.81%
tq1m	1.08	1.23	1.05	0.06	**41.46%**	8.13%
tq1n	1.32	1.19	0.71	-0.41	**28.46%**	11.38%
tq1t	0.99	1.13	1.19	0.69	**41.46%**	5.69%
tq1q	1.63	1.29	0.45	-0.95	**20.33%**	**16.26%**
tq1cc	0.83	1.08	1.20	0.52	**52.03%**	8.13%
tq1dd	1.28	1.28	0.82	-0.41	**33.33%**	8.94%
tq1ee	1.11	1.14	0.90	0.01	**36.59%**	8.13%
tq1w	1.16	1.05	0.89	0.34	**28.46%**	7.32%
tq1x	0.97	1.09	0.96	-0.01	**43.90%**	9.76%
tq1bb	1.04	1.06	0.99	0.52	**36.59%**	4.88%
tq2a	1.89	1.22	0.46	-1.03	7.32%	**17.89%**
tq2c	1.16	1.11	0.89	0.24	**32.52%**	4.88%
tq3b	1.74	1.29	0.40	-0.91	**17.89%**	11.38%
tq4a	1.36	1.05	0.73	0.09	**19.51%**	8.94%
tq4b	1.11	0.95	0.76	0.24	**27.64%**	7.32%
tq4f	1.11	1.09	0.91	0.31	**34.96%**	4.88%
tq5b	1.44	1.15	0.68	-0.29	**20.33%**	10.57%
tq5c	1.20	1.06	0.81	0.12	**27.64%**	8.13%
tq5e	0.98	1.09	0.98	0.21	**42.28%**	7.32%
tq6a	1.28	1.11	0.80	0.04	**25.20%**	8.13%
tq6e	1.32	1.06	0.83	0.16	**20.33%**	9.76%
tq6g	2.14	1.28	0.00	-1.20	9.76%	**23.58%**
tq7c	1.77	1.19	0.50	-0.82	9.76%	**17.07%**
tq7d	1.81	1.1	0.41	-0.68	8.13%	**17.89%**
tq7h	2.04	1.21	0.01	-1.10	9.76%	**28.46%**
tq8a	0.67	1.08	1.70	2.19	**62.60%**	2.44%
tq8b	0.67	1.04	1.49	1.32	**62.60%**	5.69%
tq8c	0.62	0.84	1.45	1.96	**56.10%**	3.25%
tq9a	0.41	0.78	2.18	4.87	**72.36%**	2.44%
tq9c	0.50	1.00	2.26	4.46	**71.54%**	3.25%
tq9e	0.63	0.87	1.58	2.66	**55.28%**	2.44%
tq11a	0.97	1.19	1.26	0.64	**44.72%**	7.32%
tq11d	1.03	1.22	1.17	0.43	**43.09%**	4.88%
tq11e	0.65	1.04	1.68	2.10	**62.60%**	4.88%
tq12	0.90	1.04	1.16	0.86	**43.90%**	4.88%

Note: **Bolding texts** indicate significant CFE (>15%)
Items starting with tq10 were removed for the ThyPRO-39 from ThyPRO, which were items related to impaired sex life.

Table I7. Descriptive Statistics at the Scale Level Based on Mode

Interview group (N = 81)						
Scale	Mean	SD	Skewness	Kurtosis	Floor (%)	Ceiling (%)
Goiter symptoms	2.79	2.17	0.88	0.83	**16.05%**	0.00%
Hyperthyroid symptom	3.42	2.58	0.85	0.38	13.58%	0.00%
Hypothyroid symptom	3.94	2.87	1.02	0.47	4.94%	0.00%
Eye symptoms	2.65	2.26	1.22	1.40	13.58%	0.00%
Tiredness	4.53	2.19	0.53	-0.42	0.00%	0.00%
Cognitive complaints	2.80	2.32	0.72	-0.40	**16.05%**	0.00%
Anxiety	2.57	2.14	0.86	0.44	**19.75%**	0.00%
Depressivity	3.80	2.48	0.73	0.16	7.41%	0.00%
Emotional susceptibility	4.79	2.44	0.30	-0.32	2.47%	1.23%
Impaired social life	1.36	1.89	1.68	2.53	**49.38%**	0.00%
Impaired daily life	1.20	1.52	1.59	2.42	**43.21%**	0.00%
Cosmetic complaints	2.38	2.51	1.87	4.00	**19.75%**	2.47%
Composite	20.60	10.54	0.62	-0.41	0.00%	0.00%
Electronic self-administered group (N = 98)						
Scale	Mean	SD	Skewness	Kurtosis	Floor (%)	Ceiling (%)
Goiter symptoms	2.92	2.92	1.04	0.53	**27.55%**	1.02%
Hyperthyroid symptom	4.63	3.40	0.54	-0.20	11.22%	0.00%
Hypothyroid symptom	5.03	3.54	0.53	-0.70	7.14%	0.00%
Eye symptoms	3.27	2.78	0.73	-0.45	**15.31%**	0.00%
Tiredness	5.95	2.88	0.19	-0.65	1.02%	4.08%
Cognitive complaints	4.14	2.95	0.80	0.13	9.18%	2.04%
Anxiety	4.51	3.17	0.52	-0.77	6.12%	2.04%
Depressivity	5.30	2.85	0.37	-0.56	2.04%	2.04%
Emotional susceptibility	6.10	2.40	0.05	-0.22	1.02%	2.04%
Impaired social life	2.91	2.96	1.02	0.32	**27.55%**	1.02%
Impaired daily life	2.70	3.06	1.15	0.37	**31.63%**	0.00%
Cosmetic complaints	3.41	3.08	0.84	-0.23	**16.33%**	1.02%
Composite	30.37	14.79	0.42	-0.37	1.02%	0.00%

Note: **Bolding texts** indicate significant CFE (>15%)

Table I8. Descriptive Statistics at the Item Level for Interview Mode (N = 81)

Item	Mean	SD	Skewness	Kurtosis	Floor (%)	Ceiling (%)
tq1a	1.01	0.89	0.82	0.60	**29.63%**	4.94%
tq1c	0.98	0.84	0.68	0.00	**29.63%**	6.17%
tq1h	0.80	1.01	1.12	0.41	**50.62%**	7.41%
tq1l	0.42	0.74	1.74	2.36	**70.37%**	2.47%
tq1m	1.16	1.22	0.84	-0.32	**38.27%**	9.88%
tq1n	1.12	1.11	0.89	-0.04	**33.33%**	11.11%
tq1t	0.72	0.95	1.44	1.90	**53.09%**	2.47%
tq1q	1.40	1.29	0.52	-0.94	**32.10%**	**16.05%**
tq1cc	0.53	0.76	1.5	2.01	**59.26%**	3.70%
tq1dd	1.12	1.10	0.98	0.35	**32.10%**	7.41%
tq1ee	0.89	1.08	0.97	0.01	**50.62%**	6.17%
tq1w	1.01	0.99	1.03	0.99	**34.57%**	2.47%
tq1x	0.84	1.04	1.23	0.81	**48.15%**	7.41%
tq1bb	0.80	0.98	1.34	1.48	**46.91%**	4.94%
tq2a	1.74	1.01	0.67	-0.25	4.94%	**13.58%**
tq2c	0.83	0.85	1.06	1.33	**39.51%**	2.47%
tq3b	2.04	1.28	0.04	-1.13	12.35%	**22.22%**
tq4a	1.16	1.07	0.66	-0.35	**32.10%**	9.88%
tq4b	0.84	0.77	0.44	-0.71	**37.04%**	1.23%
tq4f	0.80	0.95	1.16	0.80	**46.91%**	6.17%
tq5b	1.05	0.99	0.98	0.63	**30.86%**	7.41%
tq5c	0.83	0.82	1.00	1.43	**38.27%**	1.23%
tq5e	0.69	0.89	1.47	2.09	**50.62%**	4.94%
tq6a	1.00	1.02	1.10	0.90	**35.80%**	4.94%
tq6e	1.14	1.02	0.98	0.60	**27.16%**	7.41%
tq6g	2.33	1.32	-0.11	-1.41	6.17%	**24.69%**
tq7c	1.43	1.06	0.89	0.14	13.58%	11.11%
tq7d	1.53	1.12	0.61	-0.32	**16.05%**	11.11%
tq7h	2.17	1.26	-0.14	-1.18	9.88%	**30.86%**
tq8a	0.48	0.91	1.97	3.21	**71.60%**	4.94%
tq8b	0.36	0.75	2.17	4.08	**76.54%**	3.70%
tq8c	0.52	0.84	1.58	1.63	**65.43%**	4.94%
tq9a	0.21	0.47	2.10	3.69	**81.48%**	0.00%
tq9c	0.36	0.78	2.89	9.61	**75.31%**	0.00%
tq9e	0.63	0.86	1.24	0.69	**56.79%**	4.94%
tq11a	0.90	1.06	1.33	1.44	**43.21%**	2.47%
tq11d	1.00	1.12	1.11	0.54	**40.74%**	6.17%
tq11e	0.48	1.01	2.22	4.08	**75.31%**	3.70%
tq12	0.74	0.86	1.21	1.49	**46.91%**	2.47%

Note: **Bolding texts** indicate significant CFE (>15%)
Items starting with tq10 were removed for the ThyPRO-39 from ThyPRO, which were items related to impaired sex life.

Table I9. Descriptive Statistics at the Item Level for Self-Administration Mode (N = 98)

Item	Mean	SD	Skewness	Kurtosis	Floor (%)	Ceiling (%)
tq1a	0.98	1.03	0.89	0.12	**39.80%**	7.14%
tq1c	1.05	1.12	1.05	0.44	**37.76%**	6.12%
tq1h	0.89	1.15	1.19	0.43	**51.02%**	8.16%
tq1l	0.74	1.00	1.26	0.93	**55.10%**	4.08%
tq1m	1.28	1.25	0.76	-0.44	**33.67%**	9.18%
tq1n	1.52	1.25	0.48	-0.75	**24.49%**	12.24%
tq1t	1.09	1.15	0.99	0.13	**36.73%**	9.18%
tq1q	1.56	1.24	0.61	-0.70	**18.37%**	14.29%
tq1cc	0.98	1.24	0.99	-0.25	**52.04%**	10.20%
tq1dd	1.30	1.29	0.80	-0.51	**32.65%**	11.22%
tq1ee	1.19	1.12	0.80	-0.18	**30.61%**	11.22%
tq1w	1.16	1.05	0.73	-0.28	**29.59%**	12.24%
tq1x	0.98	1.06	0.90	-0.06	**41.84%**	9.18%
tq1bb	1.12	1.10	0.83	0.02	**34.69%**	7.14%
tq2a	1.94	1.27	0.32	-1.18	9.18%	**19.39%**
tq2c	1.45	1.23	0.70	-0.42	**23.47%**	8.16%
tq3b	1.44	1.23	0.72	-0.29	**24.49%**	4.08%
tq4a	1.36	1.01	0.87	0.50	**16.33%**	7.14%
tq4b	1.37	1.10	0.68	-0.21	**21.43%**	11.22%
tq4f	1.42	1.18	0.75	-0.21	**22.45%**	7.14%
tq5b	1.65	1.18	0.58	-0.52	**14.29%**	10.20%
tq5c	1.49	1.14	0.62	-0.40	**18.37%**	12.24%
tq5e	1.37	1.25	0.62	-0.65	**30.61%**	11.22%
tq6a	1.51	1.18	0.52	-0.63	**20.41%**	14.29%
tq6e	1.56	1.16	0.66	-0.39	**15.31%**	11.22%
tq6g	1.78	1.21	0.29	-0.9	14.29%	**18.37%**
tq7c	1.88	1.20	0.30	-0.96	10.20%	**19.39%**
tq7d	1.97	1.07	0.31	-0.89	4.08%	**23.47%**
tq7h	1.74	1.17	0.23	-0.95	14.29%	**22.45%**
tq8a	1.03	1.21	1.11	0.38	**44.90%**	2.04%
tq8b	1.03	1.26	0.96	-0.25	**48.98%**	9.18%
tq8c	0.85	1.00	1.29	1.38	**44.90%**	4.08%
tq9a	0.83	1.07	1.33	1.15	**51.02%**	4.08%
tq9c	0.87	1.26	1.26	0.31	**58.16%**	9.18%
tq9e	1.01	1.12	1.03	0.27	**40.82%**	8.16%
tq11a	1.21	1.29	0.81	-0.59	**37.76%**	**14.29%**
tq11d	1.21	1.26	0.91	-0.20	**35.71%**	7.14%
tq11e	0.98	1.19	0.98	-0.18	**48.98%**	10.20%
tq12	1.28	1.20	0.78	-0.29	**30.61%**	9.18%

Note: **Bolding texts** indicate significant CFE (>15%)
Items starting with tq10 were removed for the ThyPRO-39 from ThyPRO, which were items related to impaired sex life.

Table I10. Results of K-S tests for all scales

Scale	D value	*p*
Goiter symptoms	0.15	< 0.001**
Hyperthyroid symptoms	0.13	0.004**
Hypothyroid symptoms	0.16	< 0.001**
Eye symptoms	0.16	< 0.001**
Tiredness	0.13	0.003**
Cognitive complaints	0.19	< 0.001**
Anxiety	0.19	< 0.001**
Depressivity	0.14	0.002**
Emotional Susceptibility	0.09	0.08
Impaired social life	0.21	< 0.001**
Impaired daily life	0.24	< 0.001**
Cosmetic complaints	0.19	< 0.001**
Composite	0.09	0.09

*p < 0.05, **p < 0.01

Table I11. Comparison between SC ThyPRO-39 and TC ThyPRO-39 (Wong et al., 2018)

	SC ThyPRO-39 (this study)		TC ThyPRO-39 (Wong et al., 2018)	
Subtypes				
Thyroid nodules & non-toxic goiter	86	44.0%	289	93.8%
Thyroiditis	25	12.6%	1	0.3%
Thyroid cancer	30	15.4%	n/a	n/a
Hyperthyroidism	37	18.8%	13	4.2%
Hypothyroidism	12	6.3%	n/a	n/a
Thyroid cyst	6	2.9%	5	1.6%
# of floor effect				
Item level		35		39
Scale level		4		9
# of scales with low ICR*		5		6
# of items with low ICV**		5		8

Note: Thyroid nodules and non-toxic goiter were not distinguishable by the self-report from participants in this study, therefore the two diseases are combined in descriptive statistics.
n/a=Not applicable.
ICR=Internal Consistent Reliability
ICV=Internal Construct Validity

Table I12. Comparison of Floor Effects, Cronbach's Alpha and Item-total Correlation for SC & TC ThyPRO-39

	Simplified Chinese ThyPRO-39			Traditional Chinese ThyPRO-39		
Scale & item	Floor (%)	Cronbach's Alpha	Item-total correlation	Floor (%)	Cronbach's Alpha	Item-total correlation
Goiter symptoms	22.35%	0.815		27.9	0.803	
tq1a	35.20%		0.703	39.6		0.659
tq1c	34.08%		0.663	47.4		0.733
tq1h	50.84%		0.638	68.2		0.609
Hyperthyroid symptoms	12.29%	0.615		10.1	0.596	
tq1l	62.01%		0.485	65.9		0.446
tq1m	35.75%		0.437	41.9		0.352
tq1n	28.49%		0.493	33.4		0.533
tq1t	44.13%		0.337	44.5		0.226
Hypothyroid symptoms	6.15%	0.665		20.8	0.660	
tq1q	24.58%		0.372	59.1		0.336
tq1cc	55.31%		0.428	45.8		0.504
tq1dd	32.40%		0.565	62.7		0.560
tq1ee	39.66%		0.431	53.9		0.385
Eye symptoms	14.53%	0.758		27.6	0.680	
tq1w	31.84%		0.573	68.8		0.351
tq1x	44.69%		0.599	41.2		0.592
tq1bb	40.22%		0.590	46.8		0.554
Tiredness	0.56%	0.613		1.6	0.645	
tq2a	7.26%		0.410	15.9		0.515
tq2c	30.73%		0.528	40.9		0.630
tq3b	18.99%		0.341	32.5		0.256
Cognitive complaints	12.29%	0.843		25.7	0.883	
tq4a	23.46%		0.634	34.7		0.772
tq4b	28.49%		0.784	45.1		0.802
tq4f	33.52%		0.707	47.4		0.752
Anxiety	12.29%	0.837		27.9	0.897	
tq5b	21.79%		0.713	36.7		0.777
tq5c	27.37%		0.768	41.2		0.805
tq5e	39.66%		0.626	49.0		0.809
Depressivity	4.47%	0.687		4.9	0.649	
tq6a	27.37%		0.647	52.6		0.613
tq6e	20.67%		0.650	38.0		0.650
tq6g	10.61%		0.271	18.5		0.193
Emotional susceptibility	1.68%	0.518		8.8	0.634	
tq7c	11.73%		0.445	26.3		0.534
tq7d	9.50%		0.491	32.8		0.613
tq7h	12.29%		0.112	13.6		0.238
Impaired social life	37.43%	0.779		58.4	0.741	
tq8a	56.98%		0.665	74.4		0.637
tq8b	61.45%		0.654	71.8		0.507
tq8c	54.19%		0.541	75.6		0.566
Impaired daily life	36.87%	0.815		65.3	0.840	
tq9a	64.80%		0.752	82.5		0.696
tq9c	65.92%		0.618	85.4		0.775
tq9e	48.04%		0.647	68.5		0.677
Cosmetic Complaints	17.88%	0.745		42.9	0.830	
tq11a	40.22%		0.616	46.4		0.701
tq11d	37.99%		0.466	69.5		0.695
tq11e	60.89%		0.641	80.2		0.696
Composite	0.56%	0.928		0.7	0.929	

Appendix J. Tables for RQ 2

Content of the tables in Appendix J

Table J1. Assumption checking for Two-sample Proportion Tests in RQ 2.1

Scale	# of cases at the floor in males	# of cases at the floor in females
Goiter symptoms	5	35
Hyperthyroid symptom	5	17
Hypothyroid symptom	3	8
Eye symptoms	7	19
Tiredness	0	1
Cognitive complaints	8	14
Anxiety	7	15
Depressivity	4	4
Emotional susceptibility	2	1
Impaired social life	18	49
Impaired daily life	12	54
Cosmetic complaints	6	26
Composite	0	1

Table J2. Proportional Tests on Percentage of Floor between Gender Groups

	Sample estimates				
Scale	Percentage of floor among males	Percentage of floor among females	*df*	χ^2	*p*
Goiter symptoms	8.93%	28.46%	1	7.37	0.007
Hyperthyroid symptom	8.93%	13.82%	1	0.46	0.50
Hypothyroid symptom	5.36%	6.50%	1	< 0.001	>0.999
Eye symptoms	12.50%	15.45%	1	0.08	0.77
Cognitive complaints	14.29%	11.38%	1	0.09	0.76
Anxiety	12.50%	12.20%	1	< 0.001	>0.999
Impaired social life	32.14%	39.83%	1	0.67	0.41
Impaired daily life	21.42%	43.90%	1	7.41	0.006
Cosmetic complaints	10.71%	21.14%	1	2.18	0.14

Note: α = 0.05/9 = 0.0045 after Bonferroni correction.

Table J3. Assumption checking for Two-sample Proportion Tests in RQ 2.2

Scale	# of cases at the floor in the electronic interview group	# of cases at the floor for the electronic self-administered group
Goiter symptoms	14	31
Hyperthyroid symptom	12	13
Hypothyroid symptom	4	10
Eye symptoms	12	18
Tiredness	0	1
Cognitive complaints	13	11
Anxiety	16	10
Depressivity	7	2
Emotional susceptibility	2	1
Impaired social life	42	33
Impaired daily life	37	37
Cosmetic complaints	17	20
Composite	0	1

Table J4. Two-sample Proportion Tests on Percentage of Floor between Administration Groups

	Sample estimates				
Scale	Percentage of floor in interview	Percentage of floor in self-administration	*df*	χ^2	*p*
Goiter symptoms	16.05%	27.55%	1	2.75	0.10
Hyperthyroid symptom	13.58%	11.22%	1	0.06	0.80
Hypothyroid symptom	4.94%	7.14%	1	0.09	0.77
Eye symptoms	13.58%	15.31%	1	0.01	0.91
Cognitive complaints	16.05%	9.18%	1	1.35	0.24
Anxiety	19.75%	6.12%	1	6.43	0.01
Depressivity	7.41%	2.04%	1	1.87	0.17
Impaired social life	49.38%	27.55%	1	8.12	0.004**
Impaired daily life	43.21%	31.63%	1	2.08	0.15
Cosmetic complaints	19.75%	16.33%	1	0.16	0.69

Note: α = 0.05/9 = 0.0045 after Bonferroni correction.
** p < 0.01

Appendix K. Tables for Assumption Checking/Diagnostics for RQ 3 & 4

Content of the tables in Appendix K

Table K1. Summary of Results of Distributional Diagnostics for Response Variables of Scales with Prominent Floor Effects

Distributions	Tests	Desired Outcome	Scales			
			G	S	D	C
Normal Distribution	One-sample K-S tests	Obtain the H_0: The data is normally distributed	×	×	×	×
Poisson Distribution	Goodness-of-fit tests for Poisson	Obtain the H_0: The data can be fit to a Poisson distribution	×	×	×	×
NB Distribution	Goodness-of-fit tests for NB	Obtain the H_0: The data can be fit to a NB distribution	√	×	√	√

Notes: NB = Negative Binomial; G = Goiter Symptoms Scale; S = Impaired Social Life Scale; D = Impaired Daily Life Scale; C = Cosmetic Complaints Scale.
√ = Satisfied, × = Not satisfied.

Table K2. Results of K-S Tests for Scales with Prominent Floor Effects

Scale	D value	*p*
Goiter symptoms	0.14	< 0.001**
Impaired social life	0.20	< 0.001**
Impaired daily life	0.24	< 0.001**
Cosmetic complaints	0.19	< 0.001**

*p < 0.05, **p < 0.01

Table K3. Results of Goodness-of-fit Tests of Poisson for Scales with Prominent Floor Effects

Scale	*df*	χ^2	*p*
Goiter symptoms	11	104.48	< 0.001**
Impaired social life	10	198.96	< 0.001**
Impaired daily life	10	182.23	< 0.001**
Cosmetic complaints	11	140.50	< 0.001**

*p < 0.05, **p < 0.01

Table K4. Results of goodness-of-fit tests of NB for Scales with Prominent Floor Effects

Scale	*df*	χ^2	*p*
Goiter symptoms	10	14.06	0.17
Impaired social life	9	26.57	0.002**
Impaired daily life	9	10.26	0.33
Cosmetic complaints	10	13.45	0.20

*p < 0.05, **p < 0.01

Table K5. Summary of Assumption Checking/Diagnostic Results under One Predictor Models of Gender for Scales with Prominent Floor Effects

Assumptions /Diagnostics	Checked by	Desired Outcome	Models																						
			ML				Tobit				Poisson				NB				ZIP				ZINB			
			G	S	D	C	G	S	D	C	G	S	D	C	G	S	D	C	G	S	D	C	G	S	D	C
Linearity	Scatterplot between fitted value and standardized residuals	The residuals follow the loess line without particular variation	×	×	×	×	√	√	√	√	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Specification of Predictors			×	×	×	×	√	√	√	√	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Independence of Errors			×	×	×	×	√	√	√	√	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Homogeneity of Variance	Scatterplot between fitted value and squared root of standardized residuals	The squared roots of standardized residuals follow the loess line without particular variation	×	×	√	√	@	@	@	@	√	√	√	√	@	@	@	@	@	@	@	@	@	@	@	@
Normality of Errors	Normal Q-Q plots	The points roughly follow a positive, straight line	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	@	@	@	@	@	@	@	@
Mean = Variance	Dispersion tests	Obtain the H_0: Dispersion parameter = 1	@	@	@	@	@	@	@	@	×	×	×	×	@	@	@	@	×	×	×	×	@	@	@	@
Overdispersion (Variance > Mean)	Dispersion tests	Reject the H_0: Dispersion parameter = 1, where H_1: Dispersion parameter > 1	@	@	@	@	@	@	@	@	@	@	@	@	√	√	√	√	@	@	@	@	√	√	√	√
Zero-inflation	Zero-inflation tests	Greater observed zeros than predicted zeros	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	√	√	√	√	√	√	√	×

Note: ML = Multiple Linear Regression Models; NB = Negative Binomial Regression Models; ZIP = Zero-inflated Poisson Models; ZINB = Zero-inflated Negative Binomial Models; G = Goiter symptoms scale; S = Impaired social life scale; D = Impaired daily life scale; C = Cosmetic Complaints scale. √ = Assumption was accepted; × = Assumption was violated; @ = Assumption not applicable.

Table K6. Results of Dispersion Tests under One Predictor Models of Gender for Scales with Prominent Floor Effects

Scale	Dispersion	z	*p*
Goiter symptoms	2.01	5.21	< 0.001**
Impaired social life	3.07	5.88	< 0.001**
Impaired daily life	3.02	4.94	< 0.001**
Cosmetic complaints	2.76	5.44	< 0.001**

*p < 0.05, **p < 0.01

Table K7. Results of Zero-inflation under One Predictor Models of Gender for Scales with Prominent Floor Effects

Scale	Observed zeros	Predicted zeros	Ratio	Results
ZIP models				
Goiter symptoms	40	14	0.35	Underfitted zeros
Impaired social life	67	21	0.31	Underfitted zeros
Impaired daily life	66	29	0.44	Underfitted zeros
Cosmetic complaints	32	10	0.31	Underfitted zeros
ZINB models				
Goiter symptoms	40	21	0.53	Underfitted zeros
Impaired social life	67	40	0.60	Underfitted zeros
Impaired daily life	66	59	0.89	Underfitted zeros
Cosmetic complaints	32	34	1.03	Overfitted zeros

Table K8. Summary of Assumption Checking/Diagnostic Results under One Predictor Models of Mode of Administration for Scales with Prominent Floor Effects

Assumptions /Diagnostics	Checked by	Desired Outcome	Models																					
			ML				Tobit				Poisson				NB				ZIP				ZINB			
			G	S	D	C	G	S	D	C	G	S	D	C	G	S	D	C	G	S	D	C	G	S	D	C
Linearity	Scatterplot between fitted value and standardized residuals	The residuals follow the loess line without particular variation	×	×	×	×	√	√	√	√	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Specification of Predictors			×	×	×	×	√	√	√	√	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Independence of Errors			×	×	×	×	√	√	√	√	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Homogeneity of Variance	Scatterplot between fitted value and squared root of standardized residuals	The squared roots of standardized residuals follow the loess line without particular variation	×	×	×	×	@	@	@	@	√	√	√	√	@	@	@	@	@	@	@	@	@	@	@	@
Normality of Errors	Normal Q-Q plots	The points roughly follow a positive, straight line	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	@	@	@	@	@	@	@	@
Mean = Variance	Dispersion tests	Obtain the H_0: Dispersion parameter = 1	@	@	@	@	@	@	@	@	×	×	×	×	@	@	@	@	×	×	×	×	@	@	@	@
Overdispersion (Variance > Mean)	Dispersion tests	Reject the H_0: Dispersion parameter = 1, where H_1: Dispersion parameter > 1	@	@	@	@	@	@	@	@	@	@	@	@	√	√	√	√	@	@	@	@	√	√	√	√
Zero-inflation	Zero-inflation tests	Greater observed zeros than predicted zeros	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	√	√	√	√	√	√	√	×

Note: ML = Multiple Linear Regression Models; NB = Negative Binomial Regression Models; ZIP = Zero-inflated Poisson Models; ZINB = Zero-inflated Negative Binomial Models; G = Goiter symptoms scale; S = Impaired social life scale; D = Impaired daily life scale; C = Cosmetic Complaints scale. √ = Assumption was accepted; × = Assumption was violated; @ = Assumption not applicable.

Table K9. Results of Dispersion Tests under One Predictor Models of Mode of Administration for Scales with Prominent Floor Effects

Scale	Dispersion	z	*p*
Goiter symptoms	2.34	5.07	< 0.001**
Impaired social life	2.80	5.64	< 0.001**
Impaired daily life	2.74	5.68	< 0.001**
Cosmetic complaints	2.70	4.97	< 0.001**

*p < 0.05, **p < 0.01

Table K10. Results of Zero-inflation under One Predictor Models of Mode of Administration for Scales with Prominent Floor Effects

Scale	Observed zeros	Predicted zeros	Ratio	Results
ZIP models				
Goiter symptoms	40	10	0.25	Underfitted zeros
Impaired social life	67	26	0.39	Underfitted zeros
Impaired daily life	66	31	0.47	Underfitted zeros
Cosmetic complaints	32	11	0.34	Underfitted zeros
ZINB models				
Goiter symptoms	40	20	0.50	Underfitted zeros
Impaired social life	67	42	0.63	Underfitted zeros
Impaired daily life	66	56	0.85	Underfitted zeros
Cosmetic complaints	32	33	1.03	Within tolerance range

Table K11. Summary of Assumption Checking/Diagnostic Results under Full Models of Six Predictors for Scales with Prominent Floor Effects

Assumptions /Diagnostics	Checked by	Desired Outcome	Models																					
			ML				Tobit				Poisson				NB				ZIP				ZINB			
			G	S	D	C	G	S	D	C	G	S	D	C	G	S	D	C	G	S	D	C	G	S	D	C
Linearity	Scatterplot between fitted value and standardized residuals	The residuals follow the loess line without particular variation	√	√	√	√	√	√	√	√	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Specification of Predictors			√	√	√	√	√	√	√	√	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Independence of Errors			√	√	√	√	√	√	√	√	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Homogeneity of Variance	Scatterplot between fitted value and squared root of standardized residuals	The squared roots of standardized residuals follow the loess line without particular variation	√	√	√	√	@	@	@	@	√	√	√	√	@	@	@	@	@	@	@	@	@	@	@	@
Normality of Errors	Normal Q-Q plots	The points roughly follow a positive, straight line	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	@	@	@	@	@	@	@	@
Collinearity	VIF & Tolerance	VIF < 10 & Tolerance < 1	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
Mean = Variance	Dispersion tests	Obtain the H_0: Dispersion parameter = 1	@	@	@	@	@	@	@	@	×	×	×	×	@	@	@	@	×	×	×	×	@	@	@	@
Overdispersion (Variance > Mean)	Dispersion tests	Reject the H_0: Dispersion parameter = 1, where H_1: Dispersion parameter > 1	@	@	@	@	@	@	@	@	@	@	@	@	√	√	√	√	@	@	@	@	√	√	√	√
Zero-inflation	Zero-inflation tests	Greater observed zeros than predicted zeros	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	√	√	√	√	√	√	×	×

Note: ML = Multiple Linear Regression Models; NB = Negative Binomial Regression Models; ZIP = Zero-inflated Poisson Models; ZINB = Zero-inflated Negative Binomial Models; G = Goiter symptoms scale; S = Impaired social life scale; D = Impaired daily life scale; C = Cosmetic Complaints scale. √ = Assumption was accepted; × = Assumption was violated; @ = Assumption not applicable.

