Guideline Assessment Project II: Statistical Calibration Informed the Development of an AGREE II Extension for Surgical Guidelines

Running head:

AGREE II Calibration for Surgical Guidelines

Authors:

Sofia Tsokani, MSc,[a] Stavros A. Antoniou, MD PhD MPH FEBS,[b,c] Irini Moustaki, PhD,[d] Manuel López-Cano, MD PhD,[e] George A. Antoniou, MD PhD MSc FEBVS,[f,g] Ivan D. Flórez MD MSc PhD,[h,i] Gianfranco Silecchia,[j] Sheraz Markar,[k] Dimitrios Stefanidis, MD PhD FACS FASMBS,[l] Giovanni Zanninotto, PhD[k] Nader K. Francis, MBChB FRCS PhD,[m] George H. Hanna,[k] Salvador Morales-Conde, MD PhD,[n] Hendrik Jaap Bonjer, MD PhD FRCSC,[o] Melissa C. Brouwers PhD,[h] Dimitrios Mavridis, PhD[a,p]

Affiliations:

a. Department of Primary Education, School of Education, University of Ioannina, Ioannina, Greece
b. Surgical Department, Mediterranean Hospital of Cyprus, Limassol, Cyprus
c. Medical School, European University of Cyprus, Nicosia, Cyprus
d. London School of Economics and Political Science, London, United Kingdom
e. Abdominal Wall Surgery Unit, Department of General Surgery, Hospital Universitari Vall d'Hebron, Universitat Autònoma de Barcelona, Barcelona, Spain
f. Department of Vascular and Endovascular Surgery, The Royal Oldham Hospital, Pennine Acute Hospitals NHS Trust, Manchester, United Kingdom
g. Division of Cardiovascular Sciences, School of Medical Sciences, University of Manchester, Manchester, United Kingdom
h. Department of Health Research Methods Evidence and Impact, McMaster University, Hamilton, Ontario, Canada.
i. Department of Pediatrics, Universidad de Antioquia, Medellin, Colombia.
j. Department of medico-surgical sciences and biotecnologies, Sapienza University of Rome, Rome, Italy
k. Department of Surgery and Cancer, Imperial College, London, UK
l. Indiana University School of Medicine, Indianapolis, USA
m. Department of General Surgery, Yeovil District Hospital NHS Foundation Trust, Higher Kingston, Yeovil, United Kingdom
n. Unit of Innovation in Minimally Invasive Surgery. Department of Surgery. University Hospital Virgen del Rocio, Sevilla, Spain
o. Department of General Surgery, VU University Medical Center, Amsterdam, Netherlands
p. Paris Descartes University, Paris, France

Corresponding author:
Sofia Tsokani
University of Ioannina
451 10 Ioannina
Greece
Email: sofia.tsokani@uoi.gr

Conflicts of Interest and Source of Funding: The authors declare no competing financial interest. The authors received no funding for this work.

ABSTRACT

Objective

To inform the development of an AGREE II extension specifically tailored for surgical guidelines.

Summary background data

AGREE II was designed to inform the development, reporting and appraisal of clinical practice guidelines. Previous research has suggested substantial room for improvement of the quality of surgical guidelines.

Methods

A previously published search in MEDLINE for clinical practice guidelines published by surgical scientific organizations with an international scope between 2008 and 2017, resulted in a total of 67 guidelines. The quality of these guidelines was assessed using AGREE II. We performed a series of statistical analyses (reliability, correlation and Factor Analysis, Item Response Theory) with the objective to calibrate AGREE II for use specifically in surgical guidelines.

Results

Reliability/correlation/factor analysis and Item Response Theory produced similar results and suggested that a structure of 5 domains, instead of 6 domains of the original instrument, might be more appropriate. Furthermore, exclusion and re-arrangement of items to other domains was found to increase the reliability of AGREE II when applied in surgical guidelines.

Conclusions

The findings of this study suggest that statistical calibration of AGREE II might improve the development, reporting and appraisal of surgical guidelines.

Keywords:

clinical practice guidelines; surgery; AGREE II; methodological quality; reporting quality

Running title:

Calibrating AGREE II for Surgical Guidelines

Word count:

3872

1. INTRODUCTION

Clinical practice guidelines (CPGs) intend to inform health professionals, healthcare providers and other stakeholders about the optimal course of action or management of a specific disease or condition.[1] As such, CPGs have a direct impact on the delivery of healthcare. Recent research has indicated that 40% of surgical CPGs may not be suitable for clinical use, scoring particularly low in the domains of applicability, editorial independence and Rigour of development.[2] This highlights a substantial need for further improvement on the quality of CPGs.

A great scientific endeavor in the past few years has focused on the quality of CPGs.[3-7] The Appraisal for Research and Evaluation (AGREE) instrument constitutes a framework for developing, appraising and reporting on CPGs and it is endorsed by major agencies, such as the National Institute for Health and Care Excellence (NICE) and the World Health organization (WHO), among others.[8-10] The original AGREE instrument was updated (AGREE II) and consists of 23 Likert items consisting of seven possible answers each and comprising 6 domains; Scope and purpose (3 items), Stakeholder involvement (3 items), Rigour of development (8 items), Clarity of presentation (3 items), Applicability (4 items) and Editorial independence (2 items).[6,7] The tool concludes with 2 items that provide an overall assessment of the guideline.

The AGREE II instrument has become an established tool for supporting the development and reporting of CPGs, and assessing their methodological quality. Nevertheless, it is a generic tool that does not necessarily address specific needs in certain disciplines such as surgery. For example, it does not take into consideration the variation in practice and the diversities in surgical expertise around the world, thereby falling short of promoting health equity. Additionally, some of its items may be not relevant to surgical CPGs. For instance, the item "The guideline has been externally reviewed by experts rior to its publication" might not be sufficiently comprehensive, because surgeons with expertise in a surgical technique might pose intellectual conflicts due to personal conviction or positions. As such, AGREE II might not inform the development and reporting of surgical CPGs in the best possible way.

Each domain in the AGREE II instrument has several items and can be viewed as a sub-instrument. Ideally, all items within a domain should be indicators of the same hypothetical construct. Our aim was to explore how items within each domain were interrelated, with the objective to calibrate the AGREE II instrument for use in surgical guidelines and to provide suggestions to refine the AGREE II instrument in order to meet the requirements for CPGs in surgery.

2. MATERIALS AND METHODS

In the first part of this project, we searched MEDLINE via Ovid for clinical practice guidelines published by surgical scientific organizations with an international scope between 2008 and 2017. The methodological details of this first stage are reported elsewhere[2] Our search strategy resulted in a total of 67 eligible guidelines, which were developed under the auspices of 10 scientific organizations. Two independent authors (SAA, MLC), who had acted as coordinators and members of guideline development groups and fulfilled the criteria of a GRADE (Grading of Recommendations Assessment, Development and Evaluation) methodologist,[11] applied the AGREE II instrument **(Appendix Table 1)** on this set of guidelines.

