ESTABLISHING CATEGORIES IN THE DESIGN OF RATING SCALES FOR MA-IN-ELT THESES

Gergely A. Dávid and Katalin Piniel
School of English American Studies, Eötvös University Budapest
david.gergely@btk.elte.hu,
brozik-piniel.katalin@btk.elte.hu

Abstract:
This paper focuses on a data-driven method for developing rating scales. The authors describe a project in which a set of analytical scales were developed for the evaluation of MA-in-English Language-Teaching (ELT) theses. An important stage in designing assessment systems is how rating scale categories may be identified, especially in the case of a complex construct, as is determined by requirements of MA theses, and when the categories are not handed down from time-honoured traditions. In this project, there was a conscious effort to make scale development as broadly-based as possible by involving as many of the potential users of the scales as possible. In this article, the authors present insights into some of the key issues that arose during scale development. Although the rating scales are context specific, the steps illustrated here may be useful for those planning to embark on a similar project in their own institutions, as well as for experts involved in assessment in general.

Keywords:
Analytical rating scales, scale construction, scale construction process, assessing theses in Higher Education

1 Introduction

In masters-level higher education contexts, where foreign language majors are generally required to write their theses in the foreign language, the assessment of the theses must necessarily be based on a complex construct. Very broadly, the assessment of a thesis in a foreign language program must cover aspects of content, form and written foreign language proficiency. The present paper describes the process of developing analytical rating scales for the assessment of such written performances.

The specific context of the study is the MA-in-ELT program, revised in 2010, at a university in a non-English speaking country. This program comprises courses in language development, ELT methodology, applied linguistics, English speaking cultures, and courses on pedagogy and psychology for teachers. Exit requirements based on government regulations prescribe that the students write a thesis at the end of their studies.

Students have the assistance of a chosen supervisor while writing their theses on a topic relevant to the teaching of English. The thesis is a research paper, either theoretical or empirical, of about 70 000n (cca. 10 000-11 000 words). Those planning to write an empirical thesis are expected to use various research tools (e.g. questionnaires, interviews, observations) and approaches (e.g. case studies, classroom research) to investigate questions relevant to English language teaching, while those writing a theoretical thesis should discuss a problem regarding English language teaching by providing a synthesis of the literature and demonstrating a complex understanding of the issue at hand. The thesis is given a final grade on a scale of 1 to 5 (with scale point 5 as the best).

The aim of the authors was to design scales for this particular program. They did not intend to impose their personal constructs (Kelly, 1955) on their colleagues, or simply write the scales themselves and then train colleagues towards those concepts, but rather they wanted to construct scales in an accountable way. In particular, this meant that everyone potentially involved in the marking of theses has to take part in identifying what desirable features they would like to see in a good thesis. This entailed collecting colleagues' input systematically and comprehensively and exploring the relationships between various desirable thesis features, before going on to establishing rating categories, rather than arriving at a priori decisions about them. It was also important to have a descriptor created for every single scale point, avoiding the development of undefined and arbitrary scales. Thus, the design process began with the consideration of important issues discussed in the sections below.

The researchers were well aware that they were conducting their research in a resource-poor educational context. Therefore, they knew that double-marking by two wholly external raters was not a possibility due to the limited resources available. The researchers were also aware that involving the supervisor would not be palatable to some professionals in the field, but the only way a second rating was possible in this context was to follow department practice in which supervisors act as second raters.

2 Theoretical background

Given the manifold requirements of supervision (Shanklin & Thurrell, 1996; Swales, 2004), the researchers anticipated that, indicative of a complex thesis construct, staff would identify a large number of desirable features for the scales. This pointed to the need to explore how this complexity would be represented in the scales, with a number of categories still manageable for raters. The researchers, therefore, were to consider the issues of the complexity and heterogeneity of the assumed thesis construct as well as the issue of the scale construction process. In addition to the field of language testing, they were to look for developments in other related and relevant fields. It should be added that this paper is not about writing or the assessment of writing, as it only addresses a few aspects of assessing written performances that are relevant to theses.

2.1 Issues of complexity

The researchers felt that, for this project, they were not going to modify the thesis construct, assumed to exist in the collective awareness of staff, but rely on what experience colleagues had already acquired (Shanklin & Thurrell, 1996). They also felt that the desirable thesis features that colleagues were to identify did not deviate substantially from what is expected of MA theses internationally. An investigation of Swales (2004, pp.99-100), for example, shows that "a strong focus on the real world", the word length, the number of references and, last but not least, the following of the IMRaD structure[1] of research papers are requirements familiar to staff and compatible with current conceptualizations of academic writing (Tankó, 2011).

[1] Introduction, Method, Results and Discussion

Broad construct	Rating scale categories	Number of bands	Points awarded
Form	Format	6	0-5
	English language	6	0-5
Content	Review of the literature	6	0-5
	Analysis	11	0-10

Table 1. The Scales used in the MA in English – British Culture and History Track

A typical problem with rating systems is that the complexity of the construct is not matched by the complexity of the assessment tool. An appropriate example comes from another MA program at the same university, in which the definition of the construct is very simple with the broad and time-honoured distinction made between form and content (Table 1). Form is further specified as format and language, while content is specified as review of the literature and analysis, but beyond that the scales still allow the assessor a lot of leeway in interpreting what exactly is expected, which is probably what Weigle (2002) refers to as "general impression marking" (p.112). The lack of appropriate construct definition in the scales (Table 1) is compounded by the lack of descriptors attached to the numerical scales.

Category label	Number of bands	Points awarded
Task achievement	11	0-10
Coherence and cohesion	6	0-5
Range and accuracy	6	0-5
Appropriacy	6	0-5

Table 2. The structure of Euroexam Scales for EFL Writing until 2012

Table 2 shows writing scales by Euroexams (2007), a public language exam, reproduced here without their descriptors, where the construct appears to be appropriately differentiated. The inclusion of *coherence and cohesion* and *range and accuracy* look sensible to the expert eye because *coherence and cohesion* are well known logically related concepts as are *range and accuracy* (Tankó, 2005). Nonetheless, when the scales are actually used for rating, both combinations are suspect. The question arises about how raters would deal with scripts that are coherent but lack sufficient cohesive features, or how raters will score a paper high on accuracy, but low on range.

2.1.1 Categories, bands and psychology

Professionals in the field argue that there are limits to the number of differences one can reliably make. In the Common European Framework of Reference (CEFR) (Council of Europe, 2001, p.193), the authors state that "received wisdom is that more than 4 or 5 categories starts to cause cognitive overload, and that seven categories is psychologically an upper limit"[2]. Luoma (2004) states that "five to six criteria may be close to the maximum" (p.80). However, a closer inspection of the sources reveals that some authors actually discuss identifying the difference between bands (scale points), rather than the difference between assessment categories. Pollitt (1991) states that it is optimistic "to claim even 5 reliable

[2] The 2018 revision of the CEFR does not include a chapter that makes the same point.

bands" (p.90). Alderson, Clapham, and Wall (1995) recommend scales of "no more than about seven [scale] points, as it is difficult to make much finer distinctions" (p.111). In line with the above, Luoma (2004) suggests "the lower the number of levels [in a scale], the more consistent the decisions" (p.80), adding that in particular testing contexts scales have four to six levels. Weigle (2002) does not recommend a specific number, but only implies an upper limit to the number of points (levels) in a scale (p.123).

All statements by the sources above are driven by the notion of increasing inter-rater reliability through seeking agreement on the fewest possible number of bands. Given that rating differences can never quite be avoided, a different conceptualization of reliability, separation reliability (Linacre, 1998), actually invites the construction of rather more bands, provided, of course, that it is possible to construct meaningful descriptors for each band. In this paradigm, observed rating differences constitute the basis for the variability of information that probabilistic software uses to reliably estimate scale categories, thus suggesting to the researchers how many categories might be needed.

2.2 Heterogeneity of the Construct

The researchers also considered the breadth of the concepts staff would identify. Some of them would appear narrow, while others as fairly broad and complex. Some features would seem to denote exactly the same entity and only be a matter of different wording; some others would show a small difference in meaning, whereas further concepts would show considerable differences. The question the researchers asked was how this long list of concepts was to be combined, accountably, into a smaller set of categories.

This heterogeneity raises the issue of a differential weighting of the categories. Some colleagues would predictably want to give different weights to some categories, according to perceived breadth (inclusiveness) or importance. However, the researchers knew that differential weighting can distort measurement, and equal weighting needs to be restored, or introduced, before the performance of scales can adequately be analyzed for misfitting elements (Linacre, 2006). They were also aware that a number of pedagogical considerations might call for differential nominal weights in advance (Alderson, Clapham & Wall, 1995). They felt, however, that specifying weights in advance would possibly defeat their mission to bring out, throw light on and then operationalize staff constructs. Equal weighting was desirable and more logical, once the proposed set of categories were all to come from an unordered taxonomy.

2.2.1 The Process of Rating Scale Design

Finally, the researchers considered the process of scale development in light of the literature and what they had previously observed around them. In the testing of foreign language speaking, Fulcher (2003) identifies two broad approaches to rating scale design, being either intuitive or data-based. Fulcher's description of data-based methods includes scale design on the basis of observed test discourse (Fulcher, 1996), a method described by Upshur and Turner (1995) as empirically-derived, binary choice, boundary definition (EBB) scales. It also includes what Fulcher (2003) calls "the scaling of descriptors" (p.88), which he associates primarily with North (2000), who, in the development of the illustrative scales for the CEFR (2001), calibrated a large number of descriptors.