Table K12. Results of Diagnostics for Lack of Multicollinearity of ML Models under Full Predictor Models of Six Predictors for Scales with Prominent Floor Effects

Scale	Predictor	VIF	Increased SE	Tolerance
Goiter symptoms	Mode	1.04	1.02	0.96
	Gender	1.18	1.09	0.85
	Age	1.13	1.06	0.89
	Education level	1.08	1.04	0.93
	Duration of treatment	1.05	1.03	0.95
	Levels of control	1.04	1.02	0.96
Impaired social life	Mode	1.04	1.02	0.96
	Gender	1.18	1.09	0.85
	Age	1.13	1.06	0.89
	Education level	1.08	1.04	0.93
	Duration of treatment	1.05	1.03	0.95
	Levels of control	1.04	1.02	0.96
Impaired daily life	Mode	1.04	1.02	0.96
	Gender	1.18	1.09	0.85
	Age	1.13	1.06	0.89
	Education level	1.08	1.04	0.93
	Duration of treatment	1.05	1.03	0.95
	Levels of control	1.04	1.02	0.96
Cosmetic complaints	Mode	1.04	1.02	0.96
	Gender	1.18	1.09	0.85
	Age	1.13	1.06	0.89
	Education level	1.08	1.04	0.93
	Duration of treatment	1.05	1.03	0.95
	Levels of control	1.04	1.02	0.96

Table K13. Results of Diagnostics for Lack of Multicollinearity of Poisson Models under Full Predictor Models of Six Predictors for Scales with Prominent Floor Effects

Scale	Predictor	VIF	Increased SE	Tolerance
Goiter symptoms	Mode	1.10	1.05	0.91
	Gender	1.30	1.14	0.77
	Age	1.21	1.10	0.83
	Education level	1.15	1.07	0.87
	Duration of treatment	1.10	1.05	0.91
	Levels of control	1.05	1.02	0.95
Impaired social life	Mode	1.08	1.04	0.92
	Gender	1.16	1.07	0.88
	Age	1.16	1.08	0.86
	Education level	1.14	1.07	0.91
	Duration of treatment	1.10	1.05	0.91
	Levels of control	1.06	1.03	0.94
Impaired daily life	Mode	1.09	1.04	0.92
	Gender	1.20	1.10	0.83
	Age	1.16	1.07	0.87
	Education level	1.13	1.06	0.89
	Duration of treatment	1.09	1.05	0.91
	Levels of control	1.05	1.02	0.95
Cosmetic complaints	Mode	1.10	1.05	0.91
	Gender	1.27	1.13	0.87
	Age	1.18	1.09	0.85
	Education level	1.17	1.05	0.90
	Duration of treatment	1.11	1.05	0.90
	Levels of control	1.05	1.02	0.95

Table K14. Results of Diagnostics for Lack of Multicollinearity of NB Models under Full Predictor Models of Six Predictors for Scales with Prominent Floor Effects

Scale	Predictor	VIF	Increased SE	Tolerance
Goiter symptoms	Mode	1.08	1.04	0.93
	Gender	1.24	1.11	0.80
	Age	1.16	1.08	0.86
	Education level	1.10	1.05	0.91
	Duration of treatment	1.08	1.04	0.93
	Levels of control	1.04	1.02	0.96
Impaired social life	Mode	1.07	1.03	0.93
	Gender	1.20	1.10	0.83
	Age	1.14	1.07	0.88
	Education level	1.10	1.05	0.91
	Duration of treatment	1.07	1.03	0.93
	Levels of control	1.05	1.02	0.96
Impaired daily life	Mode	1.09	1.05	0.92
	Gender	1.20	1.10	0.83
	Age	1.14	1.07	0.88
	Education level	1.09	1.05	0.91
	Duration of treatment	1.07	1.03	0.94
	Levels of control	1.04	1.02	0.96
Cosmetic complaints	Mode	1.08	1.04	0.93
	Gender	1.23	1.11	0.81
	Age	1.15	1.07	0.87
	Education level	1.11	1.05	0.90
	Duration of treatment	1.08	1.04	0.93
	Levels of control	1.05	1.02	0.96

Table K15. Results of Dispersion Tests under Full Predictor Models of Six Predictors for Scales with Prominent Floor Effects

Scale	dispersion	z	*p*
Goiter symptoms	1.74	4.12	< 0.001**
Impaired social life	2.57	3.75	< 0.001**
Impaired daily life	2.17	4.58	< 0.001**
Cosmetic complaints	2.14	3.18	< 0.001**

*p < 0.05, **p < 0.01

Table K16. Results of Zero-inflation under Full Predictor Models of Six Predictors for Scales with Prominent Floor Effects

Scale	Observed zeros	Predicted zeros	Ratio	Results
ZIP models				
Goiter symptoms	40	22	0.55	Underfitted zeros
Impaired social life	67	35	0.52	Underfitted zeros
Impaired daily life	66	40	0.61	Underfitted zeros
Cosmetic complaints	32	20	0.62	Underfitted zeros
ZINB models				
Goiter symptoms	40	25	0.62	Underfitted zeros
Impaired social life	67	45	0.67	Underfitted zeros
Impaired daily life	66	55	0.83	Underfitted zeros
Cosmetic complaints	32	35	1.09	Overfitted zeros

Table K17. Summary of Results of Distributional Diagnostics for Response Variables of Scales without Prominent Floor Effects

Distributions	Tests	Desired Outcome	Scales							
			HE	HO	E	T	CO	A	D	S
Normal Distribution	One-sample K-S tests	Obtain the H_0: The data is normally distributed	×	×	×	×	×	×	×	√
Poisson Distribution	Goodness-of-fit tests for Poisson	Obtain the H_0: The data can be fit to a Poisson distribution	×	×	×	×	×	×	×	√
NB Distribution	Goodness-of-fit tests for NB	Obtain the H_0: The data can be fit to a NB distribution	√	×	√	√	×	×	√	√

Notes: HE = Hyperthyroid Symptoms Scale; HO = Hyperthyroid Symptoms Scale; E = Eye symptoms scale; T = Tiredness Scale; CO = Cognitive Complaints Scale; A = Anxiety Scale; D = Depressivity Scale; S = Emotional Susceptibility Scale.
√ = Satisfied, × = Not satisfied.

Table K18. Results of K-S Tests for Scales without Prominent Floor Effects

Scale	D value	*p*
Hyperthyroid symptoms	0.13	0.004**
Hypothyroid symptoms	0.16	< 0.001**
Eye symptoms	0.16	< 0.001**
Tiredness	0.13	0.003**
Cognitive complaints	0.19	< 0.001**
Anxiety	0.19	< 0.001**
Depressivity	0.14	0.002**
Emotional Susceptibility	0.09	0.08
Composite	0.09	0.09

*p < 0.05, **p < 0.01

Table K19. Results of goodness-of-fit tests of Poisson for Scales without Prominent Floor Effects

Scale	*df*	χ^2	*p*
Hyperthyroid symptoms	12	110.66	< 0.001**
Hypothyroid symptoms	12	116.94	< 0.001**
Eye symptoms	9	88.33	< 0.001**
Tiredness	11	23.84	0.01*
Cognitive complaints	10	96.63	< 0.001**
Anxiety	11	109.31	< 0.001**
Depressivity	11	46.05	< 0.001**
Emotional Susceptibility	11	16.18	0.13
Composite	51	866.29	< 0.001**

*p < 0.05, **p < 0.01

Table K20. Results of goodness-of-fit tests of NB for Scales without Prominent Floor Effects

Scale	*df*	χ^2	*p*
Hyperthyroid symptoms	11	16.35	0.13
Hypothyroid symptoms	11	25.91	0.007**
Eye symptoms	8	12.23	0.05
Tiredness	10	14.99	0.13
Cognitive complaints	9	28.69	< 0.001**
Anxiety	10	22.31	0.01*
Depressivity	10	17.02	0.07
Emotional Susceptibility	10	14.72	0.14
Composite	50	82.69	0.002**

*p < 0.05, **p < 0.01

Table K21. Summary of Assumption Checking/Diagnostic Results under One Predictor Models of Gender for Scales without Prominent Floor Effects

Assumptions /Diagnostics	Checked by	Desired Outcome	Models: ML								Models: Tobit								Models: Poisson							
			HE	HO	E	T	CO	A	D	S	HE	HO	E	T	CO	A	D	S	HE	HO	E	T	CO	A	D	S
Linearity	Scatterplot between fitted value and standardized residuals	The residuals follow the loess line without particular variation	×	×	×	√	×	×	×	×	√	√	√	×	√	√	√	√	@	@	@	@	@	@	@	@
Specification of Predictors			×	×	×	√	×	×	×	×	√	√	√	×	√	√	√	√	@	@	@	@	@	@	@	@
Independence of Errors			×	×	×	√	×	×	×	×	√	√	√	×	√	√	√	√	@	@	@	@	@	@	@	@
Homogeneity of Variance	Scatterplot between fitted value and squared root of standardized residuals	The squared roots of standardized residuals follow the loess line without particular variation	×	×	×	√	√	×	×	×	@	@	@	@	@	@	@	@	√	√	√	√	√	√	√	√
Normality of Errors	Normal Q-Q plots	The points roughly follow a positive, straight line	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
Mean = Variance	Dispersion tests	Obtain the H_0: Dispersion parameter = 1	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	×	×	×	×	×	×	×	√
Overdispersion (Variance > Mean)	Dispersion tests	Reject the H_0: Dispersion parameter = 1, where H_1: Dispersion parameter > 1	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Zero-inflation	Zero-inflation tests	Greater observed zeros than predicted zeros	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@

Assumptions /Diagnostics	Checked by	Desired Outcome	Models: NB								Models: ZIP								Models: ZINB							
			HE	HO	E	T	CO	A	D	S	HE	HO	E	T	CO	A	D	S	HE	HO	E	T	CO	A	D	S
Linearity	Scatterplot between fitted value and standardized residuals	The residuals follow the loess line without particular variation	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Specification of Predictors			@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Independence of Errors			@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Homogeneity of Variance	Scatterplot between fitted value and squared root of standardized residuals	The squared roots of standardized residuals follow the loess line without particular variation	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Normality of Errors	Normal Q-Q plots	The points roughly follow a positive, straight line	√	√	√	√	√	√	√	√	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Mean = Variance	Dispersion tests	Obtain the H_0: Dispersion parameter = 1	@	@	@	@	@	@	@	@	×	×	×	×	×	×	×	√	@	@	@	@	@	@	@	@
Overdispersion (Variance > Mean)	Dispersion tests	Reject the H_0: Dispersion parameter = 1, where H_1: Dispersion parameter > 1	√	√	√	√	√	√	√	×	@	@	@	@	@	@	@	@	√	√	√	√	√	√	√	×
Zero-inflation	Zero-inflation tests	Greater observed zeros than predicted zeros	@	@	@	@	@	@	@	@	√	√	√	×	√	√	√	√	√	√	√	×	√	√	√	√

Note: ML = Multiple Linear Regression Models; NB = Negative Binomial Regression Models; ZIP = Zero-inflated Poisson Models; ZINB = Zero-inflated Negative Binomial Models.
HE = Hyperthyroid Symptoms Scale; HO = Hyperthyroid Symptoms Scale; E = Eye Symptoms scale; T = Tiredness Scale; CO = Cognitive Complaints Scale; A = Anxiety Scale; D = Depressivity Scale; S = Emotional Susceptibility Scale.
√ = Assumption was accepted; × = Assumption was violated; @ = Assumption not applicable

Table K22. Results of Dispersion Tests under One Predictor Models of Gender for Scales without Prominent Floor Effects

Scale	dispersion	z	*p*
Hyperthyroid symptoms	2.31	6.00	< 0.001**
Hypothyroid symptoms	2.33	6.02	< 0.001**
Eye symptoms	2.18	5.32	< 0.001**
Tiredness	1.34	2.87	0.002**
Cognitive complaints	2.14	5.06	< 0.001**
Anxiety	2.32	6.16	< 0.001**
Depressivity	1.65	4.29	< 0.001**
Emotional Susceptibility	1.13	1.14	0.13
Composite	7.28	8.49	< 0.001**

*p < 0.05, **p < 0.01

Table K23. Results of Zero-inflation under One Predictor Models of Gender for Scales without Prominent Floor Effects

Scale	Observed zeros	Predicted zeros	Ratio	Results
ZIP models				
Hyperthyroid symptoms	22	3	0.14	Underfitted zeros
Hypothyroid symptoms	11	2	0.18	Underfitted zeros
Eye symptoms	26	9	0.35	Underfitted zeros
Tiredness	1	1	1.00	Within tolerance range
Cognitive complaints	22	5	0.23	Underfitted zeros
Anxiety	22	5	0.23	Underfitted zeros
Depressivity	8	2	0.25	Underfitted zeros
Emotional Susceptibility	3	1	0.33	Underfitted zeros
Composite	1	0	0.00	Underfitted zeros
ZINB models				
Hyperthyroid symptoms	22	11	0.50	Underfitted zeros
Hypothyroid symptoms	11	10	0.91	Underfitted zeros
Eye symptoms	26	24	0.92	Underfitted zeros
Tiredness	1	2	2.00	Overfitted zeros
Cognitive complaints	22	14	0.64	Underfitted zeros
Anxiety	22	17	0.77	Underfitted zeros
Depressivity	8	5	0.62	Underfitted zeros
Emotional Susceptibility	3	1	0.33	Underfitted zeros
Composite	1	0	0.00	Underfitted zeros

Table K24. Summary of Assumption Checking/Diagnostic Results under One Predictor Models of Mode of Administration for Scales without Prominent Floor Effects

Assumptions /Diagnostics	Checked by	Desired Outcome	Models: ML								Models: Tobit								Models: Poisson							
			HE	HO	E	T	CO	A	D	S	HE	HO	E	T	CO	A	D	S	HE	HO	E	T	CO	A	D	S
Linearity	Scatterplot between fitted value and standardized residuals	The residuals follow the loess line without particular variation	×	×	×	×	×	×	√	×	√	√	√	√	√	√	√	√	@	@	@	@	@	@	@	@
Specification of Predictors			×	×	×	×	×	×	√	×	√	√	√	√	√	√	√	√	@	@	@	@	@	@	@	@
Independence of Errors			×	×	×	×	×	×	√	×	√	√	√	√	√	√	√	√	@	@	@	@	@	@	@	@
Homogeneity of Variance	Scatterplot between fitted value and squared root of standardized residuals	The squared roots of standardized residuals follow the loess line without particular variation	×	×	×	×	√	×	×	×	@	@	@	@	@	@	@	@	√	√	√	√	√	√	√	√
Normality of Errors	Normal Q-Q plots	The points roughly follow a positive, straight line	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
Mean = Variance	Dispersion tests	Obtain the H_0: Dispersion parameter = 1	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	×	×	×	×	×	×	×	√
Overdispersion (Variance > Mean)	Dispersion tests	Reject the H_0: Dispersion parameter = 1, where H_1: Dispersion parameter > 1	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Zero-inflation	Zero-inflation tests	Greater observed zeros than predicted zeros	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@

Assumptions /Diagnostics	Checked by	Desired Outcome	Models: NB								Models: ZIP								Models: ZINB							
			HE	HO	E	T	CO	A	D	S	HE	HO	E	T	CO	A	D	S	HE	HO	E	T	CO	A	D	S
Linearity	Scatterplot between fitted value and standardized residuals	The residuals follow the loess line without particular variation	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Specification of Predictors			@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Independence of Errors			@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Homogeneity of Variance	Scatterplot between fitted value and squared root of standardized residuals	The squared roots of standardized residuals follow the loess line without particular variation	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Normality of Errors	Normal Q-Q plots	The points roughly follow a positive, straight line	√	√	√	√	√	√	√	√	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Mean = Variance	Dispersion tests	Obtain the H_0: Dispersion parameter = 1	@	@	@	@	@	@	@	@	×	×	×	×	×	×	×	√	@	@	@	@	@	@	@	@
Overdispersion (Variance > Mean)	Dispersion tests	Reject the H_0: Dispersion parameter = 1, where H_1: Dispersion parameter > 1	√	√	√	√	√	√	√	×	@	@	@	@	@	@	@	@	√	√	√	√	√	√	√	×
Zero-inflation	Zero-inflation tests	Greater observed zeros than predicted zeros	@	@	@	@	@	@	@	@	√	√	√	×	√	√	√	√	√	√	√	×	√	√	√	√

Note: ML = Multiple Linear Regression Models; NB = Negative Binomial Regression Models; ZIP = Zero-inflated Poisson Models; ZINB = Zero-inflated Negative Binomial Models.
HE = Hyperthyroid Symptoms Scale; HO = Hyperthyroid Symptoms Scale; E = Eye Symptoms scale; T = Tiredness Scale; CO = Cognitive Complaints Scale; A = Anxiety Scale; D = Depressivity Scale; S = Emotional Susceptibility Scale.
√ = Assumption was accepted; × = Assumption was violated; @ = Assumption not applicable

Table K25. Results of Dispersion Tests under One Predictor Models of Mode of Administration for Scales without Prominent Floor Effects

Scale	dispersion	z	*p*
Hyperthyroid symptoms	2.22	5.52	< 0.001**
Hypothyroid symptoms	2.28	6.25	< 0.001**
Eye symptoms	2.15	5.42	< 0.001**
Tiredness	1.23	2.08	0.02*
Cognitive complaints	1.99	5.35	< 0.001**
Anxiety	2.00	5.70	< 0.001**
Depressivity	1.55	3.79	< 0.001**
Emotional Susceptibility	1.07	0.61	0.27
Composite	6.31	8.72	< 0.001**

*p < 0.05, **p < 0.01

Table K26. Results of Zero-inflation under One Predictor Models of Mode of Administration for Scales without Prominent Floor Effects

Scale	Observed zeros	Predicted zeros	Ratio	Results
ZIP models				
Hyperthyroid symptoms	22	4	0.18	Underfitted zeros
Hypothyroid symptoms	11	2	0.18	Underfitted zeros
Eye symptoms	26	9	0.35	Underfitted zeros
Tiredness	1	1	1.00	Within tolerance range
Cognitive complaints	22	6	0.27	Underfitted zeros
Anxiety	22	7	0.32	Underfitted zeros
Depressivity	8	2	0.25	Underfitted zeros
Emotional Susceptibility	3	1	0.33	Underfitted zeros
Composite	1	0	0.00	Underfitted zeros
ZINB models				
Hyperthyroid symptoms	22	10	0.45	Underfitted zeros
Hypothyroid symptoms	11	10	0.91	Underfitted zeros
Eye symptoms	26	24	0.92	Underfitted zeros
Tiredness	1	2	2.00	Overfitted zeros
Cognitive complaints	22	14	0.64	Underfitted zeros
Anxiety	22	16	0.73	Underfitted zeros
Depressivity	8	5	0.62	Underfitted zeros
Emotional Susceptibility	3	1	0.33	Underfitted zeros
Composite	1	0	0.00	Underfitted zeros

Table K27. Summary of Assumption Checking/Diagnostic Results under Full Models of Six Predictors for Scales without Prominent Floor Effects

Assumptions /Diagnostics	Checked by	Desired Outcome	Models																							
			ML								Tobit								Poisson							
			HE	HO	E	T	CO	A	D	S	HE	HO	E	T	CO	A	D	S	HE	HO	E	T	CO	A	D	S
Linearity			√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	@	@	@	@	@	@	@	@
Specification of Predictors	Scatterplot between fitted value and standardized residuals	The residuals follow the loess line without particular variation	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	@	@	@	@	@	@	@	@
Independence of Errors			√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	@	@	@	@	@	@	@	@
Homogeneity of Variance	Scatterplot between fitted value and squared root of standardized residuals	The squared roots of standardized residuals follow the loess line without particular variation	√	√	√	√	√	√	√	√	@	@	@	@	@	@	@	@	√	√	√	√	√	√	√	√
Normality of Errors	Normal Q-Q plots	The points roughly follow a positive, straight line	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
Collinearity	VIF & Tolerance	VIF < 10 & Tolerance < 1	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
Mean = Variance	Dispersion tests	Obtain the H_0: Dispersion parameter = 1	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	×	×	×	√	×	×	×	√
Overdispersion (Variance > Mean)	Dispersion tests	Reject the H_0: Dispersion parameter = 1, where H_1: Dispersion parameter > 1	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Zero-inflation	Zero-inflation tests	Greater observed zeros than predicted zeros	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@

Assumptions /Diagnostics	Checked by	Desired Outcome	Models																							
			NB								ZIP								ZINB							
			HE	HO	E	T	CO	A	D	S	HE	HO	E	T	CO	A	D	S	HE	HO	E	T	CO	A	D	S
Linearity			@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Specification of Predictors	Scatterplot between fitted value and standardized residuals	The residuals follow the loess line without particular variation	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Independence of Errors			@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Homogeneity of Variance	Scatterplot between fitted value and squared root of standardized residuals	The squared roots of standardized residuals follow the loess line without particular variation	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Normality of Errors	Normal Q-Q plots	The points roughly follow a positive, straight line	√	√	√	√	√	√	√	√	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@	@
Collinearity	VIF & Tolerance	VIF < 10 & Tolerance < 1	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
Mean = Variance	Dispersion tests	Obtain the H_0: Dispersion parameter = 1	@	@	@	@	@	@	@	@	×	×	√	×	×	×	×	√	@	@	@	@	@	@	@	@
Overdispersion (Variance > Mean)	Dispersion tests	Reject the H_0: Dispersion parameter = 1, where H_1: Dispersion parameter > 1	√	√	√	×	√	√	√	×	@	@	@	@	@	@	@	@	√	√	√	×	√	√	√	×
Zero-inflation	Zero-inflation tests	Greater observed zeros than predicted zeros	@	@	@	@	@	@	@	@	√	√	√	×	√	√	√	√	√	√	×	×	√	√	√	√

Note: ML = Multiple Linear Regression Models; NB = Negative Binomial Regression Models; ZIP = Zero-inflated Poisson Models; ZINB = Zero-inflated Negative Binomial Models.
HE = Hyperthyroid Symptoms Scale; HO = Hyperthyroid Symptoms Scale; E = Eye Symptoms scale; T = Tiredness Scale; CO = Cognitive Complaints Scale; A = Anxiety Scale; D = Depressivity Scale; S = Emotional Susceptibility Scale. √ = Assumption was accepted; × = Assumption was violated; @ = Assumption not applicable

Table K28. Results of Diagnostics for Lack of Multicollinearity of ML Models under Full Predictor Models of Six Predictors for Scales without Prominent Floor Effects

Scale	Predictor	VIF	Increased SE	Tolerance
Hyperthyroid symptoms	Mode	1.04	1.02	0.96
	Gender	1.18	1.09	0.85
	Age	1.13	1.06	0.89
	Education level	1.08	1.04	0.93
	Duration of treatment	1.05	1.03	0.95
	Levels of control	1.04	1.02	0.96
Hypothyroid symptoms	Mode	1.04	1.02	0.96
	Gender	1.18	1.09	0.85
	Age	1.13	1.06	0.89
	Education level	1.08	1.04	0.93
	Duration of treatment	1.05	1.03	0.95
	Levels of control	1.04	1.02	0.96
Eye symptoms	Mode	1.04	1.02	0.96
	Gender	1.18	1.09	0.85
	Age	1.13	1.06	0.89
	Education level	1.08	1.04	0.93
	Duration of treatment	1.05	1.03	0.95
	Levels of control	1.04	1.02	0.96
Tiredness	Mode	1.04	1.02	0.96
	Gender	1.18	1.09	0.85
	Age	1.13	1.06	0.89
	Education level	1.08	1.04	0.93
	Duration of treatment	1.05	1.03	0.95
	Levels of control	1.04	1.02	0.96
Cognitive complaints	Mode	1.04	1.02	0.96
	Gender	1.18	1.09	0.85
	Age	1.13	1.06	0.89
	Education level	1.08	1.04	0.93
	Duration of treatment	1.05	1.03	0.95
	Levels of control	1.04	1.02	0.96
Anxiety	Mode	1.04	1.02	0.96
	Gender	1.18	1.09	0.85
	Age	1.13	1.06	0.89
	Education level	1.08	1.04	0.93
	Duration of treatment	1.05	1.03	0.95
	Levels of control	1.04	1.02	0.96
Depressivity	Mode	1.04	1.02	0.96
	Gender	1.18	1.09	0.85
	Age	1.13	1.06	0.89
	Education level	1.08	1.04	0.93
	Duration of treatment	1.05	1.03	0.95
	Levels of control	1.04	1.02	0.96
Emotional susceptibility	Mode	1.04	1.02	0.96
	Gender	1.18	1.09	0.85
	Age	1.13	1.06	0.89
	Education level	1.08	1.04	0.93
	Duration of treatment	1.05	1.03	0.95
	Levels of control	1.04	1.02	0.96
Composite	Mode	1.04	1.02	0.96
	Gender	1.18	1.09	0.85
	Age	1.13	1.06	0.89
	Education level	1.08	1.04	0.93
	Duration of treatment	1.05	1.03	0.95
	Levels of control	1.04	1.02	0.96

Table K29. Results of Diagnostics for Lack of Multicollinearity of Poisson Models under Full Predictor Models of Six Predictors for Scales without Prominent Floor Effects

Scale	Predictor	VIF	Increased SE	Tolerance
Hyperthyroid symptoms	Mode	1.06	1.03	0.94
	Gender	1.19	1.09	0.84
	Age	1.14	1.07	0.88
	Education level	1.09	1.05	0.91
	Duration of treatment	1.08	1.04	0.92
	Levels of control	1.05	1.02	0.96
Hypothyroid symptoms	Mode	1.04	1.02	0.96
	Gender	1.18	1.09	0.84
	Age	1.14	1.07	0.87
	Education level	1.10	1.05	0.91
	Duration of treatment	1.08	1.04	0.92
	Levels of control	1.05	1.02	0.96
Eye symptoms	Mode	1.05	1.02	0.96
	Gender	1.17	1.08	0.86
	Age	1.15	1.07	0.87
	Education level	1.09	1.04	0.92
	Duration of treatment	1.07	1.03	0.94
	Levels of control	1.06	1.03	0.95
Tiredness	Mode	1.05	1.02	0.95
	Gender	1.17	1.08	0.86
	Age	1.12	1.06	0.89
	Education level	1.08	1.04	0.92
	Duration of treatment	1.06	1.03	0.95
	Levels of control	1.05	1.02	0.95
Cognitive complaints	Mode	1.05	1.02	0.95
	Gender	1.16	1.08	0.86
	Age	1.13	1.06	0.89
	Education level	1.08	1.04	0.92
	Duration of treatment	1.06	1.03	0.94
	Levels of control	1.05	1.03	0.95
Anxiety	Mode	1.06	1.03	0.95
	Gender	1.16	1.08	0.86
	Age	1.11	1.05	0.90
	Education level	1.09	1.04	0.92
	Duration of treatment	1.06	1.03	0.95
	Levels of control	1.05	1.02	0.95
Depressivity	Mode	1.06	1.03	0.94
	Gender	1.20	1.10	0.83
	Age	1.13	1.06	0.89
	Education level	1.10	1.05	0.91
	Duration of treatment	1.06	1.03	0.94
	Levels of control	1.05	1.02	0.96
Emotional susceptibility	Mode	1.05	1.03	0.95
	Gender	1.18	1.08	0.85
	Age	1.11	1.05	0.90
	Education level	1.09	1.04	0.92
	Duration of treatment	1.05	1.03	0.95
	Levels of control	1.04	1.02	0.96
Composite	Mode	1.06	1.03	0.94
	Gender	1.19	1.09	0.84
	Age	1.13	1.06	0.89
	Education level	1.09	1.05	0.92
	Duration of treatment	1.06	1.03	0.94
	Levels of control	1.05	1.02	0.95

Table K30. Results of Diagnostics for Lack of Multicollinearity of NB Models under Full Predictor Models of Six Predictors for Scales without Prominent Floor Effects

Scale	Predictor	VIF	Increased SE	Tolerance
Hyperthyroid symptoms	Mode	1.05	1.03	0.95
	Gender	1.19	1.09	0.84
	Age	1.13	1.06	0.88
	Education level	1.08	1.04	0.92
	Duration of treatment	1.07	1.03	0.94
	Levels of control	1.04	1.02	0.96
Hypothyroid symptoms	Mode	1.04	1.02	0.96
	Gender	1.18	1.09	0.85
	Age	1.13	1.06	0.88
	Education level	1.09	1.04	0.92
	Duration of treatment	1.05	1.03	0.95
	Levels of control	1.05	1.02	0.96
Eye symptoms	Mode	1.04	1.02	0.96
	Gender	1.17	1.08	0.86
	Age	1.13	1.06	0.88
	Education level	1.08	1.04	0.93
	Duration of treatment	1.06	1.03	0.95
	Levels of control	1.05	1.02	0.96
Tiredness	Mode	1.05	1.02	0.95
	Gender	1.17	1.08	0.86
	Age	1.12	1.06	0.89
	Education level	1.08	1.04	0.93
	Duration of treatment	1.06	1.03	0.95
	Levels of control	1.05	1.02	0.96
Cognitive complaints	Mode	1.05	1.02	0.96
	Gender	1.17	1.08	0.86
	Age	1.13	1.06	0.89
	Education level	1.08	1.04	0.93
	Duration of treatment	1.06	1.03	0.95
	Levels of control	1.04	1.02	0.96
Anxiety	Mode	1.05	1.03	0.95
	Gender	1.17	1.08	0.85
	Age	1.12	1.06	0.90
	Education level	1.08	1.04	0.93
	Duration of treatment	1.05	1.03	0.95
	Levels of control	1.04	1.02	0.96
Depressivity	Mode	1.06	1.03	0.95
	Gender	1.20	1.09	0.84
	Age	1.13	1.06	0.89
	Education level	1.09	1.04	0.92
	Duration of treatment	1.06	1.03	0.95
	Levels of control	1.04	1.02	0.96
Emotional susceptibility	Mode	1.05	1.03	0.95
	Gender	1.18	1.08	0.85
	Age	1.11	1.05	0.90
	Education level	1.09	1.04	0.92
	Duration of treatment	1.05	1.03	0.95
	Levels of control	1.04	1.02	0.96
Composite	Mode	1.05	1.02	0.95
	Gender	1.18	1.09	0.77
	Age	1.13	1.06	0.89
	Education level	1.08	1.04	0.93
	Duration of treatment	1.06	1.03	0.95
	Levels of control	1.04	1.02	0.96

Table K31. Results of Dispersion Tests under Full predictor Models of All Six Predictors for Scales without Prominent Floor Effects

Scale	dispersion	z	*p*
Hyperthyroid symptoms	1.98	5.77	< 0.001**
Hypothyroid symptoms	2.17	5.94	< 0.001**
Eye symptoms	2.02	5.37	< 0.001**
Tiredness	1.15	1.51	0.07
Cognitive complaints	1.93	5.25	< 0.001**
Anxiety	1.87	6.13	< 0.001**
Depressivity	1.42	3.03	0.001**
Emotional Susceptibility	0.96	-0.39	0.65
Composite	5.49	8.48	< 0.001**

*p < 0.05, **p < 0.01

Table K32. Results of Zero-inflation under Full predictor Models of All Six Predictors for Scales without Prominent Floor Effects

Scale	Observed zeros	Predicted zeros	Ratio	Results
ZIP models				
Hyperthyroid symptoms	22	5	0.23	Underfitted zeros
Hypothyroid symptoms	11	3	0.27	Underfitted zeros
Eye symptoms	26	11	0.42	Underfitted zeros
Tiredness	1	1	1.00	Within tolerance range
Cognitive complaints	22	7	0.32	Underfitted zeros
Anxiety	22	9	0.41	Underfitted zeros
Depressivity	8	3	0.38	Underfitted zeros
Emotional Susceptibility	3	2	0.67	Underfitted zeros
Composite	1	0	0.00	Underfitted zeros
ZINB models				
Hyperthyroid symptoms	22	13	0.59	Underfitted zeros
Hypothyroid symptoms	11	10	0.91	Underfitted zeros
Eye symptoms	26	26	1.00	Within tolerance range
Tiredness	1	2	2.00	Overfitted zeros
Cognitive complaints	22	14	0.64	Underfitted zeros
Anxiety	22	16	0.73	Underfitted zeros
Depressivity	8	5	0.62	Underfitted zeros
Emotional Susceptibility	3	2	0.67	Underfitted zeros
Composite	1	0	0.00	Underfitted zeros

Table K33. Summary of Assumption Checking/Diagnostic Results for Composite Scale

Assumptions/Diagnostics	Checked by	Desired Outcome	Results
Normal Distribution	One-sample K-S tests	Obtain the H_0: The data is normally distributed	√
Poisson Distribution	Goodness-of-fit tests for Poisson	Obtain the H_0: The data can be fit to a Poisson distribution	×
NB Distribution	Goodness-of-fit tests for NB	Obtain the H_0: The data can be fit to a NB distribution	×

One Predictor Model of Gender			Models					
			ML	T	P	NB	ZIP	ZINB
Linearity	Scatterplot between fitted value and standardized residuals	The residuals follow the loess line without particular variation	×	√	@	@	@	@
Specification of Predictors			×	√	@	@	@	@
Independence of Errors			×	√	@	@	@	@
Homogeneity of Variance	Scatterplot between fitted value and squared root of standardized residuals	The squared roots of standardized residuals follow the loess line without particular variation	×	@	√	@	@	@
Normality of Errors	Normal Q-Q plots	The points roughly follow a positive, straight line	√	√	√	√	@	@
Mean = Variance	Dispersion tests	Obtain the H_0: Dispersion parameter = 1	@	@	×	@	×	@
Overdispersion (Variance > Mean)	Dispersion tests	Reject the H_0: Dispersion parameter = 1, where H_1: Dispersion parameter > 1	@	@	@	√	@	√
Zero-inflation	Zero-inflation tests	Greater observed zeros than predicted zeros	@	@	@	@	√	√

One Predictor Model of Mode of Administration			Models					
			ML	T	P	NB	ZIP	ZINB
Linearity	Scatterplot between fitted value and standardized residuals	The residuals follow the loess line without particular variation	×	√	@	@	@	@
Specification of Predictors			×	√	@	@	@	@
Independence of Errors			×	√	@	@	@	@
Homogeneity of Variance	Scatterplot between fitted value and squared root of standardized residuals	The squared roots of standardized residuals follow the loess line without particular variation	×	@	√	@	@	@
Normality of Errors	Normal Q-Q plots	The points roughly follow a positive, straight line	√	√	√	√	@	@
Mean = Variance	Dispersion tests	Obtain the H_0: Dispersion parameter = 1	@	@	×	@	×	@
Overdispersion (Variance > Mean)	Dispersion tests	Reject the H_0: Dispersion parameter = 1, where H_1: Dispersion parameter > 1	@	@	@	√	@	√
Zero-inflation	Zero-inflation tests	Greater observed zeros than predicted zeros	@	@	@	@	√	√

Assumptions/Diagnostics	Checked by	Desired Outcome	Results					
Full Predictor Model of Six Predictors			Models					
			ML	T	P	NB	ZIP	ZINB
Linearity	Scatterplot between fitted value and standardized residuals	The residuals follow the loess line without particular variation	√	√	@	@	@	@
Specification of Predictors			√	√	@	@	@	@
Independence of Errors			√	√	@	@	@	@
Homogeneity of Variance	Scatterplot between fitted value and squared root of standardized residuals	The squared roots of standardized residuals follow the loess line without particular variation	√	@	√	@	@	@
Normality of Errors	Normal Q-Q plots	The points roughly follow a positive, straight line	√	√	√	√	@	@
Collinearity	VIF & Tolerance	VIF < 10 & Tolerance < 1	√	@	√	√	@	@
Mean = Variance	Dispersion tests	Obtain the H_0: Dispersion parameter = 1	@	@	×	@	×	@
Overdispersion (Variance > Mean)	Dispersion tests	Reject the H_0: Dispersion parameter = 1, where H_1: Dispersion parameter > 1	@	@	@	√	@	√
Zero-inflation	Zero-inflation tests	Greater observed zeros than predicted zeros	@	@	@	@	√	√

Note: ML = Multiple Linear Regression Models; NB = Negative Binomial Regression Models; ZIP = Zero-inflated Poisson Models; ZINB = Zero-inflated Negative Binomial Models.
HE = Hyperthyroid Symptoms Scale; HO = Hyperthyroid Symptoms Scale; T = Tiredness Scale; CO = Cognitive Complaints Scale; A = Anxiety Scale; D = Depressivity Scale; S = Emotional Susceptibility Scale.
√ = Assumption was accepted; × = Assumption was violated; @ = Assumption not applicable.