We conducted a series of statistical methods to explore the reliability/internal consistency and unidimensionality of each domain of the AGREE instrument. Internal consistency refers to the ability of a test/instrument/scale to measure consistently the same construct. It measures how several items that are proposed to measure the same construct produce similar scores. A set of items is unidimensional if all items measure a single latent trait/hypothetical construct. Large intercorrelations among test items are indicative

of the items measuring the same construct. However, internal consistency is a necessary but not sufficient condition for unidimensionality.[12]

The AGREE II instrument is structured in predefined groups (domains) and we aimed to explore whether items of each group are 1) highly intercorrelated (internal consistency) and 2) indicators of the same construct (unidimensionality).

Ideally, we would like items within a domain to be strongly correlated and items between different domains to be weakly correlated. To this aim, we used the following methods:

- Cronbach's alpha (internal consistency)
- Polychoric correlation (internal consistency)
- Factor Analysis / Item Response Theory (IRT) (dimensionality)

2.1 Cronbach's alpha

Cronbach's alpha, denoted as α or coefficient alpha[13] is probably the most well-known measure of measuring reliability/internal consistency of an instrument. Its computation is straightforward, and it is a function of the number of items, the average covariance between pairs of items and the variance of the total score. It typically assumes values between 0 and 1 with a score of over 0.7 indicating high internal consistency. [14] One limitation is that the larger the number of items, the larger the value of Cronbach. A low value of alpha could be due to a low number of questions, poor correlations between items or heterogeneous constructs.[12] Caution is needed in that although items' intercorrelations maximize when all items measure the same construct, Cronbach's alpha cannot be used for assessing unidimensionality.[15]

2.2 Polychoric correlation coefficient

The polychoric correlation[16] is a measure of correlation for ordinal variables. When estimating polychoric correlations, we assume that the manifest ordinal items have been derived by categorizing latent normally distributed variables. Hence, we assume that there is a continuous metrical variable underlying each ordinal item. Olsson[17] developed a method for estimating the correlation between two ordinal items based on the respective underlying metrical variables. The range of correlation is from –1 to 1.

2.3 Factor analysis

Factor analysis is a statistical method that describes variability among manifest variables in terms of a potentially lower number of unobserved variables, also known as latent variables/factors. The aim is to reduce the dimensionality of a data set (sample) by finding a new smaller set of variables that however contains most of the sample's information. Caution is needed on that this is not proof that the items measure what the creators of the instrument have designed them for (reliability is necessary but not sufficient for validity. Factor analysis is typically used for continuous outcomes whereas its counterpart for categorical responses is known as IRT.

We should also bear in mind that items of the AGREE II instrument are Likert-type/ordinal and the interval differences between items are not necessarily meaningful. It is common with such items to consider the total score, assuming that the scale is similar across all items and that interval differences are meaningful. It is also common to apply factor analysis as if the Likert items are normally distributed. These assumptions are most probably not true but are widely used with Likert-type items. Instead, we conducted an exploratory factor analysis (EFA), to define any underlying structure of the included items, using polychoric correlations to express correlations among ordinal items. EFA ignores the initial grouping of items into sub-domains and tries to identify the underlying relationships among observed variables from scratch. In a nutshell, EFA will tell us how many factors are needed to account for a large proportion of variability among the observed items and will also link items to these factors.

The main output of the factor analysis is a loading matrix that informs us how much each item loads in a factor. We used "oblimin" rotation to achieve a better interpretation of the identified factors. Ideally, each factor should represent a domain and we would like items of a domain to have large loadings on the corresponding factor. We used the Kaiser-Meyer-Olkin (KMO) test to test how suitable data were for factor analysis. We used the Bartlett's test of sphericity for the hypothesis that the correlation matrix is an identity matrix; in this instance, variables are unrelated and therefore unsuitable for structure detection and factor analysis.

We are aware that our sample size was small, and this may cast doubt in the estimated polychoric correlations and subsequently on the factor analysis results. Additionally, to the EFA using the polychoric correlations, we conducted an IRT approach, which is the counterpart of factor analysis for categorical responses. IRT has become a popular methodological framework for modeling response data from assessments in health sciences. IRT models describe the interactions of persons (here we have guidelines instead of persons) and test items.[18] We used IRT to explore the discrimination of each item by examining their factor loadings. IRT estimates two parameters, the discrimination and difficulty parameters, for each item. The discrimination parameter, denoted by **a**, represents the estimated factor loadings and indicates how well the item differentiates participants with different positions on the latent dimension. The larger the discrimination parameter for an item, the higher the correlation between an/the item and the (measured construct) latent variable. We used IRT within each domain to assess unidimensionality of the corresponding items and how much each item loads on the hypothetical construct of the domain. More specifically, we explored whether variability in items of a domain of the AGREE II can be explained by a single factor (unidimensionality) and which items are responsible for deviations from unidimensionality. This is an example of a confirmatory factor analysis where we want to confirm unidimensionality of domains. All the analyses, except for IRT, were performed in R, version 3.6.3. [19] The IRT analysis was conducted using STATA, version 14.0.[20]

3. RESULTS

3.1 Cronbach's alpha

A reliability analysis was carried out on each of the six domains and the results are summarized in **Table 1**. Domain 1 *"Scope and Purpose"* consists of three questions. The Reliability Statistics table gives the Cronbach's alpha coefficient regarding Domain 1, that was α=0.572. This is a value lower than 0.7, which suggests that the 3-item questionnaire for "Scope and Purpose" was not much reliable.

Table 1. Cronbach's Alpha for each domain

Domain	Cronbach's Alpha	N of Items	Result on Internal Consistency
(1) "Scope and Purpose"	0.572	3	Poor
(2) "Stakeholder Involvement"	-1.630	3	Problematic
(3) "Rigour of Development"	0.849	8	Good
(4) "Clarity of Presentation"	0.775	3	Acceptable
(5) "Applicability"	0.726	4	Acceptable
(6) "Editorial Independence"	0.357	2	Low

The next domain to examine was the *"Stakeholder Involvement"* Domain, where Cronbach's alpha was negative and not interpretable. This might be due to the fact that a substantial proportion of guidelines scored low in the items "The guideline development group includes individuals from all relevant professional groups" and "The views and preferences of the target population have been sought", and high in the item "The target users of the guideline are clearly defined". In addition, Item 4 "*The views and preferences of the target population (patients, public, etc.) have been sought*" did not contribute to this Domain, because all responses were the same (only one guideline involved patient representatives/advocates) and therefore it is considered a constant.

Concerning the third Domain, which is designed to reflect the *"Rigour of Development"*, the Cronbach's alpha coefficient was α=0.849. Consequently, this questionnaire which consists of three questions is much reliable.

Similarly, Domain 4 *"Clarity of Presentation"* is composed of three items and the Cronbach's alpha coefficient was α=0.775; the questionnaire is reliable with an acceptable internal consistency. Moreover, the alpha coefficient for the four items consisting the fifth Domain *"Applicability"* was α=0.726, suggesting that these items comprise a scale with acceptable internal consistency. Finally, Domain 6 *"Editorial Independence"* scale is composed of two items, but Cronbach's alpha showed the questionnaire not to reach acceptable reliability, α=0.357.