Fulcher, Davidson, and Kemp (2011) have applied a different system of categories, classifying methods as either measurement-driven or performance data-based. While it is not entirely clear why the intuitive (a priori) methods would be classified as measurement-driven here, Fulcher et al. (2011) suggest that in such methods descriptors are scaled "to create a scale with a pre-defined number of levels" (p.7), whereas performance data-based methods are "grounded in performance data" (p.9). Thus, Fulcher et al. (2011) move North's CEFR-related scaling work (North, 2000) from the status of data-based scale development in Fulcher (2003) to a status of a measurement-based method (i.e., not performance data-based).

According to the researchers' experience, it is often a small group of professionals, even a single person who writes the descriptors for the many assessors who use them. Although anecdotal evidence from 'industrial lore' does not readily translate into publications that can be referred to, causing real practices to remain hidden, the intuitive approach, coupled with differential weighting, may result in unjustified combinations of categories. The researchers have first-hand knowledge that the scales in Table 1 and Table 2 were developed intuitively. The designers used their judgement but had no information whether these combinations were in fact justified.

In addition, the authors suggest that it may not be reasonable to train examiners heavily for scales whose wording is not their own. Copying, lifting from other scales, rather than developing them on the basis of empirical data has also been witnessed (Fulcher, 2003, p.93), which can jeopardize validity, as any measure developed for a particular purpose may not lead to similar attributes of validity in a different context where the purposes of assessment are most probably also different.

The validity of (scores from) such intuitively developed scales, shaped by a mere handful of people can easily be called into question. The researchers observed that the editing and drafting process may be affected by the status of the participants at work, by variable attendance at and the group dynamics of meetings, rather than theory, the wider practice and research. Therefore, the researchers' goal was to involve as many of the potential raters as possible in the data-driven scale development process. It should be more professional, as well as more valid and ethical, to research the examiners' personal constructs as an alternative (Kelly, 1955) and compile the assessment scales from the insights of as many staff as possible.

2.2.2 Scale Construction in Foreign Language Testing

Although there are publications on designing rating scales, most studies rely on analyzing performance data to arrive at scale categories (e.g., Biber & Gray, 2013). There are also examples of studies that use both performance data as well as rater cognition (e.g., Brown, 2006a; Brown, Iwashita and McNamara, 2005; Cumming, Kantor, and Powers, 2002) to inform the rating scale development process. Finally, there have been researchers who have arrived at rating scale categories by investigating previous literature, looking at documents as well as investigating performance data (e.g., Jin & Mak, 2013). Nonetheless, in the present case, the researchers were not very well served by the literature about how to determine context-specific scale categories in the absence of performance data.

In other projects, the authors focused on the vertical placement of descriptors on a scale and the differences between levels (e.g., Brown, 2006b) rather than on determining the rating categories themselves. Fulcher (1996) developed a single scale, of fluency, while Upshur and Turner (1995) designed descriptors for only two categories. North (2000) qualitatively identified a hierarchy of categories, in advance again (pp.182-183). His descriptors formed a single, very long scale after calibration. The approach of these authors may be characterized by determining the categories a priori and the development of descriptors for predetermined categories. One comparable and highly relevant venture, however, is Chalhoub-Deville's (1995), who replaced a complex set of hierarchical generic and task specific categories with three simpler, generic categories.

Fulcher et al. (2011) propose using Performance Decision Trees (PDTs), in the context of travel agency service encounters. PDTs are a series of binary yes/no decisions and, as Fulcher et al. (2011) admit, bear a strong resemblance to Upshur and Turner's (1995) EBB scales. While Fulcher et al. (2011) provide a decision tree that amounts to a single category, they do not offer a solution for determining scale categories. Nevertheless, their work highlighted some of the most important values (goals) for this project. Apart from PDTs as a technique, these include the need to base the rating scales on performance data, as much as possible, although Fulcher et al. (2011) clearly recognize that some performance data-based techniques are extremely time consuming to apply (p.9).

2.3 Rating Scale Design in Psychology, Health and Related Fields

The researchers also felt the need to collect insights from other fields such as psychology, health and social work, where rating scales are used for a variety of evaluation purposes. These scales demonstrate interesting differences, in comparison with what practitioners expect in language testing. Most importantly, many scales are not presented in a table, with descriptors and attached scale values, but are more like long series of dichotomous or polytomous test items. The reader cannot help noticing that scales so constructed usually have many more categories (items) than might be anticipated on the basis of the literature reviewed above (Council of Europe, 2001; Luoma, 2004). If five or six categories are the maximum that one can pay attention to, the rater would not be able to pay attention to 25 statements, each rated on a 4-point scale (Yanosky, Schwanenflugel, & Kamphaus, 2013); or 25 items, each rated on a 5-point scale (Kivissari, Laasonen, Leppämäki, Tani, & Hokkanen, 2012); or even 15 items, rated with a 7-point scale (Mayes et al., 2012). It appears, in these disciplines, responses from all the items together add up to an overall scale – an instance of terminological difference.

Inspiration also came from Thurstone's (1927a, 1927b) and Edwards' (1957) work. He stated that judgments of sensed differences would fall along a psychological continuum and presented the law of comparative judgment (1927a), which is essentially a formula that can be used to calculate scale values on the basis of paired (pairwise) comparisons. His work is instructive because it assumed that the desirable features to be consolidated in one category were located somewhere along a continuum of psychological distance. In addition, Thurstone's law bears a strong resemblance to the theoretical bases of more recent probabilistic measurement models. Considering the variety of potential features and their breadth, the researchers felt that constructing their scales in a merely intuitive way would not make for accountability and would be open to the various threats discussed above. The complexity of the thesis construct assumes a complex assessment tool with a number of rating

categories and which, consequently, points to developing analytical scales rather than a single holistic one.

Considerations of psychology and measurement requirements, in addition to reasons of practicality and feasibility, demand that a large number of desirable features be consolidated into a smaller number of categories in most educational contexts, when rating scales are used as assessor-oriented scales (Alderson, 1991). The central problem for this project was how many assessment points of view (and in what combination) should be allowed as a maximum so that the rating could still be done reliably, and the complexity of the thesis construct be maintained to the largest extent possible.

3 Methods

The aim of the research was to develop analytical scales for the assessment of theses in an MA-in-ELT program, using a data-driven framework, based on the considerations outlined above. The main research question guiding the study was as follows: How can a set of analytical scales be developed in an accountable way?
This question can be broken down into the following more specific ones:

1. What desirable features should good MA theses demonstrate?
2. How can a large number of features be consolidated into a smaller number of categories (as staff members were expected to identify many features)?
3. How can descriptors for each band be formulated in a way that they capture levels of performance?
4. How should the appropriacy of the scales, obtained through PDTs (EBBs), be checked?
5. How good were the scales shown to be, when tested with data from ratings?

Research comprising five phases of a mixed methods design (Tashakkori & Teddlie, 2003) was devised. Essentially, the researchers chose an empirical approach including both qualitative and quantitative data collection and analysis. The first phase was qualitative, in which they explored their colleagues' views asking them what features they thought were criterial. In the second, quantitative phase they used a questionnaire to explore the multiple relationships between the criterial features, with the purpose of consolidating them into rating categories (Glaser & Strauss, 1967). In the third, a branching approach (Upshur & Turner, 1995) was followed, leading to the writing of descriptors. In the fourth phase, the content of the emerging descriptors was scaled. Finally, in the fifth, the scales were tested on data from thesis ratings, which were analyzed statistically. For an overview of the methodological framework and approaches taken in particular phases, see Table 3.

3.1 Participants

All colleagues currently teaching in the program, thus supervisors and markers of the theses (N = 40), were asked to participate. It was an important principle, and an important feature of accountability, to try to collect input in a way that all staff could contribute. For this reason, there was less emphasis on meetings, where attendance would unavoidably vary, and more on techniques, such as questionnaires, that ensure the independence of the respondent and prevent dominance by authority figures. With the exception of a few, the majority have over ten years' experience of supervision in higher education.

3.2 Instruments and Procedures

In the following sections the procedures are presented, which is necessary in order to appreciate the thinking behind the decisions. What may seem like results of the research are, in fact, procedural results, necessary for a proper appreciation of the answers to the research questions. Final results will be discussed in the relevant section following these procedural results.