Appendix L. Tables of Model Comparisons for RQ 3

Content of the tables in Appendix L

Table L1. Results of AIC under One Predictor Models of Gender for Scales with Prominent Floor Effects

Model	Goiter symptoms	Impaired social life	Impaired daily life	Cosmetic complaints
Linear	830.340	857.645	839.996	887.138
Tobit	778.964	731.130	714.494	838.125
Poisson	824.348	877.190	819.437	912.626
NB	767.043	714.067	679.250	794.901
ZIP	771.281	744.751	735.690	872.047
ZINB	759.405	711.651	681.677	798.889
Best model	ZINB	ZINB	NB	NB
Worst	Linear	Poisson	Linear	Poisson

Table L2. Results of BIC under One Predictor Models of Gender for Scales with Prominent Floor Effects

Model	Goiter symptoms	Impaired social life	Impaired daily life	Cosmetic complaints
Linear	839.902	867.208	849.558	896.700
Tobit	778.527	740.692	724.056	847.687
Poisson	830.723	883.565	825.812	919.000
NB	776.605	723.629	688.812	804.463
ZIP	784.030	757.501	748.439	844.796
ZINB	775.342	727.588	697.614	814.826
Best model	ZINB	NB	NB	NB
Worst	Linear	Poisson	Linear	Poisson

Table L3. Results of AIC under One Predictor Models of Mode of Administration for Scales with Prominent Floor Effects

Model	Goiter symptoms	Impaired social life	Impaired daily life	Cosmetic complaints
Linear	855.273	844.876	838.213	885.550
Tobit	802.433	717.649	716.129	837.780
Poisson	875.582	836.481	808.498	907.651
NB	787.327	701.848	674.600	792.976
ZIP	794.902	725.429	717.203	864.943
ZINB	776.450	697.741	676.509	796.936
Best model	ZINB	ZINB	NB	NB
Worst	Poisson	Linear	Linear	Poisson

Table L4. Results of BIC under One Predictor Models of Mode of Administration for Scales with Prominent Floor Effects

Model	Goiter symptoms	Impaired social life	Impaired daily life	Cosmetic complaints
Linear	864.835	854.438	847.775	895.112
Tobit	811.995	727.211	725.692	847.342
Poisson	881.957	842.856	814.873	914.026
NB	796.890	711.410	648.163	802.538
ZIP	807.651	738.178	729.952	877.692
ZINB	792.387	713.678	692.446	812.873
Best model	ZINB	NB	NB	NB
Worst	Poisson	Linear	Linear	Poisson

Table L5. Results of AIC under Full predictor Models of Six Predictors for Scales with Prominent Floor Effects

Model	Goiter symptoms	Impaired social life	Impaired daily life	Cosmetic complaints
Linear	804.222	828.391	814.024	851.313
Tobit	749.748	699.555	692.464	801.007
Poisson	772.907	773.855	721.933	808.544
NB	742.711	689.119	652.005	755.569
ZIP	721.124	695.412	664.261	797.648
ZINB	718.660	678.440	646.154	751.788
Best model	ZINB	ZINB	ZINB	ZINB
Worst	Linear	Linear	Linear	Linear

Table L6. Results of BIC under Full predictor Models of Six Predictors for Scales with Prominent Floor Effects

Model	Goiter symptoms	Impaired social life	Impaired daily life	Cosmetic complaints
Linear	829.721	853.891	839.523	876.812
Tobit	775.247	725.054	717.963	826.506
Poisson	795.219	796.166	744.244	830.856
NB	768.211	714.619	677.505	781.068
ZIP	765.747	740.885	704.885	842.272
ZINB	766.471	726.250	693.965	799.599
Best model	ZIP	NB	NB	NB
Worst	Linear	Linear	Linear	Linear

Table L7. Results of LRTs under One Predictor of Gender for Scales with Prominent Floor Effects

Comparison	Scale	*df*	χ^2	*p*
Between Poisson and NB	Goiter symptoms	1	59.31	< 0.001**
	Impaired social life	1	165.12	< 0.001**
	Impaired daily life	1	142.19	< 0.001**
	Cosmetic complaints	1	119.72	< 0.001**
Between ZIP and ZINB	Goiter symptoms	1	13.88	< 0.001**
	Impaired social life	1	35.10	< 0.001**
	Impaired daily life	1	56.01	< 0.001**
	Cosmetic complaints	1	75.16	< 0.001**

Note: *p < 0.05, **p < 0.01

Table L8. Results of LRTs under One Predictor of Mode of Administration for Scales with Prominent Floor Effects

Comparison	Scale	*df*	χ^2	*p*
Between Poisson and NB	Goiter symptoms	1	90.25	< 0.001**
	Impaired social life	1	136.63	< 0.001**
	Impaired daily life	1	135.90	< 0.001**
	Cosmetic complaints	1	116.67	< 0.001**
Between ZIP and ZINB	Goiter symptoms	1	20.45	< 0.001**
	Impaired social life	1	29.69	< 0.001**
	Impaired daily life	1	42.69	< 0.001**
	Cosmetic complaints	1	70.00	< 0.001**

Note: *p < 0.05, **p < 0.01

Table L9. Results of LRTs under Full predictor Models of All Six Predictors for Scales with Prominent Floor Effects

Comparison	Scale	*df*	χ^2	*p*
Between Poisson and NB	Goiter symptoms	1	32.20	< 0.001**
	Impaired social life	1	86.74	< 0.001**
	Impaired daily life	1	71.93	< 0.001**
	Cosmetic complaints	1	54.98	< 0.001**
Between ZIP and ZINB	Goiter symptoms	1	4.46	0.03*
	Impaired social life	1	18.97	< 0.001**
	Impaired daily life	1	20.11	< 0.001**
	Cosmetic complaints	1	47.86	< 0.001**

Note: *p < 0.05, **p < 0.01

Table L10. Results of BIC-corrected Vuong Tests under One Predictor of Gender for Scales with Prominent Floor Effects

Comparison	Scale	H_a[1]	z	*p*
Between Poisson and ZIP	Goiter symptoms	ZIP > P	-2.73	0.003**
	Impaired social life	ZIP > P	-4.41	< 0.001**
	Impaired daily life	ZIP > P	-3.17	< 0.001**
	Cosmetic complaints	ZIP > P	-2.00	0.02*
Between NB and ZINB	Goiter symptoms	ZINB > NB	-0.20	0.42
	Impaired social life	NB > ZINB	0.91	0.18
	Impaired daily life	ZINB > NB	3.65	< 0.001**
	Cosmetic complaints	NB > ZINB	47.63	< 0.001**

Note: [1] H_a = Alternative Hypothesis; *p < 0.05, **p < 0.01.

Table L11. Results of BIC-corrected Vuong Tests under One Predictor of Mode of Administration for Scales with Prominent Floor Effects

Comparison	Scale	H_a[1]	z	*p*
Between Poisson and ZIP	Goiter symptoms	ZIP > P	-3.30	< 0.001**
	Impaired social life	ZIP > P	-4.02	< 0.001**
	Impaired daily life	ZIP > P	-3.39	< 0.001**
	Cosmetic complaints	ZIP > P	-2.08	0.02*
Between NB and ZINB	Goiter symptoms	ZINB > NB	-0.61	0.27
	Impaired social life	NB > ZINB	0.46	0.32
	Impaired daily life	NB > ZINB	2.88	0.002**
	Cosmetic complaints	NB > ZINB	24.61	< 0.001**

Note: [1] H_a = Alternative Hypothesis; *p < 0.05, **p < 0.01.

Table L12. Results of BIC-corrected Vuong Tests under Full predictor Models of All Six Predictors for Scales with Prominent Floor Effects

Comparison	Scale	H_a[1]	z	*p*
Between Poisson and ZIP	Goiter symptoms	ZIP > P	-1.64	0.05
	Impaired social life	ZIP > P	-2.37	0.009**
	Impaired daily life	ZIP > P	-1.67	0.048*
	Cosmetic complaints	P > ZIP	0.86	0.19
Between NB and ZINB	Goiter symptoms	ZINB > NB	-0.15	0.44
	Impaired social life	NB > ZINB	1.32	0.09
	Impaired daily life	NB > ZINB	1.88	0.03*
	Cosmetic complaints	NB > ZINB	2.82	0.002**

Note: [1] H_a = Alternative Hypothesis; *p < 0.05, **p < 0.01.

Table L13. Summary of Results of Best Models under One Predictor Models of Gender for Scales with Prominent Floor Effects

Comparison	Test applied to the comparison	Goiter symptoms	Impaired social life	Impaired daily life	Cosmetic complaints
L vs. T	AIC & BIC	T	T	T	T
L vs. P	AIC & BIC	P	L	P	L
L vs. NB	AIC & BIC	NB	NB	NB	NB
L vs. ZIP	AIC & BIC	ZIP	ZIP	ZIP	ZIP
L vs. ZINB	AIC & BIC	ZINB	ZINB	ZINB	ZINB
T vs. P	AIC & BIC	T	T	T	T
T vs. NB	AIC & BIC	NB	NB	NB	NB
T vs. ZIP	AIC & BIC	Unknown	T	T	Unknown
T vs. ZINB	AIC & BIC	ZINB	ZINB	ZINB	ZINB
P vs. NB	LRT	NB	NB	NB	NB
P vs. ZIP	Vuong test	ZIP	ZIP	ZIP	ZIP
P vs. ZINB	AIC & BIC	ZINB	ZINB	ZINB	ZINB
NB vs. ZIP	AIC & BIC	NB	NB	NB	NB
NB vs. ZINB	Vuong test	no difference	no difference	ZINB	NB
ZIP vs. ZINB	LRT	ZINB	ZINB	ZINB	ZINB
Best model		NB/ZINB	NB/ZINB	NB/ZINB	NB/ZINB

Note: L = Multiple linear model, P = Poisson regression model, T = Tobit model.

Table L14. Summary of Results of Best Models under One Predictor Models of Mode of Administration for Scales with Prominent Floor Effects

Comparison	Test applied to the comparison	Goiter symptoms	Impaired social life	Impaired daily life	Cosmetic complaints
L vs. T	AIC & BIC	T	T	T	T
L vs. P	AIC & BIC	L	P	P	L
L vs. NB	AIC & BIC	NB	NB	NB	NB
L vs. ZIP	AIC & BIC	ZIP	ZIP	ZIP	ZIP
L vs. ZINB	AIC & BIC	ZINB	ZINB	ZINB	ZINB
T vs. P	AIC & BIC	T	T	T	T
T vs. NB	AIC & BIC	NB	NB	NB	NB
T vs. ZIP	AIC & BIC	ZIP	T	T	T
T vs. ZINB	AIC & BIC	ZINB	ZINB	ZINB	ZINB
P vs. NB	LRT	NB	NB	NB	NB
P vs. ZIP	Vuong test	ZIP	ZIP	ZIP	ZIP
P vs. ZINB	AIC & BIC	ZINB	ZINB	ZINB	ZINB
NB vs. ZIP	AIC & BIC	NB	NB	NB	NB
NB vs. ZINB	Vuong test	no difference	no difference	NB	NB
ZIP vs. ZINB	LRT	ZINB	ZINB	ZINB	ZINB
Best model		ZINB/NB	ZINB/NB	NB	NB

Note: L = Multiple linear model, P = Poisson regression model, T = Tobit model.

Table L15. Summary of Results of Best Models under Full predictor Models of Six Predictors for Scales with Prominent Floor Effects

Comparison	Test applied to the comparison	Goiter symptoms	Impaired social life	Impaired daily life	Cosmetic complaints
L vs. T	AIC & BIC	T	T	T	T
L vs. P	AIC & BIC	P	P	P	P
L vs. NB	AIC & BIC	NB	NB	NB	NB
L vs. ZIP	AIC & BIC	ZIP	ZIP	ZIP	ZIP
L vs. ZINB	AIC & BIC	ZINB	ZINB	ZINB	ZINB
T vs. P	AIC & BIC	T	T	T	T
T vs. NB	AIC & BIC	NB	NB	NB	NB
T vs. ZIP	AIC & BIC	ZIP	T	ZIP	Unknown
T vs. ZINB	AIC & BIC	ZINB	ZINB	ZINB	ZINB
P vs. NB	LRT	NB	NB	NB	NB
P vs. ZIP	Vuong test	no difference	ZIP	ZIP	no difference
P vs. ZINB	AIC & BIC	ZINB	ZINB	ZINB	ZINB
NB vs. ZIP	AIC & BIC	unknown	NB	NB	NB
NB vs. ZINB	Vuong test	no difference	no difference	no difference	NB
ZIP vs. ZINB	LRT	ZINB	ZINB	ZINB	ZINB
Best model		NB/ZINB	NB/ZINB	NB/ZINB	NB

Note: L = Multiple linear model, P = Poisson regression model, T = Tobit model.

Table L16. Results of AIC under One Predictor Models of Gender for Scales without Prominent Floor Effects

Model	Hyperthyroid symptoms	Hypothyroid symptoms	Eye symptoms	Eye symptoms	Cognitive Complaints	Anxiety	Depressivity	Emotional Susceptibility	Composite
Linear	917.290	935.386	849.258	849.258	875.922	895.415	878.093	840.642	1453.257
Tobit	888.582	922.069	821.231	821.231	851.897	868.330	873.186	840.687	1449.588
Poisson	969.933	977.626	857.019	857.019	897.263	930.263	887.510	842.273	2204.749
NB	879.747	894.131	788.424	788.424	831.392	845.491	861.349	842.819	1433.771
ZIP	913.686	956.965	828.572	828.572	862.639	892.194	875.168	840.584	2170.063
ZINB	877.370	897.522	791.628	791.628	832.957	848.479	861.069	842.303	1433.647
Best model	ZINB	NB	NB	NB	NB	NB	ZINB	ZIP	ZINB
Worst	Poisson	Poisson	Poisson	Poisson	Poisson	Poisson	Poisson	ZINB	Poisson

Table L17. Results of BIC under One Predictor Models of Gender for Scales without Prominent Floor Effects

Model	Hyperthyroid symptoms	Hypothyroid symptoms	Eye symptoms	Tiredness	Cognitive Complaints	Anxiety	Depressivity	Emotional Susceptibility	Composite
Linear	926.852	944.949	858.820	875.130	885.484	904.978	887.645	850.204	1462.820
Tobit	898.144	931.631	830.793	874.896	861.549	877.892	882.249	850.249	1489.151
Poisson	976.307	984.001	863.565	867.220	903.637	936.864	893.885	848.648	2211.124
NB	889.309	903.693	797.986	863.563	840.954	855.053	870.911	852.381	1443.333
ZIP	926.435	969.714	841.231	877.376	875.389	904.943	887.917	853.334	2182.813
ZINB	893.306	913.459	807.565	873.940	848.894	864.415	877.006	858.240	1449.583
Best model	NB	NB	NB	NB	NB	NB	NB	Poisson	NB
Worst	Poisson	ZIP	Poisson	ZIP	Poisson	Poisson	Poisson	ZINB	Poisson

Table L18. Results of AIC under One Predictor Models of Mode of Administration for Scales without Prominent Floor Effects

Model	Hyperthyroid symptoms	Hypothyroid symptoms	Eye symptoms	Tiredness	Cognitive Complaints	Anxiety	Depressivity	Emotional Susceptibility	Composite
Linear	911.836	934.313	848.713	852.600	865.110	874.363	866.010	827.982	1431.420
Tobit	884.062	921.259	820.727	852.647	841.558	846.715	860.629	828.178	1428.255
Poisson	957.141	975.318	855.902	843.846	874.435	883.023	867.845	828.372	2051.032
NB	874.720	893.219	787.954	840.967	820.861	824.755	849.370	830.140	1412.603
ZIP	901.593	952.863	828.124	846.567	845.101	857.897	861.881	829.019	2005.145
ZINB	871.165	895.724	791.092	844.593	822.117	826.726	851.398	831.019	1409.251
Best model	ZINB	NB	NB	NB	NB	NB	NB	Tobit	ZINB
Worst	Poisson	Poisson	Poisson	Linear	Poisson	Poisson	Poisson	ZINB	Poisson

Table L19. Results of BIC under One Predictor Models of Mode of Administration for Scales without Prominent Floor Effects

Model	Hyperthyroid symptoms	Hypothyroid symptoms	Eye symptoms	Tiredness	Cognitive Complaints	Anxiety	Depressivity	Emotional Susceptibility	Composite
Linear	921.398	943.875	858.275	862.162	874.672	883.925	875.572	837.544	1440.982
Tobit	893.628	930.821	830.289	862.209	851.120	856.277	870.191	837.741	1437.817
Poisson	963.516	981.693	862.278	850.221	880.810	889.220	874.220	843.747	2057.407
NB	884.283	902.781	797.516	850.530	830.423	834.317	858.932	839.703	1422.165
ZIP	914.342	965.613	840.874	859.316	857.850	870.647	874.630	841.768	2017.895
ZINB	887.102	911.661	807.029	860.530	838.054	842.663	867.335	846.956	1425.188
Best model	NB	NB	NB	Poisson	NB	NB	NB	Poisson	NB
Worst	Poisson	ZIP	Poisson	Tobit	Poisson	Poisson	Linear	ZINB	Poisson

Table L20. Results of AIC under Full Models of Six Predictors for Scales without Prominent Floor Effects

Model	Hyperthyroid symptoms	Hypothyroid symptoms	Eye symptoms	Tiredness	Cognitive Complaints	Anxiety	Depressivity	Emotional Susceptibility	Composite
Linear	902.229	936.696	848.994	849.501	867.955	871.287	857.977	813.487	1415.687
Tobit	874.068	923.754	820.798	849.390	843.968	841.917	853.550	813.044	1412.274
Poisson	924.821	968.236	846.462	838.874	871.132	867.861	851.206	814.506	1907.334
NB	867.525	895.076	788.885	838.727	824.564	822.862	841.848	816.508	1399.902
ZIP	890.343	952.494	829.306	846.885	849.908	850.668	846.981	817.811	1881.056
ZINB	862.741	896.020	790.093	847.379	829.117	827.579	843.113	819.811	1400.737
Best model	ZINB	NB	NB	NB	NB	NB	NB	Tobit	ZINB
Worst	Poisson	Poisson	Linear	Linear	Poisson	Poisson	Linear	ZINB	Poisson

Table L21. Results of BIC under Full Models of Six Predictors for Scales without Prominent Floor Effects

Model	Hyperthyroid symptoms	Hypothyroid symptoms	Eye symptoms	Tiredness	Cognitive Complaints	Anxiety	Depressivity	Emotional Susceptibility	Composite
Linear	927.728	962.195	874.493	875.000	893.454	896.786	883.476	838.986	1441.186
Tobit	899.567	949.253	846.297	874.889	869.467	867.417	879.049	838.543	1437.773
Poisson	947.132	990.548	868.774	861.186	893.444	890.173	873.518	836.818	1929.646
NB	893.024	920.575	814.384	864.226	850.063	848.361	867.347	842.007	1425.401
ZIP	934.966	997.117	873.929	891.509	894.531	895.291	891.605	862.434	1925.680
ZINB	910.551	943.831	837.904	895.190	876.928	875.390	890.924	867.622	1448.548
Best model	NB	NB	NB	Poisson	NB	NB	NB	Poisson	NB
Worst	Poisson	ZIP	Linear	ZINB	ZIP	Linear	ZIP	ZINB	Poisson

Table L22. Results of LRTs under One Predictor of Gender for Scales without Floor Effect

Comparison	Scale	*df*	χ^2	*p*
Between Poisson and NB	Hyperthyroid symptoms	1	92.19	< 0.001**
	Hypothyroid symptoms	1	85.50	< 0.001**
	Eye symptoms	1	70.60	< 0.001**
	Tiredness	1	8.84	0.003**
	Cognitive complaints	1	67.87	< 0.001**
	Anxiety	1	87.00	< 0.001**
	Depressivity	1	28.16	< 0.001**
	Emotional Susceptibility	1	1.45	0.23
	Composite	1	509.43	< 0.001**
Between ZIP and ZINB	Hyperthyroid symptoms	1	38.32	< 0.001**
	Hypothyroid symptoms	1	61.44	< 0.001**
	Eye symptoms	1	38.94	< 0.001**
	Tiredness	1	8.62	0.003**
	Cognitive complaints	1	31.68	< 0.001**
	Anxiety	1	16.10	< 0.001**
	Depressivity	1	45.72	< 0.001**
	Emotional Susceptibility	1	0.28	0.60
	Composite	1	482.32	< 0.001**

Note: *p < 0.05, **p < 0.01

Table L23. Results of LRTs under One Predictor of Mode of Administration for Scales without Floor Effect

Comparison	Scale	*df*	χ^2	*p*
Between Poisson and NB	Hyperthyroid symptoms	1	84.42	< 0.001**
	Hypothyroid symptoms	1	84.10	< 0.001**
	Eye symptoms	1	69.95	< 0.001**
	Tiredness	1	4.88	0.03*
	Cognitive complaints	1	55.57	< 0.001**
	Anxiety	1	60.27	< 0.001**
	Depressivity	1	20.58	< 0.001**
	Emotional Susceptibility	1	0.23	0.63
	Composite	1	509.43	< 0.001**
Between ZIP and ZINB	Hyperthyroid symptoms	1	32.43	< 0.001**
	Hypothyroid symptoms	1	59.14	< 0.001**
	Eye symptoms	1	39.03	< 0.001**
	Tiredness	1	3.97	0.046*
	Cognitive complaints	1	24.98	< 0.001**
	Anxiety	1	33.17	< 0.001**
	Depressivity	1	12.48	< 0.001**
	Emotional Susceptibility	1	-0.0004	> 0.999
	Composite	1	482.32	< 0.001**

Note: *p < 0.05, **p < 0.01

Table L24. Results of LRTs under Full Model of Six Predictors for Scales without Floor Effect

Comparison	Scale	*df*	χ^2	*p*
Between Poisson and NB	Hyperthyroid symptoms	1	59.30	< 0.001**
	Hypothyroid symptoms	1	75.16	< 0.001**
	Eye symptoms	1	59.58	< 0.001**
	Tiredness	1	2.15	0.14
	Cognitive complaints	1	48.59	< 0.001**
	Anxiety	1	47.00	< 0.001**
	Depressivity	1	11.36	< 0.001**
	Emotional Susceptibility	1	-0.001	> 0.999
	Composite	1	509.43	< 0.001**
Between ZIP and ZINB	Hyperthyroid symptoms	1	29.60	< 0.001**
	Hypothyroid symptoms	1	58.57	< 0.001**
	Eye symptoms	1	41.21	< 0.001**
	Tiredness	1	1.51	0.22
	Cognitive complaints	1	22.79	< 0.001**
	Anxiety	1	25.09	< 0.001**
	Depressivity	1	5.87	0.02*
	Emotional Susceptibility	1	< 0.001	> 0.999
	Composite	1	482.32	< 0.001**

Note: *p < 0.05, **p < 0.01

Table L25. Results of BIC-corrected Vuong Tests under One Predictor of Gender for Scales without Floor Effect

Comparison	Scale	H_a[1]	z	*p*
Between Poisson and ZIP	Hyperthyroid symptoms	ZIP > P	-2.51	0.006**
	Hypothyroid symptoms	ZIP > P	-1.08	0.14
	Eye symptoms	ZIP > P	-1.60	0.05
	Tiredness	P > ZIP	9.98	< 0.001**
	Cognitive complaints	ZIP > P	-1.84	0.03*
	Anxiety	ZIP > P	-1.96	0.02
	Depressivity	ZIP > P	-0.56	0.29
	Emotional Susceptibility	P > ZIP	0.75	0.23
	Composite	ZIP > P	-0.70	0.24
Between NB and ZINB	Hyperthyroid symptoms	NB > ZINB	0.80	0.21
	Hypothyroid symptoms	NB > ZINB	6.26	< 0.001**
	Eye symptoms	NB > ZINB	5.30	< 0.001**
	Tiredness	NB > ZINB	1641.15	< 0.001**
	Cognitive complaints	NB > ZINB	2.60	0.005**
	Anxiety	NB > ZINB	4.92	< 0.001**
	Depressivity	NB > ZINB	1.31	0.09
	Emotional Susceptibility	NB > ZINB	1.11	0.13
	Composite	NB > ZINB	1.08	0.14

Note: [1] H_a = Alternative Hypothesis; *p < 0.05, **p < 0.01.

Table L26. Results of BIC-corrected Vuong Tests under One Predictor of Mode of Administration for Scales without Floor Effect

Comparison	Scale	H_a[1]	z	*p*
Between Poisson and ZIP	Hyperthyroid symptoms	ZIP > P	-2.44	0.007**
	Hypothyroid symptoms	ZIP > P	-1.15	0.12
	Eye symptoms	ZIP > P	-1.55	0.06
	Tiredness	P > ZIP	3.28	< 0.001*
	Cognitive complaints	ZIP > P	-1.62	0.05
	Anxiety	ZIP > P	-1.46	0.07
	Depressivity	P > ZIP	0.05	0.48
	Emotional Susceptibility	P > ZIP	1.59	0.06
	Composite	ZIP > P	-0.76	0.22
Between NB and ZINB	Hyperthyroid symptoms	NB > ZINB	0.50	0.31
	Hypothyroid symptoms	NB > ZINB	3.53	< 0.001**
	Eye symptoms	NB > ZINB	4.98	< 0.001**
	Tiredness	NB > ZINB	7.43	< 0.001**
	Cognitive complaints	NB > ZINB	2.43	0.008**
	Anxiety	NB > ZINB	3.05	0.001**
	Depressivity	NB > ZINB	2.91	0.002**
	Emotional Susceptibility	NB > ZINB	1.79	0.04*
	Composite	NB > ZINB	0.34	0.37

Note: [1] H_a = Alternative Hypothesis; *p < 0.05, **p < 0.01.

Table L27. Results of BIC-corrected Vuong Tests under Full Models of Six Predictor of for Scales without Floor Effect

Comparison	Scale	H_a[1]	z	*p*
Between Poisson and ZIP	Hyperthyroid symptoms	ZIP > P	-0.69	0.25
	Hypothyroid symptoms	P > ZIP	0.42	0.34
	Eye symptoms	P > ZIP	0.37	0.36
	Tiredness	P > ZIP	3.95	< 0.001**
	Cognitive complaints	P > ZIP	0.08	0.47
	Anxiety	P > ZIP	0.39	0.35
	Depressivity	P > ZIP	1.60	0.05
	Emotional Susceptibility	P > ZIP	2.97	0.001*
	Composite	ZIP > P	-0.10	0.46
Between NB and ZINB	Hyperthyroid symptoms	NB > ZINB	1.61	0.05
	Hypothyroid symptoms	NB > ZINB	2.52	0.006**
	Eye symptoms	NB > ZINB	2.86	0.002**
	Tiredness	NB > ZINB	4.42	< 0.001**
	Cognitive complaints	NB > ZINB	4.01	< 0.001**
	Anxiety	NB > ZINB	4.63	< 0.001**
	Depressivity	NB > ZINB	2.70	0.004**
	Emotional Susceptibility	NB > ZINB	2.97	0.001**
	Composite	NB > ZINB	1.59	0.06

Note: [1] H_a = Alternative Hypothesis; *p < 0.05, **p < 0.01.