Appendix Table 2 shows whether Cronbach's alpha changes if an item is deleted. Items are denoted as in the AGREE II instrument **(Appendix Table 1)**. The main findings concern the first and fourth Domains. In particular, for Domain 1 there was a considerable increase in Cronbach's alpha (27.8%) when Item 3 "*The population (patients, public, etc.) to whom the guideline is meant to apply is specifically described*" was omitted. A single item deletion did not have a large impact on the rest of the domains with the next larger increase occurring in Domain 4 where the Cronbach coefficient increased by 5% when Item 17 "*Key recommendations are easily identifiable*" was deleted.

3.2 Polychoric correlation coefficients

Using polychoric correlation coefficient, ρ, (**Appendix Tables 3.1-3.6)**, the results expressing the amount of association between the items of each domain are summarized below:

- Domain 1: Items 1 and 2 are correlated (ρ=0.5). Item 3 was correlated neither with Item 1 (ρ=-0.02) nor with Item 2 (ρ=-0.04).
- Domain 2: Items 4 and 6 are negatively correlated (ρ=-0.73). The rest of the items were not correlated.
- Domain 3: Items 13 and 14 seemed to be weakly correlated with the rest of items while the rest were correlated with each other.
- Domain 4: Item 15 was correlated with Item 16 (ρ=0.53) and Item 17 (ρ =0.35). Items 16 and 17 were correlated with each other (ρ=0.31).

- Domain 5: Items 20 and 21 (ρ=0.5) were correlated with each other. Item 18 was correlated with Item 19 (ρ=0.35), Item 20 (ρ=0.65) and Item 21 (ρ=0.61). Item 19 was correlated with Item 20 (ρ=0.59) but not with Item 21 (ρ=-0.12).
- Domain 6: Item 22 and Item 23 were correlated (ρ=0.43) with each other.

We estimated the correlation between items 3, 5 and 14 and the remaining items of the AGREE II instrument using the polychoric correlation coefficient and we found Item 3 was correlated with the AGREE II Items 13 (ρ=0.39), 14 (ρ=0.33), 19 (ρ=0.41), 20 (ρ=0.33), 21 (ρ=0.3), 22 (ρ=0.33). Item 5 was correlated with Item 13 (ρ=0.55) and Item 21 (ρ=0.47). Item 14 was correlated with Items 18 (ρ=0.57), 19 (ρ=0.46), 20 (ρ= 0.57), 21 (ρ=0.61), which correspond to the "Applicability" domains. **(Appendix Table 4)**.

3.3 Factor analysis – Item Response Theory

For our dataset, the KMO measure of sampling adequacy was 0.72, exceeding the recommended value of 0.6, and Bartlett's test of sphericity had a p-value of <0.001 (X^2 (253) =909.89, P<0.001) indicating that an exploratory factor analysis can be applied.

The scree plot **(Appendix Figure 1)** supports decreasing the number of factors to 4. In a scree plot, the point where the slope of the curve is clearly leveling off indicates the number of factors that should be generated by the analysis. According to the exploratory factor analysis where no items have been linked to factors, a 4-factor solution explained 56% of the total variability in the data **(Appendix Table 6)**.

On the Rotated Component Matrix **(Appendix Table 5)** the following categories of items are loading to the same factor and bringing them together in one factor should be considered. The closer to 1 the loadings, the more important they are in explaining the variation in a factor and within each factor potential relationships are indicated. In particular, items 1, 2, 7, 8, 9, 10, 11, 12, 15, 16, 17 can compose one factor. These items were also strongly correlated. Items 3, 14, 18, 20, 22 together can compose a second factor. All these items, except for Item 3, were correlated with each other. A third factor could consist of items 4, 6, 23. Component 4 can be composed of items 5, 13, 19, 21. The most items of both new created factors were correlated.

3.4 Item Response Theory

In our dataset, IRT analysis converged only for Domains 1, 3 and 4 and the results are given in **Appendix Tables 7-9**. In these tables, each Domain's item is given along with its discrimination parameter a, the 95% confidence interval and the p-value. Also, the Boundary Characteristic Curves and Item Information Functions are plotted **(Appendix Figures 2-8)**.

According to the results of the IRT analysis for Domain 1, Item 3 (Coefficient: 0.22, 95% CI -0.18 to 0.84) appears to be the less informative and it had the smallest discrimination coefficient. The same happens for Domain 3 and Items 13 (Coefficient: 0.93, 95% CI -0.17 to 2.03) and 14 (Coefficient 0.78, 95% CI 0.49 to 9.19). All Items of Domain 4 provided significant information and their discrimination coefficients were high enough. Since most of the seven-point Likert-type items were sparse, we also conducted multiple IRT models merging categories even up to two. Results did not vary significantly.

4. DISCUSSION

By making a synopsis of all statistical methods applied (Cronbach's alpha coefficients, polychoric correlation coefficients, Factor analysis and IRT analysis), we suggest tailoring the AGREE II instrument for surgical guidelines into four groups composed of the following items **(Table 2)**:

- Group 1: Items 1 ,2,
- Group2: Items 7, 8, 9, 10, 11, 12, 15, 16, 17

- Group 3: Items 4, 6, 23
- Group 4: Items 14, 18, 20, 22
- Group 5: 5, 13, 19, 21

From the conceptual perspective, there are several considerations:

- Given that the first factor includes items from three domains of the AGREE II instrument, it is fair to split it in three parts labeled just like the original domains, with each part having the relevant items (Scope and purpose having items 1 and 2, Rigour of development having items 7, 8, 9, 10, 11 and 12 and Clarity of Presentation having items 15, 16 and 17). In this way, we will have six domains and scores in these three domains will be highly correlated.
- Item 3 ("The population (patients, public, etc.) to whom the guideline is meant to apply is specifically described") was not correlated with other items of its domain and we suggest excluding it from Domain 1. Compared to guidelines that do not involve surgical interventions (e.g. guidelines on portal hypertension, where the underlying disease would have to be indicated; or guidelines on urinary tract infection, where the gender of the population of interest would need to be specified), surgical guidelines are usually straightforward regarding their target population (e.g. patients with cholecystitis; or patients with rectal cancer). Nevertheless, this information is still pertinent in several circumstances, such as guidelines on gastric cancer, where the external validity is of specific importance (Western low-risk versus Asian high-risk population). It would be therefore reasonable to maintain this parameter or to incorporate it into another item, such as Item 2, which could be formulated as follows: The health question(s) covered by the guideline *and the patient population(s) it is meant to apply to* are specifically described.
- According to Factor Analysis, the new Domain 1 should be composed by a combination of items from the domains "Scope and Purpose", "Rigour of Development" and "Clarity of Presentation", excluding Item 13 ("The guideline has been externally reviewed by experts prior to its publication."). This is also reasonable, as this item might not be completely relevant to the Domain "Rigour of Development", it is, however of conceptual importance.

  Combining the items from the Domain "Clarity of Presentation" with items from the Domain "Rigour of Development" is also justified. Specific and unambiguous recommendations and alternative management options are features of guidelines produced within high development standards. The GRADE methodology is an example of such methodology,[7] which recommends clear, concise and actionable recommendations, under consideration of associated risks and benefits, resources required, patient preferences etc.