Step	Purpose	Methodology and data	Data analysis procedures	Justification	Expected results
1	Explore potential raters' expectations of the performance	Qualitative, Raters provide descriptions of ideal performances	Constant comparative method (Maykut & Morehouse, 1994)	Qualitative approach for data collection allows for new aspects to emerge	Preliminary list of criteria for assessment
2	Investigate psychological distance between features (see Results of step 1) creating a manageable number of categories for the raters	Quantitative, Participants' responses on paired-contrasts questionnaire indicating relationship between pairs of features (Thurstone, 1927a)	Facets (Linacre, 2006), X^2 statistic Multi-Dimensional Scaling (SPSS)	Allows gathering quantitative evidence for the distance between categories	A manageable number of main criteria for assessment
3	Create levels and band descriptors for each category	Qualitative, Descriptions of ideal performances by potential raters	Performance Decision Trees (Fulcher et al., 2011)	PDTs to identify features with clear characteristics of the different levels	Band descriptors for each category
4	Scaling descriptors' content elements	Qualitative and quantitative, Field notes of the discussions of focus-group interviews. Second questionnaire data	Emerging issues Facets	Ensure transferability and credibility across raters	Refining the descriptors for each category and each band
5	Test run	Quantitative Real performance data	Many-Facet Rasch Measurement analysis with Facets (Linacre, 2006)	Many-Facet Rasch Measurement (Linacre, 2006) allows researchers to assess the functioning of the rating scale (categories, bands and raters)	Suggestions for ways to further fine-tune the rating scale

Table 3. Tabular Overview of Phases and Methodologies

3.2.1 Eliciting the desirable features of an MA-in-ELT thesis.

In the first, exploratory phase, participants were asked to write short, 100-word passages about what features they found important in good MA theses. These definitions (n = 13) were analyzed using the constant comparative method (Maykut & Morehouse, 1994), with the emphasis on the "what" (i.e., what features colleagues considered desirable) rather than on "how" or "how well" (i.e., the degree to which criterial features may be observed), hence the focus on nouns and noun phrases, rather than adjectives. In this way, 21 desirable features were identified, which included such concepts as "analytical framework", "citation conventions" or "familiarity with the literature" (full list in Table 4).

1. Analytical framework	13. Originality
2. Argumentation	14. Quality and number of sources
3. Citation conventions	15. Quality of research
4. Contribution to the field	16. Quality of writing
5. Enhanced awareness	17. Reporting of research
6. Focus	18. Research methods and procedures
7. Familiarity with the literature	19. Structure of writing
8. Formal requirements	20. Synthesis of knowledge and skills
9. Implications	21. Theoretical and experiential basis
10. Independence	
11. Interpretation of findings	
12. Layout	

Table 4. The Initial 21 Thesis Features as Identified by Staff

3.2.2 The paired contrasts questionnaire and its outcomes.

As expected, the list of 21 features could not be used to evaluate theses. The researchers deemed they were too many for the readers to reliably work with. Apart from this, the items on the list were heterogeneous, including broad concepts such as "quality of research" as well as some with a more limited scope, such as "layout", while others appeared either identical or very close, such as "structure of writing" and "argumentation". As a result, in the second phase, the goal was to consolidate the list into fewer categories. For this purpose, a long questionnaire made up of paired contrasts was devised in which each feature was contrasted with every other feature except itself. Thus, the questionnaire comprised 210 items ($(n \times (n-1))/2 = 210$ comparisons).

The instrument was expected to shed light on the psychological distance between the concepts in each pair (cf. Thurstone, 1927a, law of comparative judgments). Participants were asked to respond to each item and indicate whether they saw

- the paired concepts as identical with each other (no psychological distance between them),
- one of the concepts as part of the other (part/whole relationship, with little psychological distance in between),
- the paired concepts as near synonymous (some psychological distance between them),
- the paired concepts as different (with a considerable distance between them).

As the four choices above imply a scale, respondents' answers were assigned numerical values (*identical* - 0, *part/whole* - 1, *near synonymous* - 2, *different* -3). Out of 40 staff, 33 filled in the questionnaire.

The consistency of the questionnaire was determined statistically, using SPSS version 17.0. The instrument had a Cronbach's alpha value of 0.97. The fit characteristics of the data were also examined using Facets (Many-facet Rasch Measurement, MFRM) (Linacre, 2006). This meant checking whether some items or respondents generate unlikely and improbable responses, thereby generating variance that cannot be explained on the basis of the measurement model. The researchers identified one feature, "reporting of research", as misfitting, of the proposed 21. Out of the seven contrast pairs that showed considerable misfit, "reporting of research" was present in four. Indeed, respondents may have had difficulty establishing the relationship between "reporting of research" and the other features since all the features are, one way or other, a report on some research aspect. Likewise, the analysis identified three respondents as misfitting. This included one person, for example, who apparently only thought the features to be different from others, therefore, making no attempt to establish the links between features. As a result, the one misfitting feature and the three participants above were eliminated from further analyses.

In order to evaluate the psychological distance between the remaining 20 features, that is, to establish larger categories, a series of non-parametric tests (one-way χ^2, $p < .05$) was used. These analyses would not only serve as a basis for merging the features, but they would also lead to categories that are still informative and manageable for raters. Going through as many χ^2 distributions as questionnaire items yielded important information about which features may be justifiably combined into a category. Some contrast pairs, for example item 12 in Table 5, showed the uncertainty of the respondents and failed to show significant differences. Some other differences were significant, however, and testified to a good measure of agreement among the participants. For example, item 48 showed that, although the features of "originality" and "independence" were not seen as identical by many, the fact that 24 respondents out of 30 thought they were related convinced the researchers that these features could be brought together in the same category. Similarly convincing responses were given to item 110, where altogether 26 respondents agreed that these features were related.

The data also provided information about which features should not be combined. Building the "independence/originality" category gained further support from the counter-examples of items 81 and 123, as the responses suggested that "originality" should not be grouped with either "familiarity with the relevant literature" or with observing "formal requirements". Thus, the responses implied that "originality" is not to be expected from studying the literature or formal requirements. Following up further contrast pairs, the researchers received support even from negative evidence. For example, items 131 and 156 both showed respondents as rather divided, leading to the conclusion that "originality", and consequently "independence", should not be placed in the same category where either "synthesis of the literature" or the "use of an analytical framework" will be placed.

Item	**Contrast**	**Response choices and frequencies**				χ^2	**Sign. (*p*)**	***Df.***
		Identical	Part/ whole	Near-synonym	Different			
12	Implications of research vs. originality	--	11	5	14	4.2	0.12	2
48	Originality vs. independence	2	10	14	4	12.13	0.07	3
110	Originality vs. contribution to the field	2	12	14	2	16.4	0.001	3
81	Originality vs. familiarity with the literature	--	7	--	23	8.53	0.003	1
123	Originality vs. citation conventions	--	1	--	29	26.1	0.001	1
131	Originality vs. synthesis of knowledge and skills	--	16	1	13	12.6	0.002	2
156	Originality vs. analytical framework	--	10	1	19	16.2	0.000	2

Table 5. Example Contrasts from the Questionnaire

The researchers also used Multidimensional Scaling (MDS) to confirm the tentative categories they had at this point. The MDS analysis was based on exactly the same response data that were obtained for the questionnaire, but the 30 respondents by 210 contrast pairs was not immediately suitable for treatment with MDS. Since MDS can only accept fully crossed datasets, in which the 20 remaining features were contrasted with the same 20 features, the mean for each questionnaire item was taken across all 30 non-misfitting respondents and entered in the MDS data matrix. Thus, for questionnaire item 1, contrasting "argumentation" with "analytical framework", for example, the mean was 2.00. In this way, the authors believe, they avoided having to produce a data matrix for each respondent (30 MDS runs to be summarized), hoping the mean might still be able to show something useful. The MDS procedure was successful (Stress = 0.14, R^2 = 0.86), lending additional support to the six distinct categories that had been consolidated from the original 20 (21) identified by staff.

3.2.3 Constructing six plus one binary yes/no questions.

In phase three, following Upshur and Turner (1995), the researchers drafted six series of binary yes/no questions, or performance decision trees (PDTs, Fulcher et al. 2011), one for each consolidated category (see Appendix A). On the basis of the trees, they constructed six scales each with five bands of skill, the content of which was to be scaled in the fourth phase. It should be emphasized that because it was clear from the outset that the quality of the English language in the thesis would also have to be evaluated, "quality of the English

language" was added as a seventh category to the six categories already identified (Appendix B). With this, the authors believe, the number of categories was extended to the limit the raters could be expected to process, as is stated in the language testing literature (Alderson, 1991; Luoma, 2004).

3.2.4 The scaling of descriptors' content elements.

With the researchers as moderators, staff discussed the wording of each descriptor in focus groups. With the wording of the descriptors modified, the draft scales were included in a rating form, a second questionnaire in fact. Then, for every draft scale, a list of content elements was added, which were formulated as single stand-alone statements. There were altogether 42 such content elements. Thus, phase four was what might be referred to as the scaling of the content of the descriptors.

The researchers asked their colleagues to identify the lowest point on a 0-4 scale at which a particular content element should appear. The rationale for the task was that there was a need to check whether the resulting descriptors/scales represented a consensual build from the lowest band to the highest. This scaling effort would certainly not make the (pre)testing of the scales on real data (scores) superfluous, but until that was possible, the researchers thought judgments were needed in order to gain empirical feedback on the drafted scales, from which they should be able to predict how the scales would likely capture the departmental construct.

Twenty-four respondents marked each content element, (the researchers consciously decided to use the values 0 and 4 to avoid coincidence with the 1-5 scale, the traditional marking system in this country). Facets (Linacre, 2006) was used again, first to identify misfitting respondents and content elements. Thus, in two rounds of analyses, having excluded one respondent, the software was in a better position to unambiguously scale content elements onto the 0-4 scale.

In the final, fifth phase, the rating scales were tested with rating data. In compliance with the department requirements for double-marking, supervisors' and referees' scores were collected for all theses from five academic terms, thereby producing a dataset that comprised a total of 78 theses, rated across all seven scales by 25 colleagues who were asked to do the rating, thus forming a three-facet rating situation for the Many-Facet Rasch analysis (Linacre, 2006).