Table L28. Summary of Results of Best Models under One Predictor Models of Gender for Scales without Prominent Floor Effects

Comparison	Test applied to the comparison	Hyper	Hypo	Eye	Tiredness	Cognition	Anxiety	Depressivity	Emotion	Composite
L vs. T	AIC & BIC	T	T	T	T	T	T	T	L	T
L vs. P	AIC & BIC	L	L	L	P	L	L	L	Unknown	L
L vs. NB	AIC & BIC	NB	NB	NB	NB	NB	NB	NB	L	NB
L vs. ZIP	AIC & BIC	ZIP	L	ZIP	unknown	ZIP	ZIP	unknown	L	L
L vs. ZINB	AIC & BIC	ZINB	ZINB	ZINB	ZINB	ZINB	ZINB	ZINB	L	ZINB
T vs. P	AIC & BIC	T	T	T	P	T	T	T	unknown	T
T vs. NB	AIC & BIC	NB	NB	NB	NB	NB	NB	NB	T	NB
T vs. ZIP	AIC & BIC	T	T	T	Unknown	T	T	T	T	T
T vs. ZINB	AIC & BIC	ZINB	ZINB	ZINB	ZNB	ZINB	ZINB	ZINB	T	ZINB
P vs. NB	LRT	NB	NB	NB	NB	NB	NB	NB	P	NB
P vs. ZIP	Vuong test	ZIP	=	=	P	ZIP	=	=	=	=
P vs. ZINB	AIC & BIC	ZINB	ZINB	ZINB	unknown	ZINB	ZINB	ZINB	P	ZINB
NB vs. ZIP	AIC & BIC	NB	NB	NB	NB	NB	NB	NB	unknown	NB
NB vs. ZINB	Vuong test	=	=	NB	NB	NB	NB	=	=	=
ZIP vs. ZINB	LRT	ZINB	ZINB	ZINB	=	ZINB	ZINB	ZINB	=	ZINB
Best model		NB/ZINB	NB/ZINB	NB/ZINB	NB	NB	NB	NB/ZINB	P	NB/ZINB

Note: L = Multiple linear model, P = Poisson regression model, T = Tobit model, Hyper = Hyperthyroid symptom scale, Hypo = Hypothyroid symptom scale, Eye = Eye symptoms scale; Cognition = Cognitive Complaints scale, Emotion = Emotional susceptibility scale.
"=" means no difference between the two models.

Table L29. Summary of Results of Best Models under One Predictor Models of Mode of Administration for Scales without Prominent Floor Effects

Comparison	Test applied to the comparison	Hyper	Hypo	Eye	Tiredness	Cognition	Anxiety	Depressivity	Emotion	Composite
L vs. T	AIC & BIC	T	T	T	L	T	T	T	L	T
L vs. P	AIC & BIC	L	L	L	P	L	L	L	P	L
L vs. NB	AIC & BIC	NB	NB	NB	NB	NB	NB	NB	L	NB
L vs. ZIP	AIC & BIC	ZIP	L	ZIP	ZIP	ZIP	ZIP	ZIP	L	L
L vs. ZINB	AIC & BIC	ZINB	ZINB	ZINB	ZINB	ZINB	ZINB	ZINB	L	ZINB
T vs. P	AIC & BIC	T	T	T	P	T	T	T	P	T
T vs. NB	AIC & BIC	NB	NB	NB	NB	NB	NB	NB	T	NB
T vs. ZIP	AIC & BIC	T	T	T	ZIP	T	T	ZIP	T	T
T vs. ZINB	AIC & BIC	ZINB	ZINB	ZINB	ZINB	ZINB	ZINB	ZINB	T	ZINB
P vs. NB	LRT	NB	NB	NB	NB	NB	NB	NB	P	NB
P vs. ZIP	Vuong test	ZIP	=	=	P	=	=	=	=	=
P vs. ZINB	AIC & BIC	ZINB	ZINB	ZINB	ZINB	ZINB	ZINB	ZINB	P	ZINB
NB vs. ZIP	AIC & BIC	NB	NB	NB	NB	NB	NB	NB	ZIP	NB
NB vs. ZINB	Vuong test	=	NB	NB	NB	NB	NB	NB	NB	=
ZIP vs. ZINB	LRT	ZINB	ZINB	ZINB	ZINB	ZINB	ZINB	ZINB	=	ZINB
Best model		NB/ZINB	NB	NB	NB	NB	NB	NB	P	NB/ZINB

Note: L = Multiple linear model, P = Poisson regression model, T = Tobit model, Hyper = Hyperthyroid symptom scale, Hypo = Hypothyroid symptom scale, Eye = Eye symptoms scale; Cognition = Cognitive Complaints scale, Emotion = Emotional susceptibility scale.
"=" means no difference between the two models.

Table L30. Summary of Results of Best Models under Full Models Six Predictors for Scales without Prominent Floor Effects

Comparison	Test applied to the comparison	Hyper	Hypo	Eye	Tiredness	Cognition	Anxiety	Depressivity	Emotion	Composite
L vs. T	AIC & BIC	T	T	T	T	T	T	T	T	T
L vs. P	AIC & BIC	L	L	L	P	L	P	P	P	L
L vs. NB	AIC & BIC	NB	NB	NB	NB	NB	NB	NB	L	NB
L vs. ZIP	AIC & BIC	L	L	ZIP	unknown	unknown	ZIP	unknown	L	L
L vs. ZINB	AIC & BIC	ZINB	ZINB	ZINB	unknown	ZINB	ZINB	unknown	L	ZINB
T vs. P	AIC & BIC	T	T	T	P	T	T	P	P	T
T vs. NB	AIC & BIC	NB	NB	NB	NB	NB	NB	NB	T	NB
T vs. ZIP	AIC & BIC	T	T	Unknown	unknown	ZIP	T	unknown	T	T
T vs. ZINB	AIC & BIC	unknown	ZINB	ZINB	unknown	ZINB	unknown	unknown	T	ZINB
P vs. NB	LRT	NB	NB	NB	=	NB	NB	NB	=	NB
P vs. ZIP	Vuong test	=	=	=	P	=	=	=	=	=
P vs. ZINB	AIC & BIC	ZINB	ZINB	ZINB	unknown	ZINB	unknown	unknown	unknown	ZINB
NB vs. ZIP	AIC & BIC	NB	NB	NB	NB	NB	NB	NB	NB	NB
NB vs. ZINB	Vuong test	=	NB	NB	NB	NB	NB	NB	NB	=
ZIP vs. ZINB	LRT	ZINB	ZINB	ZINB	=	ZINB	ZINB	ZINB	=	ZINB
Best model		NB/ZINB	NB	NB/ZINB	NB	NB	NB	NB	P	NB/ZINB

Note: L = Multiple linear model, P = Poisson regression model, T = Tobit model, Hyper = Hyperthyroid symptom scale, Hypo = Hypothyroid symptom scale, Eye = Eye symptoms scale; Cognition = Cognitive Complaints scale, Emotion = Emotional susceptibility scale.
"=" means no difference between the two models.

Appendix M. Tables of Predictor Effects for RQ 4

Content of the tables in Appendix M

Table M1. Results from Main Effects of Gender for Scales with Prominent Floor Effects

Scales	Models							NCP			Chi-square test	
			Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
Goiter symptoms scale	NB models		Gender	-0.65	0.14	-4.65	< 0.001**	-0.92	-0.38	1	20.38	< 0.001**
	ZINB models	Count model coefficients	Gender	-0.48	0.12	-3.85	< 0.001**	-0.72	-0.23	1	23.65	< 0.001**
		Zero-Inflation model coefficients		1.47	0.84	1.75	0.08	-0.18	3.12			
			Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
Impaired social life scale	NB models		Gender	-0.33	0.21	-1.59	0.11	-0.74	0.07	1	2.55	0.11
	ZINB models	Count model coefficients	Gender	-0.27	0.18	-1.44	0.15	-0.63	0.10	1	3.06	0.08
		Zero-Inflation model coefficients		0.25	0.52	0.48	0.63	-0.77	1.27			
			Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
Impaired daily life scale	NB models		Gender	-0.69	0.20	-3.42	< 0.001**	-1.09	-0.30	1	11.59	< 0.001**
	ZINB models	Count model coefficients	Gender	-0.52	0.22	-2.35	0.02*	-0.96	-0.09	1	13.16	< 0.001**
		Zero-Inflation model coefficients		7.75	61.83	0.125	0.90	-113.44	128.94			
			Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
Cosmetic complaints scale	NB models		Gender	-0.30	0.15	-1.95	0.05	-0.61	0.00	1	3.80	0.05
	ZINB models	Count model coefficients	Gender	-0.30	0.17	-1.73	0.08	-0.63	0.04	1	3.81	0.05
		Zero-Inflation model coefficients		4.99	33.12	0.14	0.89	-63.84	73.82			

Table M2. Results from Main Effects of Mode of Administration for Scales with Prominent Floor Effects

Scales	Models							NCP		Chi-square test		
			Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
Goiter symptoms scale	NB models		Mode	0.05	0.14	0.31	0.76	-0.24	0.33	1	0.10	0.76
	ZINB models	Count model coefficients	Mode	0.23	0.3	1.78	0.07	-0.02	0.49	1	6.60	0.01*
		Zero-Inflation model coefficients		1.45	0.93	1.55	0.12	-0.38	3.28			
			Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
Impaired social life scale	NB models		Mode	0.53	0.19	2.84	0.005**	0.17	0.90	1	16.97	< 0.001**
	ZINB models	Count model coefficients	Mode	0.52	0.18	2.86	0.004**	0.17	0.90	1	16.97	< 0.001**
		Zero-Inflation model coefficients		-0.86	0.49	-1.75	0.08	-0.86	-1.83			
			Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
Impaired daily life scale	NB models		Mode	0.81	0.20	4.16	< 0.001**	0.43	1.20	1	16.24	< 0.001**
	ZINB models	Count model coefficients	Mode	0.95	0.23	4.09	< 0.001**	0.50	1.41	1	18.33	< 0.001**
		Zero-Inflation model coefficients		1.81	6.00	0.30	0.76	-9.96	13.57			
			Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
Cosmetic complaints scale	NB models		Mode	0.36	0.15	2.42	0.02*	0.07	0.65	1	5.73	0.02*
	ZINB models	Count model coefficients	Mode	0.37	0.16	2.32	0.02*	0.06	0.68	1	5.78	0.02*
		Zero-Inflation model coefficients		5.24	37.97	0.14	0.89	-69.17	79.65			

Note: NCP = non-centrality parameter, SE = Standard Error, LL = Lower limit of 95% CI, UL = Upper limit of 95% CI.
*p < 0.05, **p < 0.01

Table M3. Results from Overall Effects of All Six Predictors for Goiter Symptoms Scores

Models							NCP		Chi-square test	
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
NB models	Gender	-0.40	0.14	-2.89	0.004**	-0.67	-0.14	6	54.71	< 0.001**
	Mode	0.19	0.13	1.48	0.14	-0.06	0.44			
	Age	-0.02	0.005	-4.15	< 0.001**	-0.03	-0.01			
	Education	-0.24	0.06	-4.26	< 0.001**	-0.35	-0.13			
	Treatment	0.03	0.04	0.91	0.37	-0.04	0.10			
	Control	-0.13	0.05	-2.65	0.008**	-0.24	-0.03			
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
ZINB models — Count model coefficients	Gender	-0.37	0.12	-3.12	0.002**	-0.61	-0.14	6	84.39	< 0.001**
	Mode	0.31	0.12	2.58	0.0099**	0.08	0.55			
	Age	-0.01	0.005	-2.78	0.005**	-0.02	0.00			
	Education	-0.10	0.05	-1.99	0.046*	-0.20	0.00			
	Treatment	-0.002	0.03	-0.062	0.95	-0.07	0.06			
	Control	-0.09	0.05	-1.78	0.08	-0.18	0.01			
	Predictor	Estimate	SE	z	p	LL	UL			
Zero-Inflation model coefficients	Gender	1.34	1.12	1.19	0.23	-0.86	3.54			
	Mode	1.56	0.75	2.08	0.004**	0.09	3.03			
	Age	0.10	0.04	2.65	0.008**	0.03	0.17			
	Education	1.70	0.70	2.43	0.02*	0.33	0.17			
	Treatment	-0.15	0.15	-0.95	0.34	-0.44	0.15			
	Control	0.54	0.28	1.97	0.049*	0.00	1.08			

Note: Mode = Administration Mode; Treatment = Duration of treatment; Education = Education level; Control = Level of disease control. NCP = non-centrality parameter; SE = Standard Error, LL = Lower limit of 95% CI, UL = Upper limit of 95% CI.
Gender: 0=male, 1=female; Mode: 0=interview, 1=self-administered
*p < 0.05, **p < 0.01

Table M4. Results from Overall Effects of All Six Predictors for Impaired Social Life Scores

Models							NCP		Chi-square test	
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
NB models	Gender	-0.25	0.20	-1.26	0.21	-0.65	0.13	6	37.50	< 0.001**
	Mode	0.83	0.19	4.49	< 0.001**	0.47	1.19			
	Age	-0.02	0.01	-2.28	0.02*	-0.03	0.00			
	Education	-0.19	0.08	-2.34	0.02*	-0.34	-0.03			
	Treatment	0.08	0.05	1.59	0.11	-0.02	0.17			
	Control	-0.22	0.07	-3.18	0.002**	-0.37	-0.08			
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
ZINB models — Count model coefficients	Gender	-0.34	0.18	-1.93	0.05	-0.68	0.00	6	43.74	< 0.001**
	Mode	0.65	0.18	3.70	0.001**	0.31	1.00			
	Age	-0.01	0.01	-2.02	0.04*	-0.03	0.00			
	Education	-0.13	0.07	-1.94	0.05	-0.27	0.00			
	Treatment	0.10	0.04	2.32	0.02*	0.02	0.18			
	Control	-0.05	0.06	-0.83	0.40	-0.17	0.07			
	Predictor	Estimate	SE	z	p	LL	UL			
Zero-Inflation model coefficients	Gender	-0.30	0.61	-0.50	0.62	-1.49	0.89			
	Mode	-0.92	0.54	-1.70	0.09	-1.97	0.14			
	Age	0.02	0.02	0.77	0.44	-0.03	0.06			
	Education	0.24	0.24	0.995	0.32	-0.23	0.71			
	Treatment	0.10	0.14	0.74	0.46	-0.17	0.37			
	Control	1.46	0.69	2.11	0.03*	0.11	2.81			

Note: Mode = Administration Mode; Treatment = Duration of treatment; Education = Education level; Control = Level of disease control. NCP = non-centrality parameter; SE = Standard Error, LL = Lower limit of 95% CI, UL = Upper limit of 95% CI.
Gender: 0=male, 1=female; Mode: 0=interview, 1=self-administered
*p < 0.05, **p < 0.01

Table M5. Results from Overall Effects of All Six Predictors for Impaired Daily Life Scores

Models							NCP		Chi-square test	
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
NB models	Gender	-0.63	0.20	-3.22	0.001**	-1.01	-0.25	6	48.83	< 0.001**
	Mode	1.01	0.19	5.37	< 0.001**	0.65	1.38			
	Age	-0.01	0.007	-1.98	0.049*	-0.03	0.00			
	Education	-0.23	0.08	-2.90	0.004**	-0.39	-0.07			
	Treatment	0.04	0.05	0.87	0.38	-0.06	0.14			
	Control	-0.15	0.07	-2.16	0.03*	-0.30	-0.01			
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
ZINB models – Count model coefficients	Gender	-0.48	0.18	-2.68	0.007**	-0.84	-0.13	6	68.68	< 0.001**
	Mode	1.22	0.18	6.64	< 0.001**	0.86	1.58			
	Age	-0.006	0.007	-0.769	0.44	-0.02	0.01			
	Education	-0.33	0.08	-4.27	< 0.001**	-0.48	-0.18			
	Treatment	0.02	0.05	0.50	0.62	-0.07	0.12			
	Control	-0.19	0.08	-2.39	0.02*	-0.35	-0.03			
	Predictor	Estimate	SE	z	p	LL	UL			
ZINB models – Zero-Inflation model coefficients	Gender	2.63	2.58	1.02	0.31	-2.43	7.70			
	Mode	11.68	115.88	0.10	0.92	-215.44	238.80			
	Age	0.04	0.03	1.04	0.30	-0.03	0.10			
	Education	-0.88	0.70	-1.26	0.21	-2.26	0.49			
	Treatment	-0.09	0.22	-0.43	0.67	-0.53	0.34			
	Control	-0.37	0.61	-0.61	0.54	-1.57	0.82			

Note: Mode = Administration Mode; Treatment = Duration of treatment; Education = Education level; Control = Level of disease control. NCP = non-centrality parameter; SE = Standard Error, LL = Lower limit of 95% CI, UL = Upper limit of 95% CI.
Gender: 0=male, 1=female; Mode: 0=interview, 1=self-administered
*p < 0.05, **p < 0.01

Table M6. Results from Effects of All Six Predictors for Cosmetic Complaints Scores

Models							NCP		Chi-square test	
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
NB models	Gender	-0.10	0.15	-0.68	0.50	0.68	1.20	6	53.13	< 0.001**
	Mode	0.42	0.14	3.10	0.002**	1.17	1.99			
	Age	-0.03	0.01	-5.72	< 0.001**	0.96	0.98			
	Education	-0.12	0.06	-2.03	0.04*	0.78	1.00			
	Treatment	0.09	0.04	2.40	0.02*	1.01	1.17			
	Control	-0.18	0.05	-3.42	0.001**	0.75	0.93			
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
ZINB models — Count model coefficients	Gender	-0.15	0.14	-1.07	0.28	-0.44	0.13	6	70.92	< 0.001**
	Mode	0.36	0.13	2.72	0.007**	0.10	0.62			
	Age	-0.03	0.01	-4.47	< 0.001**	-0.04	-0.02			
	Education	-0.15	0.06	-2.42	0.02*	-0.27	-0.03			
	Treatment	0.11	0.04	3.02	0.002**	0.04	0.19			
	Control	-0.17	0.05	-3.21	0.001**	-0.28	-0.07			
	Predictor	Estimate	SE	z	p	LL	UL			
Zero-Inflation model coefficients	Gender	-55.33	646.43	-0.09	0.93	-1322.30	1211.65			
	Mode	-37.65	396.93	-0.10	0.92	-815.61	740.31			
	Age	2.02	17.65	0.11	0.91	-32.56	36.60			
	Education	-7.29	100.35	-0.07	0.94	-203.96	189.39			
	Treatment	9.28	134.86	0.07	0.95	-255.04	273.60			
	Control	-1.79	247.15	-0.01	0.99	-486.20	482.61			

Note: Mode = Administration Mode; Treatment = Duration of treatment; Education = Education level; Control = Level of disease control. NCP = non-centrality parameter; SE = Standard Error, LL = Lower limit of 95% CI, UL = Upper limit of 95% CI.
Gender: 0=male, 1=female; Mode: 0=interview, 1=self-administered
*p < 0.05, **p < 0.01

Table M7. Results from Main Effects of Gender for Scales without Prominent Floor Effects

Scales	Models							NCP			Chi-square test	
			Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
Hyperthyroid symptoms scale	NB models		Gender	-0.14	0.13	-1.13	0.26	-0.40	0.11	1	1.27	0.26
	ZINB models	Count model coefficients	Gender	-0.10	0.12	-0.84	0.40	-0.33	0.13	1	1.60	0.21
		ZI model coefficients		0.92	0.73	1.26	0.21	-1.25	2.56			
			Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
Hypothyroid symptoms scale	NB models		Gender	0.24	0.12	2.00	0.046*	0.00	0.48	1	3.94	0.047*
	ZINB models	Count model coefficients	Gender	0.26	0.12	2.18	0.03	-112.70	91.32	1	26.05	0.03*
		ZI model coefficients		6.83	52.06	0.13	0.90	-95.21	108.87			
			Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
Eye Symptoms scale	NB models		Gender	0.20	0.14	1.41	0.16	-0.08	0.48	1	1.96	0.16
	ZINB models	Count model coefficients	Gender	0.25	0.16	1.70	0.09	-0.04	0.54	1	2.76	0.10
		ZI model coefficients		7.65	58.43	0.13	0.90	-106.89	122.18			
			Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
Tiredness scale	NB models		Gender	0.001	0.08	0.01	0.99	1.53	1.80	1	< 0.001	0.99
	ZINB models	Count model coefficients	Gender	0.001	0.08	0.01	0.99	-0.16	0.16	1	30.62	> .999
		ZI model coefficients		14.02	25339.03	0.001	> 0.999	-49649.57	49677.62			
			Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
Cognitive complaints scale	NB models		Gender	0.03	0.13	0.23	0.82	-0.22	0.28	1	0.05	0.82
	ZINB models	Count model coefficients	Gender	-0.004	0.13	-0.03	0.98	-0.26	0.25	1	24.14	0.59
		ZI model coefficients		-0.58	1.10	-0.53	0.60	-2.73	1.57			

Scales	Models							NCP		Chi-square test		
			Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
Anxiety scale	NB models		Gender	-0.005	0.13	-0.04	0.97	-0.27	0.26	1	0.001	0.97
	ZINB models	Count model coefficients	Gender	-0.009	0.14	-0.07	0.95	-0.28	0.26	1	40.91	0.93
		ZI model coefficients		-0.12	1.55	-0.08	0.94	-3.16	2.93			
			Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
Depressivity scale	NB models		Gender	-0.11	0.10	-1.09	0.27	-0.30	0.08	1	1.19	0.28
	ZINB models	Count model coefficients	Gender	-0.16	0.10	-1.64	0.10	-0.35	0.03	1	41.89	0.047*
		ZI model coefficients		-2.44	3.66	-0.67	0.51	-9.61	4.73			
			Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
Emotional susceptibility scale	Poisson models		Gender	0.02	0.07	0.31	0.76	-0.12	0.16	1	0.09	0.76
	ZIP models	Count model coefficients	Gender	-0.01	0.07	-0.10	0.92	-0.14	0.13	1	1.62	0.20
		ZI model coefficients		-2.03	2.06	-0.99	0.32	-6.06	2.00			

Note: NCP = non-centrality parameter, SE = Standard Error, LL = Lower limit of 95% CI, UL = Upper limit of 95% CI. Gender: 0=male, 1=female; Mode: 0=interview, 1=self-administered
*p < 0.05, **p < 0.01

Table M8. Results from Main Effects of Mode of Administration for Scales without Prominent Floor Effects

Scales	Models							NCP			Chi-square test	
			Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
Hyperthyroid symptoms scale	NB models		Mode	0.30	0.12	2.54	0.01*	0.07	0.54	1	6.30	0.01*
	ZINB models	Count model coefficients	Mode	0.31	0.11	2.78	0.005**	0.09	0.53	1	7.81	0.005**
		ZI model coefficients		0.10	0.76	0.13	0.90	-1.38	1.58			
			Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
Hypothyroid symptoms scale	NB models		Mode	0.24	0.11	2.22	0.03*	0.03	0.46	1	4.85	0.03*
	ZINB models	Count model coefficients	Mode	0.28	0.11	2.55	0.01*	0.06	0.49	1	6.34	0.01*
		Zero-Inflation model coefficients		8.32	71.71	0.12	0.91	-132.23	148.87			
			Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
Eye symptoms scale	NB models		Mode	0.21	0.13	1.57	0.12	-0.05	0.47	1	2.43	0.12
	ZINB models	Count model coefficients	Mode	0.26	0.14	1.88	0.06	-0.01	0.53	1	3.30	0.07
		ZI model coefficients		7.64	59.78	0.13	0.90	-109.52	124.81			
			Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
Tiredness scale	NB models		Mode	0.27	0.07	3.66	< 0.001**	0.13	0.42	1	13.03	< 0.001**
	ZINB models	Count model coefficients	Mode	0.28	0.07	3.74	< 0.001**	0.13	0.42	1	13.41	< 0.001**
		ZI model coefficients		17.66	10433.78	0.002	0.999	-20432.18	20467.49			
			Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
Cognitive complaints scale	NB models		Mode	0.39	0.12	3.30	< 0.001**	0.16	0.62	1	10.58	0.001**
	ZINB models	Count model coefficients	Mode	0.37	0.12	3.06	0.002**	0.13	0.60	1	11.14	< 0.001**

Scales	Models							NCP		Chi-square test		
		ZI model coefficients		-0.39	1.03	-0.38	0.70	-2.41	1.62			
Anxiety scale	NB models		Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
			Mode	0.56	0.12	4.69	< 0.001**	0.33	0.80	1	20.74	< 0.001**
	ZINB models	Count model coefficients	Mode	0.48	0.13	3.79	< 0.001**	0.23	0.74	1	22.76	< 0.001**
		ZI model coefficients		-2.00	2.44	-0.82	0.41	-6.78	2.77			
Depressivity scale	NB models		Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
			Mode	0.33	0.09	3.69	< 0.001**	0.16	0.51	1	13.17	< 0.001**
	ZINB models	Count model coefficients	Mode	0.30	0.09	3.27	0.001**	0.12	0.48	1	13.61	< 0.001**
		ZI model coefficients		-1.87	2.57	-0.73	0.47	-6.91	3.16			
Emotional susceptibility scale	Poisson models		Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
			Mode	0.24	0.07	3.71	< 0.001**	0.11	0.38	1	14.00	< 0.001**
	ZIP models	Count model coefficients	Mode	0.23	0.07	3.52	< 0.001**	0.10	0.36	1	13.19	< 0.001**
		ZI model coefficients		-0.76	1.64	-0.46	0.64	-3.97	2.45			

Note: NCP = non-centrality parameter, SE = Standard Error, LL = Lower limit of 95% CI, UL = Upper limit of 95% CI.; Gender: 0=male, 1=female; Mode: 0=interview, 1=self-administered
*p < 0.05, **p < 0.01

Table M9. Results from Overall Effects of All Six Predictors for Hyperthyroid Symptoms Scores

Models							NCP		Chi-square test	
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
NB models	Gender	-0.12	0.13	-0.94	0.35	-0.38	0.13	6	23.50	< 0.001**
	Mode	0.31	0.12	2.67	0.008**	0.09	0.54			
	Age	-0.01	0.005	-2.70	0.007**	-0.02	0.00			
	Education	-0.02	0.05	-0.33	0.74	-0.12	0.09			
	Treatment	0.08	0.03	2.60	0.009**	0.02	0.14			
	Control	-0.10	0.05	-2.12	0.03*	-0.19	0.00			
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
ZINB models – Count model coefficients	Gender	-0.17	0.12	-1.37	0.17	-0.40	0.07	6	36.23	< 0.001**
	Mode	0.38	0.11	3.51	< 0.001**	0.17	0.59			
	Age	-0.01	0.004	-2.75	0.006**	-0.02	0.00			
	Education	0.01	0.05	0.16	0.88	-0.09	0.11			
	Treatment	0.09	0.03	3.08	0.002**	0.03	0.15			
	Control	-0.06	0.04	-1.43	0.15	-0.15	0.02			
	Predictor	Estimate	SE	z	p	LL	UL			
ZINB models – Zero-Inflation model coefficients	Gender	-1.94	1.93	-1.01	0.31	-5.72	1.84			
	Mode	11.86	64.51	0.18	0.85	-114.58	138.29			
	Age	-0.02	0.08	-0.26	0.79	-0.17	0.13			
	Education	2.62	1.45	1.81	0.07	-0.22	5.46			
	Treatment	0.94	0.56	1.70	0.09	-0.15	2.03			
	Control	12.25	31.20	0.39	0.69	-48.89	73.40			

Note: Mode = Administration Mode; Treatment = Duration of treatment; Education = Education level; Control = Level of disease control. NCP = non-centrality parameter; SE = Standard Error, LL = Lower limit of 95% CI, UL = Upper limit of 95% CI.
Gender: 0=male, 1=female; Mode: 0=interview, 1=self-administered
*p < 0.05, **p < 0.01

Table M10. Results from Overall Effects of All Six Predictors for Hypothyroid Symptoms Scores

Models						NCP		Chi-square test		
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
NB models	Gender	0.25	0.13	1.95	0.05	0.00	0.50	6	12.99	0.04*
	Mode	0.21	0.11	1.91	0.06	0.00	0.43			
	Age	0.001	0.004	0.21	0.83	-0.01	0.01			
	Education	-0.06	0.05	-1.24	0.22	-0.16	0.04			
	Treatment	-0.002	0.03	-0.07	0.94	-0.06	0.05			
	Control	-0.09	0.04	-1.99	0.046*	-0.18	0.00			
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
ZINB models	Count model coefficients									
	Gender	0.26	0.12	2.12	0.03*	0.02	0.50	6	26.05	< 0.001**
	Mode	0.25	0.11	2.34	0.02*	0.04	0.45			
	Age	0.003	0.004	0.62	0.53	-0.01	0.01			
	Education	-0.06	0.05	-1.33	0.18	-0.16	0.03			
	Treatment	0.002	0.03	0.07	0.95	-0.05	0.06			
	Control	-0.07	0.04	-1.65	0.10	-0.16	0.01			
	Predictor	Estimate	SE	z	p	LL	UL			
	Zero-Inflation model coefficients									
	Gender	11.76	44.20	0.27	0.79	-74.88	98.40			
	Mode	8.28	75.91	0.11	0.91	-140.50	157.06			
	Age	0.28	0.21	1.37	0.17	-0.12	0.69			
	Education	2.00	1.98	1.01	0.31	-1.88	5.87			
	Treatment	0.22	0.38	0.58	0.56	-0.53	0.97			
	Control	7.97	20.83	0.38	0.70	-32.87	48.82			

Note: Mode = Administration Mode; Treatment = Duration of treatment; Education = Education level; Control = Level of disease control. NCP = non-centrality parameter; SE = Standard Error, LL = Lower limit of 95% CI, UL = Upper limit of 95% CI.
Gender: 0=male, 1=female; Mode: 0=interview, 1=self-administered
*p < 0.05, **p < 0.01

Table M11. Results from Overall Effects of All Six Predictors for Eye Symptoms Scores

Models							NCP		Chi-square test		
		Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
NB models		Gender	0.14	0.15	0.94	0.35	-0.15	0.43	6	11.50	0.07
		Mode	0.16	0.13	1.24	0.22	-0.09	0.42			
		Age	0.003	0.005	0.57	0.57	-0.01	0.01			
		Education	-0.003	0.06	-0.05	0.57	-0.12	0.11			
		Treatment	0.06	0.03	1.70	0.09	-0.01	0.12			
		Control	-0.11	0.05	-2.18	0.03	-0.22	-0.01			
		Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
ZINB models	Count model coefficients	Mode	0.17	NA	NA	NA	NA	NA	6	24.29	< 0.001**
		Gender	0.19	NA	NA	NA	NA	NA			
		Age	0.004	NA	NA	NA	NA	NA			
		Education	0.01	NA	NA	NA	NA	NA			
		Treatment	0.05	NA	NA	NA	NA	NA			
		Control	1.08	NA	NA	NA	NA	NA			
		Predictor	Estimate	SE	z	p	LL	UL			
	Zero-Inflation model coefficients	Gender	30.66	NA	NA	NA	NA	NA			
		Mode	38.01	NA	NA	NA	NA	NA			
		Age	0.31	NA	NA	NA	NA	NA			
		Education	12.26	NA	NA	NA	NA	NA			
		Treatment	-6.44	NA	NA	NA	NA	NA			
		Control	-24.86	NA	NA	NA	NA	NA			

Note: Mode = Administration Mode; Treatment = Duration of treatment; Education = Education level; Control = Level of disease control. NCP = non-centrality parameter; SE = Standard Error, LL = Lower limit of 95% CI, UL = Upper limit of 95% CI.
Gender: 0=male, 1=female; Mode: 0=interview, 1=self-administered
*p < 0.05, **p < 0.01

Table M12. Results from Overall Effects of All Six Predictors for Tiredness Scores

Models							NCP	Chi-square test		
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
NB models	Gender	-0.05	0.08	-0.58	0.56	-0.21	0.11	6	25.27	< 0.001**
	Mode	0.25	0.07	3.40	< 0.001**	0.11	0.39			
	Age	-0.001	0.003	-0.29	0.77	-0.01	0.00			
	Education	0.01	0.03	0.27	0.79	-0.06	0.07			
	Treatment	0.02	0.02	1.20	0.23	-0.01	0.06			
	Control	-0.09	0.03	-3.38	< 0.001**	-0.15	-0.04			
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
ZINB models – Count model coefficients	Gender	-0.05	NA	NA	NA	NA	NA	6	30.62	< 0.001**
	Mode	0.03	NA	NA	NA	NA	NA			
	Age	< 0.001	NA	NA	NA	NA	NA			
	Education	0.01	NA	NA	NA	NA	NA			
	Treatment	<0.001	NA	NA	NA	NA	NA			
	Control	-0.09	NA	NA	NA	NA	NA			
	Predictor	Estimate	SE	z	p	LL	UL			
Zero-Inflation model coefficients	Gender	18.05	NA	NA	NA	NA	NA			
	Mode	19.25	NA	NA	NA	NA	NA			
	Age	0.33	NA	NA	NA	NA	NA			
	Education	-1.30	NA	NA	NA	NA	NA			
	Treatment	0.49	NA	NA	NA	NA	NA			
	Control	23.48	NA	NA	NA	NA	NA			