  Since the new Domain contains a large number of items, it can be split to 2 Domains labelled "Scope" (items 1 and 2) and "Rigour of Development and Presentation" (items 5, 7, 8-12, 15-17). The new Domain can be summarized under the label "Rigour of Development and Presentation".
- The statistical models suggested combining items 14, 18, 20 and 22 into one Domain. Indeed, these items refer to post-production considerations, including implementation and future update of the guidelines, except for Item 22 ("The views of the funding body have not influenced the content of the guideline."). This item might better fit in the new Domain labelled "Rigour of Development and Presentation" or "Stakeholder Involvement".
- The statistical models suggested that Items 4 ("The guideline development group includes individuals from all relevant professional groups."), 6 ("The target users of the guideline are clearly defined.") and 23 ("Competing interests of guideline development group members have been recorded and addressed.") be grouped in the same Domain.
- The Items 5, 13, 19, 21 could create a new domain according to the statistical models. Cronbach's alpha suggested that Item 5 ("The views and preferences of the target population (patients, public, etc.) have been sought.") did not contribute to its domain, but this was because all responses where the same (only one guideline sought the input of patients). Nevertheless, this item is a principal parameter of proper guideline development. The input of patients is important both in the preparatory steps of guideline

development (identification of topics, definition of patient-important outcomes) and when making recommendations and deciding on their strength, where risks must be weighed against benefits under consideration of patients' values and preferences.

According to Factor Analysis, Item 13 ("The guideline has been externally reviewed by experts prior to its publication") should be excluded from the Domain "Rigour of Development" as it might not be relevant to it. In our experience, the input by external reviewers is limited both prior to submission for publication and after submission in the process of peer review. This is probably due to the fact that external reviewers have not been involved in the process of guideline development (assessment of the quality of evidence and development of the evidence-to-decision framework). The input, however, by a guideline methodologist is invaluable and may be considered of specific importance in the pre-publication phase. It would be therefore reasonable to exclude or modify this item.

Item 21 ("The guideline presents monitoring and/or auditing criteria.") may also be of limited importance in some contexts and guidelines. Although monitoring of the use of guidelines is of specific importance in the context of guidelines sponsored or directly supported by policymakers (e.g. WHO or NICE), in our experience, it is difficult to assess guideline implementation or adherence to recommendations produced by discipline-specific scientific organizations.

Generally, all methods we used produced similar results. However, some limitations might be discussed. The sample of guidelines were documents focused exclusively on surgical topics and they are not representative of guidelines with a general scope with recommendations pertaining to surgery (e.g. guidelines for the management of obesity, including recommendations on bariatric surgery). A problem with all analyses was the sparseness of data. With 67 guidelines and 23 seven-point items, we expect that results cannot be conclusive. Polychoric correlations assume a normally distributed variable underlying each ordinal item. This assumption could not be tested. Although IRT takes into account the true nature of the responses and do not treat them as scale variables, with 67 guidelines and 7-point items we expect that we will not have much information for many possible patterns. Even the combinations between two seven-point Likert items are 49. Despite the assumptions made and the small number of included guidelines, we find it reassuring that all methods identified the same structure. However, results are not conclusive and are prone to change once more data are collected.

The present work is the second component of a tripartite project, with the ultimate objective to develop an AGREE II extension for surgical guidelines.[21] The first part identified parameters that were associated with guidelines quality. It has identified 3 parameters to be associated with quality: regular guideline output by a surgical organization, guideline development by a dedicated committee and adhering to the GRADE methodology.[2] The present work aimed at statistically exploring and improving the internal validity of AGREE II in the context of surgical guidelines. The third part involves summarization of findings from the previous work and other published research on the topic, and presentation to a multidisciplinary panel of surgical specialists, journal editors, guideline development bodies, GRADE representatives, and patient representatives. Under consideration of the evidence, stakeholders will be asked to provide their input through a Delphi process, which will inform the development of an AGREE II extension for surgical guidelines.

Table 2. Proposal for a modified AGREE II instrument for surgical guidelines according to statistical modeling

Domain 1	1. The overall objective(s) of the guideline is (are) specifically described.
	2. The health question(s) covered by the guideline and the patient population(s) it is meant to apply to are specifically described.
Domain 2	7. Systematic methods were used to search for evidence.
	8. The criteria for selecting the evidence are clearly described.
	9. The strengths and limitations of the body of evidence are clearly described.
	10. The methods for formulating the recommendations are clearly described.
	11. The health benefits, side effects, and risks have been considered in formulating the recommendations.
	12. There is an explicit link between the recommendations and the supporting evidence.
	15. The recommendations are specific and unambiguous.
	16. The different options for management of the condition or health issue are clearly presented.
	17. Key recommendations are easily identifiable
Domain 3	4. The guideline development group includes individuals from all relevant professional groups.
	6. The target users of the guideline are clearly defined.
	23. Competing interests of guideline development group members have been recorded and addressed.
Domain 4	14. A procedure for updating the guideline is provided.
	18. The guideline describes facilitators and barriers to its application.
	20. The potential resource implications of applying the recommendations have been considered.
	22. The views of the funding body have not influenced the content of the guideline.
Domain 5	5. The views and preferences of the target population (patients, public, etc.) have been sought.
	13. The guideline has been externally reviewed by experts prior to its publication.
	19. The guideline provides advice and/or tools on how the recommendations can be put into practice.
	21. The guideline presents monitoring and/or auditing criteria.

Numbers represent numbering of items in the original AGREE II instrument.
Items 3 of the original AGREE II instrument has been removed based on statistical modeling.

REFERENCES

1 Graham, R., Institute of, M. & Committee on Standards for Developing Trustworthy Clinical Practice, G. *Clinical practice guidelines we can trust*. (2011).

2 Antoniou, S. A. *et al.* Guideline Assessment Project: Filling the GAP in Surgical Guidelines: Quality Improvement Initiative by an International Working Group. *Annals of surgery* **269**, 642-651, doi:10.1097/sla.0000000000003036 (2019).

3 Guyatt, G. H. *et al.* GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *Bmj* **336**, 924-926, doi:10.1136/bmj.39489.470347.AD (2008).

4 Alhazzani, W. & Guyatt, G. An overview of the GRADE approach and a peek at the future. *Medical Journal of Australia* **209**, 291-292, doi:10.5694/mja18.00012 (2018).

5 Brouwers, M. C. *et al.* Development of the AGREE II, part 1: performance, usefulness and areas for improvement. *Canadian Medical Association Journal* **182**, 1045-1052, doi:10.1503/cmaj.091714 (2010).

6 Brouwers, M. C. *et al.* Development of the AGREE II, part 2: assessment of validity of items and tools to support application. *Canadian Medical Association Journal* **182**, E472-E478, doi:10.1503/cmaj.091716 (2010).

7 Schünemann H, B. J., Guyatt G, Oxman A. GRADE handbook for grading quality of evidence and strength of recommendations. *The GRADE Working Group*.

8 Brouwers, M. C., Kerkvliet, K. & Spithoff, K. The AGREE Reporting Checklist: a tool to improve reporting of clinical practice guidelines. *BMJ* **352**, i1152, doi:10.1136/bmj.i1152 (2016).