4 Results and Discussion

The research results were, first and foremost, the consolidated features as rating categories (research questions 1-2). Table 6 shows the 20 original features the researchers started from on the left with the six categories they arrived at on the right. In each category, one feature was chosen as the overall label, on the basis of it being the most relevant or the one that seemed to subsume all other features in the group. In this way, for example, the "Research methods and procedures" category was felt to be sufficiently broad to include, as requirements of good research, both an "analytical framework" and "focus", while "quality of research" was considered to be too broad to be useful as the distinguishing overall label for this category.

Regarding category 2, the researchers felt that the "Theoretical and experiential basis" label was appropriately inclusive: Students need to be familiar with the literature, which means choosing the right sources, and the right number of them, as expected at the MA level and being able to synthesize this knowledge in their literature review. The researchers chose "Interpretation of findings" as the label for category 3 as it should entail the discussion of the implications of research as well as an enhanced awareness of the field. For category 4, "Independence" seemed to be the appropriate umbrella term since originality, in the few theses where it might be observed as showing a novel contribution to the field, may be conceived of as a higher form of independence. For category 5, "Formal requirements" were an obvious choice since it includes both citation conventions and layout. For category 6, again, it was felt that "Quality of writing" sufficiently covered both the way the thesis writer deals with facts, data and ideas (argumentation) and the structural aspects of writing.

PDTs (EBBs), included in Appendix A, were used to formulate descriptors (research question 3). The scaling of descriptor content (research question 4) has been described in the procedural results above, with the result that the proposed descriptor contents were checked as to whether they had the desirable build from lowest to highest in the eyes of staff. In terms of quality assurance, that is, the procedures employed to increase validity and accountability, the scales lived up to most expectations when they were tested with real data from the ratings (research question 5). As Table 7 shows, the difficulty of most of the scales was not shown to be widely different, as the moderate distances between the measure values demonstrate, relative to the magnitude of the Standard Error (SE) values. The fit indices of the scales were also all appropriate, falling within 2 Standard Deviations (SD) from the mean. It is perhaps only the "Formal requirements" scale that consistently shows higher fit values across all four quality control indices in Facets, but the unexplained variance, responsible for the high values, does not go beyond 2 SD in this case either.

20 features in phase 1 and 2	**Six consolidated categories**
Quality of research	Category 1: Research methods and procedures
Analytical framework	
Research methods and procedures	
Focus	
Quality and number of sources	Category 2: Theoretical and experiential basis
Theoretical and experiential basis	
Familiarity with the literature	
Synthesis of knowledge and skills	
Enhanced awareness	Category 3: Interpretation of findings
Implications	
Interpretation of findings	
Contribution to the field	Category 4: Independence
Originality	
Independence	
Formal requirements	Category 5: Formal requirements
Citation conventions	
Layout	
Quality of writing	Category 6: Quality of writing
Argumentation	
Structure of writing	

Table 6. Consolidating the 20 Features into Six Categories

The values in Table 7 are not for identifying a possible cause of the high fit values for "Formal requirements" because they themselves are means from across the scale points. The statistical breakdown of the problematic scale in Table 8 shows that the distance between average logit measures for the bands is smaller than desirable (1.4 acc. to Bond & Fox, 2001) between scale points 3 and 4. This is especially true in light of the comparable expected measures. The outfit mean square (mnsq) at scale point 2 is already at 2 SD from the mean (of outfit values) and, the value of 1.7 at scale point 3 constitutes an extreme value, suggesting excess, unexplained (improbable) variation in the responses.

While it may be stated with some confidence that scale points 2 and 3 must be the source of problematic fit values for the "Formal requirements" scale, the researchers could not be certain of what caused the problem. The descriptions did not look problematic in the scales, none of the relevant descriptors were seriously debated in the meetings, nor was there any indication previously in the research that this scale would not function as expected. Therefore, it may only be suggested that some staff have deeply-seated misgivings about the appropriacy of the formal requirements of a research project in a teacher-training program (McDonough, 1996).

Scales	Measure	*S.E.*	Fit indices				Discrimination[3]	
			Infit mean square	Standardized infit (Z)	Outfit mean square	Standardized outfit (Z)	PtBis.	Discrim.
Research methods & procedures	-0.17	0.13	0.78	-2.04	0.77	-1.88	0.55	1.25
Theoretical & experiential basis	0.10	0.15	0.96	-0.27	1.16	1.12	0.49	0.97
Interpretations & findings	0.30	0.15	1.05	0.47	1.15	1.13	0.48	0.93
Independence	-0.81	0.15	0.83	-1.27	0.79	-1.12	0.51	1.14
Formal requirements.	0.16	0.15	1.24	1.93	1.41	2.53	0.42	0.69
Quality of writing	0.22	0.15	0.96	-0.36	0.93	-0.5	0.47	1.05
Quality of English	0.21	0.16	1.02	0.18	0.99	-0.09	0.4	1.00

Table 7. Overall Performance Statistics for the MA in ELT Thesis Rating Scales

A second, nonetheless very important outcome of this project was the application of a procedure that the researchers feel yielded the advantages they expected. With the exception of the first phase, appropriate coverage in the collection of data (saturation of data, Glaser & Strauss, 1967) was achieved for the project, contributing, the authors believe, to the validity of decisions based on the ratings. While only thirteen colleagues (out of $N = 40$) participated in writing the definitions, in the second phase, 33 colleagues (83% of those eligible) provided answers for the paired contrasts questionnaire. In phase four, in which 24 colleagues helped scale the descriptors, the response rate was 60%. The final analyses included all 78 theses (100%) assessed in the program so far and all 25 (100%) colleagues who have so far been asked to mark theses.

Rating scale points (bands)	Score	Points awarded for category	Distribution of responses (%)	Average logit measures	Expected average logit measures	Outfit mean-square
0	0	--	--	--	--	--
1	1	4	3	-1.82	-1.25	0.7
2	2	24	16	0.46	0.14	1.4
3	3	51	34	1.96	1.78	1.7
4	4	73	48	2.98	3.18	1.2

Table 8. Performance Statistics for the "Formal requirements" Scale

[3] S.E.: Standard error of estimate; PtBis.: Point-biserial correlation; Discrim: Discrimination as computed by Facets (Linacre 2006).

With the approach adopted, the researchers made it possible for colleagues to respond independently of each other. The researchers believe they managed to elicit opinions that might have been stifled in a series of meetings, by more vocal members of staff, had it been an intuitive approach to scale construction. The development of the "Independence" category provides an example. The inclusion of originality in the questionnaire was debated by an influential member of staff, on the grounds that originality is not an MA requirement as it is in a PhD program. While this is true, it was also clear that when each responding staff member was allowed in the questionnaire to formulate an opinion, there was a clear pattern the researchers could not ignore: 24 out of 30 respondents thought there was some form of connection between originality and independence. This pattern called for the inclusion of originality in the top band of the "Independence" scale, with the rationale that while most MA theses would not demonstrate an original contribution to the field, outstanding ones may still demonstrate originality.

5 Conclusion

In this paper, a process of developing a set of analytical scales is proposed, for theses written for an MA program. The details lend insight into the considerations involved in drawing up a context-specific rating scale and arrive at an instrument that is not a mere redistribution of existing construct-elements. The researchers have outlined a process that included a balance of both exploratory and confirmatory analyses, primarily from potential raters and colleagues in the program. Since the final product is highly context-specific, the researchers would advise against using it unchanged in other contexts. The developmental process can, however, be transferred to other institutions where staff are left to devise their own scales.

With regard to accountability, the researchers sought to include as many staff as possible in the process, and to take into account the complexity of the construct in the development of the scales. To this end, both qualitative and quantitative approaches were used. Thus, with ideas generated bottom-up, another aim was to develop an instrument with a manageable number of categories. Some of the categories might still seem rather broad, but the researchers believe the scales help regulate the rating process. The descriptors were drawn up to adequately capture levels of performance. Finally, since the ratings for the real rating process were performance data, evidence was gathered for the appropriate functioning, i.e. the reliability of the scales. Based on the initial assessment of the rating instrument, it seems to work well in the given context.

Nevertheless, the study was not without limitations. One principal issue that was beyond the researchers' control was the fluctuating number of participants in the different phases of the project. Another limitation is inherent in context specific scales: the scales developed following the data-based method are context specific with a particular domain and genre in mind; thus, score interpretations cannot be easily generalized across different contexts (Fulcher et al., 2011). As scale development is said to be an ongoing process, further research could assist in fine-tuning the rating scales (the wording of the descriptors and even perhaps the number of the categories) on the basis of rating data to be collected in the future.

Proofread for the use of English by Zsuzsanna Soproni, International Business School, Budapest and Francis J. Prescott, Department of English Language Pedagogy, Eötvös Loránd University, Budapest.

References

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp.71-86). London, UK: Macmillan.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge, UK: Cambridge University Press.

Biber, D., & Gray, B. (2013). Discourse characteristics of writing and speaking task types on the TOEFL ibt® test: a lexico-grammatical analysis. *ETS Research Report Series, 2013*(1), i-128.

Bond, T., & Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Brown, A. (2006a). An examination of the rating process in the revised IELTS Speaking Test. In P. McGovern & S. Walsh (Eds.), *IELTS research reports* 2006 (pp.41–70). Canberra & Manchester: IELTS Australia and British Council.

Brown, A. (2006b). Candidate discourse in the revised IELTS Speaking Test. In P. McGovern & S. Walsh (Eds.), *IELTS research reports 2006* (pp.71–89). Canberra & Manchester: IELTS Australia and British Council.