Note: Mode = Administration Mode; Treatment = Duration of treatment; Education = Education level; Control = Level of disease control. NCP = non-centrality parameter; SE = Standard Error, LL = Lower limit of 95% CI, UL = Upper limit of 95% CI.
Gender: 0=male, 1=female; Mode: 0=interview, 1=self-administered
*p < 0.05, **p < 0.01

Table M13. Results from Overall Effects of All Six Predictors for Cognitive Complaints Scores

Models						NCP		Chi-square test		
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
NB models	Gender	-0.02	0.13	-0.15	0.88	-0.28	0.24	6	16.88	0.097**
	Mode	0.36	0.12	3.04	0.002**	0.13	0.59			
	Age	-0.001	0.005	-0.27	0.79	-0.01	0.01			
	Education	-0.007	0.05	-0.14	0.89	-0.11	0.10			
	Treatment	0.05	0.03	1.52	0.13	-0.01	0.11			
	Control	-0.10	0.05	-2.23	0.03*	-0.20	0.01			
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
ZINB models – Count model coefficients	Gender	0.02	0.13	0.12	0.90	-0.23	0.26	6	24.14	< 0.001**
	Mode	0.36	0.12	2.92	0.004**	0.12	0.61			
	Age	-0.01	0.004	-1.17	0.24	-0.01	0.00			
	Education	-0.01	0.05	-0.19	0.85	-0.11	0.09			
	Treatment	0.05	0.03	1.58	0.11	-0.01	0.10			
	Control	-0.07	0.05	-1.50	0.13	-0.16	0.02			
	Predictor	Estimate	SE	z	p	LL	UL			
ZINB models – Zero-Inflation model coefficients	Gender	0.59	1.24	0.48	0.63	-1.84	3.02			
	Mode	-0.20	1.50	-0.13	0.90	-3.13	2.74			
	Age	-0.15	0.11	-1.41	0.16	-0.37	0.06			
	Education	0.21	0.57	0.37	0.71	-0.91	1.34			
	Treatment	0.01	0.43	0.03	0.98	-0.83	0.86			
	Control	1.19	1.45	0.82	0.41	-1.65	4.04			

Note: Mode = Administration Mode; Treatment = Duration of treatment; Education = Education level; Control = Level of disease control. NCP = non-centrality parameter; SE = Standard Error, LL = Lower limit of 95% CI, UL = Upper limit of 95% CI.
Gender: 0=male, 1=female; Mode: 0=interview, 1=self-administered
*p < 0.05, **p < 0.01

Table M14. Results from Overall Effects of All Six Predictors for Anxiety Scores

Models							NCP		Chi-square test	
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
NB models	Gender	-0.04	0.13	-0.30	0.77	-0.30	0.22	6	32.63	< 0.001**
	Mode	0.56	0.12	4.68	< 0.001**	0.33	0.79			
	Age	-0.01	0.005	-2.02	0.04*	-0.02	0.00			
	Education	0.02	0.05	0.36	0.72	-0.08	0.12			
	Treatment	0.03	0.03	0.91	0.36	-0.03	0.09			
	Control	-0.12	0.05	-2.62	0.01*	-0.21	-0.03			
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
ZINB models – Count model coefficients	Gender	0.002	0.14	0.01	0.99	-0.26	0.27	6	40.91	< 0.001**
	Mode	0.48	0.12	3.96	< 0.001**	0.24	0.72			
	Age	-0.01	0.01	-1.43	0.15	-0.02	0.00			
	Education	-0.02	0.05	-0.35	0.73	-0.12	0.09			
	Treatment	0.01	0.03	0.47	0.64	-0.04	0.07			
	Control	-0.10	0.05	-1.85	0.06	-0.20	0.01			
	Predictor	Estimate	SE	z	p	LL	UL			
ZINB models – Zero-Inflation model coefficients	Gender	0.89	1.94	0.46	0.65	-2.92	4.70			
	Mode	-1.95	2.34	-0.83	0.40	-6.54	2.63			
	Age	0.03	0.07	0.43	0.66	-0.11	0.17			
	Education	-0.73	0.85	-0.87	0.39	-2.39	0.92			
	Treatment	-0.47	0.41	-1.15	0.25	-1.26	0.33			
	Control	1.02	0.66	1.55	0.12	-0.27	2.31			

Note: Mode = Administration Mode; Treatment = Duration of treatment; Education = Education level; Control = Level of disease control. NCP = non-centrality parameter; SE = Standard Error, LL = Lower limit of 95% CI, UL = Upper limit of 95% CI.
Gender: 0=male, 1=female; Mode: 0=interview, 1=self-administered
*p < 0.05, **p < 0.01

Table M15. Results from Overall Effects of All Six Predictors for Depressivity Scores

Models							NCP		Chi-square test	
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
NB models	Gender	-0.06	0.10	-0.63	0.53	-0.25	0.13	6	30.69	< 0.001**
	Mode	0.35	0.09	4.03	< 0.001**	0.18	0.53			
	Age	-0.01	0.003	-2.67	0.008**	-0.02	0.00			
	Education	-0.05	0.04	-1.38	0.17	-0.13	0.02			
	Treatment	0.005	0.02	0.20	0.85	-0.04	0.05			
	Control	-0.09	0.03	-2.75	0.006**	-0.16	-0.02			
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
ZINB models – Count model coefficients	Gender	-0.07	0.09	-0.79	0.43	-0.26	0.11	6	41.89	< 0.001**
	Mode	0.34	0.08	4.04	< 0.001**	0.108	0.51			
	Age	-0.01	0.004	-1.88	0.06	-0.01	0.00			
	Education	-0.09	0.04	-2.31	0.02*	-0.16	-0.01			
	Treatment	0.005	0.02	0.21	0.83	-0.04	0.05			
	Control	-0.084	0.03	-2.56	0.01*	-0.15	-0.02			
	Predictor	Estimate	SE	z	p	LL	UL			
ZINB models – Zero-Inflation model coefficients	Gender	-1.53	1.52	-1.01	0.31	-4.50	1.44			
	Mode	-0.97	1.30	-0.75	0.46	-3.52	1.58			
	Age	0.09	0.06	1.50	0.13	-0.03	0.20			
	Education	-1.29	0.71	-1.82	0.07	-2.68	0.10			
	Treatment	< 0.001	0.31	0.001	0.999	-0.60	0.60			
	Control	0.99	0.90	1.10	0.27	-0.77	2.75			

Note: Mode = Administration Mode; Treatment = Duration of treatment; Education = Education level; Control = Level of disease control. NCP = non-centrality parameter; SE = Standard Error, LL = Lower limit of 95% CI, UL = Upper limit of 95% CI.
Gender: 0=male, 1=female; Mode: 0=interview, 1=self-administered
*p < 0.05, **p < 0.01

Table M16. Results from Overall Effects of All Six Predictors for Emotional Susceptibility Scores

Models							NCP	Chi-square test		
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
Poisson models	Gender	0.02	0.07	0.29	0.77	-0.13	0.17	6	37.86	< 0.001**
	Mode	0.23	0.07	3.41	< 0.001**	0.10	0.36			
	Age	-0.01	0.003	-3.15	0.002**	-0.01	0.00			
	Education	0.04	0.03	1.28	0.20	-0.02	0.10			
	Treatment	-0.02	0.02	1.01	0.31	-0.01	0.05			
	Control	-0.08	0.03	-3.19	0.001**	-0.13	-0.04			
ZIP models	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
	Gender	0.005	0.07	0.07	0.94	-0.14	0.15	6	44.39	< 0.001**
	Mode	0.22	0.07	3.30	< 0.001**	0.09	0.35			
Count model coefficients	Age	-0.01	0.003	-2.83	0.005**	-0.01	0.00			
	Education	0.02	0.03	0.76	0.44	-0.04	0.08			
	Treatment	0.02	0.02	0.79	0.43	-0.02	0.05			
	Control	-0.07	0.03	-2.88	0.004**	-0.12	-0.02			
	Predictor	Estimate	SE	z	p	LL	UL			
	Gender	-1.88	1.86	-1.01	0.31	-5.54	1.78			
	Mode	-0.92	1.91	-0.48	0.63	-4.67	2.83			
Zero-Inflation model coefficients	Age	0.004	0.006	0.73	0.46	-0.08	0.17			
	Education	-1.45	0.87	-1.65	0.10	-3.16	0.27			
	Treatment	-0.31	0.80	-0.37	0.70	-1.88	1.26			
	Control	1.86	6894	0.003	0.998	-13493.56	13530.71			

Note: Mode = Administration Mode; Treatment = Duration of treatment; Education = Education level; Control = Level of disease control. NCP = non-centrality parameter; SE = Standard Error, LL = Lower limit of 95% CI, UL = Upper limit of 95% CI.
Gender: 0=male, 1=female; Mode: 0=interview, 1=self-administered
*p < 0.05, **p < 0.01

Table M17. Results from Main effects of Gender for Composite Scores

Models							NCP		Chi-square test	
		Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
NB models		Gender	-0.01	0.09	-1.23	0.22	-0.28	0.06	1	1.52	0.22
ZINB models	Count model coefficients	Gender	-0.10	0.09	-1.17	0.24	-0.27	0.07	1	55.00	0.15
	Zero-Inflation model coefficients		16.74	6510.41	0.003	0.998	-12743	12776			

Note: NCP = non-centrality parameter; SE = Standard Error, LL = Lower limit of 95% CI, UL = Upper limit of 95% CI.
Gender: 0=male, 1=female.
*p < 0.05, **p < 0.01.

Table M18. Results from Main effects of Mode of Administration for Composite Scores

Models							NCP		Chi-square test	
		Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
NB models		Mode	0.39	0.08	4.95	< 0.001**	0.23	0.54	1	22.68	< 0.001**
ZINB models	Count model coefficients	Mode	0.40	0.08	5.25	< 0.001**	0.25	0.55	1	26.49	< 0.001**
	Zero-Inflation model coefficients		17.99	8845.55	0.002	0.998	-17319	17355			

Note: Mode = Administration Mode; NCP = non-centrality parameter; SE = Standard Error, LL = Lower limit of 95% CI, UL = Upper limit of 95% CI.
Mode: 0=interview, 1=self-administered.
*p < 0.05, **p < 0.01.

Table M19. Results from Overall Effects of All Six Predictors for Composite Scores

Models							NCP		Chi-square test	
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
NB models	Gender	-0.10	0.09	-1.20	0.23	-0.27	0.06	6	45.38	< 0.001**
	Mode	0.39	0.08	5.16	< 0.001**	0.24	0.53			
	Age	-0.01	0.003	-2.38	0.02*	-0.01	0.00			
	Education	-0.02	0.03	-0.66	0.51	-0.09	0.04			
	Treatment	0.03	0.02	1.49	0.14	-0.01	0.07			
	Control	-0.11	0.03	-3.78	< 0.001**	-0.18	-0.05			
	Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
ZINB models – Count model coefficients	Gender	-0.10	NA	NA	NA	NA	NA	6	55.00	< 0.001**
	Mode	0.40	NA	NA	NA	NA	NA			
	Age	-0.006	NA	NA	NA	NA	NA			
	Education	-0.03	NA	NA	NA	NA	NA			
	Treatment	0.03	NA	NA	NA	NA	NA			
	Control	-0.11	NA	NA	NA	NA	NA			
	Predictor	Estimate	SE	z	p	LL	UL			
ZINB models – Zero-Inflation model coefficients	Gender	18.04	NA	NA	NA	NA	NA			
	Mode	19.25	NA	NA	NA	NA	NA			
	Age	0.33	NA	NA	NA	NA	NA			
	Education	-1.30	NA	NA	NA	NA	NA			
	Treatment	0.50	NA	NA	NA	NA	NA			
	Control	23.48	NA	NA	NA	NA	NA			

Note: Mode = Administration Mode; Treatment = Duration of treatment; Education = Education level; Control = Level of disease control. NCP = non-centrality parameter; SE = Standard Error, LL = Lower limit of 95% CI, UL = Upper limit of 95% CI.
Gender: 0=male, 1=female; Mode: 0=interview, 1=self-administered
*p < 0.05, **p < 0.01

Table M20. Results from Overall Effects of All Six Predictors for the Simplified ZINB Model of Eye Symptoms Scores

Models							NCP		Chi-square test		
		Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
ZINB models	Count model coefficients	Mode	0.20	0.13	1.57	0.12	-0.05	0.45			
		Gender	0.17	0.14	1.18	0.24	-0.11	0.45			
		Age	0.00	0.01	0.85	0.39	-0.01	0.01			
		Education	0.01	0.06	0.11	0.92	-0.11	0.12			
		Treatment	0.05	0.03	1.37	0.17	-0.02	0.11			
		Control	-0.14	0.05	-2.63	0.01*	-0.25	-0.04			
		Predictor	Estimate	SE	z	p	LL	UL	6	18.17	0.006**
	Zero-Inflation model coefficients	Gender	17.10	2273.23	0.01	0.99	-4438.34	4472.55			
		Mode	12.20	645.68	0.02	0.99	-1253.3	1277.71			
		Age	1.17	1.02	1.15	0.25	-0.83	3.17			
		Education	0.07	0.08	0.87	0.39	-0.09	0.24			
		Treatment	-11.88	332.29	-0.04	0.97	-663.16	639.4			

Note: Mode = Administration Mode; Treatment = Duration of treatment; Education = Education level; Control = Level of disease control. NCP = non-centrality parameter; SE = Standard Error, LL = Lower limit of 95% CI, UL = Upper limit of 95% CI.
Gender: 0=male, 1=female; Mode: 0=interview, 1=self-administered
*p < 0.05, **p < 0.01

Table M21. Results from Overall Effects of All Six Predictors for the Simplified ZINB Model of Tiredness Scores

Models							NCP		Chi-square test		
		Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
ZINB models	Count model coefficients	Gender	0.26	0.07	3.55	0.00	0.12	0.4	6	27.30	< 0.001**
		Mode	-0.04	0.08	-0.54	0.59	-0.2	0.11			
		Age	0.00	0.00	-0.05	0.96	-0.01	0.01			
		Education	0.00	0.03	0.10	0.92	-0.06	0.07			
		Treatment	0.02	0.02	1.11	0.27	-0.02	0.06			
		Control	-0.09	0.03	-3.27	< 0.001**	-0.15	-0.04			
		Predictor	Estimate	SE	z	p	LL	UL			
	Zero-Inflation model coefficients	Gender	18.20	13490.00	< 0.001	1.00	-26415.49	26451.88			
		Mode	17.49	15860.00	< 0.001	1.00	-31060.38	31095.37			
		Age	-0.45	1.41	-0.32	0.75	-3.21	2.32			
		Education	0.07	0.11	0.64	0.52	-0.15	0.29			
		Treatment	-0.20	0.65	-0.31	0.76	-1.47	1.07			

Note: Mode = Administration Mode; Treatment = Duration of treatment; Education = Education level; Control = Level of disease control. NCP = non-centrality parameter; SE = Standard Error, LL = Lower limit of 95% CI, UL = Upper limit of 95% CI.
Gender: 0=male, 1=female; Mode: 0=interview, 1=self-administered
*p < 0.05, **p < 0.01

Table M22. Results from Overall Effects of All Six Predictors for the Simplified ZINB Model of Composite Scores

Models							NCP		Chi-square test		
		Predictor	Estimate	SE	z	p	LL	UL	df	χ^2	p
ZINB models	Count model coefficients	Gender	0.40	0.07	5.49	< 0.001**	-0.05	0.45			
		Mode	-0.10	0.08	-1.19	0.23	-0.11	0.45			
		Age	-0.01	0.00	-2.15	0.03*	-0.01	0.01			
		Education	-0.03	0.03	-0.88	0.38	-0.11	0.12			
		Treatment	0.03	0.02	1.40	0.16	-0.02	0.11			
		Control	-0.11	0.03	-3.59	< 0.001**	-0.25	-0.04			
		Predictor	Estimate	SE	z	p	LL	UL	6	51.49	< 0.001**
	Zero-Inflation model coefficients	Gender	18.23	12750.00	< 0.001	1.00	-4438.34	4472.55			
		Mode	17.53	15030.00	< 0.001	1.00	-1253.3	1277.71			
		Age	-0.45	1.32	-0.34	0.73	-0.83	3.17			
		Education	0.07	0.11	0.69	0.49	-0.09	0.24			
		Treatment	-0.20	0.61	-0.33	0.74	-663.16	639.4			

Note: Mode = Administration Mode; Treatment = Duration of treatment; Education = Education level; Control = Level of disease control. NCP = non-centrality parameter; SE = Standard Error, LL = Lower limit of 95% CI, UL = Upper limit of 95% CI.
Gender: 0=male, 1=female; Mode: 0=interview, 1=self-administered
*p < 0.05, **p < 0.01

Table M23. Results of the Proposed Method of T-tests by Liu and Wang for Comparing Gender Difference

				NCP	
Scale	*df*	t	*p*	*LL*	*UP*
Scales with significant floor effects					
Goiter symptoms	138	3.82	< 0.001**	1.03	3.32
Impaired social life	111	1.24	0.22	-0.56	2.38
Impaired daily life	113	2.88	< 0.001**	0.62	3.49
Cosmetic complaints	144	1.93	0.58	-0.05	2.58
Scales without significant floor effects					
Hyperthyroid symptoms	157	0.80	0.43	-0.80	1.87
Hypothyroid symptoms	168	-1.64	0.11	-2.42	0.24
Eye symptoms	153	-1.13	0.26	-1.70	0.48
Tiredness	174	-0.36	0.72	-1.27	0.89
Cognitive complaints	155	-0.44	0.66	-1.50	0.96
Anxiety	155	-0.005	0.97	-1.28	1.27
Depressivity	169	0.84	0.40	-0.66	1.63
Emotional Susceptibility	173	-0.14	0.89	-1.03	0.90
Composite	178	0.93	0.36	-2.99	8.15

Note: The t-values are computed with Male scores minus Female scores.
*p < 0.05, **p < 0.01.

Table M24. Results of the Proposed Method of T-tests by Liu and Wang for Comparing Two Modes of Administration

				NCP	
Scale	*df*	t	*p*	*LL*	*UP*
Scales with significant floor effects					
Goiter symptoms	138	0.65	0.52	-0.67	1.32
Impaired social life	111	-3.58	<0.001**	-3.09	-0.88
Impaired daily life	113	-2.11	0.04*	-2.25	-0.06
Cosmetic complaints	144	-2.07	0.04*	-2.11	-0.04
Scales without significant floor effects					
Hyperthyroid symptoms	157	-2.23	0.03*	-2.22	-0.13
Hypothyroid symptoms	168	-1.95	0.05	-2.17	0.02
Eye symptoms	153	-1.14	0.26	-1.48	0.40
Tiredness	174	-3.54	< 0.001**	-2.29	-0.64
Cognitive complaints	155	-3.01	0.003**	-2.35	-0.48
Anxiety	155	-4.33	< 0.001**	-2.97	-1.10
Depressivity	169	-3.55	< 0.001**	-2.44	-0.69
Emotional Susceptibility	173	-3.51	< 0.001**	-2.10	-0.58
Composite	178	-4.76	< 0.001**	-13.94	-5.71

Note: The t-values are computed with scores in Interview group minus scores in Self-administered group.
*p < 0.05, **p < 0.01

Appendix N. Figures for RQ 1

Content of the figures in Appendix N

Figure N1. Histograms of Distribution for Raw Scores of Scales

Figure N1.1. Goiter Symptoms Scale

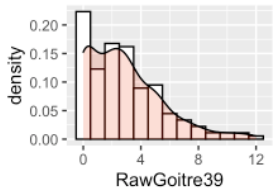


Figure N1.2. Hyperthyroid Symptoms Scale

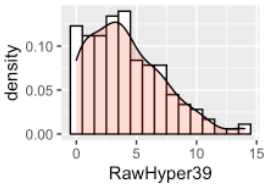


Figure N1.3. Hypothyroid Symptoms Scale

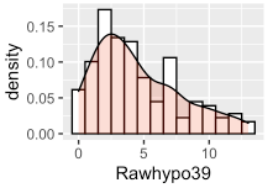


Figure N1.4. Eye Symptoms Scale

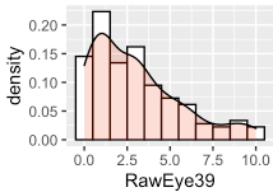


Figure N1.5. Tiredness Scale

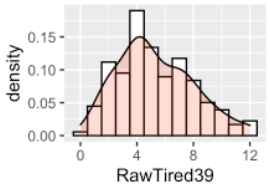


Figure N1.6. Cognitive Complaints Scale

Figure N1.7. Anxiety Scale

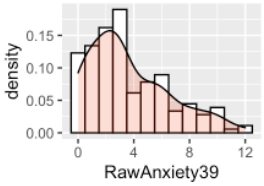


Figure N1.8. Depressivity Scale

Figure N1.9. Impaired Social Life Scale

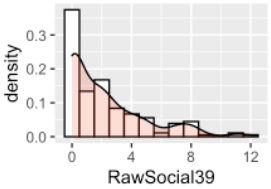


Figure N1.10. Emotional Susceptibility Scale

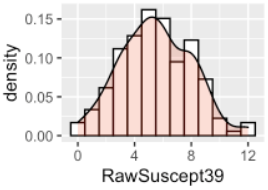


Figure N1.11. Impaired Daily Life Scale

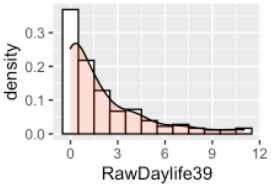


Figure N1.12. Cosmetic Complaints Scale

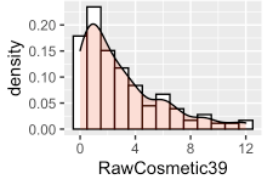


Figure N1.13. Composite Scale

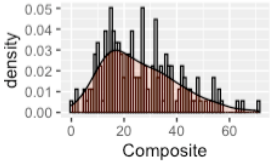


Figure N2. Histograms of Distribution for Raw Scores of Scales by Gender

Figure N3.1. Goiter Symptoms Scale

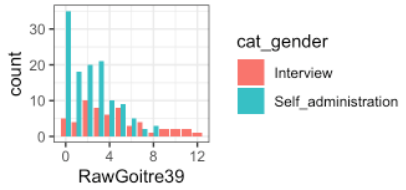


Figure N3.2. Hyperthyroid Symptoms Scale

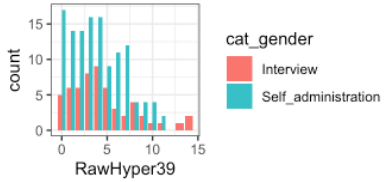


Figure N3.3. Hypothyroid Symptoms Scale

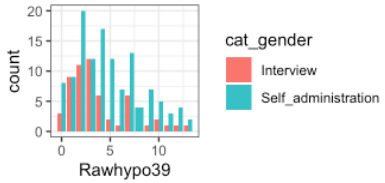


Figure N3.4. Eye Symptoms Scale

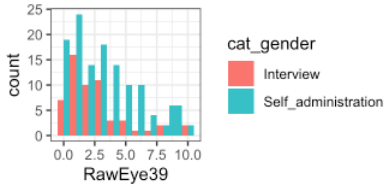


Figure N3.5. Tiredness Scale

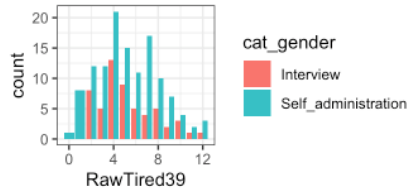


Figure N3.6. Cognitive Complaints Scale

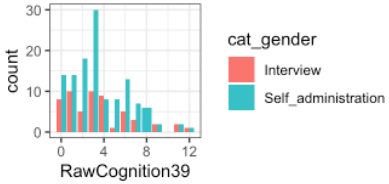


Figure N3.7. Anxiety Scale

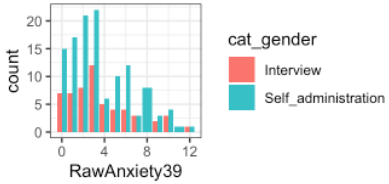


Figure N3.8. Depressivity Scale

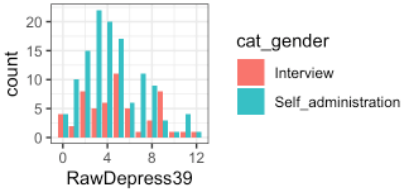


Figure N3.9. Impaired Social Life Scale

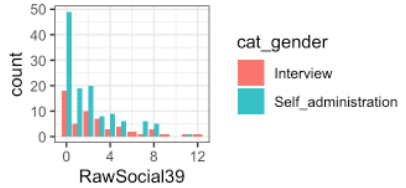


Figure N3.10. Emotional Susceptibility Scale

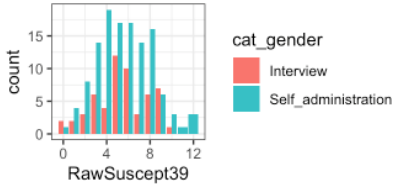


Figure N3.11. Impaired Daily Life Scale

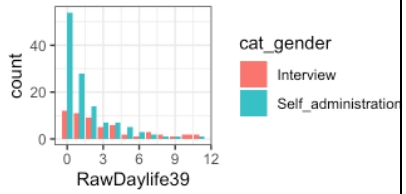


Figure N3.12. Cosmetic Complaints Scale

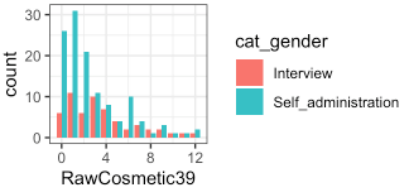


Figure N3.13. Composite Scale

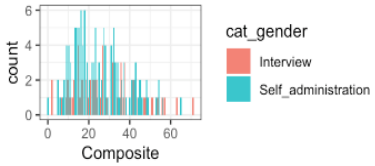


Figure N3. Histograms of Distribution for Raw Scores of Scales by Mode of Administration

Figure N2.1. Goiter Symptoms Scale

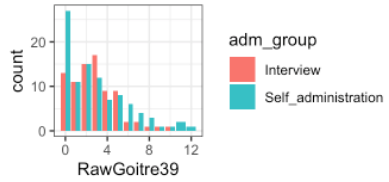


Figure N2.2. Hyperthyroid Symptoms Scale

Figure N2.3. Hypothyroid Symptoms Scale

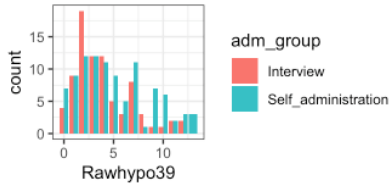


Figure N2.4. Eye Symptoms Scale

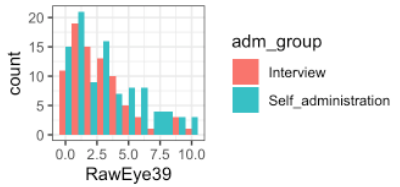


Figure N2.5. Tiredness Scale

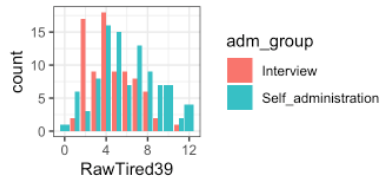


Figure N2.6. Cognitive Complaints Scale

Figure N2.7. Anxiety Scale

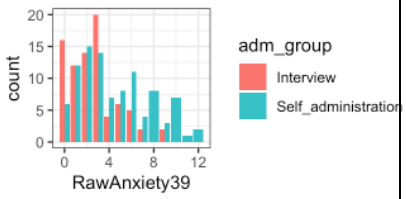


Figure N2.8. Depressivity Scale

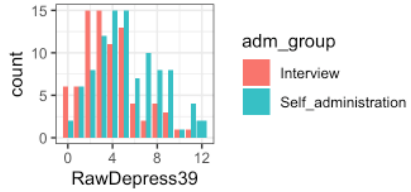


Figure N2.9. Impaired Social Life Scale

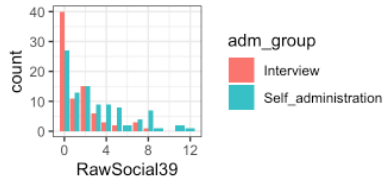


Figure N2.10. Emotional Susceptibility Scale

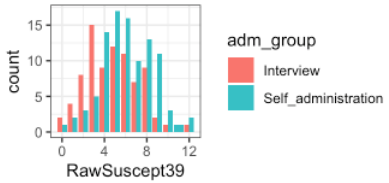


Figure N2.11. Impaired Daily Life Scale

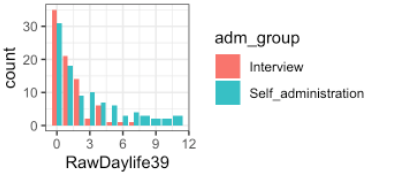


Figure N2.12. Cosmetic Complaints Scale

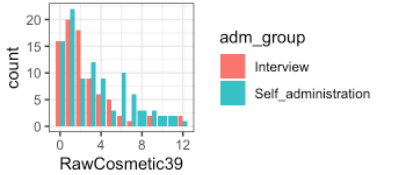
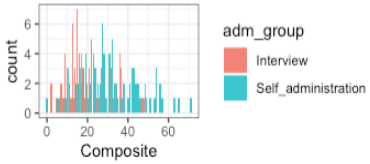


Figure N2.13. Composite Scale



Appendix O. Figures for Assumption Checking for RQ 3 & 4

Content of the tables in Appendix O

Content of the tables in Appendix O

Figure O1. Assumption Checking for One Predictor Model of Gender for Goiter Symptoms Scale

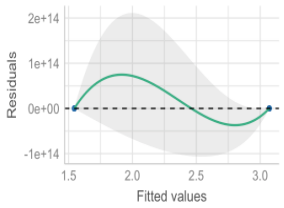
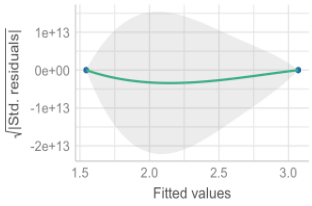
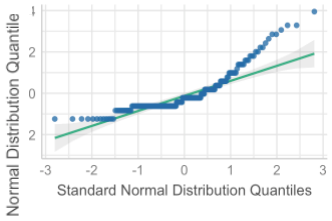
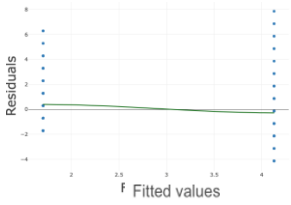
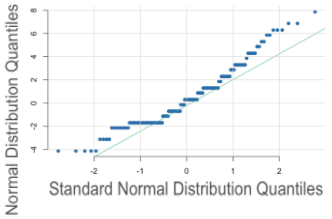
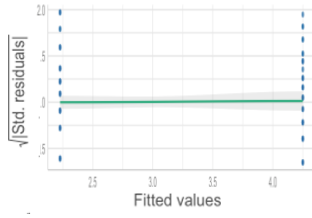
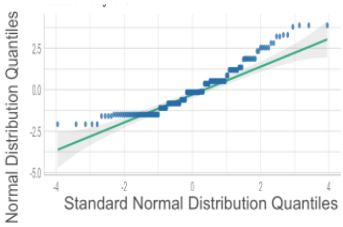
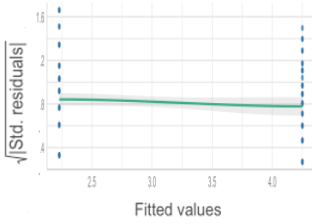
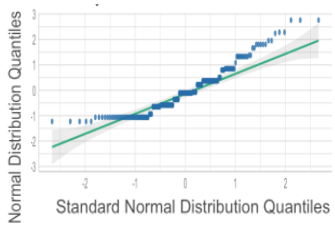
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O1.1** 	**Figure O1.2** 	**Figure O1.3** 
Tobit regression model	**Figure O1.4** 		**Figure O1.5** 
Poisson Regression Model		**Figure O1.6** 	**Figure O1.7** 
NB Regression Model		**Figure O1.8** 	**Figure O1.9** 
ZIP model			
ZINB model			