9 World Health Organization. WHO Handbook for Guideline Development. *http://apps.who.int/iris/bitstream/10665/75146/1/9789241548441_eng.pdf* (2012).

10 NICE, T. g. m. G. a. g. https://www.nice.org.uk/process/pmg6/chapter/reviewing-the-evidence. *Accessed October 2, 2017*.

11 Norris, S. L. *et al.* The skills and experience of GRADE methodologists can be assessed with a simple tool. *Journal of clinical epidemiology* **79**, 150-158.e151, doi:10.1016/j.jclinepi.2016.07.001 (2016).

12 Tavakol, M. & Dennick, R. Making sense of Cronbach's alpha. *Int J Med Educ* **2**, 53-55, doi:10.5116/ijme.4dfb.8dfd (2011).

13 Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika* **16**, 297-334, doi:10.1007/bf02310555 (1951).

14 Nunnally, J. C. *Psychometric theory*. (McGraw-Hill, 1978).

15 Bhattacherjee, A. *Social science research : principles, methods, and practices*. (2012).

16 Pearson, K. & Pearson, E. S. ON POLYCHORIC COEFFICIENTS OF CORRELATION. *Biometrika* **14**, 127-156, doi:10.1093/biomet/14.1-2.127 (1922).

17 Olsson, U. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika* **44**, 443-460, doi:10.1007/BF02296207 (1979).

18 Reckase, M. D. *Multidimensional Item Response Theory*, 2009).

19 R Core Team. R: A language and environment for statistical computing. R

Foundation for Statistical Computing,. *Vienna, Austria. URL https://www.R-project.org/* (2020).

20 StataCorp. Stata Statistical Software: Release 14. *College Station, TX: StataCorp LP* (2015).

21 AGREE II Extension for Surgical Guidelines. *Available in: https://gap-project.org* **Accessed March 30, 2020**.

APPENDIX

TABLES

Table 1. AGREE II Instrument

Domain	Items	Description
(1) Scope and Purpose	Item 1	The overall objective(s) of the guideline is (are) specifically described.
	Item 2	The health question(s) covered by the guideline is (are) specifically described.
	Item 3	The population (patients, public, etc.) to whom the guideline is meant to apply is specifically described.
(2) Stakeholder Involvement	Item 4	The guideline development group includes individuals from all relevant professional groups.
	Item 5	The views and preferences of the target population (patients, public, etc.) have been sought.
	Item 6	The target users of the guideline are clearly defined.
(3) Rigour of Development	Item 7	Systematic methods were used to search for evidence.
	Item 8	The criteria for selecting the evidence are clearly described.
	Item 9	The strengths and limitations of the body of evidence are clearly described.
	Item 10	The methods for formulating the recommendations are clearly described.
	Item 11	The health benefits, side effects, and risks have been considered in formulating the recommendations.
	Item 12	There is an explicit link between the recommendations and the supporting evidence.
	Item 13	The guideline has been externally reviewed by experts prior to its publication.
	Item 14	A procedure for updating the guideline is provided.
(4) Clarity of Presentation	Item 15	The recommendations are specific and unambiguous.
	Item 16	The different options for management of the condition or health issue are clearly presented.
	Item 17	Key recommendations are easily identifiable.
(5) Applicability	Item 18	The guideline describes facilitators and barriers to its application.
	Item 19	The guideline provides advice and/or tools on how the recommendations can be put into practice.
	Item 20	The potential resource implications of applying the recommendations have been considered.
	Item 21	The guideline presents monitoring and/or auditing criteria.
(6) Editorial Independence	Item 22	The views of the funding body have not influenced the content of the guideline.
	Item 23	Competing interests of guideline development group members have been recorded and addressed.

Table 2. Cronbach's alpha modification if item deleted

Domain	Item	Cronbach's Alpha before item's deletion	Cronbach's Alpha after item's deletion
1."Scope and Purpose"	3	0.57	0.73
3."Rigour of Development"	13	0.85	0.86
4."Clarity of Presentation"	17	0.77	0.81
5."Applicability"	21	0.73	0.73

Table 3.1. Polychoric correlation coefficients between the items of Domain 1: "Scope and Purpose"

Scope and Purpose		Item 1	Item 2	Item 3
Polychoric coefficient	Item 1	1	0.5	-0.02
	Item 2		1	-0.04
	Item 3			1

Table 3.2. Polychoric correlation coefficients between the items of Domain 2: "Stakeholder Involvement"

Stakeholder Involvement		Item 4	Item 5	Item 6
Polychoric coefficient	Item 4	1	0.17	-0.73
	Item 5		1	-0.12
	Item 6			1

Table 3.3. Polychoric correlation coefficients between the items of Domain 3: "Rigour of Development"

Rigour of Development		**Item 7**	**Item 8**	**Item 9**	**Item 10**	**Item 11**	**Item 12**	**Item 13**	**Item 14**
Polychoric coefficient	Item 7	1	0.66	0.56	0.6	0.35	0.44	0.29	0.11
	Item 8		1	0.55	0.29	0.38	0.47	0.45	0.33
	Item 9			1	0.54	0.43	0.6	0.21	0.23
	Item 10				1	0.37	0.41	-0.22	-0.16
	Item 11					1	0.39	-0.02	0.09
	Item 12						1	0.32	0.22
	Item 13							1	0.64
	Item 14								1

Table 3.4. Polychoric correlation coefficients between the items of Domain 4: "Clarity of Presentation"

Clarity of Presentation		**Item 15**	**Item 16**	**Item 17**
Polychoric coefficient	Item 15	1	0.53	0.35
	Item 16		1	0.31
	Item 17			1

.

Table 3.5. Polychoric correlation coefficients between the items of Domain 5: "Applicability"

Applicability		**Item 18**	**Item 19**	**Item 20**	**Item 21**
Polychoric coefficient	Item 18	1	0.38	0.65	0.63
	Item 19		1	0.59	-0.12
	Item 20			1	0.5
	Item 21				1

Table 3.6. Polychoric correlation coefficients between the items of Domain 6: “Editorial Independence”

Editorial Independence		**Item 22**	**Item 23**
Polychoric coefficient	Item 22	1	0.43
	Item 23		1

Table 4. (Excel file) Kendall's tau correlation coefficients between all items

Table 5. Exploratory Factor Analysis Rotated Component Matrix

The root means square of the residuals (RMSR) is 0.06, which is acceptable since this value should be close to 0. The root mean square error of approximation (RMSEA) index is 0.077(95% CI (0.054,0.102)), indicating acceptable model fit. Finally, the Tucker-Lewis Index (TLI) is 0.829 which is nearly acceptable; values of TLI over 0.9 are considered to represent a satisfactory fit.