Brown, A., Iwashita, N., &, McNamara, T. (2005). *An examination of rater orientations and test taker performance on English-for-Academic-Purposes speaking tasks (TOEFL Monograph No. 29).* Princeton, NJ: Educational Testing Service.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*(1), 16-33.

Council of Europe, (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge, UK: Cambridge University Press.

Council of Europe, (2018). *Common European framework of reference for languages: Learning, teaching, assessment. Strasbourg:* Language Policy Programme, Education Policy Division Education Department. *Retrieved January 25, 2019 from https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989.*

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 67-96.

Edwards, A. L. (1957). *Techniques of attitude scale construction.* New York, NY: Appleton-Century-Crofts.

Euroexams. (2007). Practice test. A Complete Set of the B2 Level Euro Exam Papers with Instructions, Answer Key and Audio CD. Set 3. Budapest: Euro Examination Ltd.

Fulcher, G. (1996). Does thick description lead to smart tests? *Language Testing, 13*(2). 208-238.

Fulcher, G. (2003). *Testing second language speaking.* Harlow,UK: Pearson Education Limited.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1) 5-29.

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory.* New York, NY: Aldine.

Jin, T., & Mak, B. (2013). Distinguishing features in scoring L2 Chinese speaking performance: How do they work? *Language Testing, 30*(1): 23-47.

Kelly, G. (1955). *The psychology of personal constructs* (Vols. 1-2). New York, NY: Norton.

Kivisaari, S., Laasonen, M., Leppämäki, S., Tani, P., & Hokkanen, L. (2012). Retrospective assessment of ADHD symptoms in childhood: Discriminatory validity of Finnish translation of the Wender Utah rating scale. *Journal of Attention Disorders, 16*(6), 449-459.

Linacre, J. M. (1998). *A user's guide to Facets and Facform*. [Software manual] Chicago, IL: Mesa Press.

Linacre, J. M. (2006). Facets (Version 3.59) [Computer software]. Chicago, IL: Mesa Press.

Luoma, S. (2004). *Assessing speaking*. Cambridge, UK: Cambridge University Press.

Mayes, S. D, Calhoun, S. L., Murray, M. J., Morrow, J. D., Yurich, K. K. L., Cothren, S., Purichia, H., Mahr, F., Bouder, J.N., & Petersen, C. (2012). Use of the childhood autism rating scale (CARS) for children with high functioning autism or Asperger syndrome. *Focus on Autism and Other Developmental Disabilities, 27*(1) 31-38.

Maykut, P., & Morehouse, R. (1994). *Beginning qualitative research: A philosophic and practical guide*. London, England: Falmer Press.

McDonough, S. (1996). Research methods as part of English language teacher education? Retrieved from: http://www.elted.net/uploads/7/3/1/6/7316005/v3mcdon.pdf

North, B. (2000). *The development of a common framework scale of language proficiency*. New York, NY: Peter Lang Publishing.

Pollitt, A. (1991). Response to Charles Alderson's paper: 'Bands and scores'. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp.87-94). London, UK: Macmillan.

Shanklin, T., & Thurrell, S. (1996). The thesis: theory and practice combined. In P. Medgyes & A. Malderez (Eds.), *Changing perspectives in teacher education* (pp.75-86). Oxford, UK: Heinemann.

Swales, J. (2004). *Research genres*. Cambridge, UK: Cambridge University Press.

SPSS, (2008). SPSS Statistics for Windows (Version 17.0) [Computer software]. Chicago, IL: SPSS Inc.

Tankó, Gy. (2005). *The writing handbook*. Budapest: Teleki László Foundation.

Tankó, Gy. (2011). *Professional writing.* Budapest: Eötvös University Press.

Tashakkori, A., & Teddlie, C. (2003). *Handbook of mixed methods in social and behavioral research.* London, UK: Sage Publications.

Thurstone, L.L. (1927a). A law of comparative judgment. *Psychological Review, 34,* 273-286.

Thurstone, L.L. (1927b). The method of paired comparisons for social values. *Journal of Abnormal Social Psychology, 21,* 384-400.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *English Language Teaching Journal, 49*(1), 3-12.

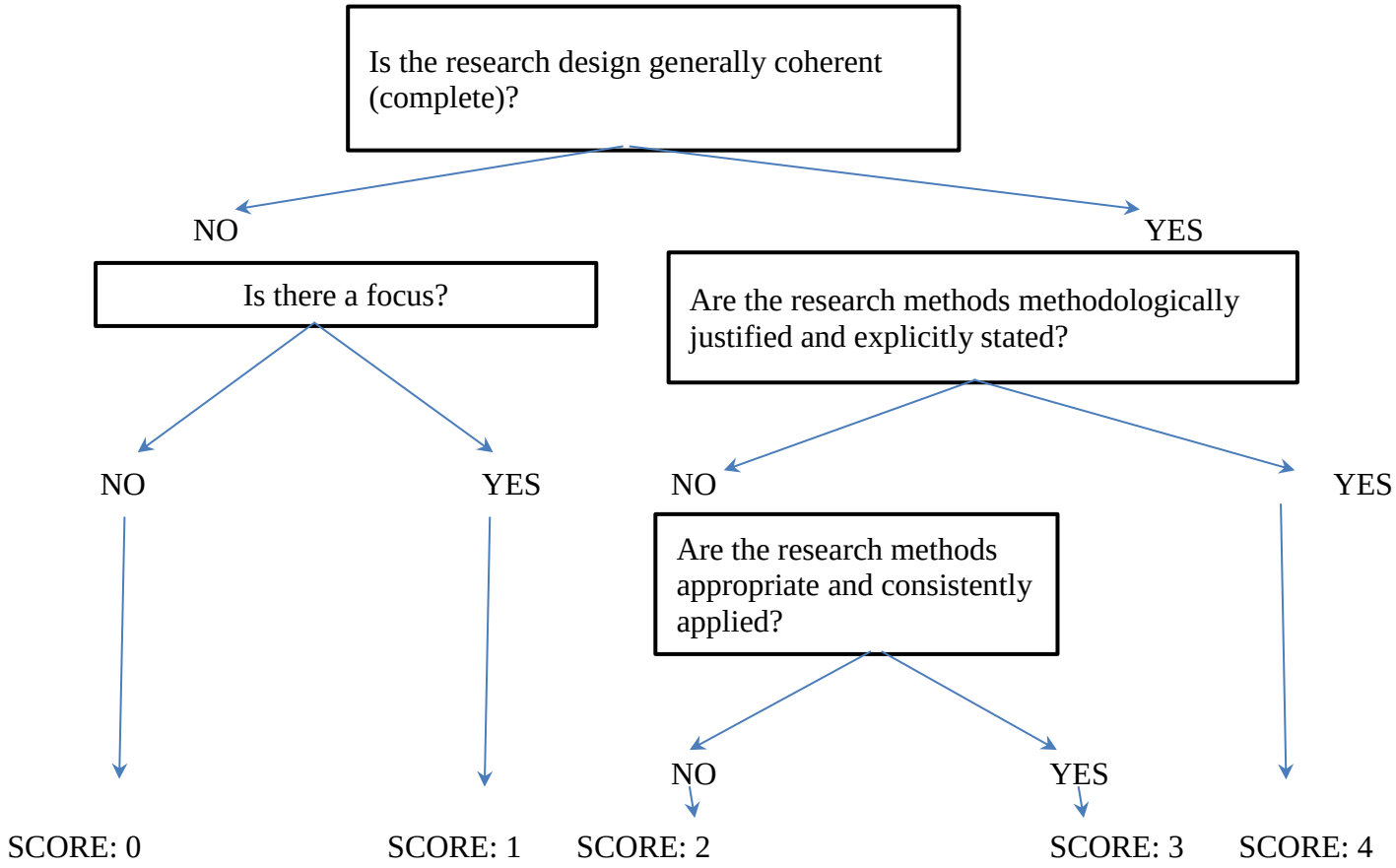
Weigle, S. C. (2002). *Assessing writing.* Cambridge, UK: Cambridge University Press.

Yanosky II, D. J., Schwanenflugel, P. J., & Kamphaus, R. W. (2013). Psychometric properties of a proposed short form of the BASC Teacher Rating Scale – Preschool. *Journal of Psychoeducational Assessment, 31*(4) 351-362.