Figure O2. Assumption Checking for One Predictor Model of Gender for Impaired Social life Scale

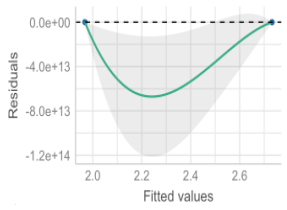
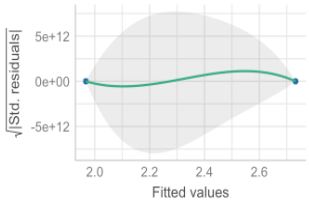
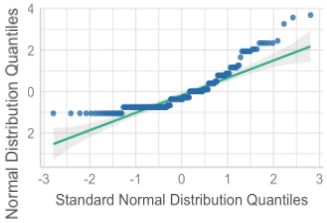
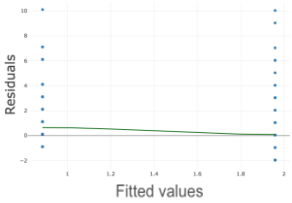
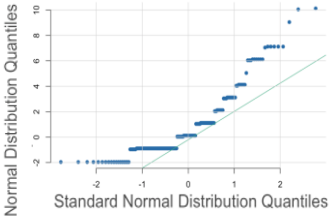
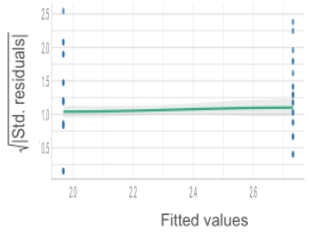
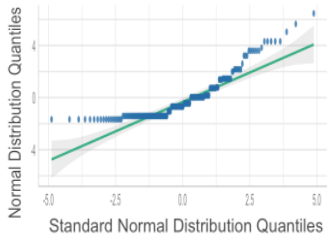
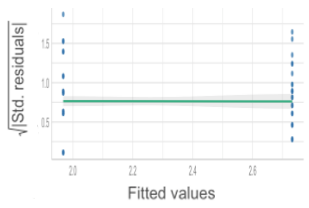
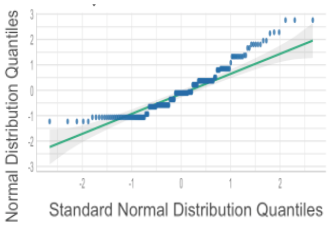
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O2.1** 	**Figure O2.2** 	**Figure O2.3** 
Tobit regression model	**Figure O2.4** 		**Figure O2.5** 
Poisson regression model		**Figure O2.6** 	**Figure O2.7** 
NB Regression Model		**Figure O2.8** 	**Figure O2.9** 
ZIP model			
ZINB model			

Figure O3. Assumption Checking for One Predictor Model of Gender for Impaired Daily Life Scale

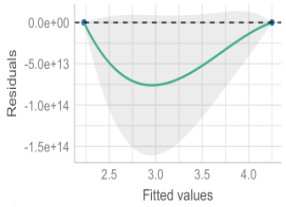
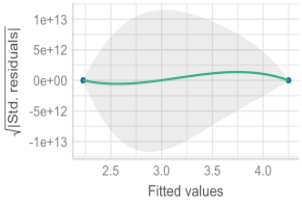
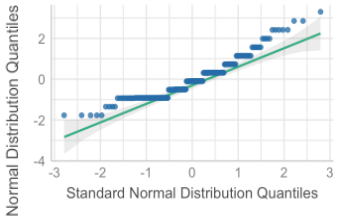
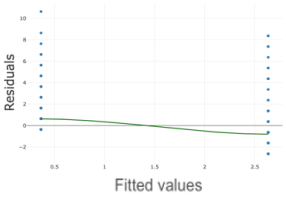
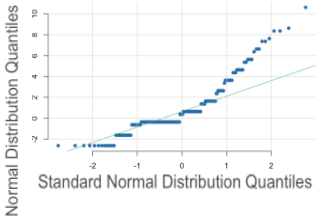
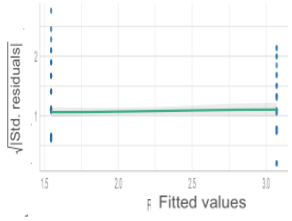
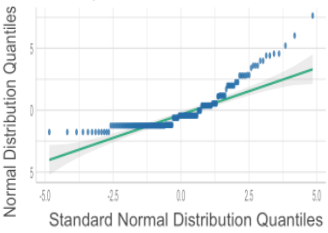
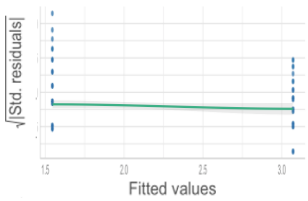
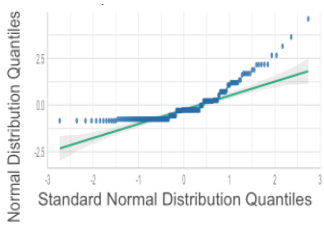
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O3.1** 	**Figure O3.2** 	**Figure O3.3** 
Tobit regression model	**Figure O3.4** 		**Figure O3.5** 
Poisson regression model		**Figure O3.6** 	**Figure O3.7** 
NB Regression Model		**Figure O3.8** 	**Figure O3.9** 
ZIP model			
ZINB model			

Figure O4. Assumption Checking for One Predictor Model of Gender for Cosmetic Complaints Scale

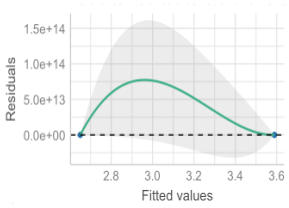
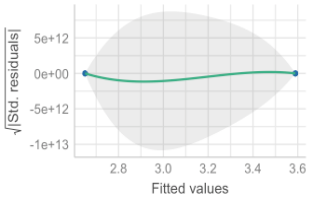
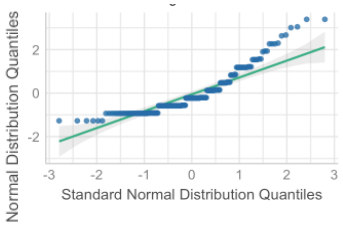
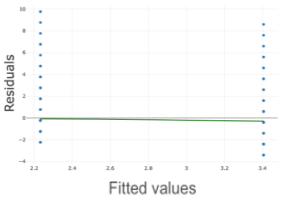
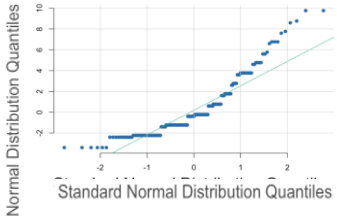
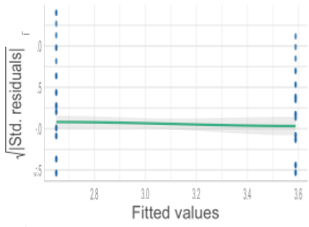
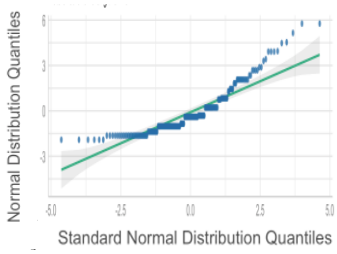
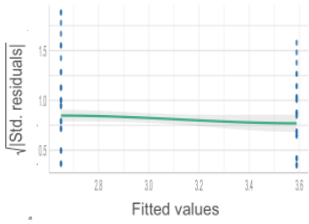
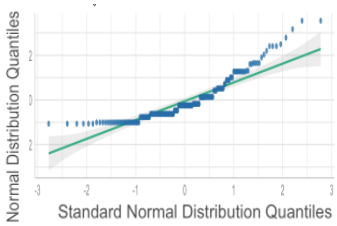
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O4.1** 	**Figure O4.2** 	**Figure O4.3** 
Tobit regression model	**Figure O4.4** 		**Figure O4.5** 
Poisson regression model		**Figure O4.6** 	**Figure O4.7** 
NB Regression Model		**Figure O4.8** 	**Figure O4.9** 
ZIP model			
ZINB model			

Figure O5. Assumption Checking for One Predictor Model of Gender for Hyperthyroid Symptoms Scale

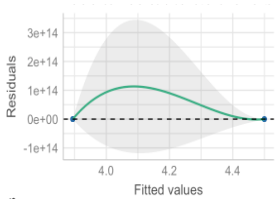
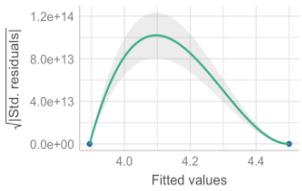
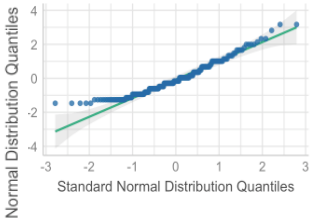
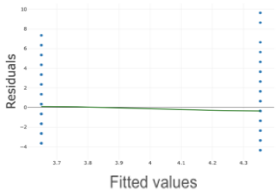
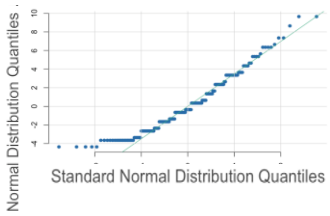
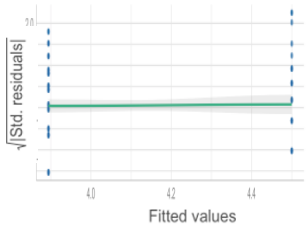
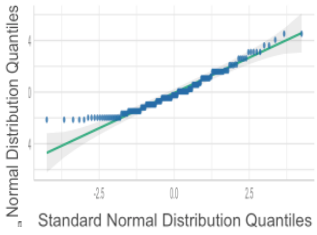
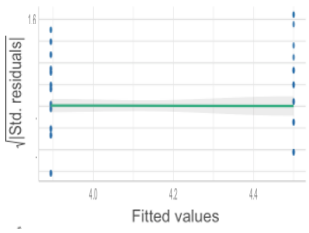
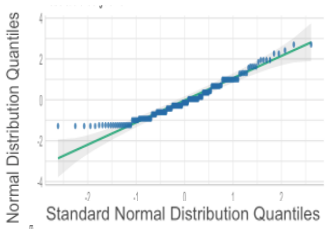
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O5.1** 	**Figure O5.2** 	**Figure O5.3** 
Tobit regression model	**Figure O5.4** 		**Figure O5.5** 
Poisson regression model		**Figure O5.6** 	**Figure O5.7** 
NB Regression Model		**Figure O5.8** 	**Figure O5.9** 
ZIP model			
ZINB model			

Figure O6. Assumption Checking for One Predictor Model of Gender for Hypothyroid Symptoms Scale

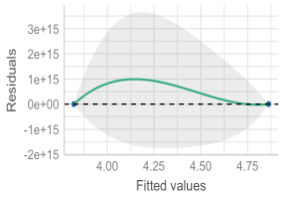
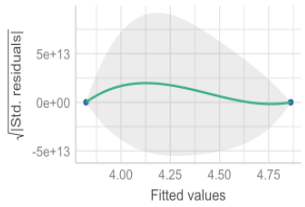
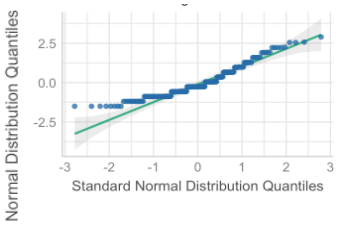
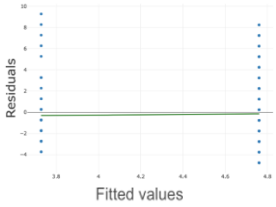
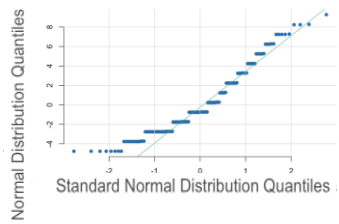
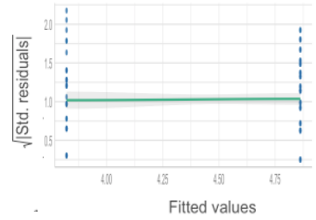
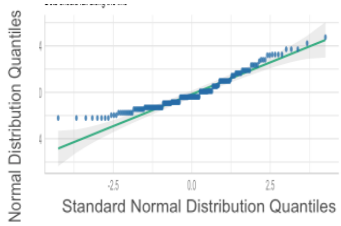
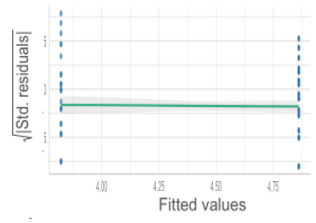
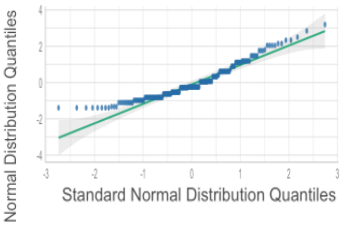
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O6.1** 	**Figure O6.2** 	**Figure O6.3** 
Tobit regression model	**Figure O6.4** 		**Figure O6.5** 
Poisson regression model		**Figure O6.6** 	**Figure O6.7** 
NB Regression Model		**Figure O6.8** 	**Figure O6.9** 
ZIP model			
ZINB model			

Figure O7. Assumption Checking for One Predictor Model of Gender for Eye Symptoms Scale

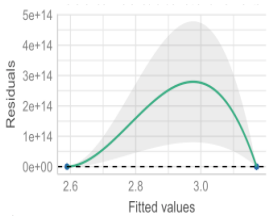
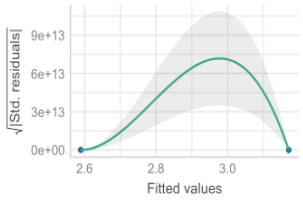
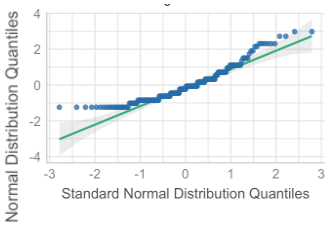
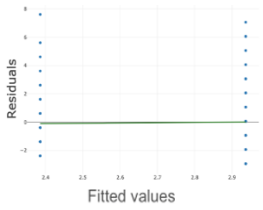
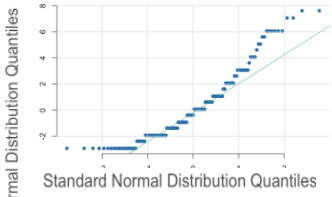
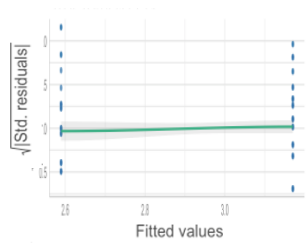
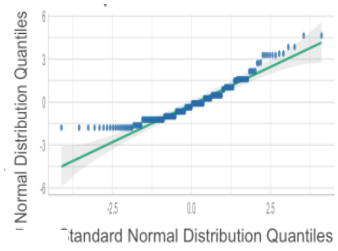
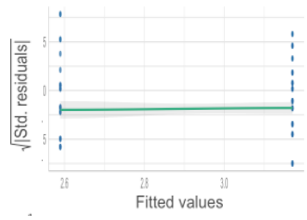
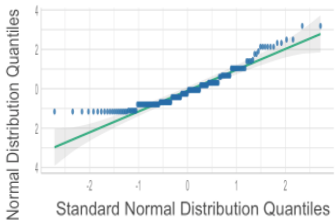
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O7.1** 	**Figure O7.2** 	**Figure O7.3** 
Tobit regression model	**Figure O7.4** 		**Figure O7.5** 
Poisson regression model		**Figure O7.6** 	**Figure O7.7** 
NB Regression Model		**Figure O7.8** 	**Figure O7.9** 
ZIP model			
ZINB model			

Figure O8. Assumption Checking for One Predictor Model of Gender for Tiredness Scale

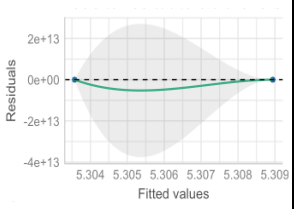
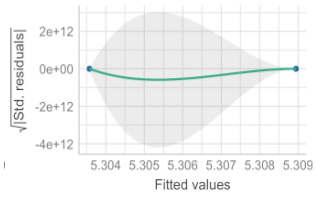
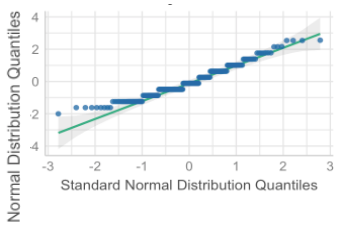
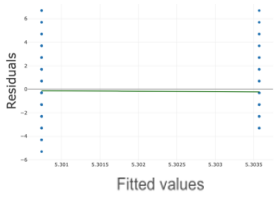
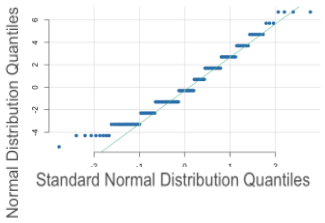
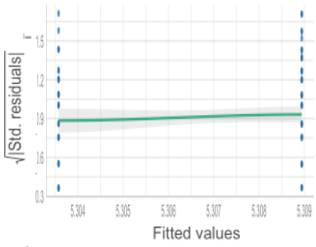
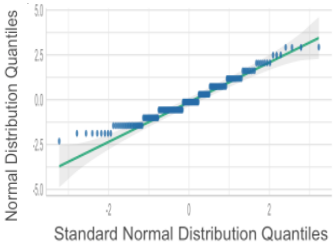
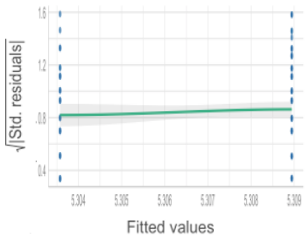
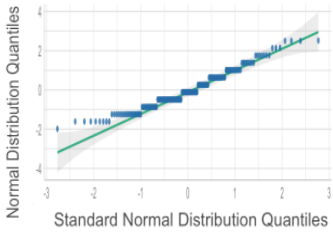
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O8.1** 	**Figure O8.2** 	**Figure O8.3** 
Tobit regression model	**Figure O8.4** 		**Figure O8.5** 
Poisson regression model		**Figure O8.6** 	**Figure O8.7** 
NB Regression Model		**Figure O8.8** 	**Figure O8.9** 
ZIP model			
ZINB model			

Figure O9. Assumption Checking for One Predictor Model of Gender for Cognitive Complaints Scale

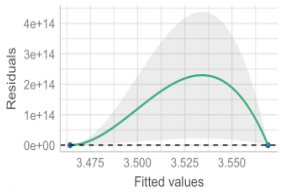
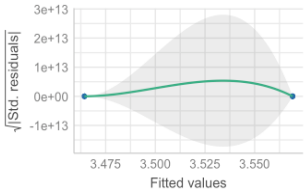
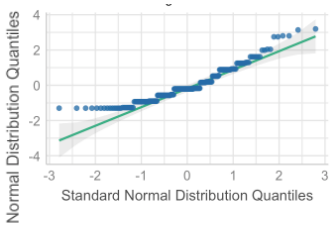
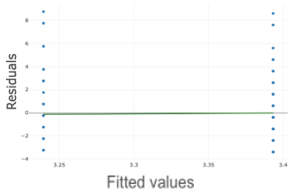
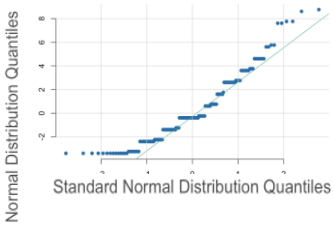
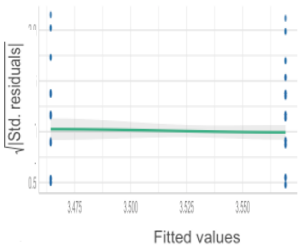
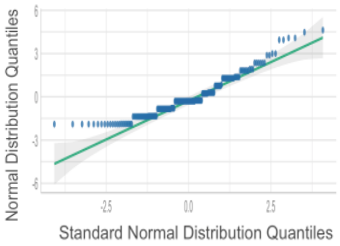
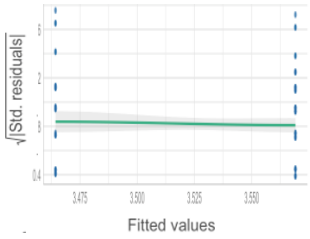
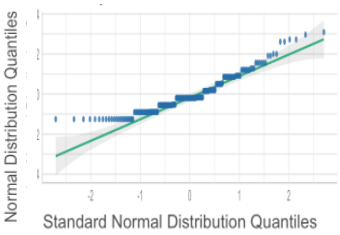
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O9.1** 	**Figure O9.2** 	**Figure O9.3** 
Tobit regression model	**Figure O9.4** 		**Figure O9.5** 
Poisson regression model		**Figure O9.6** 	**Figure O9.7** 
NB Regression Model		**Figure O9.8** 	**Figure O9.9** 
ZIP model			
ZINB model			

Figure O10. Assumption Checking for One Predictor Model of Gender for Anxiety Scale

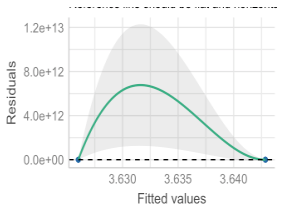
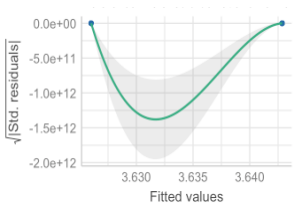
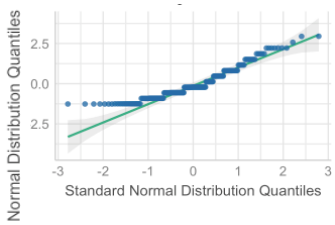
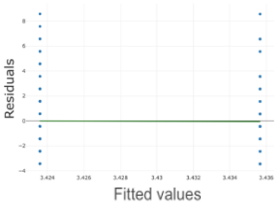
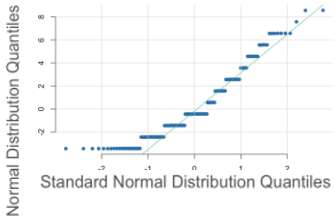
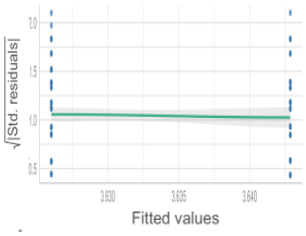
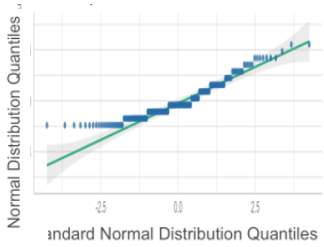
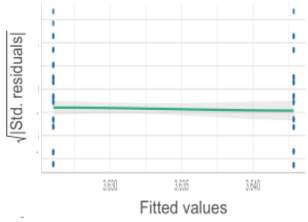
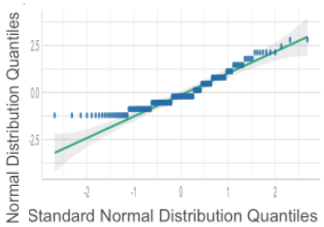
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O10.1** 	**Figure O10.2** 	**Figure O10.3** 
Tobit regression model	**Figure O10.4** 		**Figure O10.5** 
Poisson regression model		**Figure O10.6** 	**Figure O10.7** 
NB Regression Model		**Figure O10.8** 	**Figure O10.9** 
ZIP model			
ZINB model			

Figure O11. Assumption Checking for One Predictor Model of Gender for Depressivity Scale

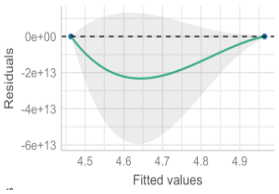
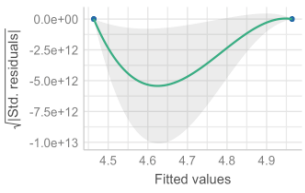
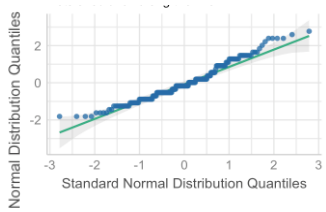
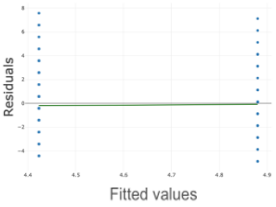
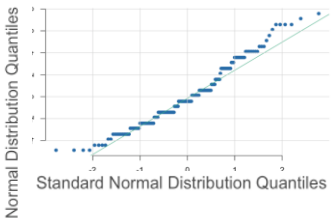
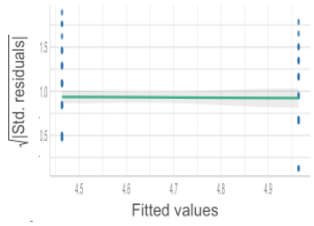
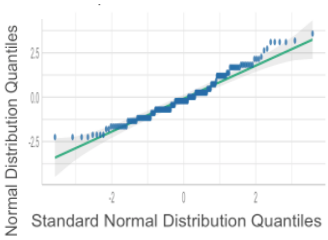
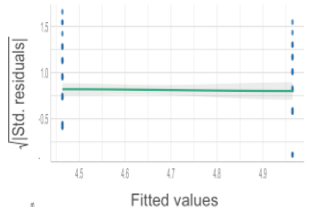
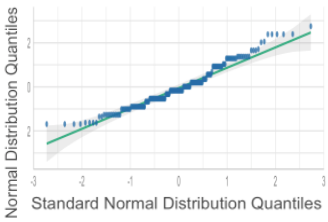
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
ML regression model	**Figure O11.1** 	**Figure O11.2** 	**Figure O11.3** 
Tobit regression model	**Figure O11.4** 		**Figure O11.5** 
Poisson regression model		**Figure O11.6** 	**Figure O11.7** 
NB Regression Model		**Figure O11.8** 	**Figure O11.9** 
ZIP model			
ZINB model			

Figure O12. Assumption Checking for One Predictor Model of Gender for Emotional Susceptibility Scale

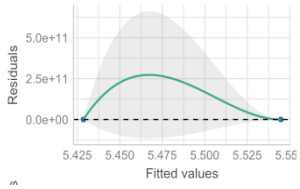
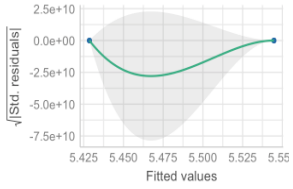
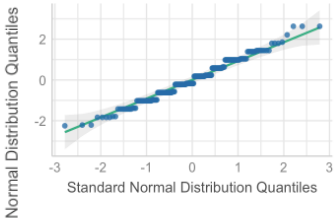
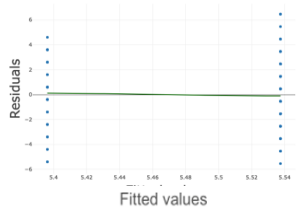
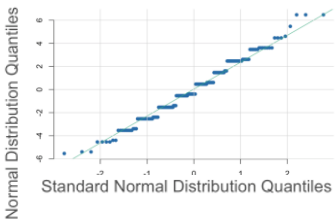
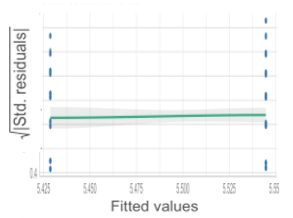
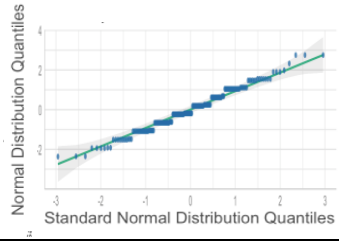
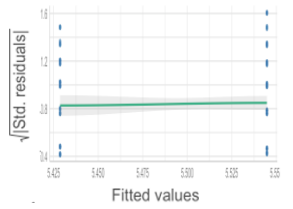
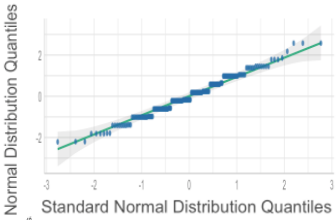
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O12.1** 	**Figure O12.2** 	**Figure O12.3** 
Tobit regression model	**Figure O12.4** 		**Figure O12.5** 
Poisson regression model		**Figure O12.6** 	**Figure O12.7** 
NB Regression Model		**Figure O12.8** 	**Figure O12.9** 
ZIP model			
ZINB model			

Figure O13. Assumption Checking for One Predictor Model of Gender for Composite Scale

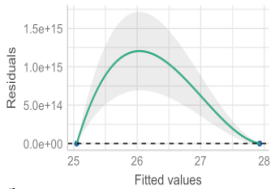
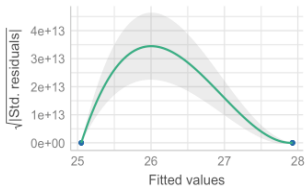
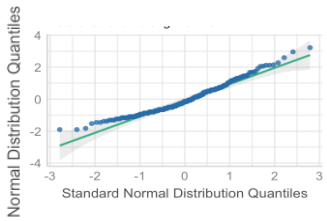
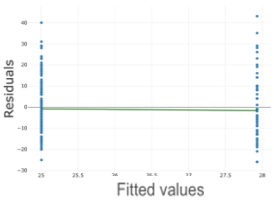
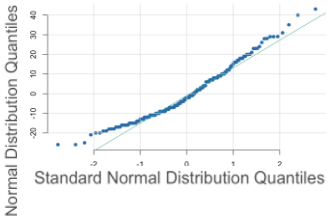
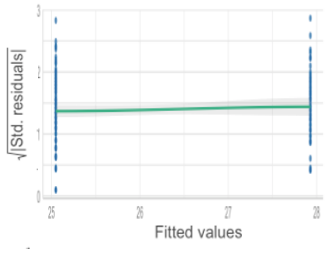
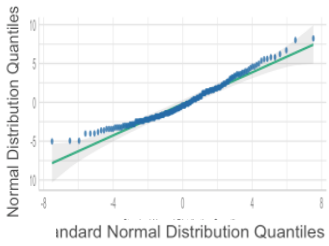
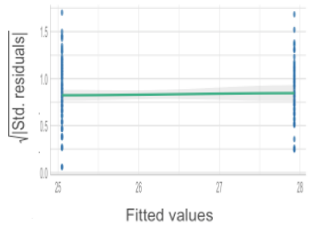
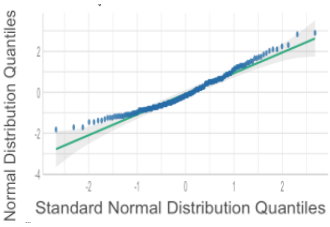
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O13.1** 	**Figure O13.2** 	**Figure O13.3** 
Tobit regression model	**Figure O13.4** 		**Figure O13.5** 
Poisson regression model		**Figure O13.6** 	**Figure O13.7** 
NB Regression Model		**Figure O13.8** 	**Figure O13.9** 
ZIP model			
ZINB model			

Figure O14. Assumption Checking for One Predictor Model of Mode of Administration for Goiter Symptoms Scale