Rotated Component Matrix					
Domain	**Items**	**Factor loadings**			
		1	**2**	**3**	**4**
Scope and Purpose	Item 1	**0.397**	-0.407		
	Item 2	**0.472**	-0.313	0.413	
	Item 3		**0.304**		
Stakeholder Invovement	Item 4			**0.723**	
	Item 5				**0.693**
	Item 6			**-0.943**	
Rigour of Development	Item 7	**0.733**			
	Item 8	**0.536**			
	Item 9	**0.695**			
	Item 10	**0.759**		-0.301	
	Item 11	**0.462**			
	Item 12	**0.579**			
	Item 13				**0.894**
	Item 14		**0.683**	-0.342	0.405
Clarity of Presentation	Item 15	**0.663**			
	Item 16	**0.648**			
	Item 17	**0.381**			
Applicability	Item 18		**0.644**		
	Item 19		0.33		**0.427**
	Item 20		**0.815**		
	Item 21		0.407		**0.453**
Editorial Independence	Item 22		0.**764**		
	Item 23			**0.731**	

Table 6. Exploratory Factor Analysis

Factors	Items	Sum of Squared loadings	Proportion of Variance	Cumulative Proportion of Variance	Proportion Explained	Cumulative Proportion
1	1,2,7,8,9,10,11,12,15,16,17	4.2	18%	18%	32%	32%
2	3,14,18,20,22	3.47	15%	33%	27%	59%
3	4,6,23	2.58	11%	45%	20%	79%
4	5,13,19,21	2.69	12%	56%	21%	100%

Table 7. Item Response Theory Analysis for Domain 1 "Scope and Purpose"

Domain 1: Scope and Purpose				
Item	**Item Content**	**a (SE)**	**p-value**	**95% Conf. Interval**
1	The overall objective(s) of the guideline is (are) specifically described.	7.1 (7.5)	0.34	(-7.60, 21.80)
2	The health question(s) covered by the guideline is (are) specifically described.	1.67 (0.49)	<0.001	(0.79, 2.55)
3	The population (patients, public, etc.) to whom the guideline is meant to apply is specifically described.	0.32 (0.26)	0.21	(-0.18, 0.84)

a: Estimated loadings; SE: Standard Error, CI: Confidence Interval

Table 8. Item Response Theory Analysis for Domain 3 "Rigour of Development"

Domain 3: Rigour of Development				
Item	**Item Content**	**a (SE)**	**p-value**	**95% CI**
7	Systematic methods were used to search for evidence.	2.03 (0.48)	<0.001	(1.09, 2.96)
8	The criteria for selecting the evidence are clearly described.	2.17 (0.53)	<0.001	(1.13, 3.21)
9	The strengths and limitations of the body of evidence are clearly described.	6.00 (2.03)	0.003	(2.03, 9.98)
10	The methods for formulating the recommendations are clearly described.	2.83 (0.73)	<0.001	(1.39, 4.27)
11	The health benefits, side effects, and risks have been considered in formulating the recommendations.	1.83 (0.39)	<0.001	(1.05, 2.60)
12	There is an explicit link between the recommendations and the supporting evidence.	3.23 (0.66)	<0.001	(1.94, 4.52)
13	The guideline has been externally reviewed by experts prior to its publication.	0.93 (0.56)	0.097	(-0.17, 2.03)
14	A procedure for updating the guideline is provided.	0.78 (0.37)	0.037	(0.4938, 9.19)

a: Estimated loadings; SE: Standard Error, CI: Confidence Interval

Table 9. Item Response Theory Analysis for Domain 4 "Clarity of Presentation"

Domain 4: Clarity of Presentation					
Item	**Item Content**	**a (SE)**	**p-value**	**95% CI**	
15	The recommendations are specific and unambiguous.	3.59 (1.71)	0.036		(0.23, 6.95)
16	The different options for management of the condition or health issue are clearly presented.	2.85 (1.00)	0.004		(0.88, 4.81)
17	Key recommendations are easily identifiable.	1.38 (0.36)	<0.001		(0.67, 2.10)

a: Estimated loadings; SE: Standard Error, CI: Confidence Interval

Table 10. Cronbach's alpha for new item groups

**If Item 6 is reversed, Cronbach's Alpha is 0.76*

Group	**Cronbach's Alpha (# items)**	**Items**
New Domain 1	0.73 (2)	1, 2
New Domain 2	0.91 (9)	7, 8, 9, 10, 11, 12, 15, 16, 17
New Domain 3	-0.93 *(3)	4, 6, 23
New Domain 4	0.79 (4)	14, 18, 20, 22
New Domain 5	0.69 (4)	5, 13, 19, 21

FIGURES

Figure 1. Scree plot

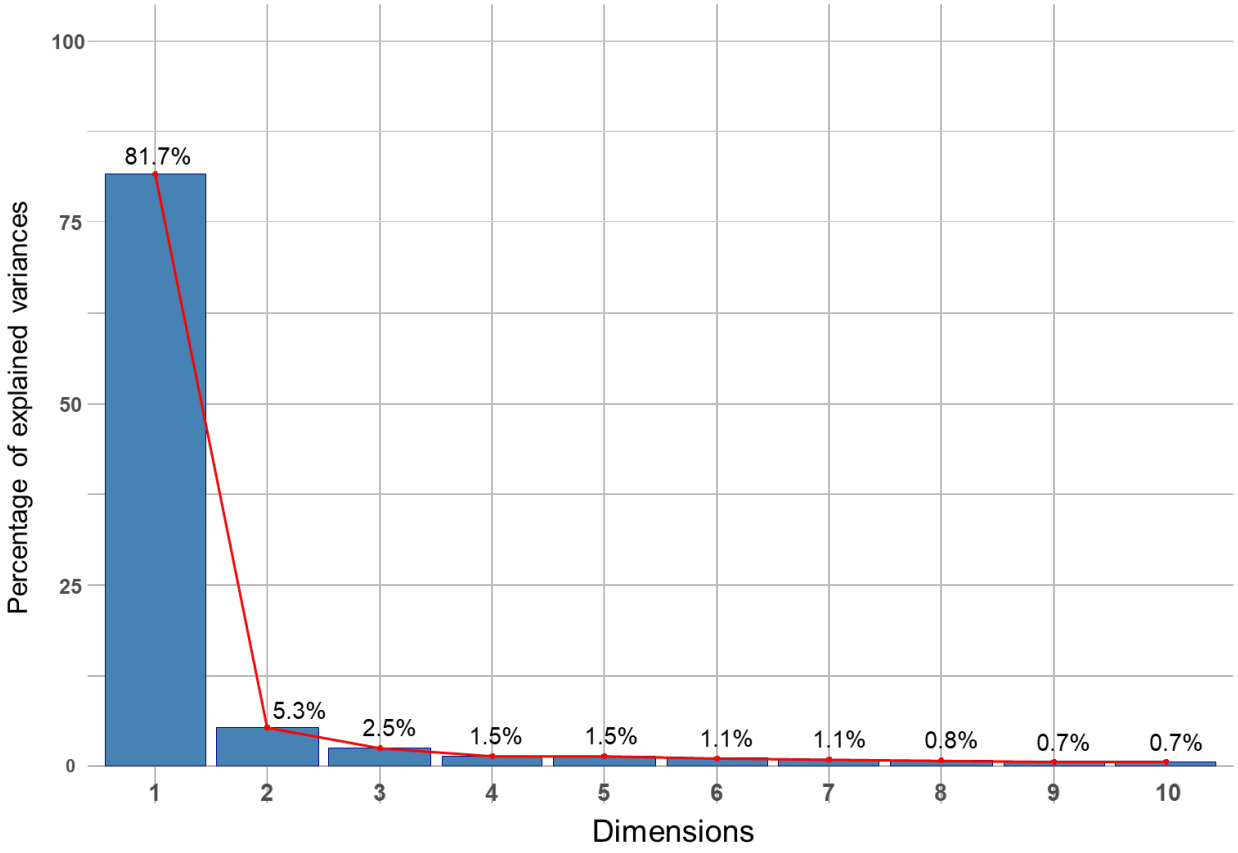

Figure 2. Item Information functions for each item of the "Scope and Purpose" domain