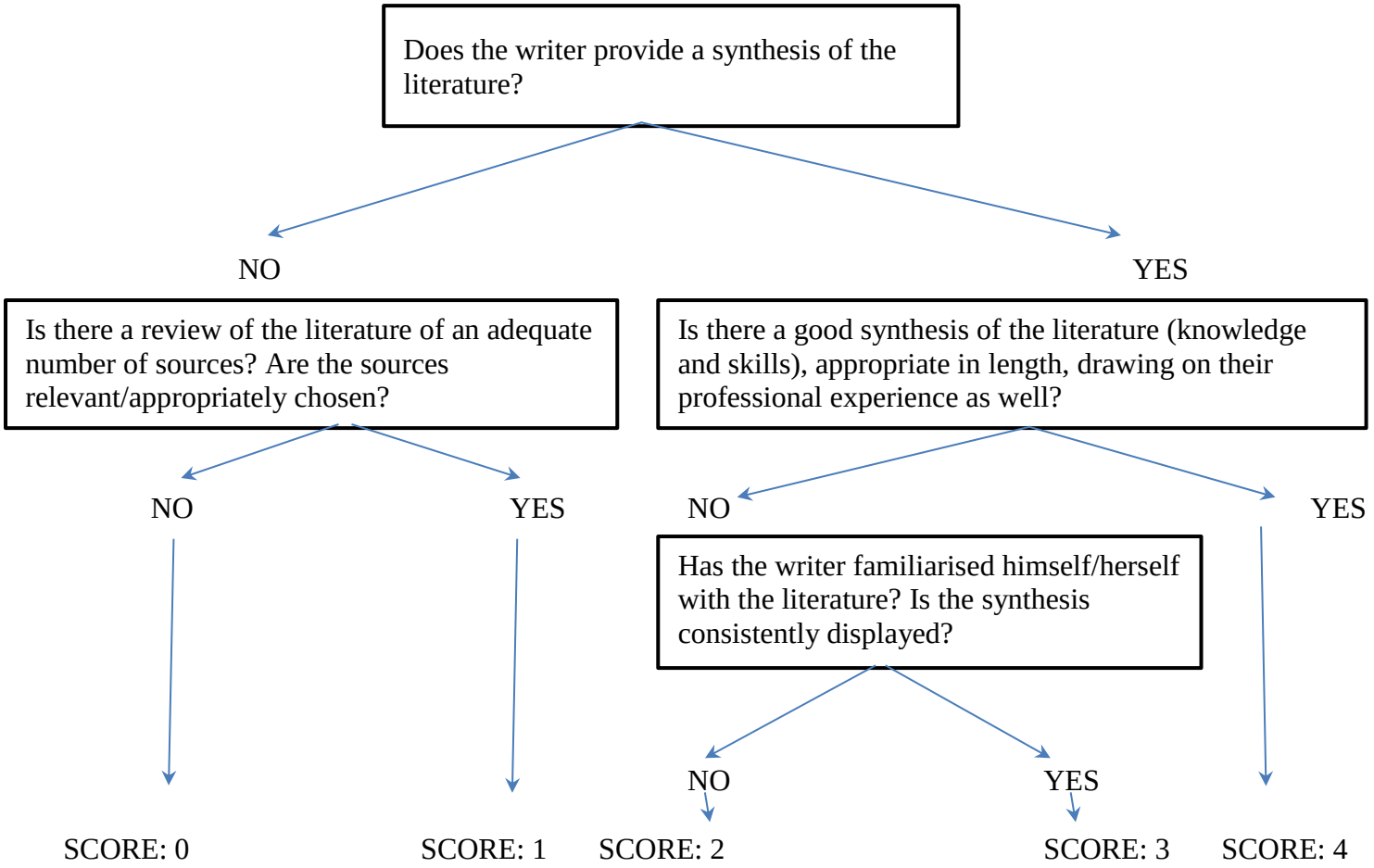
APPENDIX A

Descriptors of the five bands for each category in the rating scales

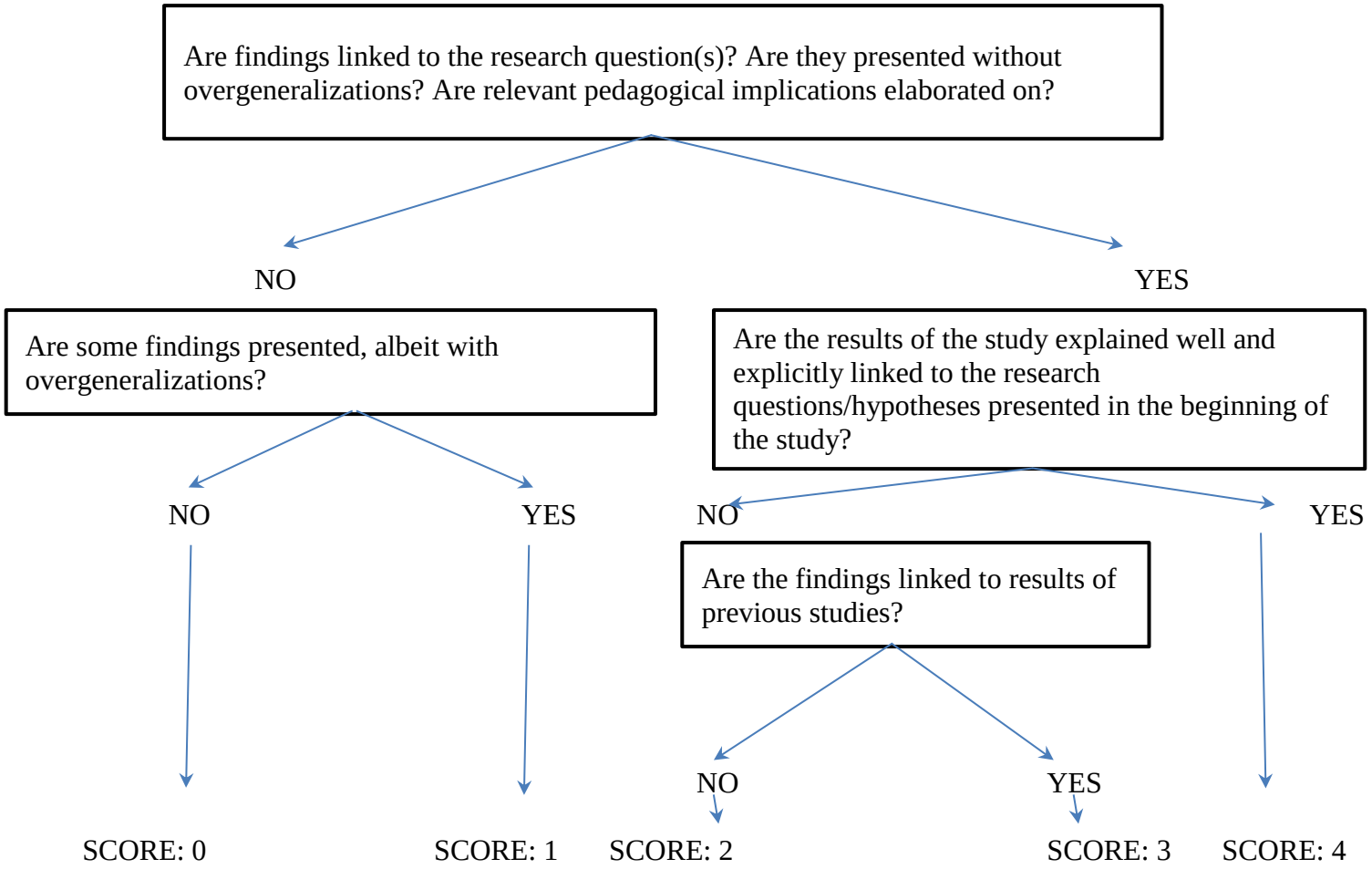
RESEARCH METHOD AND PROCEDURES



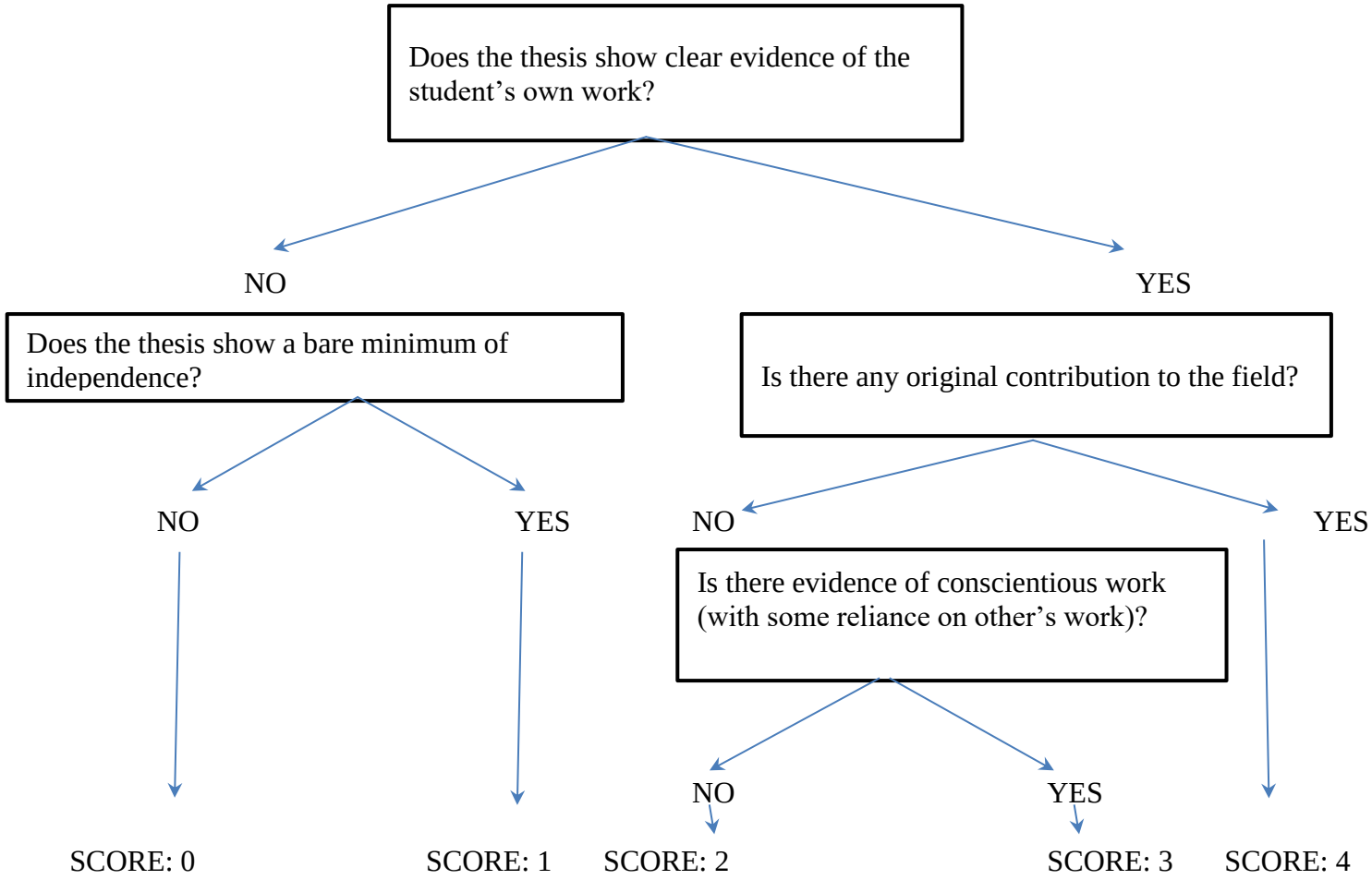
THEORETICAL AND EXPERIENTIAL BASIS



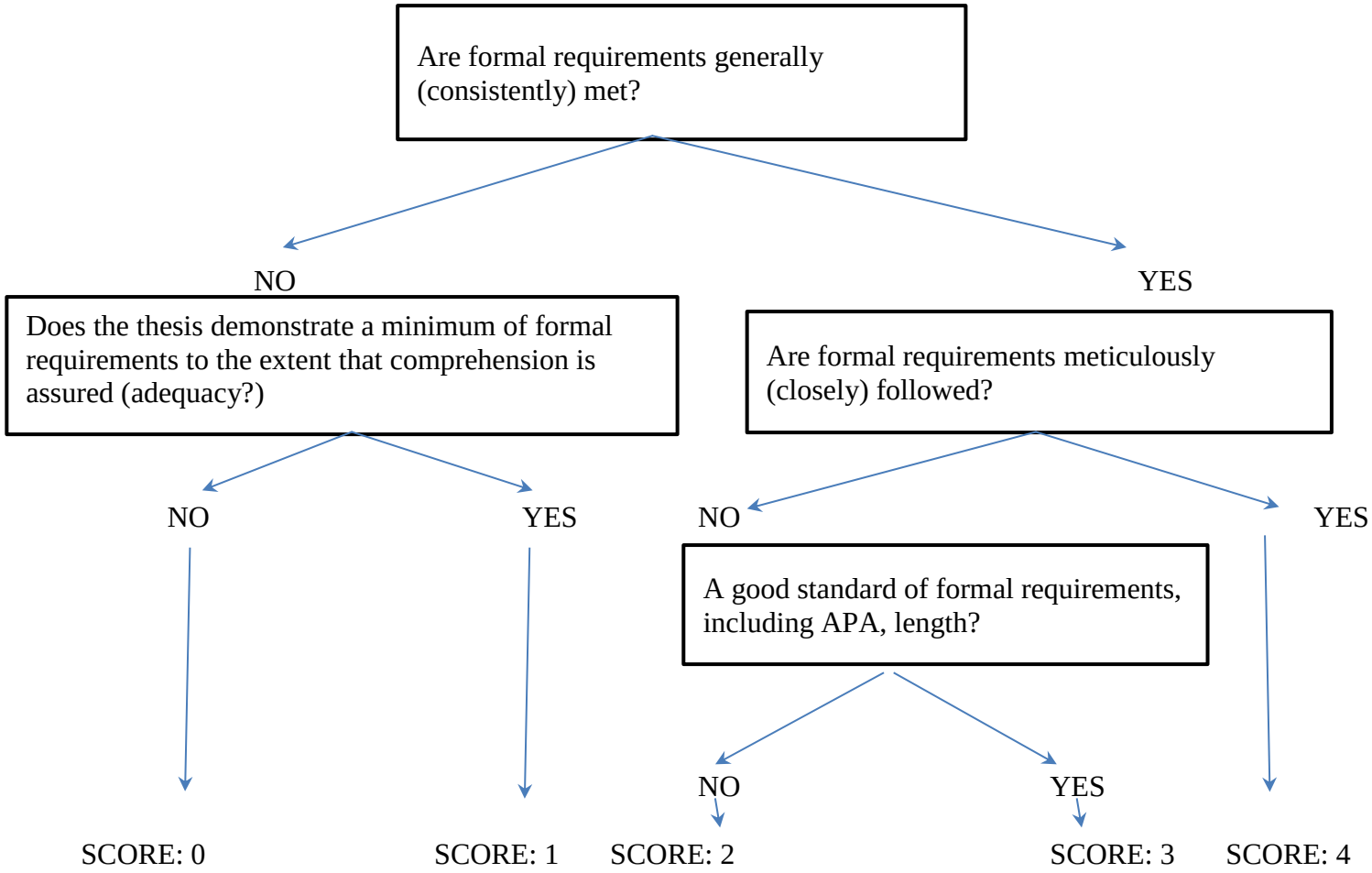
INTERPRETATION OF FINDINGS



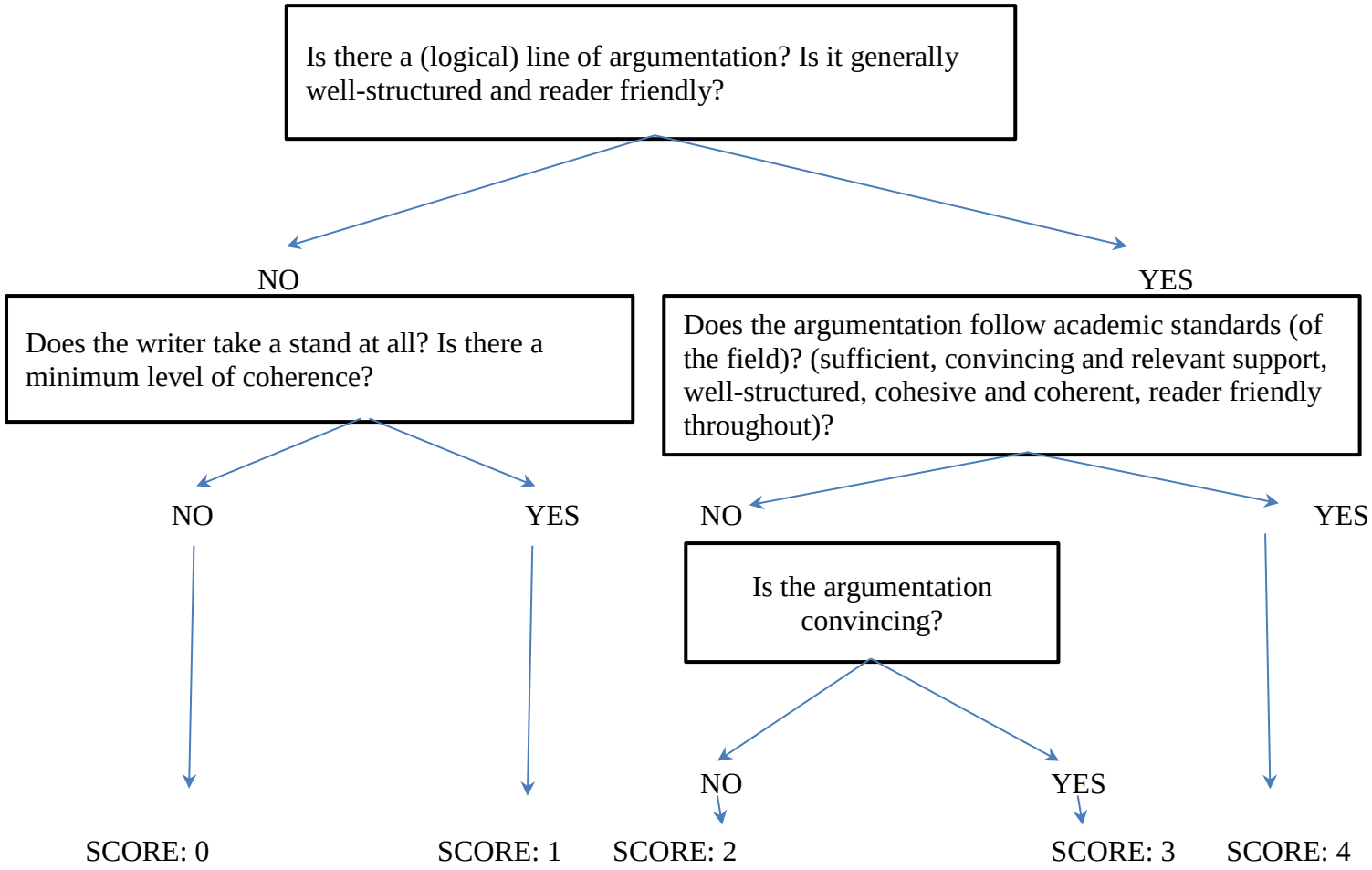
INDEPENDENCE



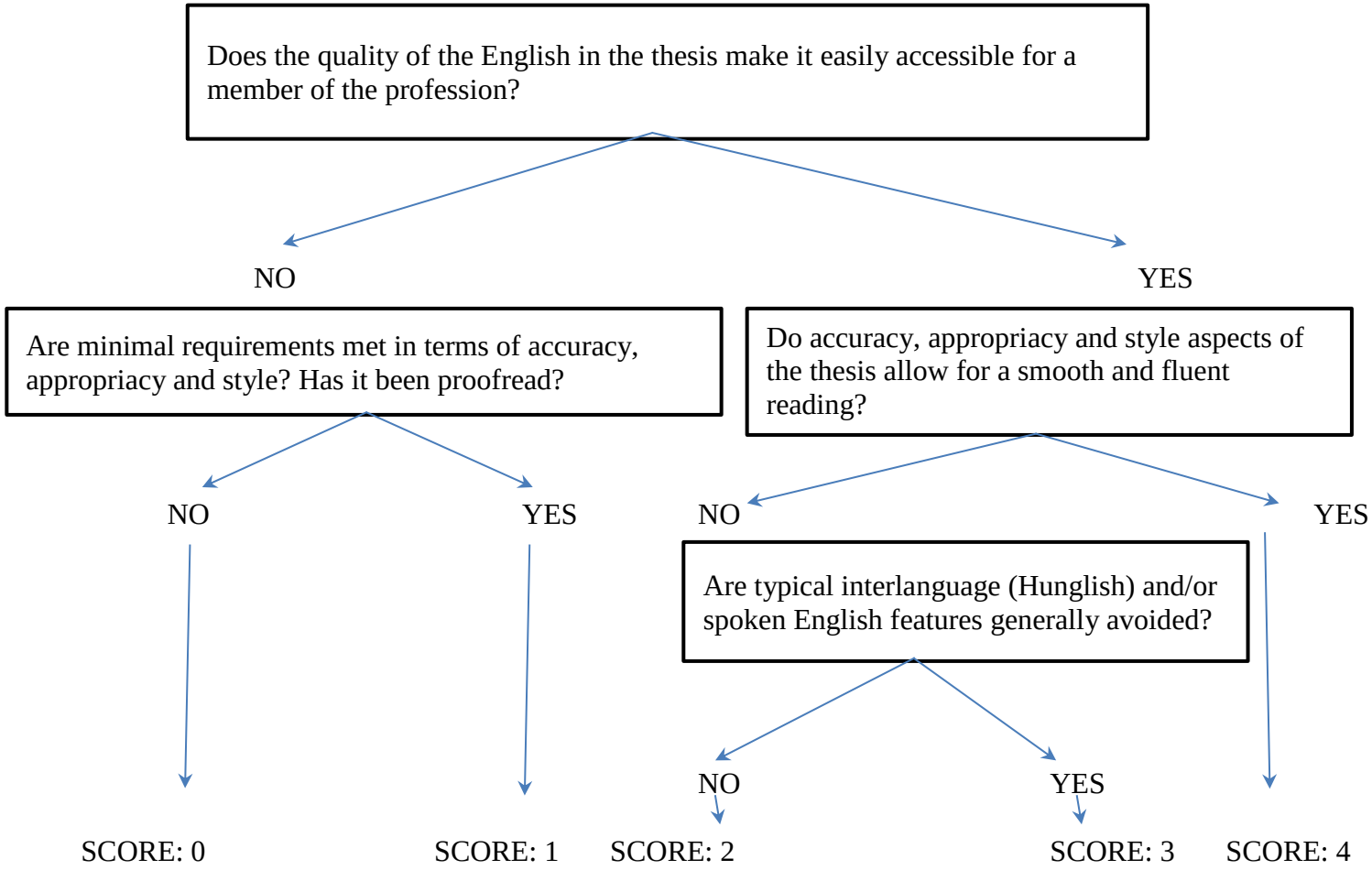
FORMAL REQUIREMENTS



QUALITY OF WRITING



QUALITY OF ENGLISH LANGUAGE USE



APPENDIX B.
Descriptors of the five bands for each category in the rating scales

Research method and procedures ● Analytical framework ● Focus ● (Quality of research design)	
4	Research design is fully coherent (complete).
3	Research design is generally coherent (complete). The research methods are appropriate, justified, explicitly stated, and consistently applied.
2	The reader perceives a clear focus. The research design is essentially coherent (complete), but there are lapses in explicit formulations and application of principles (incl. appropriacy of data collection).
1	Research design is generally not adequately justified by the author or it is not clearly stated. Relevant aspects of research can only be inferred from the text. The reader can identify the focus.
0	Research design is not justified or clearly stated. There is no proper question; if there is, the design is not appropriate. The thesis lacks a clear focus.

Theoretical and experiential basis ● Quality and number of sources ● Familiarity with the literature ● Synthesis of knowledge and skills	
4	There is an excellent synthesis of the relevant literature (knowledge, experience and skills), appropriate in length, drawing on their professional experience as well. There is evidence of the writer's critical judgment, explicitly and appropriately formulated. There is clear evidence of the writer's critical judgment.
3	There is a good synthesis of the relevant literature (knowledge, experience and skills), appropriate in length or somewhat longer than necessary. An honest, faithful description of the literature (well-selected, representative sources), albeit a little dense. There is some evidence of the writer's critical judgment.
2	Although the writer familiarized himself/herself with the literature, the synthesis is poor. Reporting takes place, but it is inconsistent, or partial (unsatisfactory, non-representative sources) or otherwise unconvincing. No evidence of critical judgment.