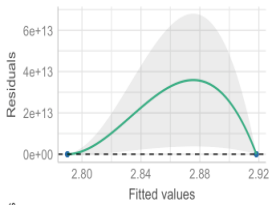
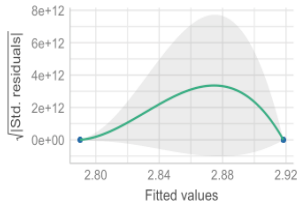
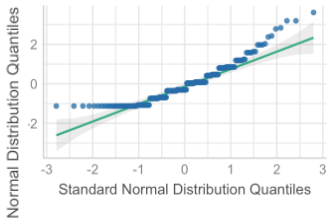
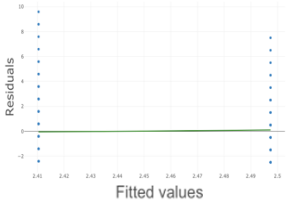
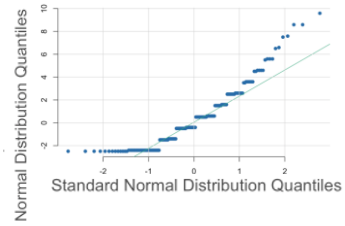
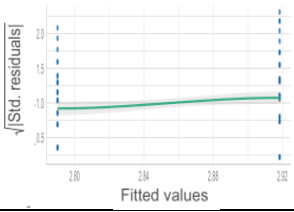
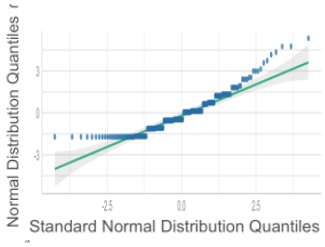
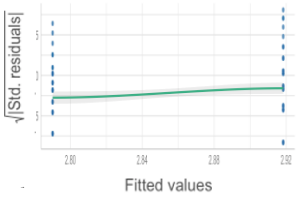
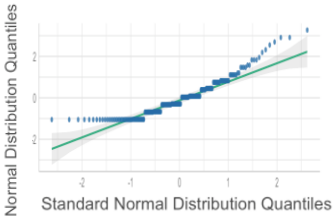
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O14.1** 	**Figure O14.2** 	**Figure O14.3** 
Tobit regression model	**Figure O14.4** 		**Figure O14.5** 
Poisson regression model		**Figure O14.6** 	**Figure O14.7** 
NB Regression Model		**Figure O14.8** 	**Figure O14.9** 
ZIP model			
ZINB model			

Figure O15. Assumption Checking for One Predictor Model of Mode of Administration for Impaired Social Life Scale

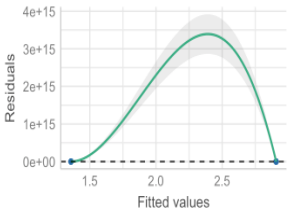
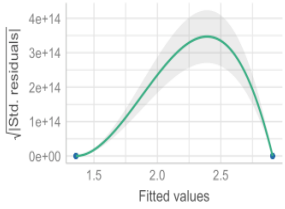
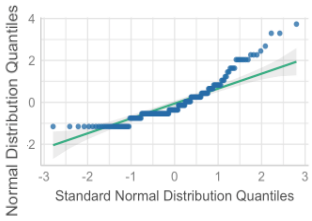
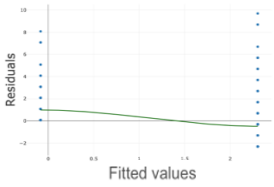
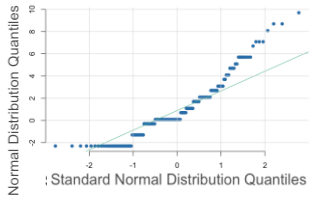
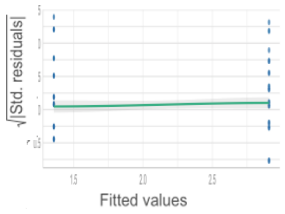
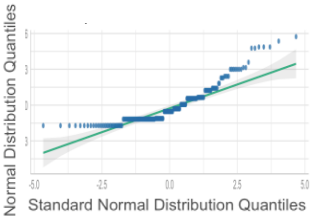
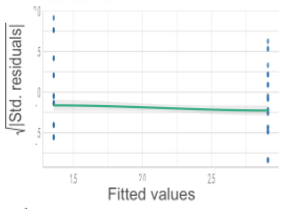
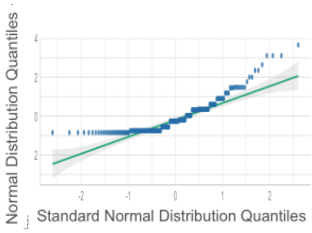
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O15.1** 	**Figure O15.2** 	**Figure O15.3** 
Tobit regression model	**Figure O15.4** 		**Figure O15.5** 
Poisson regression model		**Figure O15.6** 	**Figure O15.7** 
NB Regression Model		**Figure O15.8** 	**Figure O15.9** 
ZIP model			
ZINB model			

Figure O16. Assumption Checking for One Predictor Model of Mode of Administration for Impaired Daily Life Scale

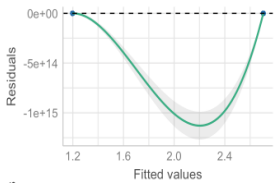
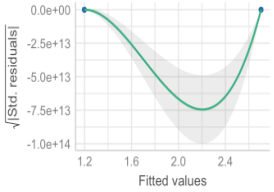
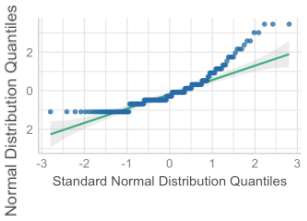
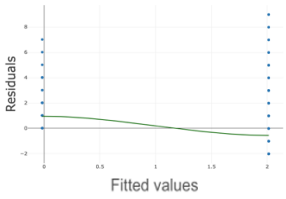
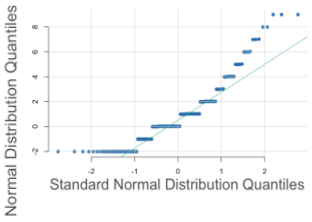
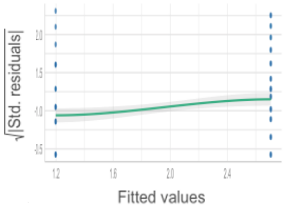
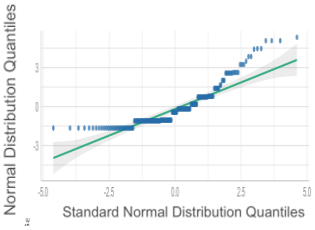
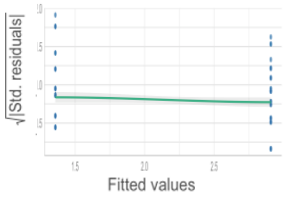
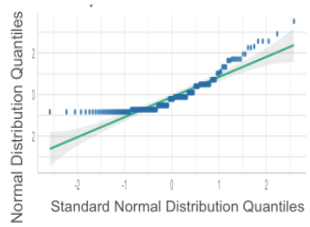
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O16.1** 	**Figure O16.2** 	**Figure O16.3** 
Tobit regression model	**Figure O16.4** 		**Figure O16.5** 
Poisson regression model		**Figure O16.6** 	**Figure O16.7** 
NB Regression Model		**Figure O16.8** 	**Figure O16.9** 
ZIP model			
ZINB model			

Figure O17. Assumption Checking for One Predictor Model of Mode of Administration for Cosmetic Complaints Scale

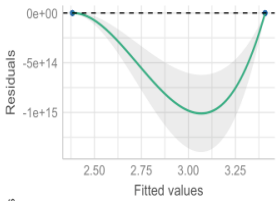
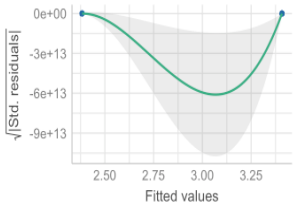
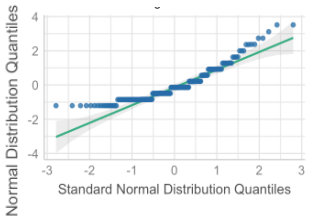
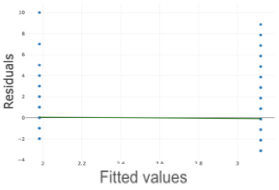
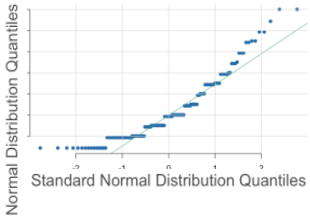
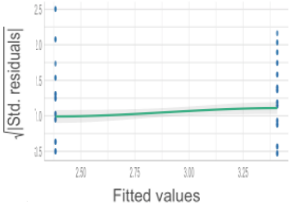
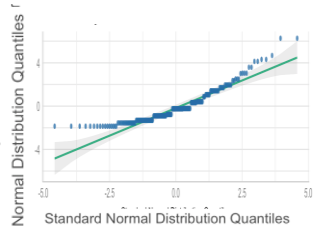
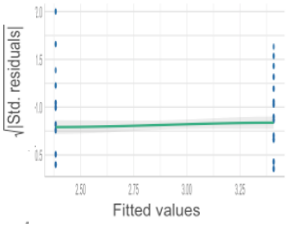
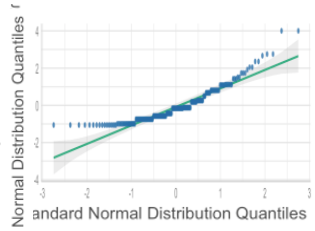
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O17.1** 	**Figure O17.2** 	**Figure O17.3** 
Tobit regression model	**Figure O17.4** 		**Figure O17.5** 
Poisson regression model		**Figure O17.6** 	**Figure O17.7** 
NB regression model		**Figure O17.8** 	**Figure O17.8** 
ZIP model			
ZINB model			

Figure O18. Assumption Checking for One Predictor Model of Mode of Administration for Hyperthyroid Symptoms Scale

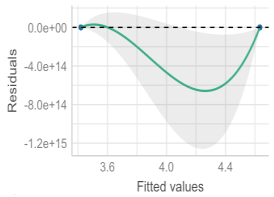
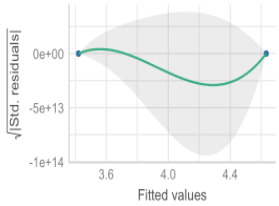
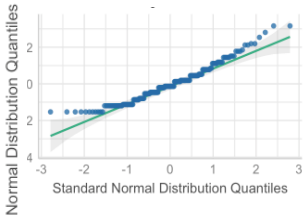
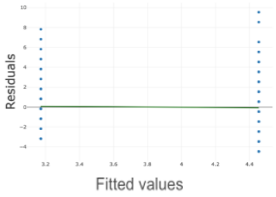
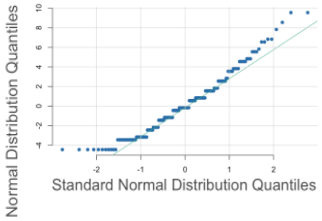
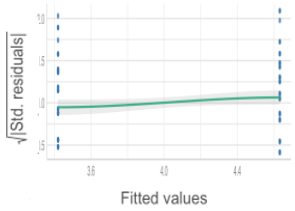
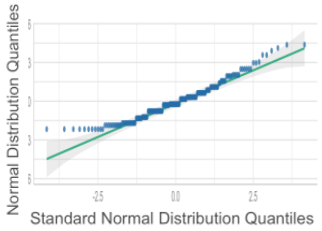
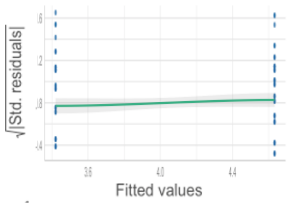
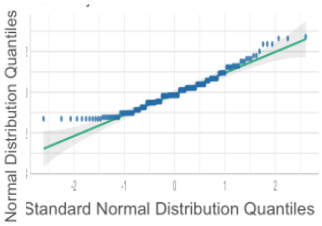
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O18.1** 	**Figure O18.2** 	**Figure O18.3** 
Tobit regression model	**Figure O18.4** 		**Figure O18.5** 
Poisson regression model		**Figure O18.6** 	**Figure O18.7** 
NB regression model		**Figure O18.8** 	**Figure O18.9** 
ZIP model			
ZINB model			

Figure O19. Assumption Checking for One Predictor Model of Mode of Administration for Hypothyroid Symptoms Scale

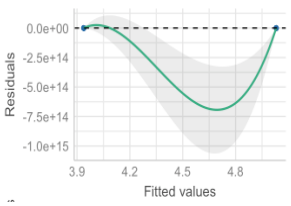
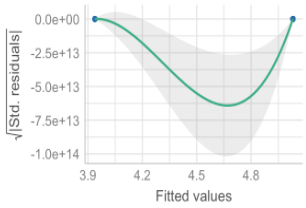
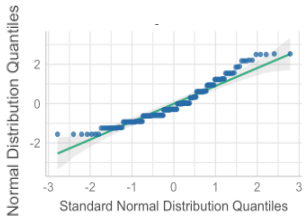
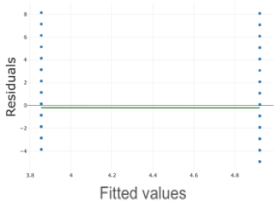
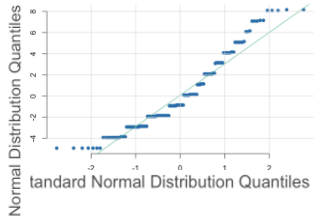
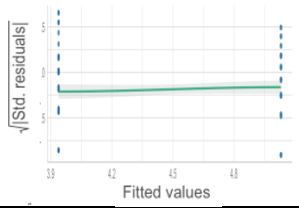
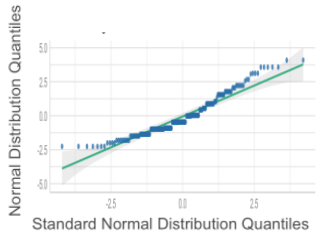
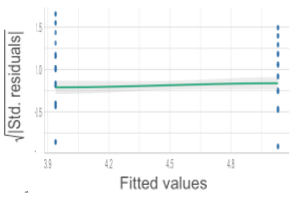
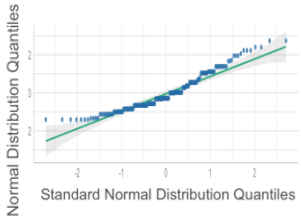
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O19.1** 	**Figure O19.2** 	**Figure O19.3** 
Tobit regression model	**Figure O19.4** 		**Figure O19.5** 
Poisson regression model		**Figure O19.6** 	**Figure O19.7** 
NB regression model		**Figure O19.8** 	**Figure O19.9** 
ZIP model			
ZINB model			

Figure O20. Assumption Checking for One Predictor Model of Mode of Administration for Eye Symptoms Scale

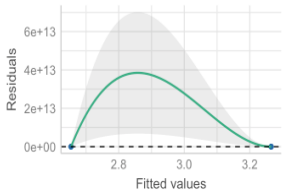
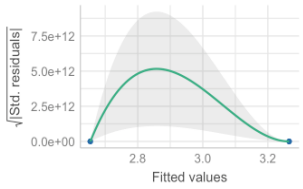
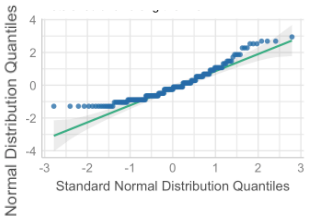
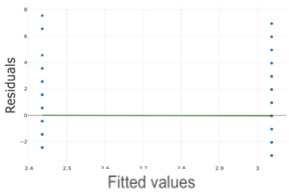
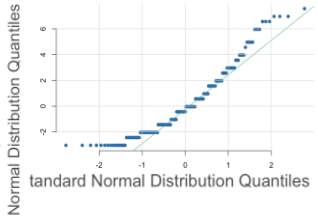
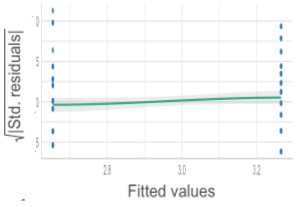
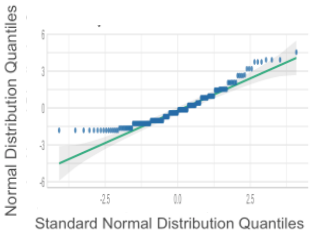
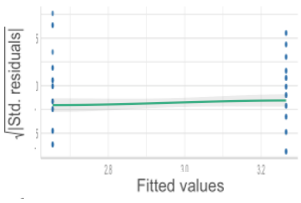
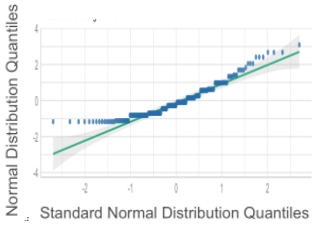
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O20.1** 	**Figure O20.2** 	**Figure O20.3** 
Tobit regression model	**Figure O20.4** 		**Figure O20.5** 
Poisson regression model		**Figure O20.6** 	**Figure O20.7** 
NB regression model		**Figure O20.8** 	**Figure O20.9** 
ZIP model			
ZINB model			

Figure O21. Assumption Checking for One Predictor Model of Mode of Administration for Tiredness Scale

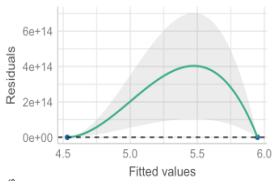
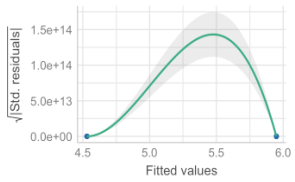
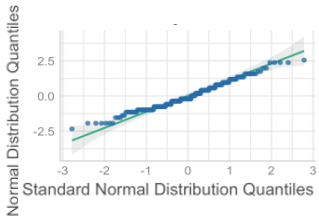
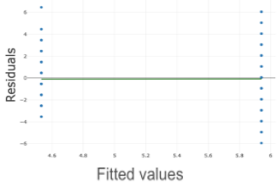
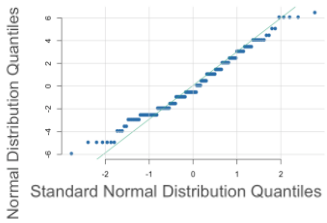
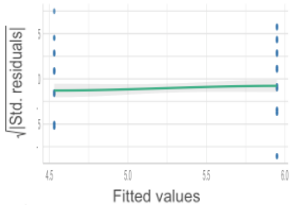
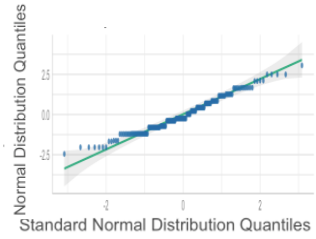
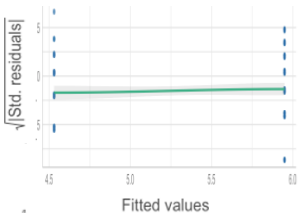
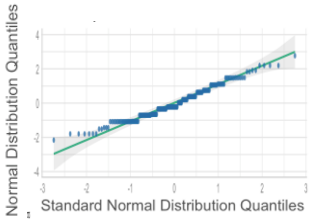
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O21.1** 	**Figure O21.2** 	**Figure O21.3** 
Tobit regression model	**Figure O21.4** 		**Figure O21.5** 
Poisson regression model		**Figure O21.6** 	**Figure O21.7** 
NB regression model		**Figure O21.8** 	**Figure O21.9** 
wq ZIP model			
ZINB model			

Figure O22. Assumption Checking for One Predictor Model of Mode of Administration for Cognitive Complaints Scale

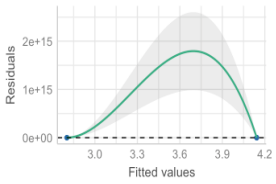
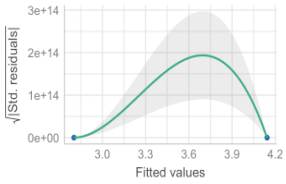
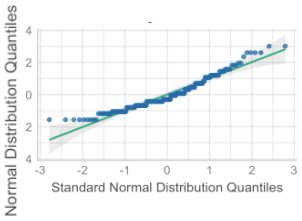
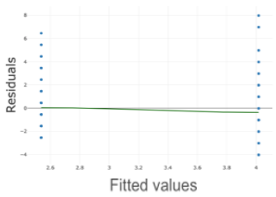
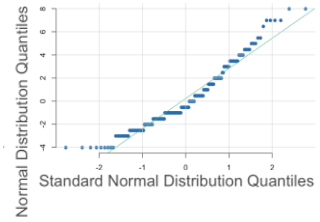
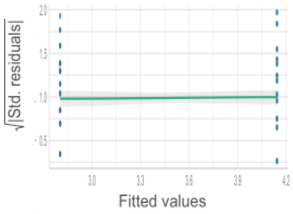
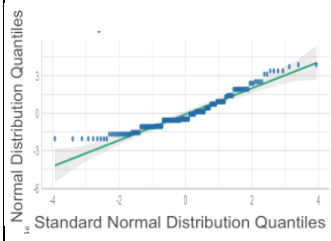
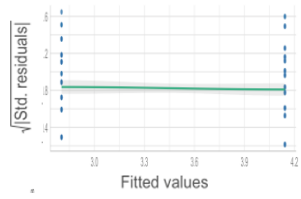
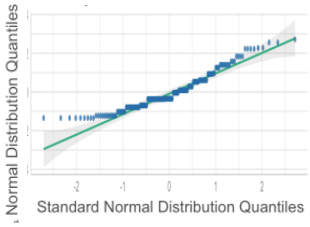
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O22.1** 	**Figure O22.2** 	**Figure O22.3** 
Tobit regression model	**Figure O22.4** 		**Figure O22.5** 
Poisson regression model		**Figure O22.6** 	**Figure O22.7** 
NB regression model		**Figure O22.8** 	**Figure O22.9** 
ZIP model			
ZINB model			

Figure O23. Assumption Checking for One Predictor Model of Mode of Administration for Anxiety Scale

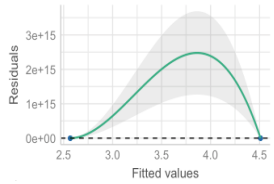
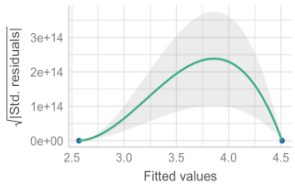
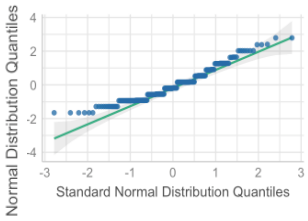
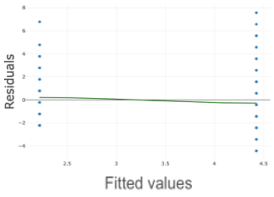
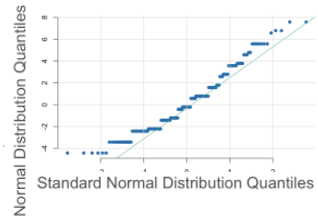
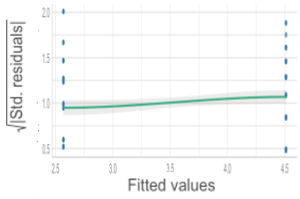
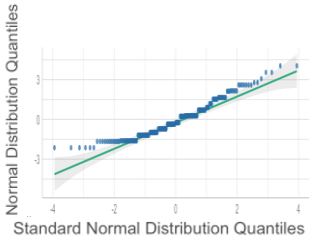
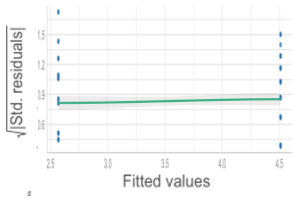
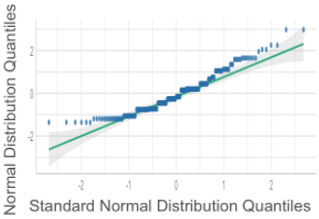
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O23.1** 	**Figure O23.2** 	**Figure O23.3** 
Tobit regression model	**Figure O23.4** 		**Figure O23.5** 
Poisson regression model		**Figure O23.6** 	**Figure O23.7** 
NB regression model		**Figure O23.8** 	**Figure O23.9** 
ZIP model			
ZINB model			

Figure O24. Assumption Checking for One Predictor Model of Mode of Administration for Depressivity Scale

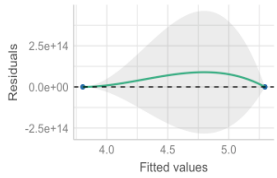
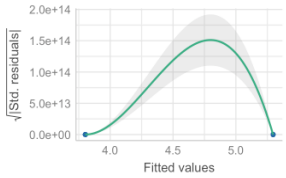
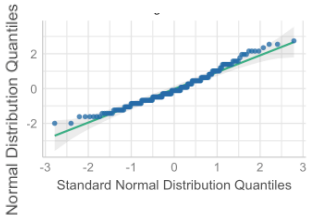
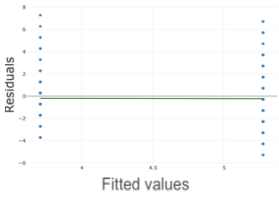
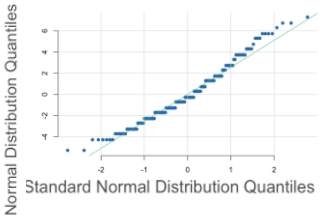
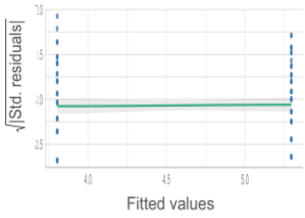
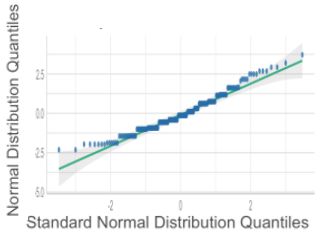
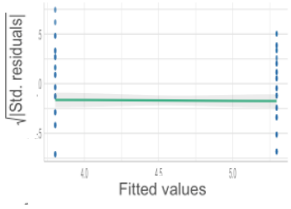
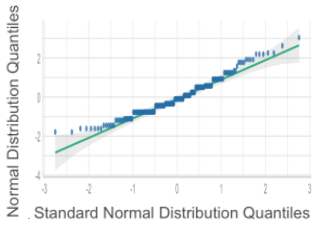
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O24.1** 	**Figure O24.2** 	**Figure O24.3** 
Tobit regression model	**Figure O24.4** 		**Figure O24.5** 
Poisson regression model		**Figure O24.6** 	**Figure O24.7** 
NB regression model		**Figure O24.8** 	**Figure O24.9** 
ZIP model			
ZINB model			

Figure O25. Assumption Checking for One Predictor Model of Mode of Administration for Emotional Susceptibility Scale

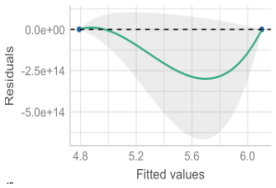
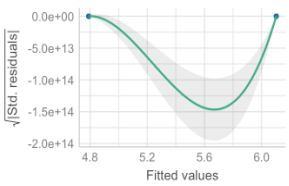
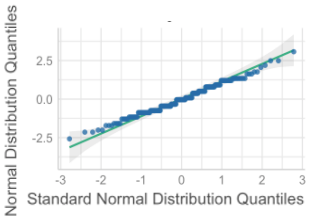
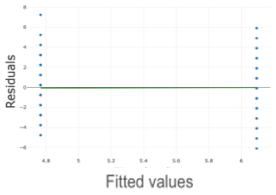
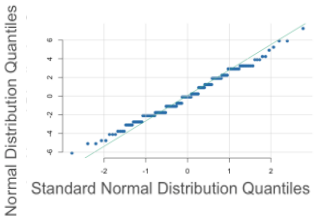
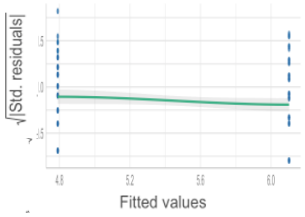
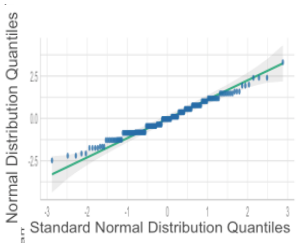
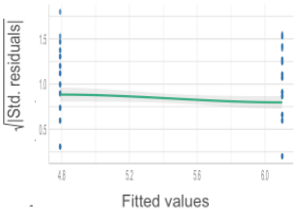
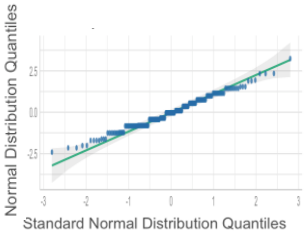
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O25.1** 	**Figure O25.2** 	**Figure O25.3** 
Tobit regression model	**Figure O25.4** 		**Figure O25.5** 
Poisson regression model		**Figure O25.6** 	**Figure O25.7** 
NB regression model		**Figure O25.8** 	**Figure O25.9** 
ZIP model			
ZINB model			

Figure O26. Assumption Checking for One Predictor Model of Mode of Administration for Composite Scale

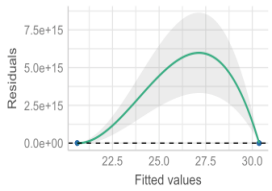
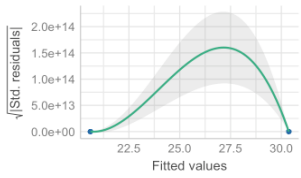
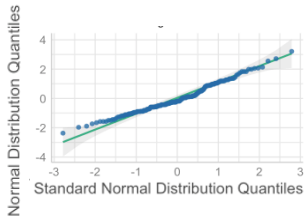
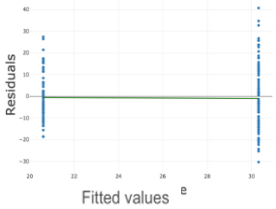
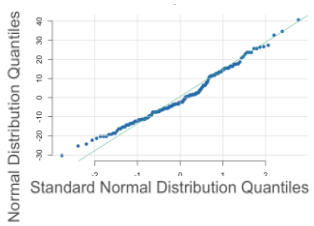
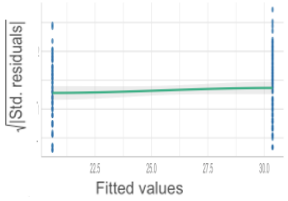
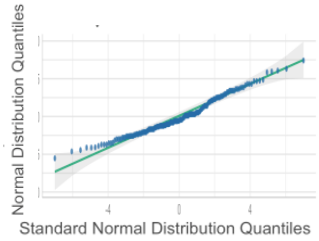
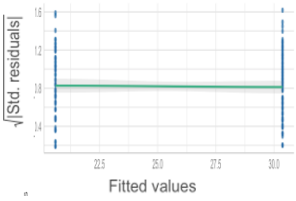
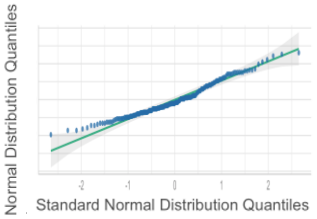
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O26.1** 	**Figure O26.2** 	**Figure O26.3** 
Tobit regression model	**Figure O26.4** 		**Figure O26.5** 
Poisson regression model		**Figure O26.6** 	**Figure O26.7** 
NB regression model		**Figure O26.8** 	**Figure O26.9** 
ZIP model			
ZINB model			

Figure O27. Assumption Checking for Full Predictor Model of All Six Predictors for Goiter Symptoms Scale