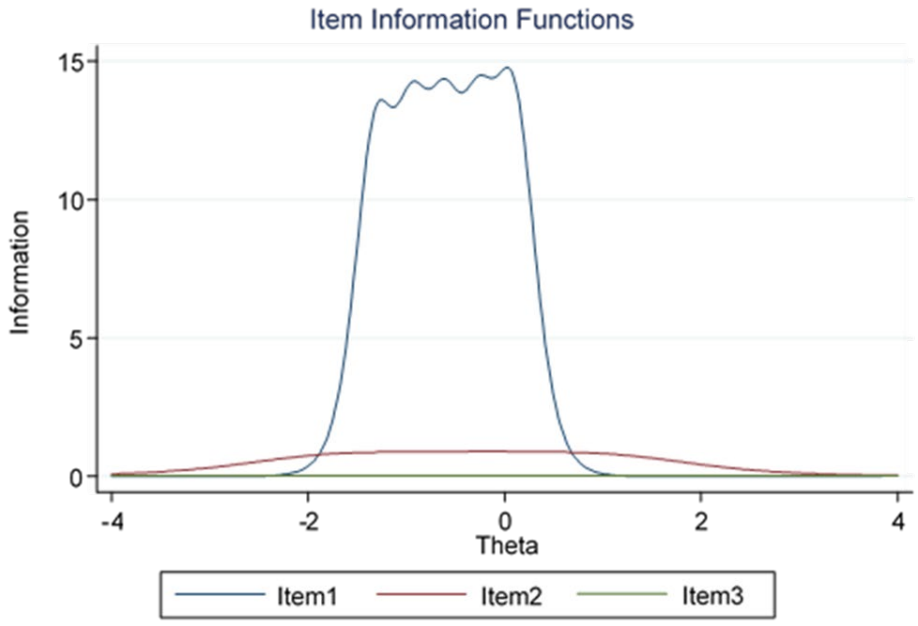

Figure 3. Boundary Characteristic Curves for each (k=1,2,3) item of the "Scope and Purpose" domain

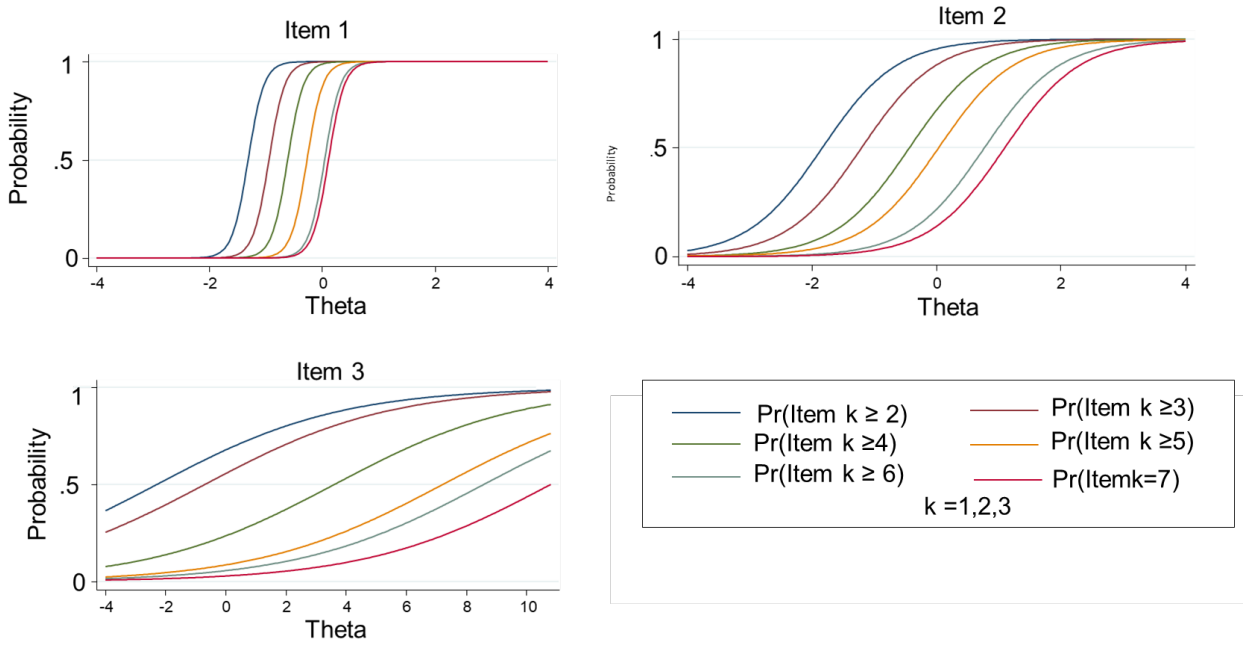

Figure 4. Item Information functions for each item of the "Rigor of Development" domain

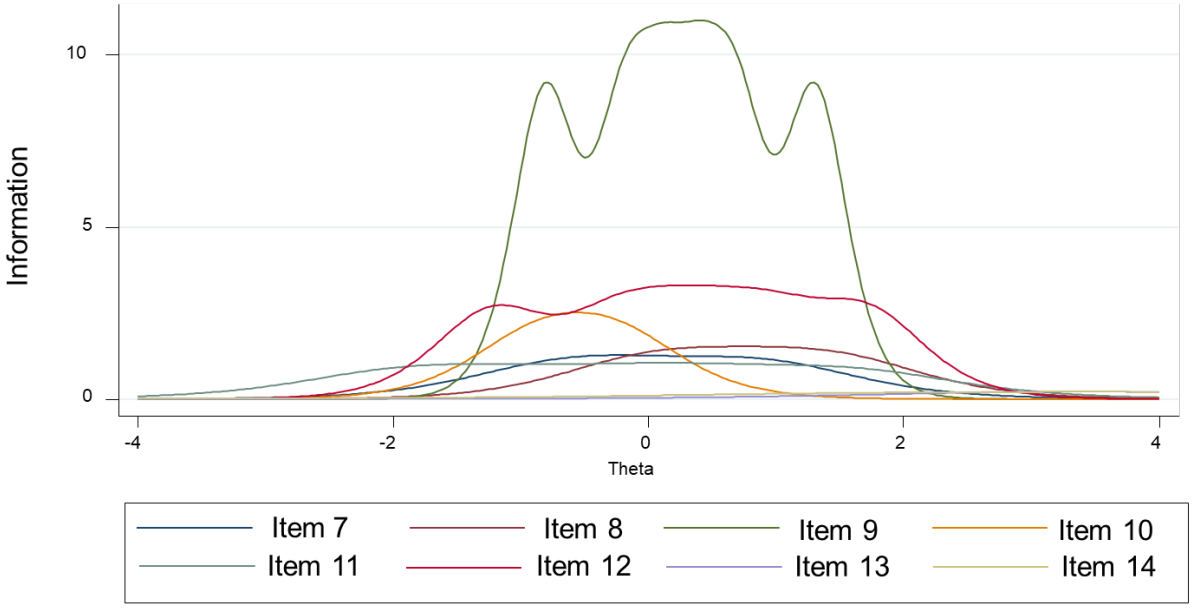

Figure 5. Boundary Characteristic Curves for each (k=11,12,13,14) item of the "Rigor of Development" domain