1	There is a literature review of an adequate number of sources, but there is little or no synthesis. The writer only verbalizes the literature. The reader wonders whether the thesis writer has adequately familiarized himself/ herself with the literature.
0	Unaccountable/untraceable sources, or too few sources selected. The relevant literature is not reviewed. No review section.

Interpretation of findings ● Discussion ● Implications ● (Enhanced awareness of professional development)	
4	Findings are linked to the research question(s) presented without overgeneralizations. The results of the study are explicitly linked to the research

	questions/hypotheses presented in the beginning of the study. They are linked to those of previous research by others. Relevant pedagogical implications are elaborated. Clear evidence of an enhanced awareness of the field/subject. Explanations are convincing.
3	Findings are linked to the research question(s) presented without overgeneralizations. The results are not explicitly linked to the research questions/hypotheses, but they may/may not be linked to those of previous research by others. Plausible explanations. Evidence of some awareness of the field/subject.
2	Findings are rather implicitly linked to the research question(s)/hypotheses. They are not linked to those of previous research by others. There is an attempt to explain the results of the study. Explanations may not be plausible. Possible presence of overgeneralizations. Relevant pedagogical implications are elaborated on.
1	Findings are not linked to the research questions/hypotheses, and/or results are mainly overgeneralizations, and the findings are not linked to those of previous studies. Pedagogical implications are superficial.
0	The thesis does not provide an interpretation of the findings. It is a mere description of the data. The results are not linked to those of previous studies/experience (no reflection), and no pedagogical implications are discussed.

Independence ● Contribution to the field ● Originality	
4	Besides independence, the thesis displays some original elements, however small in scope, which may be considered an original contribution to the field.
3	A good piece of independent work, although there is no originality in the thesis.
2	There is evidence of hard and conscientious work, but little independence is demonstrated.
1	The thesis demonstrates some elements of independence, as a bare minimum. There is heavy reliance on ideas by others.