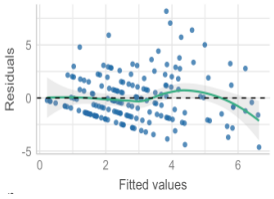
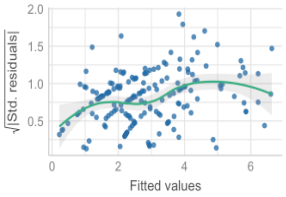
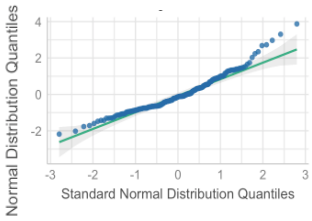
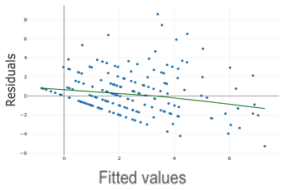
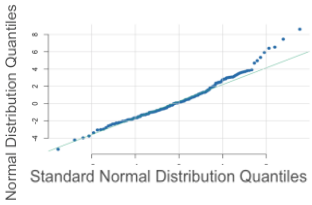
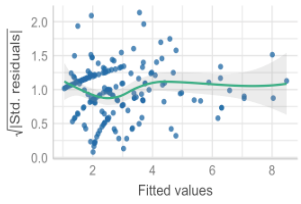
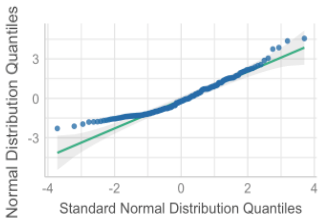
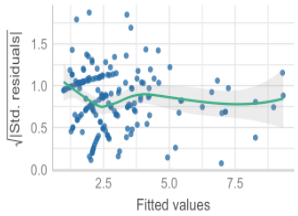
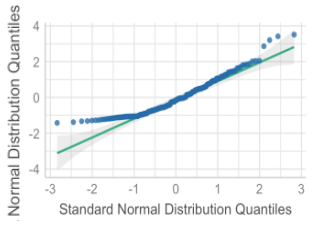
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O27.1** 	**Figure O27.2** 	**Figure O27.3** 
Tobit regression model	**Figure O27.4** 		**Figure O27.5** 
Poisson regression model		**Figure O27.6** 	**Figure O27.7** 
NB regression model		**Figure O27.8** 	**Figure O27.9** 
ZIP model			
ZINB model			

Figure O28. Assumption Checking for Full Predictor Model of All Six Predictors for Impaired Social Life Scale

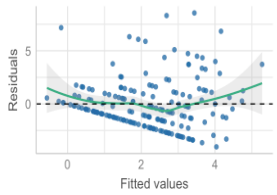
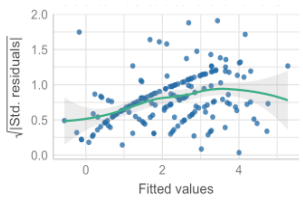
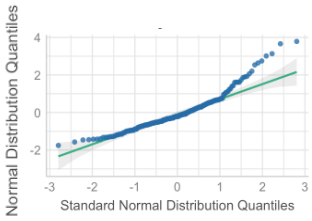
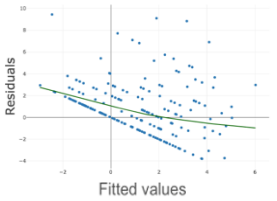
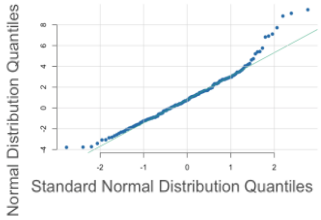
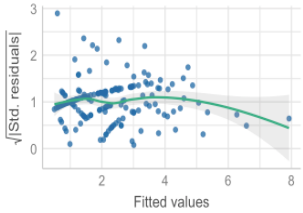
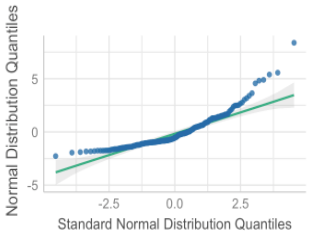
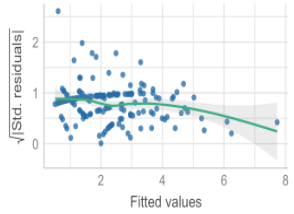
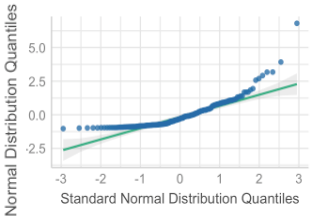
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O28.1** 	**Figure O28.2** 	**Figure O28.3** 
Tobit regression model	**Figure O28.4** 		**Figure O28.5** 
Poisson regression model		**Figure O28.6** 	**Figure O28.7** 
NB regression model		**Figure O28.8** 	**Figure O28.9** 
ZIP model			
ZINB model			

Figure O29. Assumption Checking for Full Predictor Model of All Six Predictors for Impaired Daily Life Scale

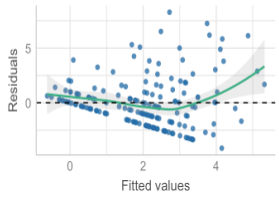
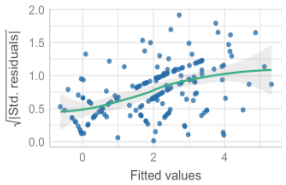
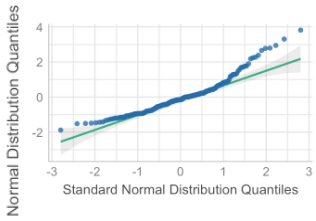
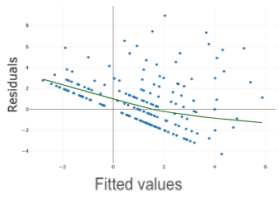
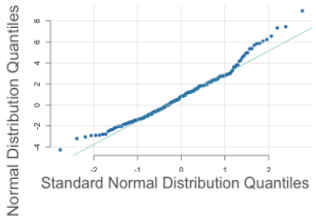
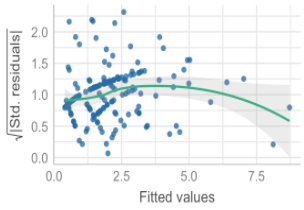
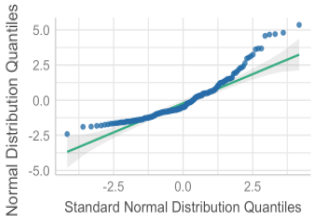
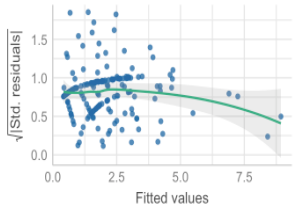
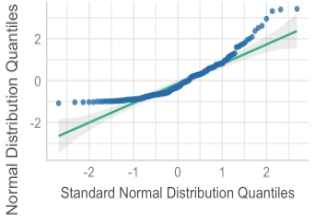
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O29.1** 	**Figure O29.2** 	**Figure O29.3** 
Tobit regression model	**Figure O29.4** 		**Figure O29.5** 
Poisson regression model		**Figure O29.6** 	**Figure O29.7** 
NB regression model		**Figure O29.8** 	**Figure O29.9** 
ZIP model			
ZINB model			

Figure O30. Assumption Checking for Full Predictor Model of All Six Predictors for Cosmetic Complaints Scale

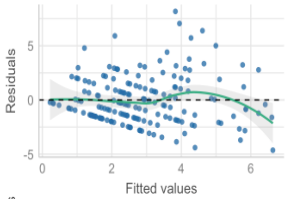
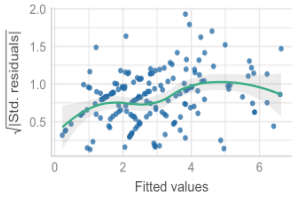
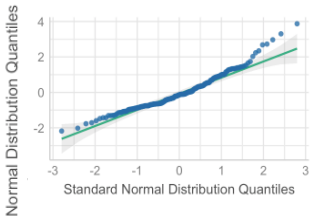
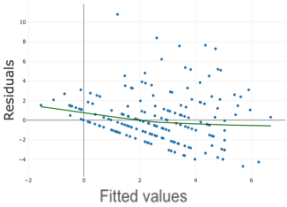
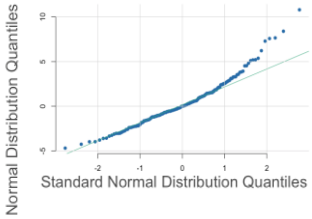
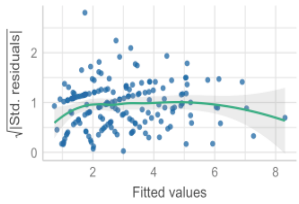
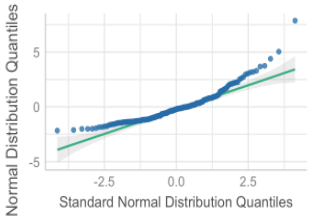
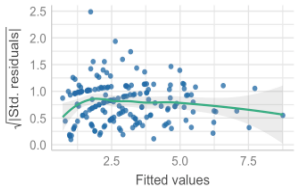
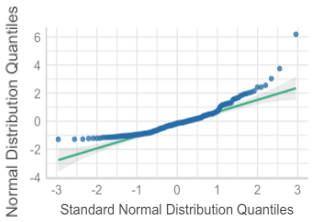
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O30.1** 	**Figure O30.2** 	**Figure O30.3** 
Tobit regression model	**Figure O30.4** 		**Figure O30.5** 
Poisson regression model		**Figure O30.6** 	**Figure O30.7** 
NB regression model		**Figure O30.8** 	**Figure O30.9** 
ZIP model			
ZINB model			

Figure O31. Assumption Checking for Full Predictor Model of All Six Predictors for Hyperthyroid Symptoms Scale

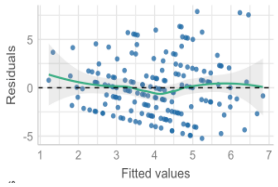
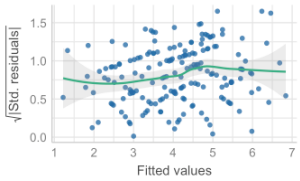
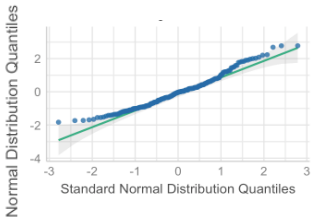
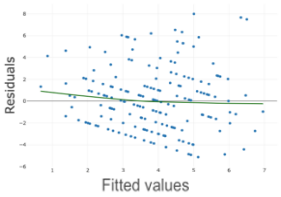
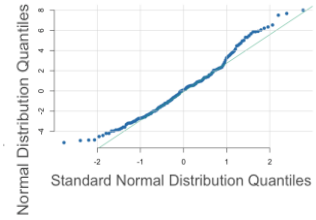
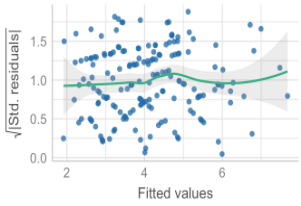
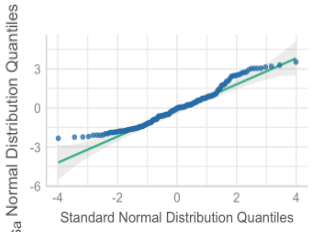
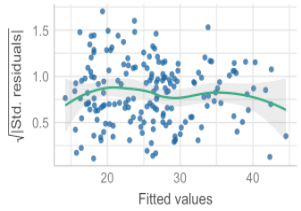
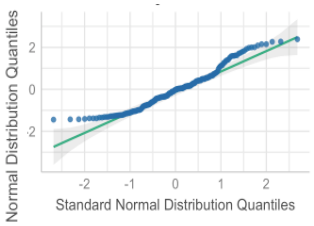
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O31.1** 	**Figure O31.2** 	**Figure O31.3** 
Tobit regression model	**Figure O31.4** 		**Figure O31.5** 
Poisson regression model		**Figure O31.6** 	**Figure O31.7** 
NB regression model		**Figure O31.8** 	**Figure O31.9›** 
ZIP model			
ZINB model			

Figure O32. Assumption Checking for Full Predictor Model of All Six Predictors for Hypothyroid Symptoms Scale

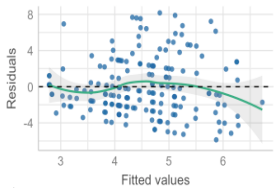
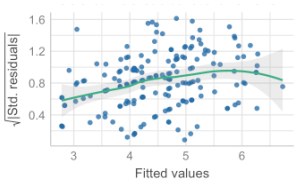
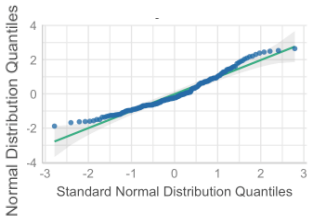
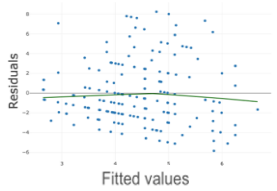
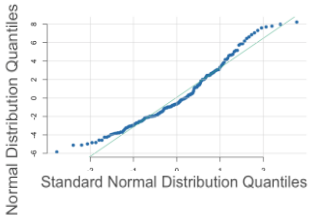
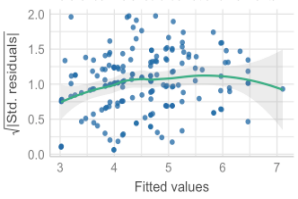
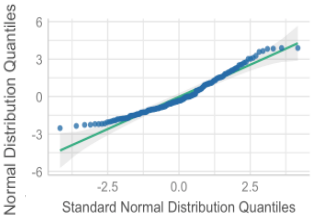
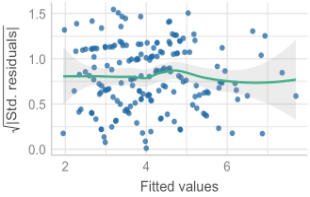
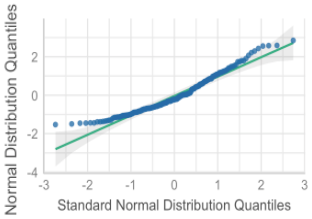
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O32.1** 	**Figure O32.2** 	**Figure O32.3** 
Tobit regression model	**Figure O32.4** 		**Figure O32.5** 
Poisson regression model		**Figure O32.6** 	**Figure O32.7** 
NB regression model		**Figure O32.8** 	**Figure O32.9** 
ZIP model			
ZINB model			

Figure O33. Assumption Checking for Full Predictor Model of All Six Predictors for Eye Symptoms Scale

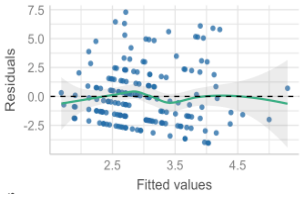
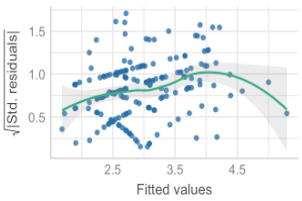
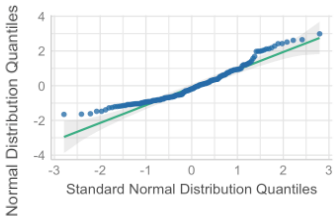
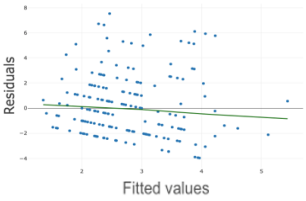
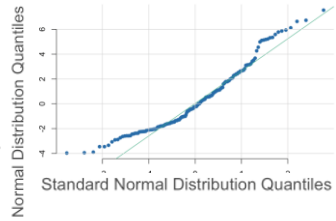
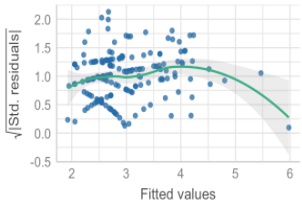
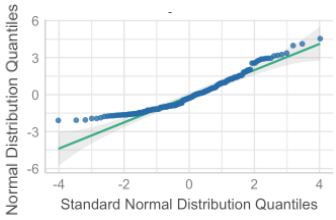
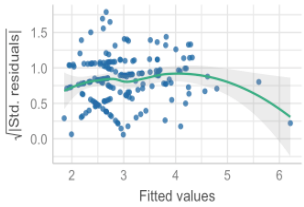
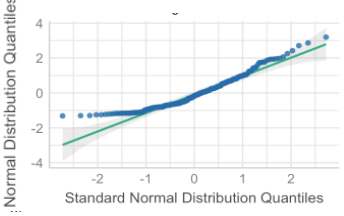
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O33.1** 	**Figure O33.2** 	**Figure O33.3** 
Tobit regression model	**Figure O33.4** 		**Figure O33.5** 
Poisson regression model		**Figure O33.6** 	**Figure O33.7** 
NB regression model		**Figure O33.8** 	**Figure O33.9** 
ZIP model			
ZINB model			

Figure O34. Assumption Checking for Full Predictor Model of All Six Predictors for Tiredness Scale

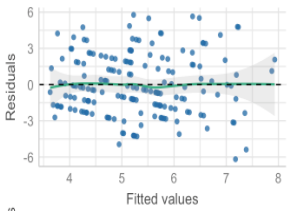
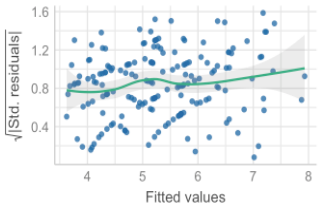
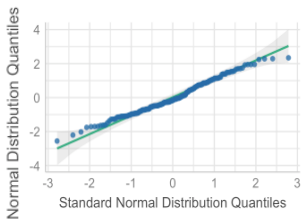
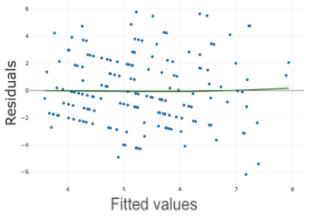
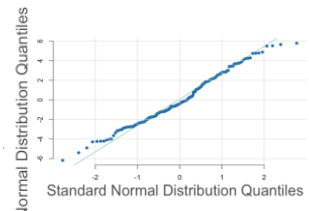
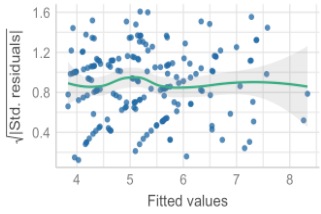
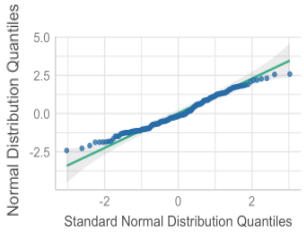
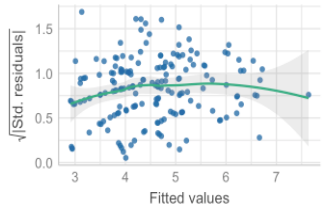
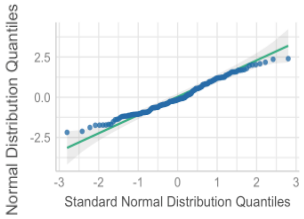
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O34.1** 	**Figure O34.2** 	**Figure O34.3** 
Tobit regression model	**Figure O34.4** 		**Figure O34.5** 
Poisson regression model		**Figure O34.6** 	**Figure O34.7** 
NB regression model		**Figure O34.8** 	**Figure O34.9** 
ZIP model			
ZINB model			

Figure O35. Assumption Checking for Full Predictor Model of All Six Predictors for Cognitive Complaints Scale

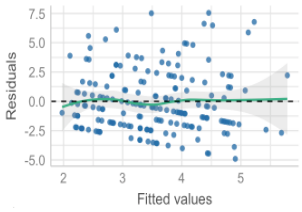
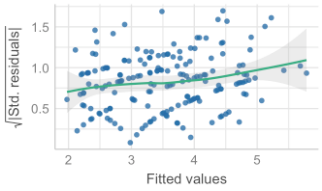
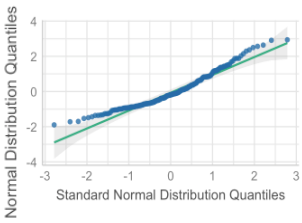
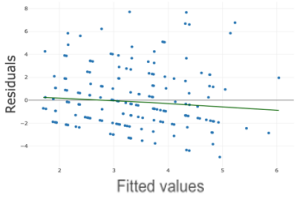
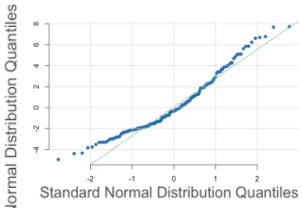
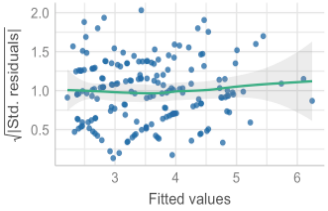
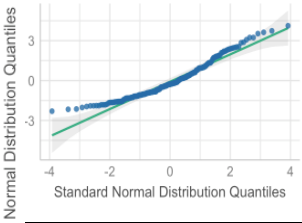
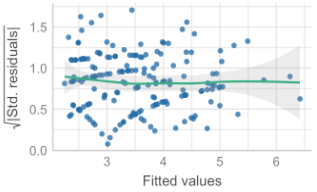
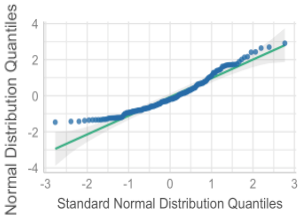
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O35.1** 	**Figure O35.2** 	**Figure O35.3** 
Tobit regression model	**Figure O35.4** 		**Figure O35.5** 
Poisson regression model		**Figure O35.6** 	**Figure O35.7** 
NB regression model		**Figure O35.8** 	**Figure O35.9** 
ZIP model			
ZINB model			

Figure O36. Assumption Checking for Full Predictor Model of All Six Predictors for Anxiety Scale

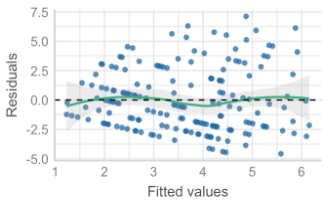
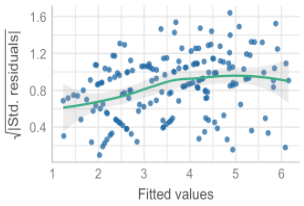
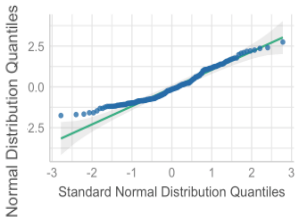
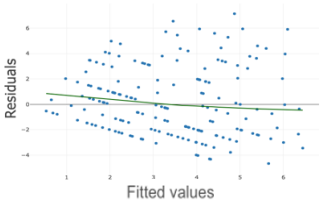
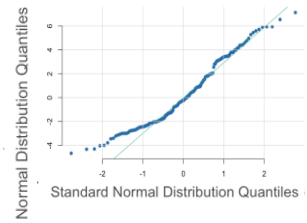
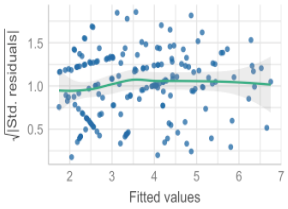
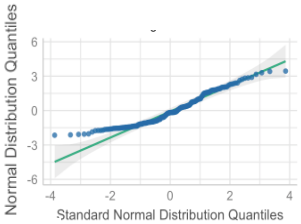
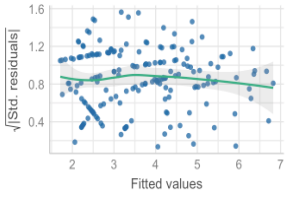
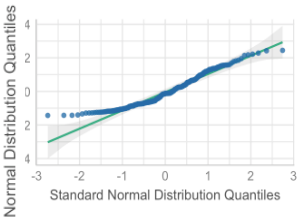
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O36.1** 	**Figure O36.2** 	**Figure O36.3** 
Tobit regression model	**Figure O36.4** 		**Figure O36.5** 
Poisson regression model		**Figure O36.6** 	**Figure O36.7** 
NB regression model		**Figure O36.8** 	**Figure O36.9** 
ZIP model			
ZINB model			

Figure O37. Assumption Checking for Full Predictor Model of All Six Predictors for Depressivity Scale

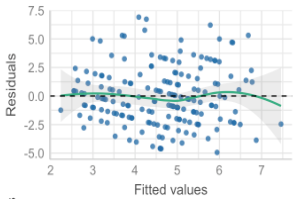
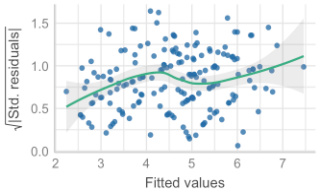
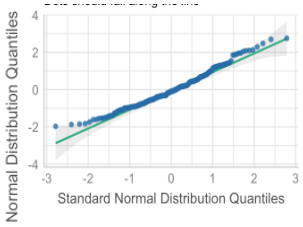
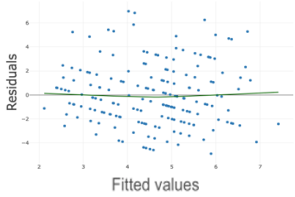
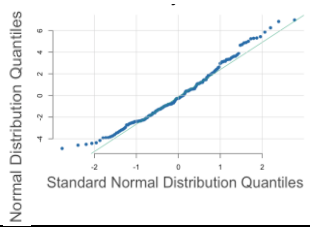
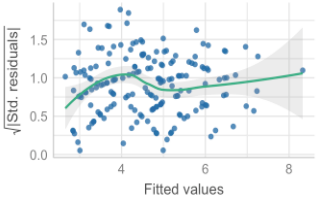
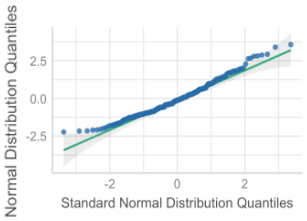
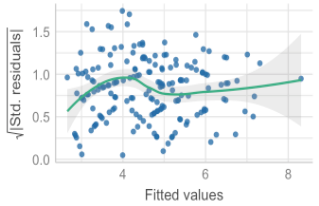
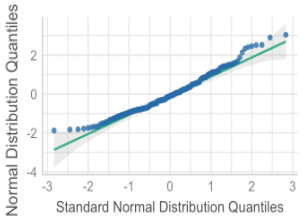
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O37.1** 	**Figure O37.2** 	**Figure O37.3** 
Tobit regression model	**Figure O37.4** 		**Figure O37.5** 
Poisson regression model		**Figure O37.6** 	**Figure O37.7** 
NB regression model		**Figure O37.8** 	**Figure O37.9** 
ZIP model			
ZINB model			

Figure O38. Assumption Checking for Full Predictor Model of All Six Predictors for Emotional Susceptibility Scale

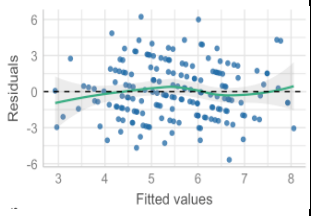
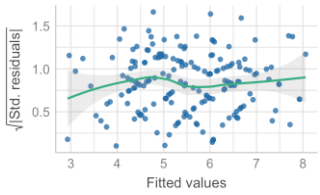
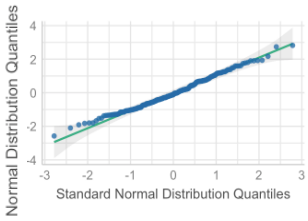
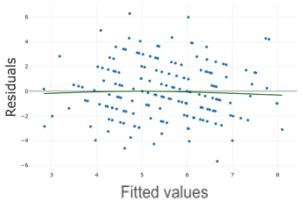
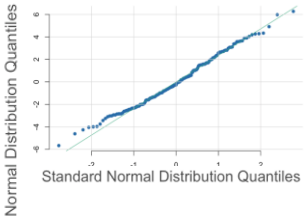
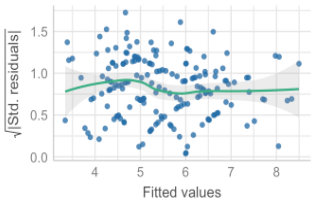
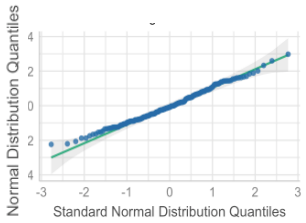
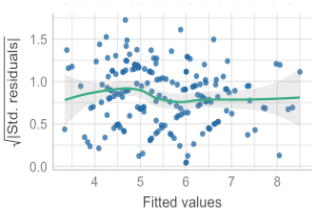
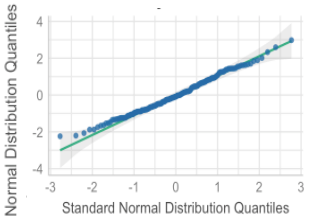
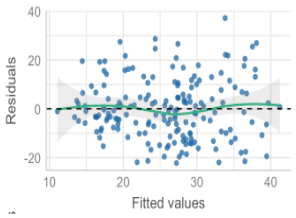
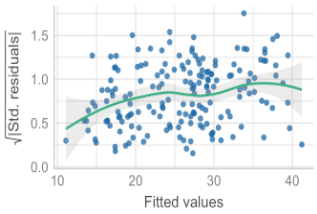
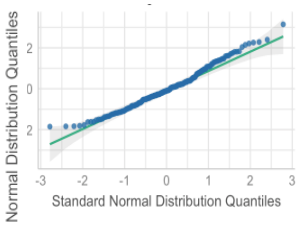
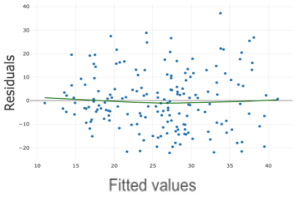
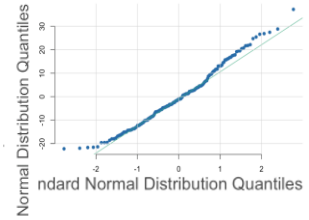
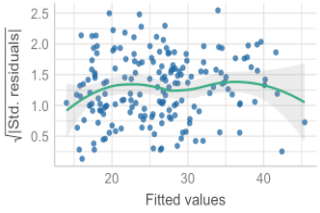
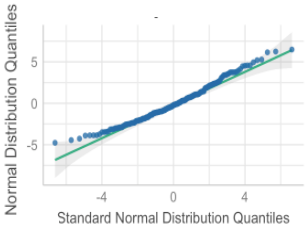
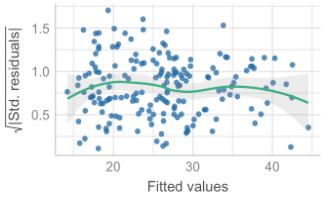
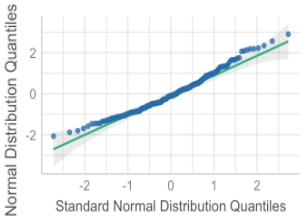
Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O38.1** 	**Figure O38.2** 	**Figure O38.3** 
Tobit regression model	**Figure O38.4** 		**Figure O38.5** 
Poisson regression model		**Figure O38.6** 	**Figure O38.7** 
NB regression model		**Figure O38.8** 	**Figure O38.9** 
ZIP model			
ZINB model			

Figure O39. Assumption Checking for Full Predictor Model of All Six Predictors for Composite Scale

Type of distributional model	Scatterplot between Fitted Value and Residuals	Scatterplot between Fitted Value and Squared Root of Standardized Residuals	Normal Q-Q plot
	Check for Linearity	Check for homogeneity of variance	Check for normality of errors
ML regression model	**Figure O39.1** 	**Figure O39.2** 	**Figure O39.3** 
Tobit regression model	**Figure O39.4** 		**Figure O39.5** 
Poisson regression model		**Figure O39.6** 	**Figure O39.7** 
NB regression model		**Figure O39.8** 	**Figure O39.9** 
ZIP model			
ZINB model			