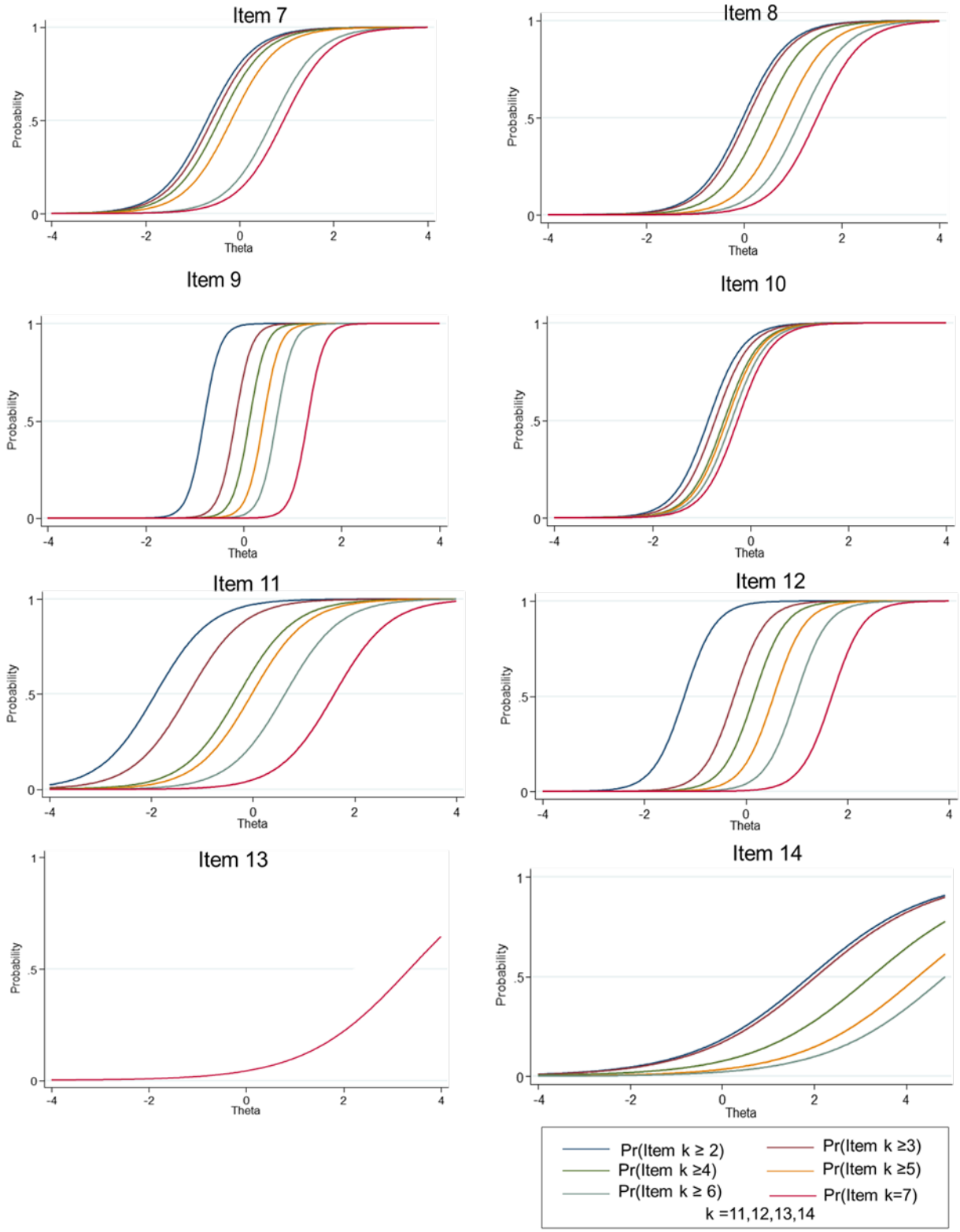

Figure 6. Item Information functions for each item of the "Clarity of presentation" domain

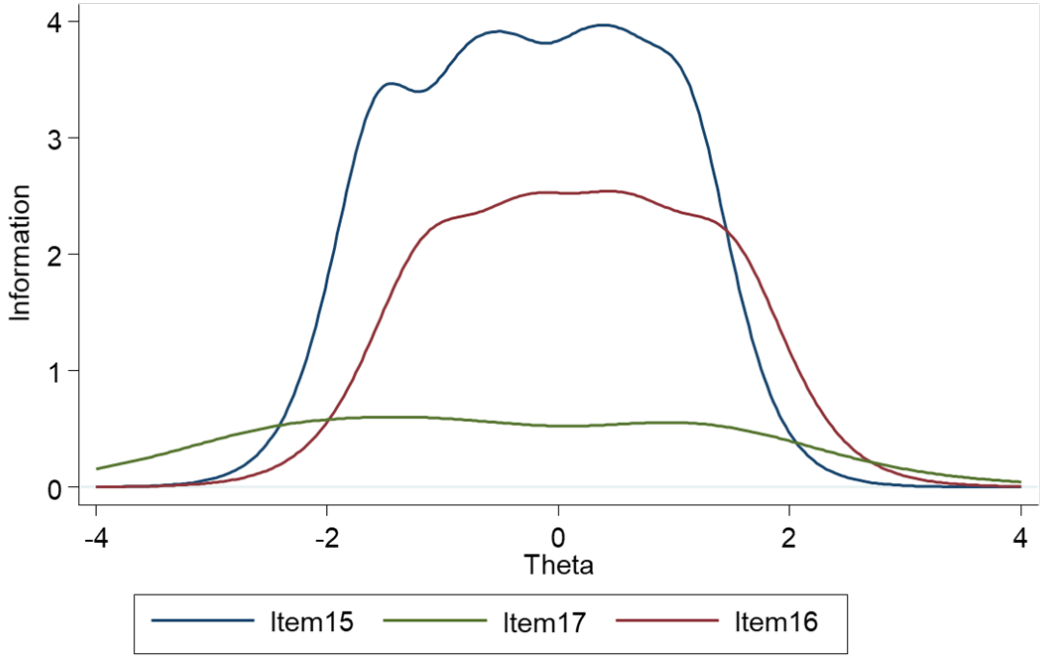

Figure 7. Boundary Characteristic Curves for each (k=11,12,13,14) item of the "Clarity of Presentation" domain