0	The thesis demonstrates an overall reliance on others' ideas and work (albeit falling short of plagiarism). Superficial copying, "regurgitation" of ideas by others, without much insight. Lack of imagination.

Formal requirements ● Layout ● Citation conventions	
4	All formal requirements are thoroughly followed.
3	All formal requirements are generally and consistently met (but not thoroughly). Nevertheless, the thesis demonstrates a good standard of formal requirements, including citation conventions (APA), layout and length.
2	Most formal requirements are met. Some problems appear in citation conventions (APA), and/or layout (e.g. paragraphing) and/or length.

1	Many formal requirements are *not* met. The thesis only demonstrates the minimum.
0	Formal requirements are not met at all. A likely case of plagiarism*

*If the there is a case of plagiarism, the thesis will be failed.

Quality of writing ● Argumentation (in the whole of the text) ● Structure of writing	
4	The argumentation follows the academic standards of the field, sufficiently, convincingly, logically, and in a relevant way. It has some palpable persuasive power. It is well-structured, cohesive and coherent, reader friendly throughout.
3	The argumentation is somewhat idiosyncratic, but it is still convincing. The thesis is generally well-structured, coherent and reader friendly.
2	There is some clear argumentation, but there are flaws in it: The argumentation is debatable. The thesis is adequately (but not very well) structured. Nonetheless, it is still coherent.
1	The thesis is not adequately structured. Although the writer does take a stand, the argumentation is not convincing. The ideas are connected but the argument fails to convince. There are unsubstantiated claims. The writer manages to establish coherence, but it is not without problems.
0	Argumentation is absent, or it is completely unconvincing. There is only description. The thesis does little more than verbalize the results. The reader struggles with an obvious lack of coherence.

Quality of English language use	
4	A high degree of accuracy, appropriacy, and the academic style of the thesis allow for a smooth and fluent reading.
3	Infrequent lapses in accuracy and/or appropriacy and/or academic style do not impede fluent reading.
2	Frequent lapses in accuracy and/or appropriacy and/or academic style result in occasional lapses in fluency.
1	The quality of language does not allow for smooth and fluent reading. Reader struggles to appreciate professional content.
0	The thesis does not meet minimum requirements in terms of accuracy, appropriacy and academic style.